DSIE'18

13th Doctoral Symposium
in Informatics Engineering

Proceedings

of the

13th Doctoral Symposium in Informatics Engineering

Editors:
A. Augusto de Sousa
Eugénio de Oliveira

https://paginas.fe.up.pt/~prodei/dsie18/

Faculty of Engineering
University of Porto

Porto, Portugal

30 - 31 January 2018

COPYRIGHT

1^{a} Edição/ 1st Edition 2018
ISBN: 978-972-752-230-9
Editors: A. Augusto Sousa and Eugénio Oliveira
Proceedings Design: João Dias
Graphical Design/Website: Bruno Tavares and João Dias
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto

DSIE’18 SECRETARIAT:
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal
Telephone: +351 22 508 21 34
Fax: +351 22 508 14 43

E-mail: dsie18@fe.up.pt
Symposium Website: http://www.fe.up.pt/dsie18

FOREWORD

STEERING COMMITTEE

Google's CEO recently stated "motto", that the company will be "AI First" can surely be rephrased and enlarged to characterize our present and future world as "Computer Sciences First".

To have chosen to pursue advanced studies in Computer Science and Informatics Engineering was indeed a good choice for those who are contributing to these proceedings. Modern world is literally being shaped by computers and, above all, by flexible, adaptable, evolving, friendly, secure and intelligent software.

DSIE - Doctoral Symposium in Informatics Engineering, now in its 13th Edition, is an opportunity for the PhD students of the FEUP Doctoral Program in Informatics Engineering (ProDEI) together with MAPi to show up and prove they are ready for starting their respective theses work.

DSIE meetings have been held since the scholar year 2005/06 and the main goal has always been to provide a forum for discussion on, and demonstration of, the practical application of a variety of scientific and technological research issues, particularly in the context of information technology, computer science and computer engineering. DSIE symposium comes out as a natural conclusion of mandatory ProDEI course called "Methodologies for Scientific Research" (MSR), this year also available to MAPi students, leading to a formal assessment of the PhD students first year's learned competencies on those methodologies.

The above mentioned specific course (MSR) aims at giving students the opportunity to learn the processes, methodologies and best practices related to scientific research, particularly in the referred areas, as well as to improve their own capability to produce adequate scientific texts. With a mixed format based on a few theoretical lessons on the meaning of a scientific approach to knowledge, together with multidisciplinary seminars, the course culminates with the realization of the DSIE meeting. DSIE is seen as a kind of laboratory test for the concepts learned by students. In the scope of DSIE, students are expected to simultaneously play different roles, such as authors of the submitted articles, members of both scientific and organization committees, and reviewers, duly guided by senior lecturers and professors.

DSIE event is then seen as the opportunity for the students to be exposed to all facets of a scientific meeting associated with relevant research activities in the above mentioned areas. Although still at an embryonic stage, and despite some of the papers still lack of maturity, we already can find some interesting research work or promising perspectives about future work. At this moment, it is not yet essential, nor often possible, for most of the students in the first semester of their PhD, to produce sound and deep research results. However, we hope that the basic requirements for publishing an acceptable scientific paper have been fulfilled.

Each year DSIE Proceedings include papers addressing different topics according to the current students' interest in Informatics. This year, the tendency is on Machine Learning, Data Analysis and Information Extraction, Multi-agent Systems, Bioinformatics and Gamification, Computer Vision, Distributed Computing, Encryption, Systems and Networks security.

The complete DSIE'18 meeting lasts two days and also includes two invited talks by an academic researcher and a more industry related researcher.

Professors responsible for ProDEI program's current edition, are proud to participate in DSIE'18 meeting and would like to acknowledge all the students who have been deeply involved in the success of this event that, hopefully, will contribute for a better understanding of the themes addressed during the above referred course, the best scientific research methods and the good practices for writing scientific papers and conveying novel ideas.

Porto, January 2018

Eugénio Oliveira and Augusto Sousa (ProDEI)

FOREWORD

ORGANIZING AND SCIENTIFIC COMMITTEES

The chairs of the Organizing and Scientific Committees of the Doctoral Symposium in Informatics Engineering (DSIE'18) warmly welcome you to the DSIE 13th edition. With a great honour, we have accepted the invitation to be a part of these committees. Organizing an event, like the DSIE, confirmed to be both a challenging and practical task, in which all the persons involved have certainly derived great value.

The joint effort of our colleagues from the Doctoral Programme in Informatics Engineering (ProDEI) and the Doctoral Programme in Computer Science of the Universities of Minho, Aveiro, and Porto (MAPi), was instrumental in making this event a success. We believe that these efforts are reflected in the quality of the communications realized and the organization in general.

Our first acknowledgment goes to our supervisors, Professor Eugénio Oliveira and Professor Augusto Sousa. We would like to thank them for their time and their efforts in making this conference possible and for providing us with all invaluable concepts.

We would like to thank all the senior members of the Scientific Committee for their involvement, the junior members for their collaboration, and the invaluable support of Pedro Silva and Sandra Reis (DEI) from the Informatics Engineering Department of Faculty of Engineering - University of Porto.

And, above all, we thank you for being a part of DSIE'18!

Porto, January 2018

João Dias and Tiago Neto (Organization Committee Chairs)

Carla Abreu and Yassine Baghoussi (Scientific Committee Chairs)

CONFERENCE COMMITTEES

STEERING COMMITTEE

A. Augusto Sousa

Eugénio Oliveira

ORGANIZING COMMITTEE CHAIR

João Dias

Tiago Neto

Organizing COMMITTEE

André Santos

António Guerra

Arnaldo Pereira

Bruno Tavares

Carla Abreu

Joana Patrícia Bordonhos Ribeiro

João Costa

João Figueira Silva

João Neto

João Pedro Dias

José Ornelas

Kemilly

Marcelo Santos

Marisa Reis

Mohamed Yassine Zarouk

Ricardo Macedo

Saulo Carpio

Simão Reis

Tiago Neto

Vitor Enes

Yassine Baghoussi

SCIENTIFIC COMMITTEE CO-CHAIRS

Hugo Barbosa

Telmo Morais

SENIOR SCIENTIFIC COMMITTEE

Ali Shoker

Ana Paiva

Bruno Lima

Filipe Correia

Henrique Lopes Cardoso

Hugo Sereno Ferreira

João José Pinto Ferreira

João Mendes Moreira

João Pascoal Faria

José Maria Fernandes

Leonel Dias

Rosaldo Rossetti

Rui Maranhão

Sérgio Nunes

JUNIOR SCIENTIFIC COMMITTEE

André Santos

António Guerra

Arnaldo Pereira

Bruno Tavares

Carla Abreu

Joana Patrícia Bordonhos Ribeiro

João Costa

João Figueira Silva

João Neto

João Pedro Dias

José Ornelas

Kemilly

Marcelo Santos

Marisa Reis

Mohamed Yassine Zarouk

Ricardo Macedo

Saulo Carpio

Simão Reis

Tiago Neto

Vitor Enes

Yassine Baghoussi

SPONSORS

DSIE'18 – Doctoral Symposium in Informatics Engineering is sponsored by:

With support of:

CONTENTS

INVITED SPEAKERS

SESSION 1 - MULTI-AGENT SYSTEMS

SESSION 2 - DATA ANALYSIS

SESSION 3 - MACHINE LEARNING

SESSION 4 - BIOINFORMATICS, EDUCATION AND GAMIFICATION

SESSION 5 - ADVANCED INFORMATION EXTRACTION AND MACHINE LEARNING

SESSION 6 - COMPUTER VISION, BLOCKCHAIN AND ENCRYPTION

SESSION 7 - System Security and Networks

SESSION 8 - DISTRIBUTED COMPUTING

INVITED SPEAKERS

INVITED SPEAKER

HUGO PEIXOTO, PHD

Hugo Peixoto is an IT Specialist in Centro Hospitalar do Tâmega e Sousa in Penafiel where he is also a Project Manager for the team of Drug and Complementary Diagnostic Exams Monitorization. He is an Invited Auxiliary Professor at the University of Minho in Portugal where he lectures on Knowledge Extraction and Electronic Health Records to Biomedical Engineering Masters Degree. From 2009 he is member of the research team of Information Systems Technologies in the Algoritmi Center in the University of Minho. He also took part in the organization of international workshops (FiCloud 2017), reviews of articles in international conferences and has member of the scientific committee of international conferences. He as 11 indexed papers in Scopus, 52 citations in scopus (h-index=4).

INVITED SPEAKER

LUÍS PAULO REIS, PHD

Luís Paulo Reis is an Associate Professor at the University of Minho in Portugal and Director of LIACC – Artificial Intelligence and Computer Science Laboratory where he also coordinates the Human-Machine Intelligent Cooperation Research Group. He is an IEEE Senior Member and vice-president of the Portuguese Association for Artificial Intelligence. During the last 25 years, he has lectured courses, at the University, on Artificial Intelligence, Intelligent Robotics, Multi-Agent Systems, Simulation and Modelling, Machine Learning, Educational/Serious Games and Computer Programming. He was the principal investigator of more than 10 research projects in those areas. He won more than 50 scientific awards including winning more than 15 RoboCup international competitions. He supervised 18 PhD and 101 MSc theses to completion. He organized more than 50 scientific events and belonged to the Program Committee of more than 250 scientific events. He is the author of more than 300 publications in international conferences and journals (indexed at SCOPUS or ISI Web of Knowledge)

SESSION 1

Multi-agent Systems

Multidimensional Byzantine Approximate Agreement in Cyber-Physical Systems with Trust
Arnaldo Pereira

Improving the efficiency of air traffic controllers in simulated environments
Tiago Neto

Multidimensional Byzantine Approximate Agreement in Cyber-Physical Systems with Trust

Arnaldo Pereira

Universidade do Porto, Portugal
arnaldop@fc.up.pt

Abstract. A widely used strategy to operationalize the control of complex distributed systems is the consideration of entities empowered to make decisions. Connecting the physical and computational worlds, cyber-physical systems have been demonstrating that they are an effective way to deal with highly dynamic systems, particularly in industrial environments. The use of multi-agent systems have been a widely used way to implement control in cyber-physical systems. The embedding of trust mechanisms in these systems, operating in highly dynamic industrial environments, increases the resilience, which is a crucial issue in distributed systems. However, the presence of agents with Byzantine failures requires the systems to be enriched with strategies to prevent this possibility from inhibiting negotiation and trust methods' effectiveness. This paper proposes the implementation of a mechanism to overcome those problems, using a strategy of multidimensional approximate agreement. The experimental results show that the system is feasible and that the proposed mechanisms play together, scaling in linear way.

Keywords: Cyber-Physical Systems, Control Systems, Multi-Agent Systems, Trust, Multidimensional Agreement.

1 Introduction

In the age of globalization, consumers freely access producers around the world, regardless of the physical location of any of these actors. In addition, consumers are increasingly demanding the low cost, high quality and customization of the products they seek to acquire. Naturally, this forces companies across the globe to fierce competition to offer their products with higher quality and lower costs. To cope with these structuring changes in world markets, companies need to embrace new paradigms, new techniques and technologies, and new business models, sustained mainly by digitalization, massive exchange, and intensive processing of information [1].

Current requirements that are imposed by the needs of the customers to companies playing in an open economy can be tackled by the strategy of using Cyber-Physical Systems (CPS) [2]. In embedded systems, the core is the use of software programs running in stand-alone devices. Extending the concept of embedded systems, the idea of CPS is to enable the interconnection of interacting computational and physical elements, creating a network of programs and devices. Notoriously, CPS approach can be used to improve the safety, reliability, adaptability, functionality, efficiency and usability of complex large-scale systems. The application areas are multiple, from the chemical industry, logistics, manufacturing, building automation and energy, to mention just a few [3].

Several technological solutions are being advocated as promising to implement CPS solutions and to some extent being already applied. Multi-Agent Systems (MAS) [4, 5] are being used to provide distributed intelligence to the system's components while the Service-oriented Architecture (SOA) principles provide seamless vertical and horizontal system integration. A MAS is a set of agents that represent the objects of the system and that try to achieve their objectives by operating over its local knowledge and communicating when local knowledge is not enough to fulfil their purposes [4]. The capability of a system to maintain stability even in the presence of hazardous condition is a very desirable feature for control systems operating in demanding conditions like the ones presented in industrial environments. The presence of disturbances can be minimized by the simple use of MAS as these are recognized for its inherent robustness, since decentralization provides more redundancy, by eliminating the presence of central nodes typical in centralized solutions [6].

Literature reports on the problematic of security, safety and robustness when considering the use of multi-agent systems. Just to give some examples, [7], in the context of supply chain networks, describes the implementation of a multi-agent system to improve the resilience and [8], referring to cyber-physical systems applied to power systems, presents a hybrid framework for robust and resilient control design. The dangers posed by the Stuxnet worm to industrial systems are described in the work of Karnouskos [9]. Vila et al., using the JADE framework, explore the security requirements of a MAS application [10], and Bibu proposes a method for tracing irregular behaviors within the organization, by monitoring a set of selected events [11]. Cavalcante et al. survey architectures and security models for MAS [12], and [13] explores the integration of trust, learning and risk management mechanisms, in order to increase the resilience of CPS. Finally, with respect to Byzantine failures, they may occur untimely in distributed systems and are characterized by their unforeseen character being able to constitute an obstacle to the accomplishment of the desired operations by the system [14].

Nowadays, trust is of paramount importance for Artificial Intelligence (AI) solutions. As stated by the Gartner's top strategic predictions for 2018, until 2022, 50% or more will consume more false information than true, and by 2020, AI-driven creation of fake content will outpace AI's ability to detect it [15]. The importance of trust and reputation in distributed systems, particularly to support negotiation, is underlined in the literature. Reputation can be defined as "*the opinion or view on someone about something*" [16] and trust as "*a belief an agent has that the other party will do what it says it will or reciprocate, given an opportunity to defect to get higher payoffs*" [17]. Examples of application are given by Urbano et al., that combine trust and sanctions in negotiation processes [18], and Wong and Sycara in the context of MAS [19].

Concerning the problem of the agreement between agents, some work may be mentioned. In [20] the multidimensional agreement is observed from the point of view of a negotiation between the agents, but without considering the possibility of

Byzantine failures and [21] uses an asynchronous particle swarm optimization (PSO) algorithm to solve the agreement problem between agents over a real value. Mendes et al. [22, 23] presents a solution to solve multidimensional agreement in Byzantine systems, but without consider the application to MAS. The agreement between agents is desirable in several situations. For instance, when mobile entities like robots, controlled by agents, with some of them experimenting Byzantine failures, have the joint task to converge to a location that needs to be agreed based on asynchronous communication between them, we are in a situation where multidimensional approximate agreement can apply. Considering that non-faulty robots' agents starts sending the robots' position (a two or three-dimensional vector), it is possible to guarantees convergence to a location inside the convex hull defined by the starting points. Another interesting application relates to voting situations in which each voter votes in a weighted way in an array of options [23].

The motivation of this work is to focus on the integration of an algorithm of approximate Byzantine multidimensional agreement into a multi-agent system with embedded trust strategies.

The rest of the paper is organized as follows: Section 2 introduces a dynamic approach integrating trust and approximate agreement to achieve more resilient multi-agent cyber-physical systems. Section 3 presents the experimental implementation and achieved results. Finally, section 4 rounds up the paper with the conclusions and future work.

2 Integration of Trust and Multidimensional Byzantine Approximate Agreement in Multi-Agent Systems

The proposed approach considers that each agent has embedded a trust-based method to permit reliable cooperation among all the participants of the system, by allowing assessing the level of trust that an entity has over another entity. Previous cooperation experiences induce a mechanism that reinforces the trust levels evolution by the use of negative or positive feedbacks. The formulas that govern the evolution of trust level, T, that an agent has relatively to agent i in the instant (or round) k is given by the equations (1) that correlates positive and negative feedbacks:

$$\begin{cases} T_i(k) = (1+\alpha) \times T_i(k-1), \text{if positive feedback} \\ T_i(k) = (1-\beta) \times T_i(k-1), \text{if negative feedback} \\ 0 \le T_i(k) \le 1 \end{cases} \tag{1}$$

Obviously, the higher the value of T_i, the greater the confidence in agent i.

The positive reinforcement contribution, α, produces an increase of the trust level and the negative reinforcement contribution, β, conduces to a reduction of the trust level. It should be noted that these parameters, mimicking what happens in human reality, are not of equal magnitude, since it is considered that the increase of trust is slower than its decrease. The generation of these parameters is done dynamically and is based on the learning capabilities embedded in the agents that takes into account the significance (or importance) and the trend. Taking into account the events that occurred in the most recent interactions, the trend measures the degree of success or failure of such operations in a given business perspective. Therefore, this parameter evaluates the degree of compliance with the assumed contract. The significance parameter evaluates the degree of the economic damage caused by the breach of contract, penalizing more failed operations involving more money. The expressions regulating α and β are represented in (2), were $\gamma_1, \gamma_2, \gamma_3$ and γ_4 are weight parameters:

$$\begin{cases} \alpha = \gamma_1 \times \text{significance} + \gamma_2 \times \text{trend} \\ \beta = \gamma_3 \times \text{significance} + \gamma_4 \times \text{trend} \end{cases} \tag{2}$$

The multidimensional Byzantine approximate agreement (MBAA) problem for asynchronous systems was firstly proposed in [22] and extensively revisited in [23]. To allow the use of those ideas in the context of multi-agent systems some adaptation need to be done in order to allow a concrete implementation of MBAA in MAS. Let us start by considering a network of $\mathrm{n} \geq 2$ agents fully connected, exchanging messages, with d-dimensional vectors of reals, to solve the approximate agreement task. A first assumption is to consider that the messages are delivered in the order they were sent, or at least that the agent that receives the messages has a mechanism to endure that property. Also, consider that $\mathrm{f} \geq 1$ agents are Byzantine, but fulfilling the condition $\mathrm{n} > (\mathrm{d}+2)\mathrm{f}$. Each agent reliably identify the sender of any received message and every message needs to be marked with a tag with a round number, even considering that the messages are exchanged asynchronously. Furthermore, communication channels guarantees that all messages are eventually delivered (i.e. the communication is reliable). Now, the starting point for each non-faulty agent is a d-dimensional input vector with components that are real numbers.

To prevent that Byzantine agents convey different contents to different agents in a single round of communication, more assumptions are needed. A reliable broadcast protocol satisfy some important properties. If a non-faulty agent never reliably broadcasts a particular message, then no other non-faulty agent will ever receive that message (non-fault integrity). Non-faulty liveness states that if a non-faulty agent does reliably broadcast all other non-faulty agents eventually receive the message. Global uniqueness states that if two agents reliably receive in the same round a message from the same sender, then the message content is the same. Finally, global liveness states that after the end of the protocol, for two non-faulty processes, if they receive a message from the same sender decorated with the same tag round, then the content of the message is the same. Another condition that must be verified is that, at the end of each round, each pair of processes has $\mathrm{n}-\mathrm{f}$ values in common. In systems with reliably broadcast a witness technique can be implemented to ensure that (see reference [24] for more details).

After each round, each non-faulty agent has a set $\mathtt{C}$ of vectors from $\mathbb{R}^{\mathrm{d}}$, corresponding to the messages sent by the other agents. The polytope of $\mathtt{C}$ is the convex hull of points of $\mathtt{C}$ (smallest convex set that contains the points of $\mathtt{C}$). Considering collections of $\mathtt{C}$ of size $|\mathtt{C}|-\mathrm{f}$, each of the collections generates a convex hull and the intersection of all these hulls is called the safe area. The MBAA may now be presented as follows:

```
Algorithm 1 agent.AsyncAgree()
SendBroadcast(p, 0, $v_0$)
W←Content(Wit) from ReceiveWitness(0)
U←$\mathrm{Safe}_f$(W)
v←barycenter of U
R←$\left\lceil \log_2\left(\frac{\sqrt{d}}{\varepsilon} \times \max\{|x[m]-y[m]| : x, y \in U\}\right) \right\rceil$
r←r+1
while r≤R do
    SendBroadcast(p, r, v)
    upon W←Content(Wit) from ReceiveWitness(r) do
        U←$\mathrm{Safe}_f$(W)
        v←barycenter of U
        if r=R then
            SendBroadcastHalt(p, r)
        r←r+1
    end upon
    upon ReceiveBroadcastHalt(q, s) do
        r = R + 1
    end upon
return v
```

In round 0, each agent p broadcast its initial value v_0. Content(Wit) is the subset of messages received by the agent in a specific round, selected by the witness technique. R is the estimation of the number of rounds needed to achieve convergence. Each agent keeps a vector v with the current value that will converge to the agreed final value.

3 Results

To test the proposed approach considering a cyber-physical system, an electrical smart grid case study was used. The smart grids are considered the next generation of power grids, adding a flow of information to the flow of energy, allowing new functionalities like the use of smart meters and the negotiation of energy costs in an automated way [25].

3.1 Experimental Implementation and Test Conditions

The case study is shown in Fig. 1, which represents an electrical smart grid. Several elements constitutes the smart grid, generically divided into two large groups of actors. On the one hand, we have the energy producers (e.g. wind turbines, photovoltaic panels, public energy suppliers) and on the other, the energy consumers (e.g. domestic consumers, electric vehicles).

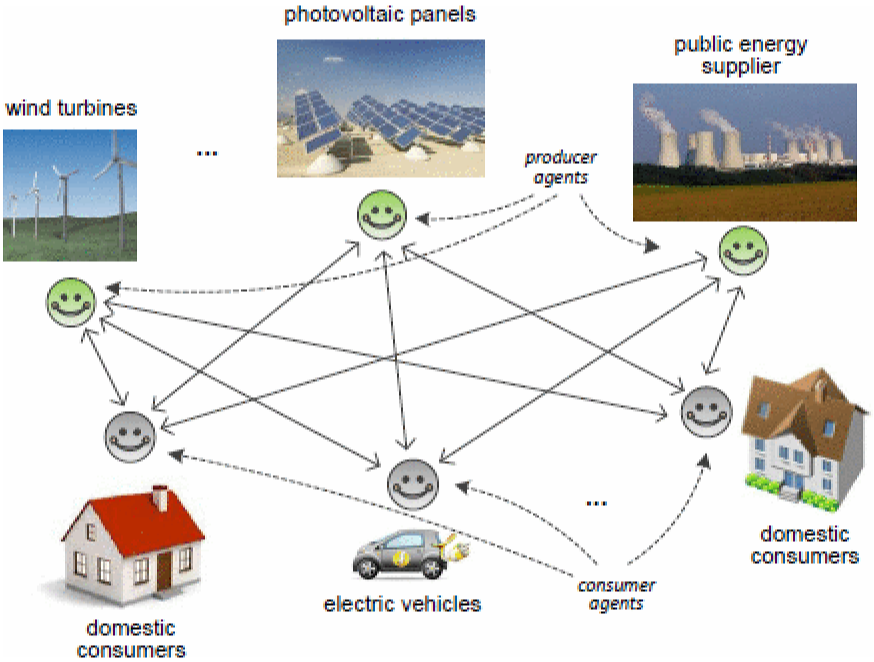


Fig. 1. Experimental case study scenario.

The MAS solution used in [13] to manage this complex system was adapted for the experimental purposes of this paper as described next. Producers and consumers are represented by agents who negotiate for the purchase and sale of energy slots. To simplify, consider that each producer sells each slot of energy that produces to a value that in each negotiation round varies between 1.0 and 1000.0. In a first phase, the consumers calculate the score (the smaller the better), as shown in equation (3), for each producer i, based on the price and the trust level T:

$$\mathrm{score_i} = \mathrm{price_i} \times (1 + 10^{-3} - \mathrm{T_i}) \tag{3}$$

Taking as the initial value the pair $(\mathrm{price}, \mathrm{trust})$ of the producer agent chosen, the consumers communicate with each other in order to reach an approximate agreement between them. Finally, the producer chosen by each consumer is the one whose values most closely match the values of the agreement between the consumers.

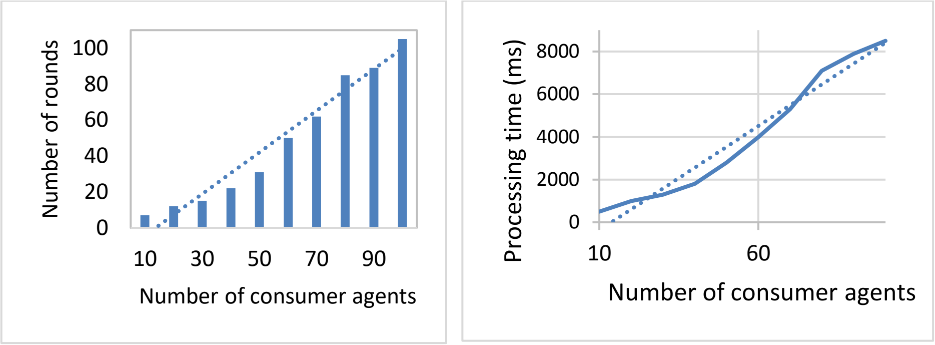
Regarding implementation, agents were created using the JADE framework [26] and the tests were done using the components described in Table 1.

Table 1. Hardware, operating system and software configuration.

Component	Description / Version
Processor	Intel Core i3-7100U
Total memory	4 GB
Operating system	Windows 10 x64
Java	1.8 update 151
JADE	4.5.0 (June 8, 2017)

3.2 Analysis of the Results

For the following, consider 10 producer agents, six of which fail 10% of the time and the remaining 25% of the time. Consider also that each consumer agent calculates the levels of confidence in the producers using the fixed values $\alpha=0.01$ for the positive reinforcement and $\beta=0.08$ for negative reinforcement, and consider $\varepsilon = 50.0$ for the approximate agreement calculations. As the agreement is about a pair of values of the form $(price, trust)$, $d=2$. Fig. 2 presents the evolution of the number of rounds and the processing time when considering 10, 20, ..., 100 consumer agents, and considering for each case that 10% of the agents are Byzantine.



(a)

(b)

Fig. 2. Evolution of the number of rounds and of the processing time against the number of consumer agents.

From the observation of the results, we can affirm that the number of rounds and the processing time increase linearly with the number of consumer agents, which can be problematic if the number of agents is very large.

4 Conclusions

The importance of having resilient distributed control systems is notorious. Empowering distributed systems with trust mechanisms allows them to be more resilient to disruptive events. Since multi-agent systems are one way to implement such systems, it is necessary to incorporate trust on them. Equally important is the possibility of agents' agreement on sets of real values, even in the presence of Byzantine failures.

This paper presented a strategy to integrate multidimensional Byzantine approximate agreement in a multi-agent system with embedded trust mechanisms. The experimental results showed that the system is feasible and can be applied successfully to complex industrial use cases, such as the smart grids scenario.

Possible limitations may emerge with the uncontrolled increase in the number of agents. The testing of these limits, together with the inclusion of more system robustness mechanisms, constitute future work.

References

1. H. Bauer, C. Baur, G. Camplone, and et. al., "Industry 4.0: How to Navigate Digitization of the Manufacturing Sector", Technical report, McKinsey Digital, 2015.
2. S. Karnouskos, "Cyber-Physical Systems in the SmartGrid", Proceedings of the 9th IEEE International Conference on Industrial Informatics (INDIN'11), pp. 20-23, 2011.
3. E. Lee, "Cyber Physical Systems: Design Challenges", Technical Report n. UCB/EECS-2008-8, University of California, 2008.
4. J. Ferber, Multi-Agent Systems, "An Introduction to Distributed Artificial Intelligence", Addison-Wesley, 1999.
5. P. Leitão, "Agent-based distributed manufacturing control: A state-of-the-art survey," Engineering Applications of Artificial Intelligence, vol. 22, pp. 979–991, Oct. 2009.
6. L. Raju, R. Rathnakumar, S. Ponnivalavan, L. D. Thavam and A. A. Morais, "Micro-grid grid robustness management using multi agent systems," 2017 International Conference on Power and Embedded Drive Control (ICPEDC), Chennai, 2017, pp. 38-43.
7. A. Smith, J.M. Vidal, "A Practical Multiagent Model for Resilience in Commercial Supply Networks", 12th International Workshop on Agent-Mediated Electronic Commerce (AMEC), 2010.
8. Z. Quanyan, T. Basar, "Robust and Resilient Control Design for Cyber-physical Systems with an Application to Power Systems," Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC), pp.4066-4071, 2011.
9. S. Karnouskos, "Stuxnet Worm Impact on Industrial Cyber-physical System Security", Proceedings of the 37th Annual Conference on IEEE Industrial Electronics Society (IECON'11), pp. 4490-4494, 2011.
10. X. Vila, A. Schuster, A. Riera, "Security for a Multi-Agent System based on JADE", Computers in Industry, 26(5), pp. 391-400, 2007.
11. G.D. Bibu, "Security in the Context of Multi-Agent Systems", Proceedings of 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'11), pp. 1339-1340, 2011.
12. R. Cavalcante, I. Bittencourt, A. Silva, M. Silva, E. Costa, R. Santos, "A Survey of Security in Multi-agent Systems", Expert Systems Applications, vol. 39, n. 5, pp. 4835-4846, 2012.
13. A. Pereira, N. Rodrigues, J. Barbosa and P. Leitão, "Trust and risk management towards resilient large-scale Cyber-Physical Systems," 2013 IEEE International Symposium on Industrial Electronics, Taipei, Taiwan, 2013, pp. 1-6.

14. M. Castro and B. Liskov, “Practical Byzantine fault tolerance”, OSDI '99 Proceedings of the third symposium on Operating systems design and implementation, 1999, pp. 173-186.
15. Gartner, “Top Strategic Predictions for 2018 and Beyond: Pace Yourself, for Sanity’s Sake”, 2017.
16. J. Sabater, C. Sierra, “REGRET: a Reputation Model for Gregarious Societies”, Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems, C. Castelfranchi and L. Johnson (eds.), pp. 475-482, 2002.
17. P. Dasgupta, “Trust as a Commodity”, D. Gambetta (ed.), Trust: Making and Breaking Cooperative Relations, Blackwell, pp. 49-72, 1998.
18. J. Urbano, H.L. Cardoso, A. Rocha, E. Oliveira, “Trust and Normative Control in Multi-agent Systems: An Empirical Study”, Highlights on Practical Applications of Agents and Multi-Agent Systems, Series: Advances in Intelligent and Soft Computing, 156, pp. 207-214, 2012.
19. H. Chi Wong, K. Sycara, “Adding Security and Trust to Multi-agent Systems”, Applied Artificial Intelligence, 14 (9), pp. 927-941, 2000.
20. E. Crawford and M. Veloso, “Negotiation in Semi-cooperative Agreement Problems,” 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, NSW, 2008, pp. 252-258.
21. V. Gazi and R. Ordonez, “Particle swarm optimization based distributed agreement in multi-agent dynamic systems,” 2014 IEEE Symposium on Swarm Intelligence, Orlando, FL, 2014, pp. 1-7.
22. H. Mendes, M. Herlihy, “Multidimensional approximate agreement in Byzantine asynchronous systems”. In: Proceedings of the 45th Annual ACM Symposium on Theory of Computing, pp. 391–400. STOC’13. ACM, New York, NY, USA (2013).
23. H. Mendes, M. Herlihy, N. Vaidya, V. K. Garg, “Multidimensional agreement in Byzantine systems” Distributed Computing, v 28, pp 423–441, 2015.
24. I. Abraham, Y. Amit, D. Dolev, “Optimal resilience asynchronous approximate agreement”. In: Higashino, T. (ed.) Principles of Distributed Systems. Lecture Notes in Computer Science, vol. 3544, pp. 229–239. Springer, Berlin (2005).
25. X. Fang, S. Misra, G. Xue, and D. Yang, “Smart Grid - The New and Improved Power Grid: A Survey,” IEEE Communications Surveys & Tutorials, vol. 14, n. 4, pp. 944-980, 2012.
26. F. Bellifemine, G. Caire, D. Greenwood, “Developing Multi-Agent Systems with JADE”, Wiley, 2007.

Improving the efficiency of air traffic controllers in simulated environments

Tiago Neto

Artificial Intelligence and Computer Science Lab., University of Porto, Portugal
ei12109@fe.up.pt,
WWW: http://paginas.fe.up.pt/ ei12109

Abstract. In the last decade, Europe has seen an intense grow in the number of flights in its airspace. The biggest concern with such a staggering growth is the lack of airports' capacity to manage all this air traffic. Because building or rebuilding airports is not always a solution given the price needed, there is a necessity of creating fast, stable and efficient tools to help air traffic controllers dealing with the extra demand of resources. The present work is inserted in a platform that is using *Microsoft Flight Simulator X* as a engine for simulating air traffic control and aircrafts. Since several mechanics are already implemented, such as the aircraft flow since it departs until it lands or runways' management, the present work intends to enable the platform to deal with congestion near airports as well as to mitigate some discrepancy existent in some mechanics between reality and the platform. To managing airspace congestion, a holding pattern system in which aircrafts can wait for their permission to land without disorganizing the airspace was implemented and a conflict-free permission to land system was introduced. To improve the airports' operations rate, it was introduced an algorithm capable of creating groups of runways that can be used simultaneously and choose the one who fits better accordingly to the demand of operations existent. In short, the introductions made brought an improvement of 25% in the landing rate as well as minimizing several non measurable discrepancies. The algorithm responsible for choosing active runways has shown to be working well, since it can adapt to different demands as well to different airports' configurations.

Keywords: Air Traffic Control, Holding Management, Airspace, Multi-Agent Platform

1 Introduction

Since 1905, with the start of the modern aviation by the Wright brothers, the world has witnessed the creation and improvement of aircrafts capable of transporting passengers or cargo in a fast and efficient way. The commodity of being able to connect two different points of the world in a short period of time, when compared with other means of transport, made air transportation one of the most essential ways of traveling. The growth of this way of transportation has been a constant as one can induce by analyzing the data given by *NAV Portugal*, where an increase of 10.6 percent in the number of flights under *IFR (Instrument Flight Rules)* is visible [1]. On a wider scale, the EuroControl has shown that every year new records of the number of flights are established.[2] The safeness of aviation rests on the air traffic controllers that provide a 24/7 service capable of helping aircrafts since they leave the hangar at the origin airport until they park on the destiny hangar.

With an increase of aircrafts on airports not followed by an improvement in airports' facilities, *ATCs (Air Traffic Controllers)* have seen their responsibility increase to values never seen since they now have at hands more aircrafts to guide. All four dimensions (3 spatial and Time) have an enormous relevance in the modern aviation world, the ability of managing not only the airspace around the airports but anticipating possible setbacks and automating processes capable of helping *ATCs* in their job are incredible important.

The present work is built in an already existent multi-agent platform which objective is to simulate multiple missions in a realistic way. Although the current state of the platform allows it to simulate the realistic flow of aircrafts since they departure until they arrive, several mistakes are being made on how aircrafts are managed in the airspace and are given permission to land. This mistakes heavily decreases the performance of a airport as well as jeopardize the aircrafts safety. To improve the platform, it is proposed the creation of an holding pattern management system and a creation of an algorithm to improve the use of the facilities, namely the runways, of the airport.

The present article is subdivided in 5 sections. Excluding the first one, *Introduction*, the article is subdivided in *Literature Review*, where some methods to improve airspace management are described, in *Better Modeling of Airspace Controlling* where the model is explained, followed by the *Results* and *Conclusions.*

2 Literature Review

With air congestion becoming a reality as a result of the maladjusted growth of airports when compared with the aviation growth, ATCs have seen an increase of aircrafts near airports. In an attempt to improve/prevent air congestion, several approaches of aircraft sequencing are studied in order to achieve a better landing rate and multi-airport operations.

The authors of [3] point out that the aircraft landing problem must be addressed as an optimum sequencing of aircraft and their division by different available runways at airports. With the objective of minimizing the time difference between the actual and intended landing times, the authors make use of integer programming. However, recognizing the complexity of the (NP-hard) problem, a hybrid meta-heuristic algorithm was developed to be able to obtain valid results in reasonable times. Thus, and taking into account the mandatory distances between different types of aircraft, both approaches developed were able to achieve optimum results the tests with one hundred aircraft and high quality solutions for tests with five hundred. Unlike the approach described in [3], the authors of the paper written in 2008 [4] try to take into account the costs involved in delaying an aircraft's landing. By introducing a method to ensure that no airline company would gain more than others and with an heuristic algorithm, all airlines achieved satisfactory results. The temporal complexity of the model used allows it to be used in real life situations.

In addition, the authors of [5] opted for the introduction of position changes in the sequencing of aircraft, with the aim of improving / optimizing the utilization rate for landing and departure. Only designed to be used in single-lane airports, the model proposed by the authors ensures fairness for the various airlines, i.e an aircraft can only switch to positions that are three positions away (e.g first to fourth, in an FCFS *(First-Come-First-Served)* queue). As with the above-described approaches, this requires horizontal and vertical separation of aircraft in order to avoid wake-vortex phenomena. The present approach also takes into account possible counterparts that may exist in exchanges of positions. The authors concluded after testing that it is possible to sequence multiple aircraft with polynomial complexity and because of that the present approach can be used in real cases.

Upon entering in one of the last stages of flight, an aircraft can be between 8 to 80km away from the airports. While in this stage, the aircrafts are often ordered to perform multiple maneuvers in order to wait for their time to land, however, those maneuvers cost money to airlines as congest the airspace, those solutions are only used in a last resource. As a solution, landing schedules that can improve landing and departure rates of the airport are used. This planning problem is known as *Time slot allocation*. Planning the landings not only reduces possible delays but also helps maintain the airspace safety. Unlike [6], whose attempt to resolve went through the attempt to minimize the total waiting time of all the aircraft or to minimize the landing time of the last aircraft, [7] aims to minimizing the time between two consecutive landings with a view to minimizing the number of times it is necessary to resort to the waiting maneuvers.

3 Better modeling of airspace controlling

The described work is implemented on the top of an already existent platform created by [8]. Prior to the present implementation, the platform was able to simulate, among others, multiple tasks done by aviation workers like pilots and *ATCs*. Despite the mapping of several tasks done by pilots and *ATCs* were done correctly, there were others such as choosing active runways or assigning runways to aircrafts that were defective. Considering all the runways available for landing as well as assigning all aircrafts that requested landing permission to the nearest runways instead of trying to minimizing their time in the air, the procedures that had been done were not just incorrect but entailed serious safety problems for aircrafts as minimum required distances were not being respected. The present article focus on this problems as one can read below. For better understanding, it is indispensable to remark that although airports have a certain number x of runways, the airport has 2 * x number of logical runways, as runways can be used both sides.

3.1 Choosing active runways

Currently, on real airports, active runways are chosen by *ATCs* that have into considerations factors as wind direction or expected operations. Since the search for a reliable source of group of runways that were used at the airports proved to be ineffective, an algorithm capable of taking into account the configuration of the various runways, generating plausible groups of runways, was created. It is important to remark that a group consists of one or more runways.

The developed algorithm has three distinct stages. On a first stage, it uses previously collected approach and departure routes for each logical runway to build polygons. Next, it test collisions between all created polygons. Lastly, the algorithm makes uses of the information collected by the second stage and creates non repeatable groups.

Creation of polygons Since approach and departure routes differ from one logical route to another, the current stage has to create polygons for each pair of logical route - operation to be used.

Although both polygons, approach and departure, are created between the first and last point of the respective route, an opening of 15 shaped like a semi-cone is added to the departure polygon. However and since the aircraft can either go to right or left after take off, the semi-cone is created at the time of the collision test in order to take into account the location of the other polygon that is being tested (see Fig.1 and Fig.2)

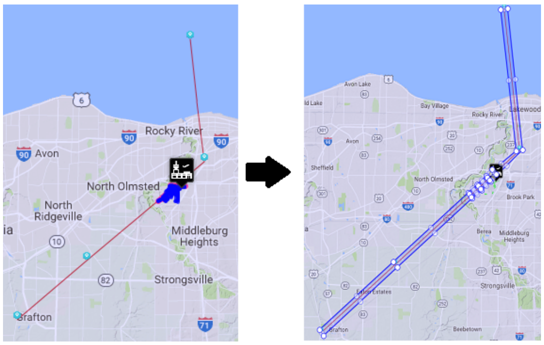


Fig. 1: Approach route and approach polygon of the runway 6L, KCLE

Testing collisions of polygons The ending of the first stage, initiates the collision tests between all polygons. To store the information in a usable way,

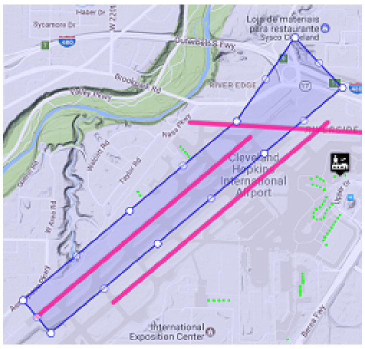

Haber Dr
Sycamore Dr
W 220th
Loja de materiais para restaurante Sysco Cleveland
Brookpark Rd
Valley Pkwy
RIVER EDGE
Cleveland Hopkins International Airport
W Area Rd
Taylor Rd
International Exposition Center
Berea Fwy

Fig. 2: Departure polygon of the runway 6L, KCLE

a graph is created where the existing nodes represent a pair logical runway - operation and the adjacencies between two nodes represent the possibility of both pairs being used simultaneously.

Creation of groups After the comparison between the various polygons it is necessary to transform the graph into groups of runways. To do that, the algorithm picks a random node and adds it to an empty group. Then and sequentially, a connected node is added. As nodes are being added, each new node has to be connected to all the nodes that were already added. If this is not the case, a second group is created containing the same nodes belonging to the first one with the only exception of nodes that have no adjacencies to the new node. As adjacencies are being used, they are deleted in order to prevent infinite loops. After every connected node is added to a group, the algorithm chooses another node which still has adjacencies to test. This process ends when all adjacencies are deleted.

After the determination of all the groups and based on the knowledge of the demand for operations that the airport will have in the next interval of time, the platform changes the active group of runways, if necessary, to allow a greater flow of aircrafts. Since the determination of the best group requires attention to several attributes, a decision algorithm SAW *(Simple Additive Weighting)*, where to each attribute is given a weight and where the sum of all the attributes weight should be equal to 1.It is possible to see in Eq. 1, the way in which it is calculated the fitness for each group. *ac, ar, dc* means, respectively, number of active runways that the group has, number of runways to be used for landing and number of runways to be used for departure. While Kr has a constant weight throughout different demands,K_{app} and K_{dep} vary according to the different needs of the airport. Before being normalized so that the sum of all weights is equal to one unit, K_{app} and K_{dep} correspond to the ratio between the number of flights to land or departure over the number of flights in the time interval to be considered.

$$P_{grouprunway} = K_r \frac{cur_{ac} - worst_{ac}}{best_{ac} - worst_{ac}} + K_{app} \frac{cur_{ar} - worst_{ar}}{best_{ar} - worst_{ar}} + K_{dep} \frac{cur_{dc} - worst_{dc}}{best_{dc} - worst_{dc}} \quad (1)$$

The group to be used is the the group who has better fitness, or in case of a draw, the first one is picked.

3.2 Airspace management near airports

Since it was necessary to organize the airspace near the airport, it was essential to build a module capable of controlling the aircraft that were in the final phase of their flight. Although a simplified form of airspace organization already existed, it did not ensure neither the safety of the aircraft or some of the rules imposed by the FAA *(Federal Aviation Administration)*. Given the rudimentary predecessor of the management described below, it was needed to implement from scratch the holding pattern maneuver so aircrafts could comply with the rules imposed by the US airspace control entity. The possibility of performing this type of maneuver, however, does not ensure a safe and organized management of aircraft in the airspace.

In a way to organize the airspace, a holding management queue, equal to the one being used nowadays by real *ATCs*, is added. The essential about this method is that all waiting aircrafts are now performing holding patterns around the same fixed point (latitude, longitude) but separated by 1000 or 1500ft as the *FAA* rules. With this method, we were able to eliminate all aircrafts that were dispersed across the airspace and agglomerate all of them in specific spots.

From Holding Queue to Landing With airports being able to choose which group of runways would be best for their current and future demand and their airspace well organized, a way of improving their landing rate in order to prevent the congestion is needed. With this in mind, it was implemented a new way of sequencing aircrafts in their way to the runway.

The conservative version of the algorithm only allowed one aircraft to be on the approach route to a specific logical runway. According with *FAA*[9], several aircrafts can be on their way to the runway as long as vertical and horizontal distances as well as time distances are fulfilled. Hereupon, it was necessary the development of an algorithm that would be able to predict the localization of several aircrafts in any given time. (see Algorithm 1)

4 Results

4.1 Creation of an algorithm capable of clustering runways

Scenarios To be able to test the algorithm implemented in Sec.3, several airports with different runways' configurations were chosen. For a first test scenario, it was chosen the airport *KCMH (John Glenn Columbus International Airport)*

Algorithm 1 Algorithm responsible for allowing aircrafts to initiate landing protocol

```
procedure CanInitiateLanding(AircraftData airc)
    Retrieve Location Of Next Aircraft
    Check if runway is receiving open

    Check if there are any aircraft in its route

    if there are aircrafts landing then
        Get both aircrafts estimated landing time
        minimumRequiredDistance = Check if distance between last aircraft and next are above minimum required
        time = Time In Minutes For Next Aircraft catch Last
        if minimumRequiredDistance && lastAircraftEstLandingTime > nextAircraftLandingTime + time then
            aircraft can leave
    else
        aircraft can leave
```

which has two parallel runways that can be used simultaneously since the distance between runways is more than the required. Next, *KCAK (Akron-Canton Regional Airport)* was picked as it has two intersecting runways. Finally, as a third scenario, *KCLE (Cleveland-Hopkins International Airport)* was chosen. This last airport has, unlike the others, three runways were none of which are crossed.

Results obtained To evaluate the results given by the algorithm, it was taken into consideration four attributes: number of groups created, number of groups containing only runways for departures or arrivals and groups containing runways to be used with both operations. In the Table 1, the results obtained for the various scenarios, KCMH, KCAK and KCLE respectively, are shown.

Table 1: Acquired Results for the three scenarios

Scenario	Groups Created	Approach Groups	Departure Groups	Mixed Groups
#1	12	4	4	4
#2	8	4	4	0
#3	12	6	6	0

As mentioned above in Sec.3, as one physical runway has both sides, there are two usable logical runways. However, the assigning of one logical, makes the its pair automatically unavailable. It is important to remark that the algorithm created more groups than those showed, however, the algorithm eliminates repeated groups as well as groups who are contained in another.

Being the only scenario where more than one runway that can be used simultaneous, the first scenario is the only airport where mixed groups can exist since both runways can be used to perform different operations. As both physical runways can be used to perform the same operations, four groups were created for the other attributes. In relation to the second scenario and because we are facing a cross-arrangement, there is an impossibility of having mixed groups. Nevertheless, 4 groups of approach and departure were created. Lastly, even thought *KCLE* have three physical runways, the airport used for last scenario, does not have any mixed group since all runways were not separated each other by the minimum required by *FAA*. In short, the algorithm created works well and it is able to adapt to different configurations of runways.

Having all possible groups created, there was a necessity to choose which group was the better in a particularly condition. In order to culminate this necessity, it was created, as described in the section above, a function capable of determining how good a certain group is for a certain airport demand. To test it, three different scenarios were created: one where the demand for departures and arrivals are the same, a second where there is only demand for departures and finally, a scenario where all aircrafts are asking for landing. The table represented in Table.2 shows for each Airport-Scenario, the estimated values of how optimal the best and worst group of runways is.

Table 2: Results obtained for best group of runways

	Scenario 1	Scenario 2	Scenario 3
KCMH	0.65	1	1
	0.55	0.1	0.1
KCAK	0.65	1	1
	0.65	0.1	0.1
KCLE	0.65	1	1
	0.65	0.1	0.1

Concerning the first airport *KCMH*, where it is possible to have mixed operation simultaneously, the function developed was able to choose a mixed group as best for a situation where the airport has arrivals and departures. Regarding the other scenarios, the function was able to pick a group containing groups with only departure or arrival runways. About the other airports and for the first scenario, because only one runway can be used at at a time, the worst group is evaluated with the same rating as the best one since the runway is meant to be used for both operations. For the latter scenario, the results were equal to the ones obtained in the first airport, where the function was capable of picking the groups of runways with the operation needed.

Through the Table 3, it is possible to notice that different demands result in a differentiation of groups of runways to be used. As for a scenario where both operations are needed, the algorithm is capable of choosing groups containing both operations and for extreme cases of demand (only departure or arrival),

Table 3: Best and worst groups for each Airport-Scenario

	Scenario 1	Scenario 2	Scenario 3
KCMH	10R departure 28R arrival	28R departure 10R departure	10L arrival 28L arrival
	10L departure 10R departure	10L arrival 28L arrival	10L departure 28L departure
KCAK	23 departure	23 departure	23 arrival
	10 arrival	10 arrival	10 departure
KCLE	28 departure	28 departure	28 arrival
	6L arrival	6R arrival	6L departure

the airport is able to opt for groups consisting of only runways with the operation needed. Through the same table, it is possible to notice that the groups created contains x runways, being x the number of runways that can be used simultaneously to that specific airport.

Although the results were satisfactory, different attributes needed to be included in the function that calculates the fitness of each group to allow the algorithm to be more dynamic.

4.2 Management of Holding Queues in Congested Airspace

To test the implemented approaches, each one was tested in three different and fully mapped(using real routes and locations for waiting) airports. For each airport, all the runways were closed and eight aircrafts were created and ordered to ask for landing permission.

On a first scenario, *KCMH* airport, it was possible to notice a quite significant improvement in both landing rate and average time that the last aircraft had to wait in the waiting queue. When compared to the sequential version which got a landing rate of 44.12 aircrafts per hour, the conservative version got a modest 13.4 aircraft per hour. With the conservative version , the last aircraft in the queue had to wait in average 8 minutes which is 3.25 times more than the sequential version. On a second scenario, *KCLE* airport, where only one runway can be used simultaneously, was used. Despite having a different configuration, the sequential version continues to be the best approach since it got a landing rate of 20.1 instead of 6.15 gotten by the conservative version. It should also be noted that the value of landing rate is not the same of the first scenario since the distance of the holding queue and the number of available runways are different. The last aircraft went from waiting over an hour, in the conservative version, to twenty minutes. Finally, for a last scenario, the *KCAK* airport was used and similarly to *KCLE*, the sequential version got a landing rate of 21.7 aircrafts per hour where the conservative version got a modest 6 aircrafts per hour. The last aircrafts waited in average eighteen minutes in order to get clearance to land instead of one hour and nine minutes when using the sequential version.

By carrying out the tests described above it has become clear that the use of the sequential approach represents an improvement in the landing rate. However, although the used version contributes immensely to the improvement of the landing rate, the length of the approach route as well as the configuration of the tracks and the aircraft speed are able to significantly change it.

5 Concluding remarks

The present work has introduced improvements to the current state of the platform by making it more dynamic and auto-sufficient. To do so, an adaptable algorithm capable of clustering runways that can be used together and a fitness function to determine which group is the best to the current demand seen at the airport were implemented. Furthermore, as an well succeed attempt to improve the landing rate of an airport, it was created an algorithm capable of sequencing aircrafts that minimize the time between landings. While the algorithm responsible for the grouping of runways has no limitation, the fitness function does need to take into consideration more attributes such as direction of wind and the length of the runways. On the other hand, the described way of controlling the clearances to land given need to be workout to accommodate emergency landing as well as redirections of aircrafts if expected landing times are too long. In short, the features developed brought an improvement of 314% in the landing rate and therefor a shorter time for aircrafts to wait for their permission to land.

References

1. E. NAV Portugal, "Movimentos ifr totais na riv de lisboa," jan 2017. Available online at https://www.nav.pt/nav/quem-somos/dados-de-tr%C3%A1fego (Acessed in June 2017).
2. Eurocontrol, "Eurocontrol seven-year forecast," sep 2016. Available online at http://www.eurocontrol.int/sites/default/files/content/documents/official-documents/forecasts/seven-year-flights-service-units-forecast-2016-2022-september-2016.pdf (Acedido em Junho 2017).
3. A. Salehipour, M. Modarres, and L. M. Naeni, "An efficient hybrid meta-heuristic for aircraft landing problem," *Computers & Operations Research*, vol. 40, no. 1, pp. 207 – 213, 2013.
4. M. Soomer and G. Franx, "Scheduling aircraft landings using airlines preferences," *European Journal of Operational Research*, vol. 190, no. 1, pp. 277 – 291, 2008.
5. H. Balakrishnan and B. G. Chandran, "Algorithms for scheduling runway operations under constrained position shifting," *Oper. Res.*, vol. 58, pp. 1650–1665, Nov. 2010.
6. A. M. Bayen, C. J. Tomlin, Y. Ye, and J. Zhang, "An approximation algorithm for scheduling aircraft with holding time," in *2004 43rd IEEE Conference on Decision and Control (CDC) (IEEE Cat. No.04CH37601)*, vol. 3, pp. 2760–2767 Vol.3, Dec 2004.
7. J. Martyna, "Runway scheduling with holding pattern and service priority," *Automatics/Automatyka*, vol. 16, no. 2, p. 137, 2013.
8. D. C. Silva, *Cooperative Multi-Robot Missions: Development of a Platform and a Specification Language.* PhD thesis, Faculty of Engineering, University of Porto, Porto, Portugal, 2011.
9. F. A. Administration, "Pilot and air traffic controller guide to wake turbulence."

SESSION 2

Data Analysis

Metamorphic Virus Detection through Data Compression
Simão Reis

Outlier Identification in Multivariate Time Series: Boilers Case Study
Joana Ribeiro

Automated Fare Collection Data For Measuring Socio-economic Impacts On The Transport Supply*
Yassine Baghoussi

Metamorphic Virus Detection through Data Compression

Simão Reis

IEETA / University of Aveiro, simao.paulo@ua.pt

Abstract. Malware is a malicious software that enforces unintended behaviour by the user on a computer system, threatening its security. Malware is commonly obfuscated to evade detection systems that use signature based techniques. The aim of this work was to detect the presence of metamorphic malware (which changes itself each time it is propagated). This is done by determining the distance between evolutions using a new compression based metric where target file is compressed using a set of files as reference. The principle is the following: if the objects x and y are very similar in terms of information content, x can significantly be compressed given y as reference, and vice versa. Moreover, if x and y are very different compressing x given y should not differ greatly from compressing x alone. The resulting malware detector achieved an accuracy of 96% over the tested family of metamorphic worms.

Keywords: Kolmogorov Complexity, Algorithmic Entropy, Data Compression, Similarity Measures

1 Introduction

Malware, short for malicious software, is any software used to disrupt computer operation, gather sensitive information, or gain access to private computer systems. Metamorphic malware modifies its internal structure at each infection, while remaining functionally equivalent [1].

Malware detection is a current area of research and given that malware obfuscation techniques are constantly evolving, so must the detection methods.

Previous work [2], developed a metamorphic worm called MWOR that carries its own morphing engine. Since the morphing engine could work as a signature, MWOR can morph its own engine. It was build to evade HMM (Hidden Markov Model) detection based techniques. Previous work [1], [3] initiated a study of entropy based similarity check. Their experimentation was conducted over the MWOR virus family. Concretely, they use data compression and Levenshtein Distance [4]. The second one is computationally hard for very large strings. Although they solve this problem by segmenting the strings into smaller ones they somehow lose context from the past segments.

Therefore, we propose an algorithm to measure similarity between files using on-line adaptive data compression and information distance, in order to label files as malware, given other metamorphosis of the same malware as reference. Although metamorphosis changes the structures of a given malware program, it should still be closer to other evolutionary stages of the same program than other programs that are functionally different.

The rest of this paper is organized as follows. Sections 2 and 3 presents a brief malware theoretical background. Sections 4 and 5 presents a theoretical background of information distance concept. Section 6 briefly presents CondComp, a compression software developed to build models using multiple files as reference. Section 7 present our proposed malware detector. Section 8 presents our experimentation. Finally, Section 9 presents our conclusions and future work.

2 Malware Classification

Computer viruses, worms, spyware, trojan horses, rootkits, are all types of malware [1], [5], [6]. In this section we pretend to describe each type of malware, each one defined by their functional behaviour.

Viruses are harmful software that append to executable files and have the ability to replicate themselves. When the infected file is run, the virus code is executed. In the case of metamorphic viruses it can evolve into new variants and modify itself. Virus propagate through the network or media devices like USB pen drives.

Worms are malicious software that can replicate themselves and propagate trough the network. Contrary to virus, they are standalone and don't require external executable code to run.

Trojans emulate the behaviour of an authentic software, hijacks the users credentials and gains control of the system.

Spyware tries to collect confidential information of user like credentials, web activity, keys pressed, etc. They don't present self-replicating abilities like viruses and worms.

Rootkits try to hijack a system by gaining administrator access of a system and tries to hide all traces of its behaviour that the system has been compromised. They fall into two categories: the ones that gain control of user mode privileges and the ones that gain control of kernel mode privileges. Both cases are irreparable.

3 Malware Variants

Malware can be classified by their structural behaviour as polymorphic or metamorphic [5].

Polymorphic malware maintains its code intact but is programmed to look different each time it replicates. It is constituted by the malicious encrypted code and a decryptor module. In each mutation the same payload (malicious code) is encrypted with a different process and has attached the correspondent decryptor module.

Metamorphic malware uses different types of obfuscation techniques to reprogram itself into a new code similar to the original. They can mutate while traversing the network. Multiple techniques used by metamorphic malware involve: (i) disassembling, (ii) permutation, (iii) expansion, (iv) assembling, etc.

4 Kolgomorov Complexity

Considering the string `"abababababababababab"`. We could also write a symbolic representation as `"10ab"` (ten times ab), this one being shorter. The first is called representation by enumeration, the second is called representation by comprehension.

There are many more examples we can find. Consider the program in C:

```
x[0] = 0; x[1] = 1; x[2] = 2; x[3] = 3; x[4] = 4;
```

Alternatively, we could express the same program with a loop instruction:

```
for (i = 0; i <= 4; i++) x[i] = i;
```

Both statements are functionally equivalent, setting all the positions of the array of length 5, but the second program with just one statement instead of five, and most importantly a shorter representation (written with less characters).

A string s can be represented by explicitly writing all of it's characters or by writing a program p that generates that string (note that writing all characters is also the result of an algorithm that generates s). That program is a symbolic representation of the string, where all steps to write the string are implemented within it. As in the first example string `"10ab"` tells that we must read the string as being the sub-string `"ab"` five times in a row.

Formally: if we have a set of descriptions of s, $d_i(s)$, the Kolgomorov Complexity $K(s)$ is the length of the description of minimal length (i.e. it uses the fewest symbols possible to represent the string) [7]

$$K(s) = \min_i |d_i(s)|. \tag{1}$$

The complexity of a string is it's minimal description, either by enumerating all its characters or either by writing the minimal program that generates that string. A description with length equal to the Kolgomorov Complexity of the string, it's called Kolgomorov random, as there are no more patters in the string that can be used to describe it in a shorter way.

5 Information Distance

The Kolgomorov Complexity of a string gives us its self information, the number of symbols needed to represent the string. But what we pretend is to measure the distance between strings instead of measuring and individual string.

From the definition of the Kolgomorov Complexity, we can define the concept of Information Distance (ID) or conditional Kolgomorov Complexity $K(x|y)$ of string x given y as input (or reference) [8]. This can be interpreted as the length of the shortest program that generates x from y, $p_i : y \rightarrow x$,

$$\mathrm{ID}(x, y) = K(x|y) = \min_i |p_i|. \tag{2}$$

The Levenshtein Distance is an example of a information distance metric. Consider the strings `"cat"` and `"rat"`. To transform the first one to the second one we just substitute character `"c"` by character `"r"`, one operation. The Levenshtein Distance of `"cat"` from `"rat"` is one. Other operations involve inserting or removing characters. Each operation has the same cost as it changes the same number of characters per operation. The Levenshtein Distance is the minimum number of operations to transform a string to another.

But two problems prevails. Consider two strings of 1000000 bits with a distance of 1000 bits (1000 operations in terms of Levenshtein Distance) and two strings of 10000 bits with 1000 bits of distance also. In the first case, they are closer to each other than in the second case [9] as in they have a less percentile of different bits (0.001%) and in the second case 10%. The first problem is that Information Distance is an absolute measurement. Therefore we must introduce the Normalized Information Distance (NID), a relative metric. It uses the self information as a factor:

$$\mathrm{NID}(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}. \tag{3}$$

The NID of the first string pair is 0.0001 and 0.1 in the second pair, now clearly distinguishing the distance of both pairs. The max terms are too see which string compresses better the other.

The second problem with this metric is that the Kolgomorov Complexity is non-computable. An alternative could be the use of the original size of the file as the factor, but as said in [9], the triangle inequality is not satisfied, and thus, it is not a metric.

Thus an approximation of the NID may be calculated using the length of compressed data as the complexity measurement, to use data compression as an approximation of the Kolgomorov Complexity or self information:

$$|C(s)| \sim K(s). \tag{4}$$

In [10], the metric Normalized Compression Distance (NCD) is defined as:

$$\mathrm{NCD}(x, y) = \frac{|C(xy)| - \min\{|C(x)|, |C(y)|\}}{\max\{|C(x)|, |C(y)|\}}. \tag{5}$$

xy is the concatenation of string x and y.

6 Conditional Compressor

An adaptive compression software was used, the Conditional Compressor (CondComp) developed by Institute of Electronics and Informatics Engineering of Aveiro (IEETA). CondComp measures information distance between files based on the compression obtained from a target file. It allows the use of a set of files as reference. It uses Markov Models (conditional probabilities) to build models of occurrence of characters of 8 bits and as such to perform nth-order entropic compression. This means that instead of the usual entropic compression where one has the probabilistic model of symbol occurrences from a given alphabet, it uses conditional probabilities, building the probabilities of a 8 bits character occurring after the occurrence of another sequence of n characters. It is adaptive because if the current models present bad results, they are discarded and new ones are build on-line in place.

The parameters of compression are a 4-tuple (R, t_k, r_k, γ) where R is the set of files to be used as reference, t_k is the search depth on the target file, r_k is the search depth of each reference file and γ is the ease to maintain the current model.

7 Malware Detector

Our developed malware detector follows the structure of a classic classifier but using the NCD as the similarity metric.

A malware detector D is a function

$$D : P \to \{\text{malicious}, \text{benign}\}, \tag{6}$$

where P is the set of executable programs. It scans a program $p \in P$ and classifies it as malicious or benign (binary classification).

The three relevant outcomes are: (i) false positives, (ii) false negatives and (iii) hit ratio. False positives is when a benign file is classified as malicious. False negative is when a malicious file is classified as benign. Hit ratio (or true positive) is when a malicious file is correctly detected.

As CondComp builds models based on a set of files R, we don't use the NCD, as traditional compression tools only use the self information of a file to compress it, but a new more powerful compression metric that approximates more to the definition of conditional Kolgomorov Complexity. We call this metric the Normalized Conditional Compression Distance, and define it as:

$$\text{NCCD}^p(x) = \frac{|C(x|R)|}{|C(x)|} = \frac{\text{CondComp}^{t_k,r_k,\gamma}(x|R)}{\text{CondComp}^{t_k,r_k,\gamma}(x)} \tag{7}$$

$\text{CondComp}^{t_k,r_k,\gamma}$ is the returned value by CondComp with parameters t_k, r_k, γ.

The algorithm proposed have the goal of measuring the NCCD of the target file x from a set of malware file references R and see if the distance is bigger than a specified threshold t. If the distance is less than or equal to t, x is labeled

Algorithm 1 Malware classification algorithm

```
procedure MalwareClassifier($f, p, t$)
    $d \leftarrow NCCD^{t_k, r_k, \gamma}(f)$                ▷ Distance from malware
    if $d \leq t$ then
        return True
    else
        return False
```

as malware. Otherwise its too far from being a malware. Algorithm 1 describes this process in pseudo code.

To find the best parameters and threshold to our algorithm, a training algorithm was developed. It first calculates the NCCD of a set of viruses V and a set of benign files B having another set of viruses V' with $V \cap V' = \emptyset$ as reference. Next, the threshold t is set as the biggest NCCD obtained from evaluating the NCCD of each element of V. Then it is tested how many files from B gave a NCCD greater than the threshold and the success rate is saved. After that it's calculated the average distance of the files in set B from the set of references V'. If the success rate and average distance are greater than all previous ones the current combination of parameters is saved as the best one. If the success rate is the same it is preferred the case with the largest distances to have more error margin. It is shown at Algorithm 2.

CondComp was developed in the C/C++ programming language. Our malware detector algorithm was implemented and tested as a Python programming language script. The detector calls CondComp binary to calculate the compression degree of each file.

8 Tests and Results

8.1 Virus data

It was used as a set of virus data a family of metamorphic worms called MWOR[1]. These worms were designed to defeat statistical based classifiers that rely on assembly opcode based analysis [1]. MWOR family has the ability to insert arbitrary amounts of dead code into the generated malicious code [1]. MWOR worms also copy benign code and place it within themselves [1].

Both in the reference virus files as in the target virus files, it was chosen multiple non-adjecent evolutions of the generation 4.0 of MWOR2, in such way that they wouldn't be the same as required by the Algorithm 2. The worms used as data reference are shown in Table 8.1, while the target worm files used are shown in Table 8.1.

[1] `http://cs.sjsu.edu/~stamp/viruses`

Algorithm 2 Malware classification training algorithm

```
procedure MalwareTraining($V, B, P$)
    $l \leftarrow |B'|$                                  ▷ Number of tests
    $b_r \leftarrow 0$                                   ▷ Best success rate
    $b_a \leftarrow 0$                                   ▷ Best average distance margin
    $b_t$                                                ▷ Best threshold
    $b_p$                                                ▷ Best parameter combination
    for $p := t_k, r_k, \gamma \in P$ do
        $V' \leftarrow \emptyset$                        ▷ Virus NCCD samples
        $B' \leftarrow \emptyset$                        ▷ Benign NCCD samples
        for $v \in V$ do
            $V' \leftarrow V' \cup \{NCCD^p(v)\}$
        for $b \in B$ do
            $B' \leftarrow B' \cup \{NCCD^p(b)\}$
        $t \leftarrow \max(V')$                          ▷ Threshold
        $s \leftarrow 0$                                 ▷ Success count
        $d \leftarrow 0$                                 ▷ Distance measure
        for $b \in B'$ do
            if $b > t$ then
                $s \leftarrow s + 1$
                $d \leftarrow d + b - t$
        $r \leftarrow s \div l$                          ▷ Success rate
        $a \leftarrow d \div l$                          ▷ Average distance
        if $r \geq br \wedge a > b_a$ then
            $b_r \leftarrow r, b_a \leftarrow a, b_p \leftarrow p, b_t \leftarrow t$
    return $(b_p, b_t)$
```

Table 1. Reference dataset of MWOR binaries from one generation.

1	MWOR2/worm_gen.dc4.0/MWOR_5	6	MWOR2/worm_gen.dc4.0/MWOR_55
2	MWOR2/worm_gen.dc4.0/MWOR_15	7	MWOR2/worm_gen.dc4.0/MWOR_65
3	MWOR2/worm_gen.dc4.0/MWOR_25	8	MWOR2/worm_gen.dc4.0/MWOR_75
4	MWOR2/worm_gen.dc4.0/MWOR_35	9	MWOR2/worm_gen.dc4.0/MWOR_85
5	MWOR2/worm_gen.dc4.0/MWOR_45	10	MWOR2/worm_gen.dc4.0/MWOR_95

Table 2. Virus dataset of MWOR binaries from one generation.

1	MWOR2/worm_gen.dc4.0/MWOR_0	6	MWOR2/worm_gen.dc4.0/MWOR_50
2	MWOR2/worm_gen.dc4.0/MWOR_10	7	MWOR2/worm_gen.dc4.0/MWOR_60
3	MWOR2/worm_gen.dc4.0/MWOR_20	8	MWOR2/worm_gen.dc4.0/MWOR_70
4	MWOR2/worm_gen.dc4.0/MWOR_30	9	MWOR2/worm_gen.dc4.0/MWOR_80
5	MWOR2/worm_gen.dc4.0/MWOR_40	10	MWOR2/worm_gen.dc4.0/MWOR_90

8.2 Benign data

MWOR2 generator produces Linux based worms [1], and injects into the code benign code of Linux files [2]. Therefore we use Linux command binaries as the benign data set. The files used are specified in Table 8.2.

Table 3. Benign data set of linux binaries.

1	/usr/bin/gcc	7	/sbin/ifconfig
2	/bin/ls	8	/bin/ping
3	/bin/cat	9	/bin/cp
4	/bin/dir	10	/usr/bin/apt-get
5	/bin/mkdir	11	/usr/bin/sudo
6	/bin/mv		

8.3 Classification Tests

Using the training algorithm, it was tested many combinations of the tuple $p = (R, t_k, r_k, \gamma)$, with $t_k \leq 6$, $r_k \leq 6$, $\gamma \in \{0.9, 0.8, ..., 0.1\}$ and R with one to eight worm references. In the end the combination that showed the best training results was $b_p = (|R| = 8, t_k = 4, r_k = 3, \gamma = 0.9)$ with a success rate $r = 1.0$, an average distance $d = 0.11198$ and a threshold $b_t = 0.79064$.

Using the best parameter combination and the classification algorithm, all WORM2 generations (0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0) and benign files from /bin/ directory and /usr/bin/ directory are classified. The success rates are shown in Table 8.3.

Table 4. Similarity Results

Data set	Success rate
MWOR2/*	99.42857%
/bin/* + /usr/bin/*	96.77246%

Having 100 evolutions from each WORM2 generation, giving a total of 700 tests only 4 resulted in false negative and all were from generation 4.0, the same used in the reference set. There was still a significant amount of false positives in the benign files, probably because of the MWOR ability to insert benign code at his metamorphosis and because the variety of the benign files is larger than the worm files.

9 Conclusion

Despite the very small quantity of training samples, the results were highly positive, most of the worms were labeled as malware. The downside is the significant amount of false positives obtained in the set of Linux command files. More accurate results could have been obtained with by testing with more benign samples.

Information based metrics have applications beyond the virus detection in the realm of classification problems. For example, detect the authorship of a text. Given texts from the same author as reference, other texts written by the same author should be significantly more compressed than texts not written by him.

So we hope in the future to apply NCCD in other fields like authorship detection. In the metamorphic virus realm, future work could involve testing more parameter combinations, mainly use different reference set and better compare our results with other previous works.

Acknowledgments

This research work was supported by the projects PEst-OE/EEI/UI0127/2014 from both IEETA (Institute of Electronics and Informatics Engineering of Aveiro) and FCT (Foundation for Science and Technology).

References

1. Jared Lee, Thomas H Austin, and Mark Stamp. Compression-based analysis of metamorphic malware. *International Journal of Security and Networks*, 10(2):124–136, 2015.
2. Sudarshan Madenur Sridhara and Mark Stamp. Metamorphic worm that carries its own morphing engine. *Journal of Computer Virology and Hacking Techniques*, 9(2):49–58, 2013.
3. Donabelle Baysa, Richard M Low, and Mark Stamp. Structural entropy and metamorphic malware. *Journal of computer virology and hacking techniques*, 9(4):179–192, 2013.
4. Robert A Wagner and Michael J Fischer. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173, 1974.
5. Kirti Mathur and Saroj Hiranwal. A survey on techniques in detection and analyzing malware executables. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(4), 2013.
6. P Vinod, R Jaipur, V Laxmi, and M Gaur. Survey on malware detection methods. In *Proceedings of the 3rd Hackers' Workshop on computer and internet security (IITKHACK'09)*, pages 74–79, 2009.
7. Lance Fortnow. Kolmogorov complexity. In *Aspects of Complexity (Short Courses in Complexity from the New Zealand Mathematical Research Institute Summer 2000 Meeting, Kaikoura)*, volume 4, pages 73–86, 2001.
8. Charles H Bennett, Péter Gács, Ming Li, Paul MB Vitányi, and Wojciech H Zurek. Information distance. *IEEE Transactions on information theory*, 44(4):1407–1423, 1998.
9. Paul MB Vitányi, Frank J Balbach, Rudi L Cilibrasi, and Ming Li. Normalized information distance. In *Information theory and statistical learning*, pages 45–82. Springer, 2009.
10. Sebastian Klenk, Dennis Thom, and Gunther Heidemann. The normalized compression distance as a distance measure in entity identification. *Advances in Data Mining. Applications and Theoretical Aspects*, pages 325–337, 2009.

Outlier Identification in Multivariate Time Series: Boilers Case Study.

Joana Ribeiro[1]

[1] Universidade de Aveiro, Portugal
joanapribeiro@ua.pt

Abstract. The existence of abnormal values in data sets of day-to-day actions is common. Usually denoted as outliers, the search for these values is frequently performed to exclude them from the study. However, thinking in fraud detection or disease diagnosis, outliers can often represent a goal of study. In this work, a methodology to detect outliers in multivariate time series by translating the multitype variables into strings is presented. After that, common classification algorithms can be applied to identify outliers.

An application is performed into a real data set regarding the energy field. This data represents boiler operations and the main goal is to identify his faults. A major part of the work concerns the data set processing steps that enable the application of common machine learning algorithms. In the end, besides the boiler malfunctions, normal operation cycles were also identified. We aim that the studied methodologies improve the real-time fault identification of the operating devices allowing safer appliances.

Keywords: Outliers, Time Series, Multivariate.

1 Introduction

With the increasing capacity of technology, the most varied types of information from real-life events are nowadays available in databases. With this, a greater difficulty in find machine learning algorithms capable of dealing with new data interest aspects, such as time, also appeared. Instead of unchangeable variables considered in the most frequent case studies, such as petal length and width in the Iris flower data set, data is now seen as a vast and complex set of variable evolutions through time representative of changing behaviors. Machine learning algorithms should consider not only the variables values but also the relation between them and time.

Abnormal values are usual in real-life events. Frequently denoted as outliers, they are different from what is considered to be normal in each specific case study. In some situations, these values are not common data set noise but, instead, they represent changes or malfunctions in the systems [1] – [6]. Considering environmental changes or fraud detection, outliers can have their own interest of study. Machine learning algorithms are useful to detect them, improving human intervention, for example when predicting components faults.

The data set in study concerns boilers operations, which may not always perform their essential function of hot water supply. The goal is to identify the abnormal boiler operations through the classification of operation cycles. Since we are considering real data, normal operation is more frequent than failure. This implies a low frequency of outliers in the data set, which usually translates into a classification preference for the most frequent class since most machine learning algorithms are statistically based.

The main difficulty of the data set stands in the variety of variables from numeric to strings, along with the time factor. The proposed method is the value-trend approach: a way of representing time series by strings. This data transformation allows to apply some common machine learning algorithms and solve the time series classification problem.

In Section 2, the main characteristics of the data set are presented along with the first processing steps. Section 3 concerns the value-trend approach along with its application to the data set. In Section 4, as an example of application, one classification algorithm is constructed considering some set variations and the performance measures are investigated. Conclusions and future work are presented in Section 5.

2 The Data

As stated before, the data in study concerns several different boilers operating in customer's houses whose main function is to supply hot water. This request can be made as domestic water (hot water - HW - cycle), such as hot water for bathing or kitchen, or as central heating (central heating - CH - cycle), in which hot water closed circulation along the house allows to heat wall-hung heating devices. There is an additional mode called boost cycle, which provides a quicker supply of hot water. These cycles are considered normal operations and are the most frequent in the data.

Boilers have sensors collecting data of, for example, temperatures, flow and number of heating requests made, and state variables such as open or close valves, requests for heating, among others. Therefore, the data set is constituted by continuous and discrete variables. Note that the state variables can be transformed in discrete values such as binary. Every such variable always assumes two opposite values: On/Off, Yes/No, HW/CH, among others, and so, these variables were coded as 1 and 0, respectively. The set is then constituted by 40 discrete and 29 continuous variables, along with 4 string variables concerning the boiler model and the gateway software that will be considered for the classification task without any processing.

One last very important variable is the fault code. Boiler malfunctions are associated with faults identified by the appliance software and automatically tagged with a fault code. This software can identify 39 different fault codes. The set is constituted by labeled data: a target variable corresponding to the fault code; but also data concerning appliance malfunctions that were not identified, resulting in unlabeled observations. The goal is to apply classification algorithms able of identify these latter faults. Novelty detection [7] is one possible approach to outlier identification. It consists in obtaining a model of normal operation and attribute the outlier label to all behaviors that do not fit in such patterns, considering some threshold.

Boilers normal operation cycles are also not identified by the appliance software and consequently, they are not labeled in the set. To apply classification algorithms, we must attribute a class to all the observations of (at least) the train set. A few variables let us recognize if there has been any hot water or central heating request. Therefore, 100 observations of each HW, CH and boost cycles were manually labeled considering expected behaviors, despite there are not consensual rules to decide if a cycle is of normal operation or not.

Through data visualization of normal operation cycles and identified faults, it was possible to exclude variables with no labeling influence, such as those with no value variations from label to label, resulting in a reduction to 40 variables. Later, a statistical study of means and percentage of missing values due to new sensors not existing in the available models, allowed to exclude 9 more variables, resulting in a total of 31 variables instead of the original 74.

The data was collected within one year and four months concerning 1563 appliances of about 5 different models and stored in Mongo[1] database. Although data is received every millisecond, only the variables for which there has been any change in the value or state are recorded, and so, data matrices are full of empty entries which do not always correspond to missing data. Consider the data matrix reduced example of Table 1 concerning three variables during 60 milliseconds of boiler operation. Investigating the variable "Water Temp.", the value 20 has been recorded at 01:30:00.000. Since there were no value changes for this variable in the next milliseconds, the matrix entries are empty (lines 2,3 and 4 of the second column) although the appliance is always recording. At 01:30:00.400 a new value has been detected: 22, and so, a new entry was recorded for the “Water Temp.” variable.

Table 1. Data recording example.

Time	Water Temp.	Gas	Fan
01:30:00.000	20		
01:30:00.100		1	
01:30:00.200			1
01:30:00.300			0
01:30:00.400	22		
01:30:00.500	23	0	

To represent the realistic continuity of values in the matrices, the first approach would be to copy the last saved value until a new entry happens, for each column of the matrix, one by one. However, this leads to a fake translation of the boiler operation. There is one specific fault of data loss due to connectivity gateway problems. When the connectivity gateway fails, data is not transmitted. So, copying the data would not allow to identify this fault, since there would be no missing values. Also, a

[1] https://www.mongodb.com/

subset of distinct faults is due to different sensors that stop working. For example, if some temperature sensor breaks, the temperature is not measured until the sensor is fixed. By copying the last saved value until a new entry happens, it would give the idea of a stable temperature over time. However, the temperature could have change due to some heat request. With this data copying method, the sensor fault would be impossible to identify, since no missing data would exist.

Through data analysis it was possible to conclude that state variables always register a new entry around every 4 minutes, even if there has been no state change. So, the continuity of values was accomplished by verifying if, for any predictive variable, the last saved value was within less than 4 minutes. If that happens, then the last value of each column is copied until a new value appears. This will allow the identification of the data loss fault, since it happens for more than 4 minutes. However, in the case of sensor faults it was not possible to avoid the impossibility of identification. There is an exception that happens when, for example in the temperature sensor fault, the variable concerning that temperature do not have saved values in the matrix time range. Therefore, there would not exist values to copy through that column. So, in few cases, sensor faults can be identified by data loss (empty columns in the matrix), together with other data patterns.

3 Value-Trend Approach

The main difficulty of the proposed problem is in the several specificities of the multivariate time series that common machine learning algorithms have difficulty to deal with. Some variables are represented in the format of strings, like the boiler name, and others as numerical, such as temperatures. Also, there is a big importance in capture each variable behavior and relate all the variables between them and time. Several algorithms have been proposed to deal with time series data [8] – [15], but we could not find any able to solve our specific problem due to the high specificity of the data.

The algorithm studied in this section is responsible for perform a representation of the information present in a time series in a simpler and unified form, enabling the implementation of common machine learning algorithms. It was devised by Eamonn Keogh and Jessica Lin in 2002 [17], consisting of two main steps: the transformation of the time series into sets of vectors through the piecewise aggregate approximation algorithm (PAA) and the conversion of those vectors into a set of letters by the symbolic aggregate approximation algorithm (SAX). An additional phase where a trend analysis is performed as an improvement of the SAX algorithm is also discussed. This data transformation is made considering each predictive variable individually.

3.1 Piecewise Aggregate Approximation

The PAA algorithm is used to divide the time series into a vector of equally sized segments [18], [19]. By definition, this algorithm transforms any time series X of length m into n segments of time, resulting in a vector $X = (x_1, x_2 \ldots, x_n)$ of temporal segments, where n is any arbitrary integer such that $n \leq m$. Then, considering each of the segments, the average value of each variable is calculated: for each temporal segment i, with $i\ = \{1,2,\ldots,n\}$, and each predictive variable j, the mean value w_{ji} is calculated.

The number of segments to be considered can vary between only one segment, and in this case the entire time series is equal to as many univariate vectors as the number of variables under study, up to the number of value registrations considered for the time series, taking each vector the same dimension as the time series.

In the specific case of the data set under study, after applying the PAA algorithm, we will have the same number of vectors as the number of predictive variables. Each vector will have dimension equal to the number of segments considered to represent each of the time series. Thus, each time series is represented by a set of multidimensional vectors, all with the same dimension. Note that we have been considering the partition of the time series into n segments and, so, we already have a dimension reduction.

Consider the time series present in the left of Fig. 1, represented as a matrix of m lines and k columns, corresponding to m time registrations of k variables. After applying the PAA algorithm, we obtain the matrix of the right side of the same Figure. Each w_{ji} represents the mean value of the $j-th$ variable in the $i-th$ segment, with $j\ = \{1,2,\ldots,k\}$ and $i\ = \{1,2,\ldots,n\}$. Note that we have been considering the partition of the time series into n segments and, so, we already have a dimension reduction.

$$\mathbf{TS} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mk} \end{bmatrix} \Rightarrow \mathbf{PAA} = \begin{bmatrix} < w_{11}, w_{12}, \ldots, w_{1n} > \\ < w_{21}, w_{22}, \ldots, w_{2n} > \\ \cdots \\ < w_{k1}, w_{k2}, \ldots, w_{kn} > \end{bmatrix}^{\top}$$

Fig. 1. Application of the PAA algorithm in matrix format.

3.2 Symbolic Aggregate Approximation

The SAX algorithm consists in assigning a letter to each mean value previously obtained with the PAA algorithm [16], [17].

An extensive and rigorous analysis performed in [20] has shown that time series data, after being normalized by Z-score, usually follows a Gaussian distribution. This detail enables a conscious attribution of letters to the mean values since allows the partition of the Gaussian probability density function into equally spaced breakpoints. This partition must be performed according to the number of letters that we want to associate with the mean values. Relating a letter to each area obtained by partitioning the probability density function, a letter is assigned to each mean value.

Fig. 2 shows a time series with one normalized variable, where six segments were considered for the PAA algorithm. Then, four breakpoints were defined in the Gaussian probability density function, resulting in the partition of possible mean values into five intervals. Subsequently, a letter of the alphabet was associated with each mean value. For example, the mean value in the second segment is above the first breakpoint and below the second one defined in the Gaussian function. So, this segment has been associated with the letter 'b'. Moreover, since the variable mean value in the fourth segment is above the last breakpoint of the probability density function, this segment was assigned with the letter 'e'. Proceeding successively in this way, the time series represented as a vector of mean values, from the PAA algorithm, is now transformed into a string. In this case, the resulting string representative of the time series is 'cbcedd'.

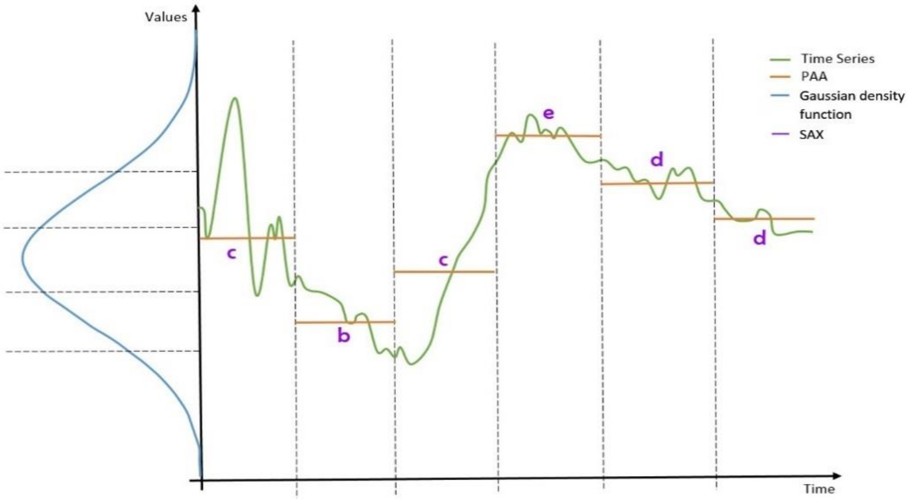


Fig. 2. Application of SAX algorithm in a univariate time series.

3.3 Trend Analysis

A study carried out in [21] suggests that an additional phase should be performed after the application of the SAX algorithm. The idea is to associate to each temporal segment not only the letter representative of the mean value but also its tendency of growing, decreasing or stability. In fact, we can easily obtain two different time series with the same SAX representation but, when associating its trends, becoming distinct strings. So, the objective is to capture the trend in each temporal segment constructed in the PAA algorithm.

The attribution of the trend information is made through its association to a straight line. The search for the line that best fits each variable behavior is performed using the least squares method, that is, we intend to obtain the model $y = ax + b$ such that $\sum_{t-1}^{s}\left(y_t - (ax_t + b)\right)^2$ is minimized, where x denotes the time variation, with $t = 1, \ldots, s$ considering the s existing values registrations of the variable in the segment under transformation, and y represents the value of the variable in each time point t. Note that this step is performed for each segment, similarly to the steps taken in the SAX algorithm. Then, after the line is obtained, its slope is used to assign a second letter to the variable under study in this time segment: 'G' for growth, 'D' for decay or 'S' for stable.

Let us illustrate the process with the following example of a univariate time series (see Fig. 3). After the application of the SAX algorithm with seven temporal segments and seven letters to represent the mean values, we obtain the sequence 'cegedcc', which is represented in purple color. Then, a straight line, represented by the green color, was adjusted to the variable in each temporal segment. According to the line slope, a second letter was attributed to each temporal segment. This last attribution is represented with a gray color. In the end, the complete time series is represented by the string 'cDeGgDeDdGcDcS'.

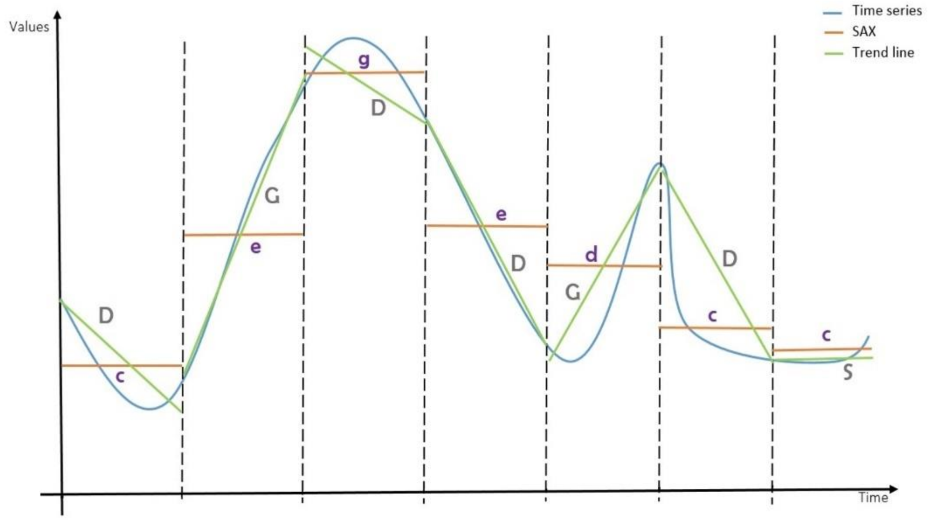


Fig. 3. Value-Trend approach applied to a univariate time series.

In conclusion, Fig. 4 shows the main processing steps starting from a multivariate time series with k variables and m time registrations. Considering the time series partition into n segments, the final representation of the matrix is a set of k strings with $2n$ length each.

$$\mathbf{TS} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mk} \end{bmatrix} \Rightarrow \mathbf{PAA} = \begin{bmatrix} < w_{11}, w_{12}, \dots, w_{1n} > \\ < w_{21}, w_{22}, \dots, w_{2n} > \\ \dots \\ < w_{k1}, w_{k2}, \dots, w_{kn} > \end{bmatrix}^{\top}$$

$$\Downarrow$$

$$\mathbf{SAX} = \begin{bmatrix} < s_{11}, s_{12}, \dots, s_{1n} > \\ < s_{21}, s_{22}, \dots, s_{2n} > \\ \dots \\ < s_{k1}, s_{k2}, \dots, s_{kn} > \end{bmatrix}^{\top} \wedge \mathbf{Trend} = \begin{bmatrix} < t_{11}, t_{12}, \dots, t_{1n} > \\ < t_{21}, t_{22}, \dots, t_{2n} > \\ \dots \\ < t_{k1}, t_{k2}, \dots, t_{kn} > \end{bmatrix}^{\top}$$

$$\Downarrow$$

$$\mathbf{VTA} = \begin{bmatrix} s_{11}t_{11}s_{12}t_{12}\dots s_{1n}t_{1n} \\ s_{21}t_{21}s_{22}t_{22}\dots s_{2n}t_{2n} \\ \dots \\ s_{k1}t_{k1}s_{k2}t_{k2}\dots s_{kn}t_{kn} \end{bmatrix}$$

Fig. 4: Processing steps of the value-trend approach.

4 Implementation and Evaluation

After processing the data, we apply the value-trend approach followed by the classification methods. All the previous and next steps were performed using Matlab software. First, the value-trend approach was implemented. Then, as an example of application, some decision tree models were constructed and evaluated taking some variations of the data set into consideration.

Main SAX code for Matlab is available for download[2]. The output of the SAX algorithm for each predictive variable is a string representing the mean value of the variable in each segment. The length of this string is equal to the number of considered segments. Some modifications were made to the original code so we can also obtain the trend approximation, which is made by fitting the variable behavior in each segment into one straight line attending the minimization of the mean square error. The segments are the same that were considered in the SAX algorithm. For this purpose, "polyfit" Matlab function was used to obtain the slope of the wanted straight line. The modified algorithm output is the string from the value-trend representation of each predictive variable.

The choice of the number of segments was based on each time series length. For data relative to more than 10 minutes, the time series is divided into segments of 2 minutes. For time series with length between 2 and 10 minutes, segments of 30 seconds have been formed. In the case of time series with less than 2 minutes, the number of segments is equal to 20 times the time series length in minutes.

The appliances software can identify 39 different faults, each fault related with a specific boiler component failure and possible maintenance solutions documented. Thus, a total of 42 classes were considered for the time series classification problem, corresponding to the 3 normal operation cycles and the 39 existing fault codes.

When trying to understand the appliances faults, we found that the causes approximate to some simple logic rules. For example, consider that there is an order to use hot water but the temperature at the output is not increasing, despite it is below the set point. This fault can be traduced into: hot water order → temperature not increasing and temperature bellow the set point ⇒ fault. So, the decision tree model was the first one making sense to be used for the classification problem as it is also based on simple logic rules. The construction of the decision tree model was made considering 10-fold cross validation and the Matlab command "fitctree". Three different splitting algorithms – "pull left by purity", "principal component-based partitioning" and "one versus all by class" - were used to construct the decision rules, all of them available in Matlab. Also, Matlab skills were used to optimize some parameters of each model, such as the minimum number of observations in each leaf node or the maximum number of splits. The choice of the best parameters values is based on the minimization of the cross validation loss.

After testing the splitting rules, "one versus all by class" was the one that minimize the cross validation loss. A new decision tree model was then constructed considering this splitting rule and its optimized parameters. In this next phase, the cross validation

[2] https://cs.gmu.edu/~jessica/sax.htm

technique is considered with 5 folds instead of 10. Some measures of performance are presented in Table 2, remarking that the obtained accuracy is considerably low (38.46%).

Table 2. Performance measures for the optimized decision tree model in the 42-class problem.

Measures	
Classified Observations	99.83%
Accuracy	38.46%

The number of labeled occurrences per fault vary between 1 to 133 in a total of 1185 classified cycles. As is possible to see in Fig. 5, this is clearly an unbalanced data set. We note that the low accuracy result is related to the multiclass problem with 42 classes. Therefore, a new approach was made: the labels were reduced to the classes that present more occurrences than the mean value of occurrences - 11 classes - plus the normal cycles - 3 classes each with 100 occurrences - in a total of 14 classes.

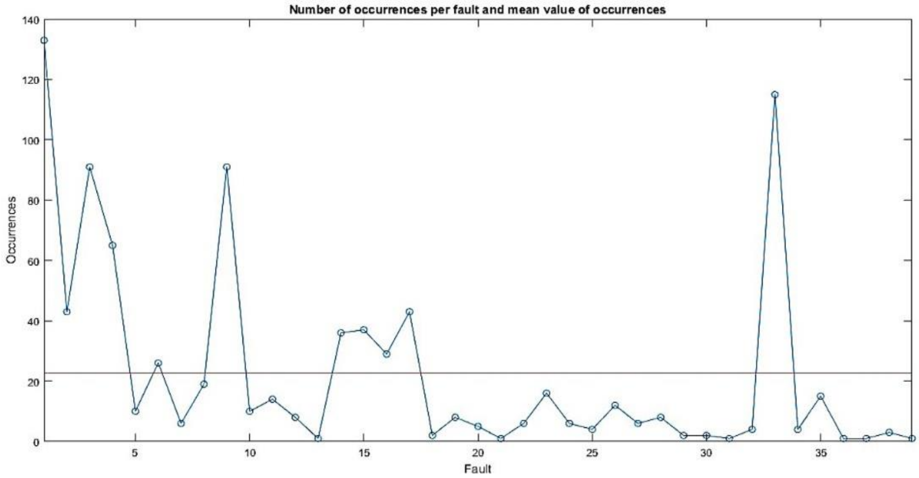


Fig. 5. Number of occurrences per fault and mean value of occurrences (horizontal line).

The decision tree models were constructed once again considering 10 folds cross validation, the previously tested splitting rules and making use of the optimization of parameters. Once again, the "one versus all by class" splitting rule is the one that minimizes the cross validation loss. A new decision tree model was constructed considering this splitting rule and its optimized parameters. The cross validation technique was performed with 5 folds and some measures were obtained (consider Table 3). However the accuracy remains low, the results were better than considering the 42-class problem. Also, in the 14-class problem, the observations were all classified, when before only 99.83% of the observations were attributed to one class.

Table 3. Performance measures for the optimized decision tree model in the 14-class problem.

Measures	
Classified Observations	100%
Accuracy	44.30%

In the following step, the classes were divided into two classes: fault or normal cycle. This allows to evaluate the appliances with more confidence, since it is not always needed to know which fault occurred but just the frequency of faults along time, so we can conclude about the appliance lifetime. Therefore, a new decision tree model was constructed and optimized, also considering the previous splitting rules and 10 fold cross validation. The "pull left by purity" splitting rule was the one that minimizes the cross validation loss. So, a new decision tree was constructed considering this splitting rule and its optimized parameters along with the technique of cross validation with 5 and 10 folds. As expected, much better results for the accuracy measure were obtained compared to the previous approaches (consider Table 4).

Table 4. Performance measures for the optimized decision tree model in the 2-class problem.

Measures	Classified Obs.	Sensitivity	Specificity	Accuracy	Precision	F1
5 folds	100%	79.00%	95.48%	91.31%	85.56%	82.15%
10 folds	100%	60.33%	99.55%	89.62%	98.22%	74.75%

The specificity was in both cases higher than the sensitivity. Since in this study the "positive" class corresponds to the normal cycles and the "negative" class to the faults, the algorithm is better at classifying faults than normal cycles. The decision tree model trained with 5 folds have classified the true class more times than the 10 folds decision tree models, according to the accuracy results, while the precision of the results were better in the 10 folds trained model - and so, this is a more consistent model. F1-measure reflects the combination of these two last metrics. Consequently, better results were obtained when considering the 5-folds cross validation technique, which is expected since this problem suffer from number of observations. 100% of the observations were classified by both models.

5 Conclusions

The main difficulty of this work was to find an algorithm able of classify time series with predictive variables in the form of strings, continuous and discrete values. The final solution was to transform the time series to allow the application of common machine learning algorithms using the value-trend approach, where each predictive variable was coded as one string. In this way, it was possible to express all the different types of predictive variables into just one (strings). Also, the behaviors previously represented in a set of matrix lines were transformed into a single string (one line matrix). Therefore, each time series became a single matrix of strings, and so, a multivariate classification problem of string type variables.

Since the behavior over time is the most important feature to the problem at hand, the trend variation was also considered to extract information from the time series. This is a novel approach to the existent SAX algorithm.

The main goal of this work was the construction of an algorithm able to identify faults, and so, the 5 folds decision tree model is the more useful model. Also, not only the faults were identified but also cycles of normal operation.

Future work includes the prediction of faults in a space time of one week before. Also, it is important to distinguish the faults of the sensors that were not possible to obtain to the continuity of values approach. The novelty detection family of methods should also be put into comparation with the presented approach.

References

1. Phuong, T. V., Hung, L. X., Cho, S. J., Lee, Y.-K., Lee, S.: An Anomaly Detection Algorithm for Detecting Attacks in Wireless Sensor Networks. In: Mehrotra S., Zeng D.D., Chen H., Thuraisingham B., Wang FY. (eds.) Intelligence and Security Informatics. ISI 2006. Lecture Notes in Computer Science, vol 3975. Springer, Berlin, Heidelberg (2006).
2. Chatzigiannakis, V., Papavassiliou, S., Grammatikou, M., Maglaris,,B.: Hierarchical Anomaly Detection in Distributed Large-Scale Sensor Networks. 11th IEEE Symposium on Computers and Communications (ISCC'06), 2006, pp. 761-767.
3. Janakiram, D., Reddy, V. A., Kumar, A. V. U. P.: Outlier Detection in Wireless Sensor Networks using Bayesian Belief Networks. 1st International Conference on Communication Systems Software & Middleware, New Delhi, 2006, pp. 1-6.
4. Subramaniam, S., Palpanas, T., Papadopoulos, D., Kalogeraki, V., Gunopulos, D.: Online outlier detection in sensor data using non-parametric models. In: Proceedings of the 32nd international conference on Very large data bases (VLDB '06), Umeshwar Dayal, Khu-Yong Whang, David Lomet, Gustavo Alonso, Guy Lohman, Martin Kersten, Sang K. Cha, and Young-Kuk Kim (Eds.). 2006, VLDB Endowment 187-198.
5. Zhang, Y., Meratnia, N., Havinga, P. J. M.: Outlier detection techniques for wireless sensor networks: A survey. Vol. 12, no. 2, pp. 159-170, 2010.
6. Fawzy, A., Mokhtar, H., Hegazy, O.: A Heuristic Approach for Sensor Network Outlier Detection. 2017.
7. Pimentel, M. A. F., Clifton, D. A., Clifton, L., Tarassenko, L.: A review of novelty detection. In: Signal Processing, Volume 99, 2014, Pages 215-249.
8. Cui, Z., Chen, W., Chen, Y.: Multi-Scale Convolutional Neural Networks for Time Series Classification. 2006.
9. Wang, L., Wang, Z., Liu, S.: An effective multivariate time series classification approach using echo state network and adaptive differential evolution algorithm. In: Expert Syst. Appl. 43, C., January 2016, pp. 237-249.
10. Kasetty, S., Stafford, C., Walker, G. P., Wang, X., Keogh, E. J.: Real-Time Classification of Streaming Sensor Data. In: Conference: 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2008), November 3-5, 2008, Dayton, Ohio, USA, Volume 1.

11. Smyl, S., Kuber, K.,: Data Preprocessing and Augmentation for Multiple Short Time Series Forecasting with Recurrent Neural Networks. In: Conference: 36th International Symposium on Forecasting, Santander, 2016.
12. Kadous, M. W., Sammut, C.: Classification of multivariate time series and structured data using constructive induction. In: Machine Learning, Feb. 2005, Volume 58, Issue 2–3, pp 179–216.
13. Górecki, T., Łuczak, M.: Multivariate time series classification with parametric derivative dynamic time warping. In: Expert Syst. Appl. 42, 5, April 2015, pp. 2305-2312.
14. Formisano, E., Martino, F., Valente, G.: Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. In: Magnetic Resonance Imaging, Volume 26, Issue 7, 2008, pp. 921-934.
15. Mohammed, K.: Learning Comprehensible Descriptions of Multivariate Time Series. 1970.
16. (Oct. 2016). Sax-vsm, [Online]. Available: https://jmotif.github.io/sax-vsm_site.
17. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: A novel symbolic representation of time series. Data Mining and Knowledge Discovery, vol. 15, pp. 107–144, 2 Oct. 2007.
18. Alsallakh, B., Bögl, M., Gschwandtner, T., Miksch, S., Esmael, B., Arnaout, A., Thonhauser, G., Zöllner, P.: A visual analytics approach to segmenting and labelling multivariate time series data. Vienna University of Technology, TDE Thonhauser Data Engineering GmbH, University of Leoben, EuroVis Workshop on Visual Analytics, 2014.
19. Wang, Q., Megalooikonomou, V.: A Dimensionality Reduction Technique for Efficient Time Series Similarity Analysis. Information Systems, 33, pp. 115–132, 2008.
20. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery (DMKD '03). ACM, New York, NY, USA, pp. 2-11, 2003.
21. Esmael, B., Arnaout, A., Fruhwirth, R. K., Thonhauser, G.: Multivariate time series classification by combining trend-based and value-based approximations. In: Proceedings of the 12th International Conference on Computational Science and Its Applications, Volume Part IV, Springer-Verlag, 2012, pp. 392–403.

Automated Fare Collection Data for Measuring Socio-economic Impacts on the Transport Supply*

Yassine Baghoussi[a†]

January 28, 2018

Abstract

Automated Fare Collection (AFC) systems are being used increasingly by the public transit agencies. Although their main purpose is to store transaction data for financial accounting processes, they also produce a large amount of data which can be useful to transit analysts as well as to researchers, from the daily operations of the transit system to the long-term evaluation of the network. This paper explores the potential of AFC data for studying and assessing the impact of a major change in transit fare structure that took place in the greater Paris metropolitan area and presents a specific method applied to measuring the socio-economic impacts of this innovative approach on the transport supply.

1 INTRODUCTION

Smart card fare collection systems are now implemented all over the world. The concept is well advanced in Europe, especially in France [6]. The smart card fare systems were introduced to facilitate the revenue collection. However, they are also designed to store large amount of data, [1]. These data have potential applications. The information collected can be used in different ways by transit planners and researchers. Various uses are possible at three levels of management, [6], namely the strategical (e.g. long-term planning), the tactical (e.g. service adjustments and network development), and the operational (e.g. ridership statistics and performance indicators). Recent studies have demonstrated how much interest exists in using smart card data for transit planning.

**This work was supported by the STIF (The transport organization authority in ile-de-France)

[†a] Faculty of Engineering, University of Porto, Porto, Portugal

Concerning the impact of fare change, from the user standpoint, a large number of studies have analyzed passengers' responses to fare change using stated-preference (SP) surveys, revealed-preference (RP) surveys, or a combination of both [5]. Passengers' responses to fare change are generally referred to as price elasticity of the demand, represented as a measure that gives the percentage change in quantity demanded in response to a one percent change in price. Price elasticity, on the other hand, is a factor in predicting transit ridership and the change in the global revenue due to fare change [4]. A literature review indicates that the price elasticity of travel demand varies greatly in terms of variables such as time span, transit mode, original fare level, income level, journey distance, data paradigm, type and direction of price change, demographic and geographic conditions [2].

The paper presents the data analysis pursued to achieve the results discussed in [12]. The objective is to describe the adopted methodology to interrogate the potential of AFC to capture the mobility behavior of individual passengers, especially so in order to assess the trend of the demand under constant supply and, distinctively, its response to a significant change in transport supply. We study a case of transit fare scheme in greater Paris, in which a number of subscriptions between concentric zones were replaced by a unique, flat fare subscription on September the 1st, 2015. Our approach combines a computer treatment with a geographical and socioeconomic analysis.

The paper is organized as follows. Section II provides a brief overview of measurement methods for passenger trips in transit networks.Section III presents the study case: the transit system in greater Paris, with its pricing system. Section IV introduces the data collection system together with our technical treatment. Section V deals with the assessment of demand trend between successive years under constant supply, whereas Section VI provides first-round results on the assessment of demand respond to the major change in the fare scheme. Lastly, section VII concludes by summing up the main findings and pointing to ongoing work.

2 AFC AND OTHER METHODS

One important premise of this work to pursue the proposed goals is that AFC data conveys relevant information about the mobility of people. Among other pieces of such information we have time and position in the gate of validation. Combining this information with the timetables of trains, we can construct a leg data table. A leg is generated systematically using two successive validations made by a user using tap-in and tap-out information at the access station and egress station respectively. Thus we can measure the quantitative and qualitative evolution of user's visits to the stations.

This is a technical-based investigation with (i) a big influence of technology challenges which highly relates to eventual maintenance of the system and (ii) the inability to satisfy the demand in some stations with high number of visitors, that is to say the information is influenced by the delay caused. Nevertheless, using AFC data represents a novel and efficient way if compared to the regular SP/RP surveys [1]. This is specially true if we consider the finance purpose as highlighted elsewhere [7], in which authors argue that online process compared to face-to-face surveys had the additional advantages of being cheap. This is a similar case compared to AFC. However, in online surveys the conveying mobility information is still not guaranteed due to the penetration rate of such an approach; it is not always guaranteed it will reach all the population. Additionally, time constraints and privacy issues are also very important concerns, needing appropriate treatment.

The data collection methods used in the submission of a survey includes several questions which can lead to the desired information through the analysis of people feedback . In their research [8], authors tried to understand how the activities of people are conducted before, during and after their journeys so as to have a perspective about it. Such a perspective involves analyzing the activity of the subject on several occasions over time which consists in submitting the same survey each time, which would demand additional funding.

As for AFC, the information given from tap-in and tap-out can be used for the same purpose which is to understand the mobility of people and will be substantiated by the subsequent details.

3 Transit system in greater Paris

Before we can further into detailing the proposed approach, it is important to understand briefly how the transportation system in greater Paris is structured, and how its Navigo system works.

3.1 Navigo system: AFC data of urban rail in the greater Paris

The two main systems of urban rail in the greater Paris metropolitan area according to [10] are: (i) the semi-closed Metro system which includes 14 lines. This system is equipped with tap-in gates at access only; (ii) the closed RER (*Réseau Express Regional*, which is the Regional Express Network) equipped with both tap-in and tap-out gates for external transfers with 5 RER lines. The AFC system of this area implemented by STIF , namely "SIDV", allows the usage

[1] Several questions submitted face-to-face or using online applications.

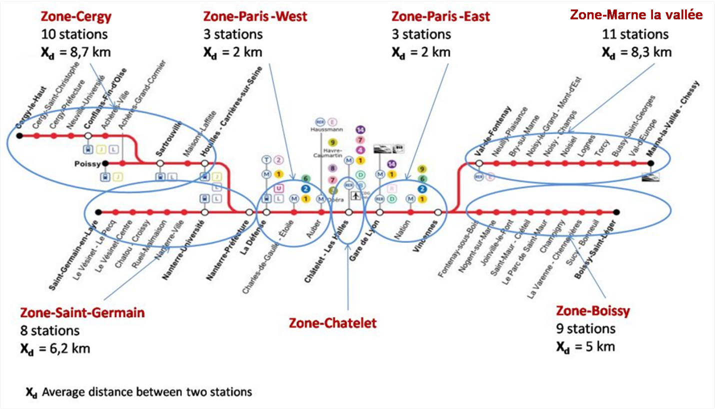


Figure 1: Line A of RER in the greater Paris, including correspondences (RER, Transilien, metro and tram).

of the smart card through the network and stores anonymous passenger information including the smartcard number which is anonymized (with anonymous number that is maintained during 3-month periods), the date, the validation instants at tap-in / tap-out gates, the gate IDs and name of access/egress stations. During the peak period on workdays, more than 90% of the trips by public transport are home-work or home-study trips using network subscription hence the smartcard. Although, the information of validation data is collected systematically, it remains incomplete due to the absence of validations in the exit, except in some cases such as the RER (for instance the RER A) and trains. Thus we focus on the RER A during this study.

3.2 Line RER A and related data set

The line A of the RER network in the greater Paris, often simply called RER A line (depicted in Figure 1 is one of the world's busiest lines, and the busiest line in Europe with around one million passengers each day. It contains 46 stations in total 109 km and is structured around a central trunk into which five branches are grafted: two eastward branches to northeast terminal Marne-la-Vallée and southeast terminal Boissy, and three westward branches to northwest terminal Cergy, central-west terminal Poissy and southwest terminal Saint-Germain, respectively. The central trunk between Vincennes and La Défense stations passes through the largest underground train hub in France, Châtelet-Les Halles in the center of the city, and serves the major business district La Défense. It serves 2 million jobs, representing 41% of regional tissue.

3.3 Navigo Pass: Re-zoning of greater Paris and the new fixed-price scale scheme

The greater Paris covers an area of 12.012 km2 and accommodates a population of 11.5 million [9] as well as 5.6 million jobs representing 29% from the GDP of France. It is characterized by a dense transit system and a great diversity of urban transit modes and services, including: 16 metro lines of over 200 km; 1.500 bus lines of over 25,000 km. The network is composed of 14 train lines of over 1.500 km; 8 tramway lines. In terms of daily mobility, it counts over 41 million daily journeys, 39% of these journeys are made by foot, 38% by car and 20% by public transport. As we move away from the city center, the private car is over-represented compare to the other modes with 2/3 of daily journeys in outer suburbs. Public transport counts 8.3 million journeys, 51% of them are made between home and work or study place [9].

Since 1975, transit pricing of Paris region was based on a system of concentric zones. The region was divided into homocentric areas. The journey price was depending on the number of zones that the passenger crossed. In September 2015, STIF introduced a flat fare system with the creation of a single price for the Navigo Pass.

4 Tap-in Tap-out data from Navigo validation system

Navigo Pass allows the access to a rail stations' gates by passing a card, which contains user's identity data, near an electronic reader tap-in/tap-out. These passes are used to access vehicles of the RATP, the SNCF (within the Transilien network), and companies under the aegis of the Syndicate of transports in ile-de-France (STIF).

In the analysis below, AFC data spanning the working days of one week (i.e. weekends excluded), from three different years, are considered: Monday-Friday from October 2013, 2014 and 2015. This data gathers validations: information recorded by AFC system aforementioned above and legs which are systematically generated from the validations data using the tap-in and tap-out details. The validation data is exploited using a specific dynamic O-D (Origin-Destination) matrix inference scheme devised by [11]. This scheme extends previous work by [3] to infer rail transit O-D matrix and to generate leg choice from Oyster smartcard data in London, based on data processing and analysis methods supported by technologies of database management systems (DBMS) and geographic information systems (GIS).Finally, the legs data table now contains the date, the validation instants at tap-in & tap-out gates, the gate IDs, the access & egress station name , duration & distance of the O-D and the carrier information : bus or rail system (Metro, RER) including the lines i.e RER A, B etc.

As for validation and leg, journey data table of all the network was included in our raw data. The method of SIDV for journey construction is to extract all O-D pairs related to each individual from leg data table. One journey contains all the information about the first and last validation of each individual after the conclusion of several legs. One individual can have multiple legs and journeys.

For the current study, a specific scheme has been pursued for the data training which involves five steps as following: (1) extracting the data related to RER A from the dataset of the entire network Table 1; (2) creation of the average day-matrix for each year data to generate average day-matrix of 2015-hat; (3) focus on the rush hour of the morning (RHM) from 4:30 to 11 am according to a survey data made in France called Enquete Globale de Transport (EGT), a questionnaire submitted every 10 years in France in which citizens such as employees and students are asked about their daily transport activities including the time of their going out for work (see Figure 2); (4) generating an O-D matrix which contains the number of legs for each O-D using the leg data table for both 2015 and 2015-hat; and finally, (5) discuss the traveled distance.

As aforementioned, the AFC data assembles anonymous passenger information. Dealing with anonymous users cannot allow the process of Panel methodology for specially, a qualitative analysis of the public transport users. The methodology of the panel can be applied with different techniques for mobility analysis such as:

- The basic technique used consists on doing a personal and well-structured interviews: face-to-face or by phone. The advantage of this methodology is to avoid the drawbacks related to each technique and provide a diachronic approach of the traveler to understand how the journey is conducted.
- More advanced technique aims to use AFC data, in case that the anonymizing process has not been applied to the passenger cards, which may assist in obtaining both qualitative and quantitative data. In our case of study, the profession was missing in the analysis for more accuracy in term of the third step in the scheme aforementioned above.By applying step (3) we try to focus on the employees and students. This step remains imprecise because RHM sample does not totally exclude other types of users (non-student and non-employee).

Table 1: NUMBER AFTER THE EXTRACTIONS

Year	Table	Monday	Tuesday	Wednesday	Thursday	Friday
2013	Validation	1273964	1317334	1238193	1353485	1304540
	Leg	940730	979194	924376	1001201	954489
	Journey	940534	978993	924199	1000986	954322
2014	Validation	1261058	1303917	1251918	1270060	1273338
	Leg	929489	962662	923193	927330	924776
	Journey	929449	962656	923186	927326	924774
2015	Validation	1353514	1391997	1374069	1413155	1205928
	Leg	1013397	1057082	1031331	1061110	925967
	Journey	1013297	1057082	1031331	1061109	925966

5 AFC to estimate trends of traffic state under constant supply

The extraction of data related to RER A from the dataset of the entire network was based on the validation information i.e. the data that belong to any validation in RER A, this was done using carrier information and station name as a conditional variables. Then, from each validation we take two values to produce RER A legs extraction: User ID and leg ID which is a foreign key that relates each leg with its two validations, thus these two validations have the same user ID and leg ID .The total number of validations is high compared to the number of legs. Each journey is also matched with all the legs that are related to it using journey ID that exists in both leg and journey of the same user ID. As results, we have RER A data tables of validations, legs and journeys.

The generation of the average day-data was done using several matrices. First, a matrix O-D for each day of the week starting from Wednesday and ending to Friday excluding Monday and Tuesday. The reason of this exclusion is due to habitually problems in the urban rail transit of greater Paris. The problem was detected after analyzing all the days-matrix, a high decrease was observed in the number of legs with origin/destination from several stations of the central trunk of RER A in 2015. The average day-matrix is the mean between all the days-matrix. After computing the average day-matrix for both year 2013 and 2014, the evolution shows a 0.55% decrease in the total number of legs between the two years respectively. From 2013 to 2014, the evolution 2 shows a high decrease of the legs number related to station Charles de Gaulle Etoile,Châtelet-Les Halles and Gare de lyon with 0.20% of decrease in all central zone. In the westward, we observe a decrease of 0.36% while we note 0.01% increase in the eastward.

According to the evolution between 2013 and 2014, we generated a new data denoted by 2015-hat. For instance, if the number of legs between station A and

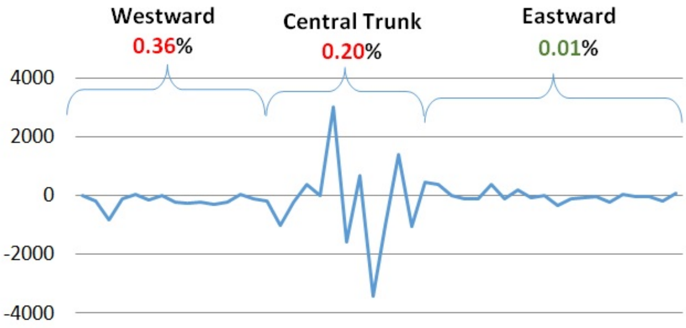


Figure 2: Evolution of the total number of legs between 2013 and 2014 For each stations

Table 2: THE NUMBER OF LEGS IN THE AVERAGE DAY

	2013	2014	2015-hat	2015
Legs	957013	928429	919762	998788

station B has increased from 2013 to 2014 with x%, this means that the number of legs between A and B will increase with x% between 2014 and 2015-hat.To generate the leg data table of an average day 2015-hat, we subtract the average day-matrix 2014 denoted by X from average day-matrix 2013 denoted by Y to compute X-Y denoted by Z. The matrix Z is used as input by an algorithm-based method, this method consists on testing the value of each O-D in the matrix Z and adding a specific legs from Thursday's leg data table if the value is positive ; or deleting a specific legs from the same data table if the value is negative i.e. let's suppose that the number of legs of Chatelet-Cergy as an O-D is equal to p^+ in the average day-matrix and the number of legs of the same O-D is equal to p^- in Thursday-matrix. Thursday 2014 was chosen as reference day-data to generate the 2015-hat in the following function : sequence(p^+-p^-, length(legs as Chatelet-Cergy))

6 AFC to estimate demand response to supply change

The average day-data contains the legs from the whole day Table 2. We determined a strategy to focus only in the rush hour of the morning Table 3. This decision aims to select only the legs related to the users that use the public transport to go for work or study. We assume that the majority of those selected users are actually citizens of the greater Paris.

```
Result: Average day-data of 2015-hat is the new T
T = Thursday-data 2014;
function SEQUENCE(x,y)
    $i = |y/x|$;
        if $x > 0$ then
        while $i \leq length(T)$ do
            p = select $i^{th}$ leg from T;
            i = (i + $|y/x|$) + 1;
            Insert p into T;
        end
        else
        while $i \leq length(T)$ do
            p = select $i^{th}$ leg from T;
            i = (i + $|y/x|$) + 1;
            Delete p from T;
        end
        end
end function
```

Algorithm 1: BODY OF THE FUNCTION SEQUENCE

Table 3: Number of legs in RHM

	2015-hat	2015
Average day	919762	998788
RHM	376890	403428

The data included some incoherent legs and thus some data cleaning is required. An incoherent leg contains an impossible O-D using only one mode of transport which means that the user has missed a validation somewhere in the exit of the first mode.

Let us suppose the following two successive validations of a specific user: the tap-in was made at (RER B) Antony and the tap-out at (RER A) Auber. However, going from Antony to Auber we normally need to make a transit in Chatelet Les halles.

The correct form of this leg is retraced as follows:

- Leg 1: (RER B) Antony to (RER B) Chatelet Les halles
- Leg 2: (RER A) Chatelet Les halles to (RER A) Auber

The reconstitution process has been dedicated to all lines RER B, RER D and RER E which have Chatelet Les halles, Gare de Lyon and Val de Fontenay as

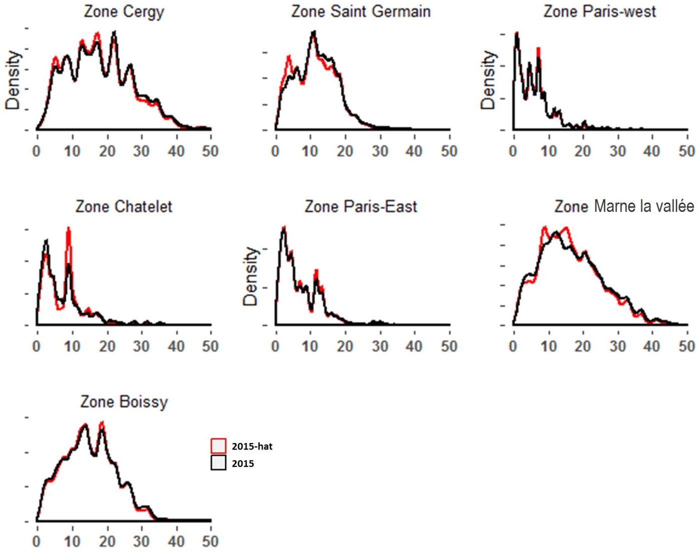


Figure 3: The effects on the distances involved on the RER A

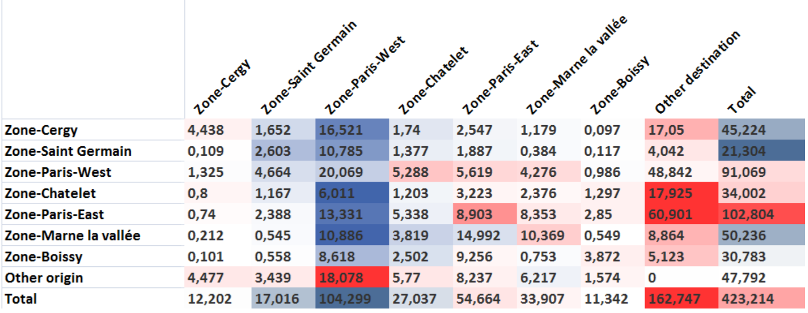
common stations with RER A respectively. The data contains some incoherent legs, including Metro and other modes which were not taken into account in the process of the reconstitution. The goal behind this correction/reconstitution is to give much efficacy to our results by recovering lost legs and it is also considered as a big step of data cleaning. Table 4 represents the number of reconstituted legs in each average day.

According to 2014 (the last year of the concentric fare transit system) and 2015 data, we note that the number of journeys between the center of Paris and suburbs are the one that increased the most by 2.8%.

Figure 4 represents a matrix of O-D between the RER A defined zones, with the indication of the decrease and the increase of the flux between each zone in the 2015 and 2015-hat , the increase goes to red and the decrease goes to blue.The values corresponds to the total number of legs in 2015 divided by 1000. The first effect noted is the increase of O-D in the eastward region, particularly northeast branch Marne-la-Vallée.

Table 4: THE NUMBER OF RECONSTITUTED LEGS

	RER B	RER D	RER E
2015-hat	116536 (12,7%)	97048 (10,6%)	39505 (4,3%)
2015	129703 (13%)	91384 (9,1%)	50272 (5%)



	Zone-Cergy	Zone-Saint Germain	Zone-Paris-West	Zone-Chatelet	Zone-Paris-East	Zone-Marne la vallée	Zone-Boissy	Other destination	Total
Zone-Cergy	4,438	1,652	16,521	1,74	2,547	1,179	0,097	17,05	45,224
Zone-Saint Germain	0,109	2,603	10,785	1,377	1,887	0,384	0,117	4,042	21,304
Zone-Paris-West	1,325	4,664	20,069	5,288	5,619	4,276	0,986	48,842	91,069
Zone-Chatelet	0,8	1,167	6,011	1,203	3,223	2,376	1,297	17,925	34,002
Zone-Paris-East	0,74	2,388	13,331	5,338	8,903	8,353	2,85	60,901	102,804
Zone-Marne la vallée	0,212	0,545	10,886	3,819	14,992	10,369	0,549	8,864	50,236
Zone-Boissy	0,101	0,558	8,618	2,502	9,256	0,753	3,872	5,123	30,783
Other origin	4,477	3,439	18,078	5,77	8,237	6,217	1,574	0	47,792
Total	12,202	17,016	104,299	27,037	54,664	33,907	11,342	162,747	423,214

Figure 4: Matrix RER A (2015) of the increase and the decrease of the flux between each zone in the 2015 and 2015-hat. The increase goes to red and the decrease goes to blue.

The graphs in Figure 3 represent the density of distances .The long distances (more than 20km) are the distances that have increased the most in the northern branches of the RER A Marne-la-Vallée and Cergy. While in the other side, the short distances (less than 10km) have decreased. In addition to this , the mean duration spent in the RER A has increased by two minutes in the suburbs and remained constant in Paris central area (Paris-west , Chatelet and Paris East).

The weighted distance (distance between stations multiplied by the number of O-D of each zone) is an indicator through which we note that the users of Marne-la-Vallée branch make the longest traveled distance.

After the flat fare application, the results show that the number of legs with long distances in the RER A increased. The same trend is observed in the evolution of the duration, specially for users who spend more than 45 minutes in RER A. According to the O-D matrix of the RER A stations, we can see that the number of legs increased in the east side of Paris central area as well as in the eastward and northwest zone (Cergy). A decrease in the number of legs is observed on the west side of Paris central area and some stations in the southwest. This study was made few months after the fare reform; the users behavior vis-a-vis public transport takes time to change.

The frequency of people entering RER A stations (origin) shows an increase of 7.54% in Paris central area in 2015, which is the opposite of the eastward and the westward which has a decrease of 2.88% and 5.64% respectively. The frequency of leaving stations (destination) shows an increase of 11.49% and 0.48% in Paris central area while in the westward the frequency decreased by 1.00%.

7 CONCLUSION AND FUTURE WORK

Wrapping up, according to the present results we note an increase of 8.6% in the total number of legs between 2015 and 2015-hat. As well the number of cards in circulation has increased by 9.6%. The long distances (more than 20km) and the number of users entering to the RER A stations has increased in the westward as well as in the eastward, specially the northeast zone of Marne-la-vallée.

As for the next steps in this research, we aim at the socio-economic assessment of the flat fare application. Therefore, future studies will also target the users who benefit the most of the flat fare reform, taking into account their social profile according to their income. In order to achieve this purpose, the further methodological approach is to combine AFC data with income data. This will allow us to combine the mobility data of users, as entering into station (inflows), with the income of the origin commune, as well as combine the mobility data of the users leaving the station (outflows) with the income of the destination commune. Additionally, the methodology herein proposed will be applied to a larger dataset so as to strengthen results and allow us make more robust analysis.

ACKNOWLEDGMENT

The author gratefully acknowledges and thanks the STIF chair for their support in carrying out this work.

References

[1] Bagchi, M. and White, P.R., 2005, “The potential of public transport smart card data”, Transport Policy 12(2005) 464-474.

[2] Farber S. et al., 2014, "Assesing social equity in distance based transit fares using a model of travel behavior", Transport Research Part A 67 (2014), 291 - 303.

[3] Zhao, J, Frumin, M, Wilson, N, Zhao, Z. 2013. Unified estimator for excess journey time under heterogeneous passenger incidence behavior using smartcard data. Transp. Res. Part C Emerg. Technol. 34: 70–88.

[4] Pelletier, M., et al., 2010, “smart card data use in public transit: A literature review”, Transportation Research part C 19 (2011) 557-568

[5] Sharaby, Y., et al., 2012, “The impact of fare integration on travel behaviour and transit ridership”, Transport Policy 21(2012) 63-70.

[6] Marie-Pier Pelletier , Martin Trépanier, and Catherine Morency , 2009, "Smart Card Data in Public Transit Planning: A Review ", CIRRELT-2009-46 5-19.

[7] WBCSD , 2016, "Methodology and indicator calculation method for sustainable urban mobility".

[8] N. Lecomte, R. Patesson, P. Steinberg, "Le panel des voyageurs".

[9] EGT, 2012 "La mobilité en ile-de-France".

[10] Aguiléra, V, Allio, S, Benezech, V, Combes, F, Milion, C. 2014. Using cell phone data to measure quality of service and passenger flows of Paris transit system. Transp. Res. Part C Emerg. Technol. 43: 198–211.

[11] Xie, X, Leurent, F, Aguiléra, V. 2015. Commuter's individual travel time estimation: using AFC data of Paris transit system. COST Action TU1004 Model. Public Transp. Passeng. Flows Era Intell. Transp. Syst. Paris Fr. May 11th-12th 2015.

[12] Zoi Christoforou and Lydia Ladjouze and Fabien Leurent, 2016, Assessing Transit Pricing Policies by Combining Large Scale Smart-card Data and Surveys,ETC Conference

SESSION 3

Machine Learning

Ensemble Weighting for Quantile Regression
Marisa Reis

A Study of Novelty Detection in Data Streams Using Different Unsupervised Approaches
Kemilly Dearo Garcia

Predicting winning teams for regular season of the NBA
João Figueira Silva, Jorge Miguel Silva, Eduardo Pinho and Carlos Costa

Ensemble Weighting for Quantile Regression

Marisa Reis

University of Porto, Porto, Portugal
INESC TEC, Porto, Portugal

Abstract. In this paper, we propose an ensemble method to effectively combine multiple quantile estimators, called Ensemble Weighting (EW). The EW consists of labelling each training instance with the algorithm which offered the best performance. It builds a classification task on those labels. In the prediction phase, knowledge of the location of a new sample is used to decide which algorithm should provide or contribute to the prediction. We demonstrate the effectiveness of the EW through experimental validation on 8 regression dataset. Results show that the proposed combination model is capable of improving the overall performance.

1 Introduction

Conditional quantile estimation has been one of the most important statistical methods and has widespread applications [1]. In statistics, regression is typically concerned with finding a real-valued function m such that its values $m(x)$ correspond to the conditional mean of y. Formally, let Y be a real value random variable and X be a set of explanatory variables. Assuming that the true underling relationship between Y and X can be modelled by:

$$Y_i = m(X_i) + \varepsilon_i \tag{1}$$

where ε_i are independent and identically distributed (i.i.d) from a distribution with mean zero and are independent of the predictors. This constitutes the standard regression setting. Based on this formulation, we consider a supervised quantile regression setting, so the conditional quantile of Y given $X = x$ as the form

$$q_{\tau,t} = m_\tau(x_t) + \varepsilon_{\tau,t}, \qquad F_{\varepsilon_{\tau,t}}(0) = \tau \tag{2}$$

where $F_{\varepsilon_{\tau,t}}$ is the cumulative distribution function of the errors, and where the functions $m(.)$ are distinct for each quantile.

The idea of quantile estimation arise from the observation that the median (i.e. $q_{0.5}(Y|X)$) minimizes the expected absolute loss - $L(y, \hat{q}_{0.5}) = |y - \hat{q}_{0.5}|$. Generally, the τ-quantile can be estimated using the pinball loss function, defined as

$$L_\tau(y, q_\tau) = \begin{cases} \tau(y - q_\tau) & \text{if} \quad (y - q_\tau) \geq 0 \\ (\tau - 1)(y - q_\tau) & \text{if} \quad (y - q_\tau) < 0 \end{cases} \tag{3}$$

where y is the observed output. More formally, the conditional quantile q_τ is the solution of the minimization problem:

$$q_\tau(Y|\mathbf{X}) \in \underset{m}{\operatorname{argmin}}\, \mathbb{E}\left[L_\tau(Y, m(\mathbf{X}))|\, \mathbf{X}\right]. \tag{4}$$

In [2], Koenker and Basset introduced regression quantile estimation by minimizing 3 with model form $m(x) = x'\beta$. It assumes that $\{(\mathbf{X}_t, Y_t)\}_{t=1,\cdots,n}$ are i.i.d such that $Y_t = \mathbf{X}_t\boldsymbol{\beta} + \varepsilon_t$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown parameters. This method is commonly known as **l**inear **q**uantile **r**egression (LQR).

As in standard regression estimation, the subject of model selection and combination has been studied in considering quantile regression [3]. Suppose we have a pool of J candidates estimators of the conditional quantile function $q_\tau(x)$, denoted $\{\hat{q}_{\tau,j}\}_{j=1}^{J}$. Our goal is to combine these estimators for an optimal performance. Since the best candidate often depends on τ the optimal search is performed for each level of interest τ.

The main rational on those methods is that different candidate methods typically have distinct relative performances. This is due to the estimator characteristics and dependence on the quantile value τ. Combining the advantages of such of such candidates aiming at a global improvement is a desirable task.

Our work integrates quantile regression and model combination. Here we propose an ensemble to effectively combine multiple quantile estimators, called Ensemble Weighting (EW). The EW consists of labelling each training instance with the algorithm which offered the best performance. It builds a classification task on those labels, being the number of classes the size of the ensemble pool. Prediction for a test instance x_i is performed by dynamically weighting the predictions of the baseline estimators $\hat{y}_{i1}, ..., \hat{y}_{iJ}$ based on the predicted class for test instance x_i.

The remaining of the paper is organized as follows. Section 2 provides a brief overview of related work. In Section 3, we introduce the proposed method, namely Ensemble Weighting for Quantile Regression (EW). Section 4 reports the experimental results on 8 benchmark datasets for regression, including comparison with a state-of-the-art method. Finally, conclusions are presented in Section 5.

2 Related Work

Ensemble or model combination typically refers to methods that generate and combine several models, either in classification or regression problems. An ensemble learning can be divided into two phases: the generation phase and the integration phase [4,5] . The first phase tackles the generation of appropriate baseline models. The integration phase involves the aggregation of the prediction baseline results. Additionally, integration can be performed statically or dynamically. The optimal selection solution found for the validation set is fixed and used as a single weight of the baseline model (static) or the decision differs for within the test instances (dynamic) [6].

In [7] an Adaptive Quantile Regression by Mixing (AQRM) is proposed. To determine the optimal combination weights, the authors consider a weighting strategy based on the negative exponential of asymmetric pinball loss obtained from the in sample dataset. This approach fits the static integration, and we used it for comparison purposes, as it handles quantile ensemble.

For dynamic selection, a method of partitioning the input samples is required [6]. In [8], the feature space regions where each classifier has best classification performance are found. A clustering step splits the correct and incorrect classified training samples from each classifier. In the selection step, the most accurate classifier in the vicinity of the input sample is nominated to provide the final decision of the committee.

In [9], Woods et al. proposed to select the single classifier that shows the best performance in the closest neighbourhood defined by an arbitrarily set number of neighbouring train samples.

Very recently, a locally linear ensemble for regression (LLER) is proposed in [10]. LLER decomposes the data into several locally regions and builds a linear model for each of the regions. It aims at dealing with the non-linear structure of the data by decomposing the feature space into a number of locally linear regions.

3 Proposed method

Let $F = \{f_1, f_2, ..., f_J\}$ be a set of available regressors. The n-dimensional feature space and the training data are denoted as $\mathbb{R}^n$ and X, respectively. Let n_0 be an integer such that $1 \leq n_0 \leq n$. The Ensemble Weighting (EW) algorithm for combining different learning algorithm is as follows.

1. Randomly partition the data $\{X, y\}$ into K-folds.
2. Consider a single fold $Z^{(2)} = \{y_l, x_l\}_{l=n_0}^{n}$ for evaluation, and the remaining $(k-1)$ folds $Z^{(1)} = \{y_l, x_l\}_{l=1}^{n_0}$ for training.
3. Based on $Z^{(1)}$, train all available regressors f_j for each $j = 1, ..., J$.
4. Obtain the predicted values $\hat{\mathbf{y}}_i = [\hat{y}_{i1}, ..., \hat{y}_{iJ}]$, for each of the fitted algorithms using the remaining data $Z^{(2)}$.
5. Based on the predictions $\hat{\mathbf{y}}_i$, compute the new target feature, as follows:

   $$c_i = \underset{l}{\operatorname{argmin}}(L(y_i, \hat{y}_{i1}), \ldots, L(y_i, \hat{y}_{iJ})) \tag{5}$$

   where c_i reflects which algorithm performed better for a given scenario x_i.
6. Repeat Steps 2-5 for the remaining folds, so that every observation $x \in X$ has a classification $c \in \{1, ..., L\}$.
7. Fit a classifier C with target c_i and input features X.

The final estimator is given

$$\hat{y}_i\left(x\right) = \hat{\mathbf{y}}_i \boldsymbol{\omega}_i\left(x_i\right) \tag{6}$$

where $\boldsymbol{\omega}_i = [\omega_{i1}, ..., \omega_{iL}]^T$ represent a weighting vector where each ω_{ij} indicates the weighting of the jth regressor in the decision fusion. $\omega_{ij} = 0$ indicates exclusion of the jth regressor in the ensemble, where $\omega_{ij} = 1$ indicates inclusion. The quantile level τ is omitted for simplicity of the notation. A classification method exists for each of the quantile levels $c_i^\tau := c_i$. The pinball loss function defined Eq.3 is used.

Steps $1 - 7$ comprise the training part of the ensemble algorithm. In the operational phase, an unknown sample, x, is classified based on the region in the feature space, such that the best algorithm is chosen to predict (see Figure 1). So that, knowledge of the location of a new sample in feature space can be used to localize the domain of the system analysis within each algorithm can be applied again.

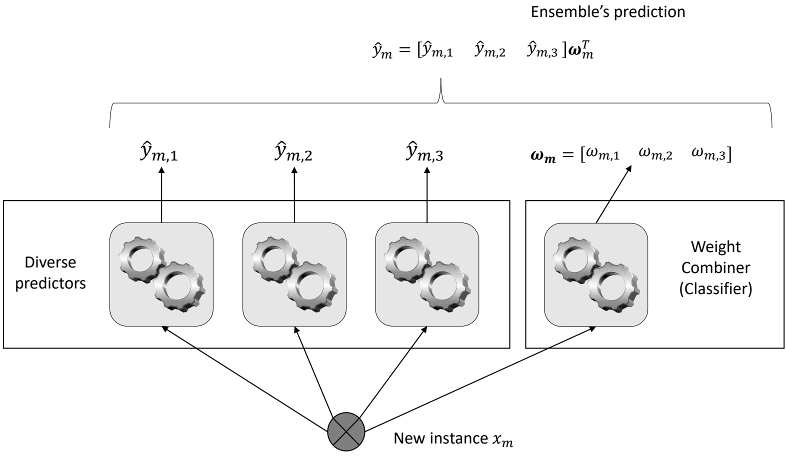

Ensemble's prediction
$\hat{y}_m = [\hat{y}_{m,1} \quad \hat{y}_{m,2} \quad \hat{y}_{m,3}]\boldsymbol{\omega}_m^T$
$\hat{y}_{m,1}$
$\hat{y}_{m,2}$
$\hat{y}_{m,3}$
$\boldsymbol{\omega_m} = [\omega_{m,1} \quad \omega_{m,2} \quad \omega_{m,3}]$
Diverse predictors
Weight Combiner (Classifier)
New instance x_m

Fig. 1. Ensemble weighting prediction scheme.

We examine two branches of the EW algorithm regarding the weighing strategy: (i) EW_{Hard}, a single algorithm (class) is considered with the highest probability; and (ii) EW_{Soft}, where the probability is served as weights in the ensemble. Additionally, we consider two classifiers for step 7 of the algorithm: (i) DT - Decision Trees, and (ii) KNN - K-Nearest Neighbors.

4 Experiments

In this section we describe the experiments performed to access the effectiveness of the proposed method.

4.1 Experimental Setup

We consider three baseline quantile regression estimators: (i) Linear Quantile Regression [2]; (ii) gradient boosting trees [11]; and (iii) kernel density estimate method [12]. Additionally, we considered the state-of-the art Adaptive Quantile Regression by Mixing (AQRM) [7] for comparison.

In our experiment, we compare these four algorithms with the proposed solution on 8 standard regression data sets. We used the 8 input Kin dataset (http://www.cs.utoronto.ca/~delve/data/kin/desc.html), generated from a realistic simulation of the forward kinematics of an 8 link all-revolute robot arm. Combinations of the following attributes are considered in datasets: (i) output: highly nonlinear (n) vs. fairly linear (f); and (ii) predicted value : medium noise (m) vs. high noise(h). Resulting in the kin8fh, kin8fm, kin8nh and kin8nm datasets.

The remaining 4 datasets can also be found on-line in https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html for data sets housing, mg and cpusmall and in http://archive.ics.uci.edu/ml/datasets.html for airfoil data set. Table 1 details for each data set the number of features and attributes.

Table 1. The 9 data sets for regression used in our experiment.

dataset	# attributes	# features
airfoil	1503	6
housing	506	13
cpusmall	8192	12
mg	1385	6
kin8fn	8192	8
kin8fm	8192	8
kin8nh	8192	8
kin8nm	8192	8

For each dataset, all input variables were normalized to fit the range $[-1, 1]$, and target values to lie between $[0, 1]$.

The forecasting skill was evaluated by splitting a given dataset into an in-sample period, used for the initial parameter estimation and model selection, and an out-of-sample period, used to evaluate forecasting performance. For all datasets we consider a portion of 80% in-sample and 20% for out-sample. The hyperparameters optimization was performed using Bayesian optimization [13] approach with 3-fold as cross validation on the in-sample period.

Point predictions, i.e. .5% quantile in this work, were evaluated with the classical Root Mean Square Error (RMSE), which is calculated as

$$RMSE = \left(\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 / N \right)^{1/2} \tag{7}$$

, with N as the number of test samples, and y_i and $\hat{y}_i$ the observed and predicted values of the target variable, respectively.

Quantile predictions were evaluated using a calibration diagrams (also called reliability diagrams). The difference between empirical and nominal probabilities is the bias b of the quantile forecasts and are usually computed for each quantile nominal proportion τ as follows:

$$b_{t+k}^{(\tau)} = \tau - \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{y_i < \hat{q}_i^{(\tau)}} \tag{8}$$

where $\mathbb{1}$ is the indicator function evaluating if the observed y_i lies below the quantile forecast $\hat{q}_i^{(\tau)}$, over the evaluation set $(i = 1, \cdots, N)$. These diagrams allow one to summarize the reliability assessment of various quantile forecast with different nominal proportions, and assess if a given method tends to systematically underestimate (or overestimate) the uncertainty [14]. In an ideal scenario zero bias is achieved. In our experiment, this is performed for a set of quantiles $\tau_i = \{0.1, 0.2, ..., 0.5, 0.9\}$ over the range $(0, 1)$ with 0.1 equally spaced quantiles. Overall, 9 quantiles were considered.

4.2 Results

The experimental results based on the $RMSE$ of the studied algorithms on 8 data sets are shown in Table 2. The bold numbers in the table represents that the algorithm in the corresponding row has the best performance on the data set on the corresponding column.

Figure 2 provides the calibration results for the 8 algorithms on the regression datasets.

4.3 Discussion

The objective of the dynamic ensemble is to combine several quantile regression at multiple quantile levels τ. Therefore, the calibration test error is treated as the most important comparison criterion for quantile regression ensemble. Figure 2 depicts the difference from the perfect calibration (i.e., perfect match between nominal and empirical probabilities represented by the red line). Underestimation is observed when negative deviation from perfect calibration is present.

The experiment findings are summarized below.

- As is revealed by the experimental results in Table 2, the proposed algorithm gets the best performance in 5 of the 8 datasets.
- As to the comparison with the state-of-the art AQRM method the proposed algorithm has lower RMSE error. This validates the dynamic ensemble integration as a competitive solution.
- When considering the variants hard and soft weighting of the EW, similar errors were obtained. This holds for RMSE and calibration results (Figure 2)

Table 2. Experimental results based on $RMSE$.

models	datasets			
	airfoil	housing	cpusmall	mg
LQR	0.1727	0.0985	0.1409	0.1616
GBT	0.1435	0.0977	0.1253	0.1515
KNN	0.0782	0.1340	0.1122	**0.0780**
AQRM	0.1085	0.0964	0.1117	0.1047
$EW_{Soft}(DT)$	0.0771	**0.0945**	0.1050	0.0880
$EW_{Hard}(DT)$	**0.0765**	0.0963	0.1082	0.0882
$EW_{Soft}(KNN)$	0.0905	0.0980	**0.1040**	0.0906
$EW_{Hard}(KNN)$	0.0905	0.0980	**0.1040**	0.0906
	datasets			
	kin8fn	kin8fm	kin8nh	kin8nm
LQR	0.0442	0.0704	0.1452	0.1440
GBT	0.0586	0.0757	0.1426	0.1493
KNN	0.0415	0.0718	**0.1170**	**0.0813**
AQRM	0.0418	0.0700	0.1293	0.1117
$EW_{Soft}(DT)$	0.0410	0.0716	0.1310	0.1078
$EW_{Hard}(DT)$	0.0409	0.0722	0.1311	0.1078
$EW_{Soft}(KNN)$	**0.0374**	**0.0699**	0.1235	0.0928
$EW_{Hard}(KNN)$	**0.0374**	**0.0699**	0.1235	0.0928

- The proposed algorithm goes beyond the single best baseline estimator for 5 out of 8 tested data sets.
- The KNN quantile estimator shows very good performance on nonlinear data sets as kin8nm and kin8nh with RMSE of 0.1170 and 0.0813 respectively. But, when looking for the quantile calibration skill it exhibits poorly calibrated quantile estimates.
- All methods are shown to be well calibrated for quantile 50%, with deviance $\hat{b}^{0.5}$ close to zero.
- In terms of calibration GBT presents the best performance, being the deviation from the ideal (red line) smaller along all nominal quantile probabilities.
- The remaining quantile models underestimate quantiles bellow 50% and overestimate up to quantile 50%.

5 Conclusion

In this paper, a new dynamic ensemble for quantile regression (EW) is proposed.

EW is evaluated on 8 standard and publicly available data sets, commonly used in regression. We consider three different quantile regressors as baseline estimators. We have made efforts to optimize the baseline algorithms and compare them with the proposed solution. We notice that even if all the baseline regressors have been optimized, dynamic integration by local accuracy is still capable of improving overall performance.

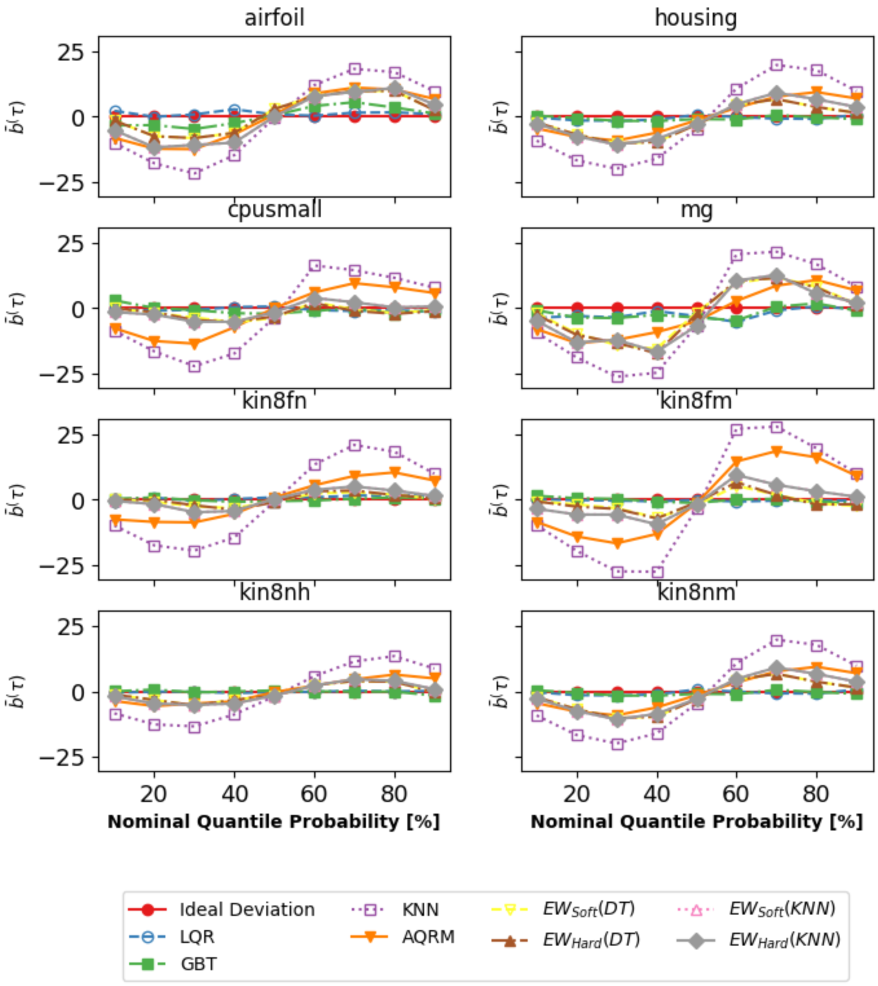


Fig. 2. Calibration results.

6 Acknowledgment

The work of Marisa Reis was financially supported by the InteGrid project (*Demonstration of INTElligent grid technologies for renewables INTEgration and INTEractive consumer participation enabling INTEroperable market solutions and INTErconnected stakeholders*), which received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under grant agreement No. 731218.

References

1. Yu, K., Lu, Z., Stander, J.: Quantile regression: Applications and current research areas (2003)
2. Koenker, R., Bassett, G.: Regression quantiles. Econometrica **46**(1) (1978) 33–50
3. Shan, K., Yang, Y.: Combining regression quantile estimators. Statistica Sinica (2009) 1171–1191
4. Rooney, N., Patterson, D., Anand, S., Tsymbal, A.: Dynamic integration of regression models. Proceedings of the International Workshop on Multiple Classifier Systems. Lecture Notes in Computer Science, vol. 3181 (2004) 164–173
5. Mendes-Moreira, J., Soares, C., Jorge, A.M., Sousa, J.F.D.: Ensemble approaches for regression. ACM Computing Surveys **45**(1) (2012) 1–40
6. Ruta, D., Gabrys, B.: Classifier selection for majority voting. Information Fusion **6**(1) (2005) 63–81
7. Shan, K., Yang, Y.: Combining Regression Quantile Estimators. Statistica Sinica **19** (2009) 1–27
8. Liu, R., Yuan, B.: Multiple classifiers combination by clustering and selection. Information Fusion **2**(3) (2001) 163–168
9. Woods, K., Kegelmeyer, W., Bowyer, K.: Combination of multiple classifiers using local accuracy estimates. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(4) (1997) 405–410
10. Kang, S., Kang, P.: Locally linear ensemble for regression. Information Sciences **432** (2018) 199–209
11. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Annals of Statistics **29**(5) (2001) 1189–1232
12. Reis, M., Garcia, A., Bessa, R.J.: A scalable load forecasting system for low voltage grids. In: PowerTech, 2017 IEEE Manchester, IEEE (2017) 1–6
13. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F.: Efficient and Robust Automated Machine Learning. Advances in Neural Information Processing Systems 28 (2015) 2944–2952
14. Pinson, P., Nielsen, H.A., Møller, J.K., Madsen, H., Kariniotakis, G.N.: Nonparametric probabilistic forecasts of wind power: Required properties and evaluation. Wind Energy **10**(6) (2007) 497–516

A Study of Novelty Detection in Data Streams Using Different Unsupervised Approaches

Kemilly Dearo Garcia, João Mendes-Moreira and André Ponce de Leon F. de Carvalho

Universidade do Porto, Portugal,
Universidade de São Paulo, Brazil,
kemilly.dearo@usp.br, jmoreira@fe.up.pt, andre@icmc.usp.br

Abstract. In data streams the data distribution can change over time, consequently classes unknown by the classifier may arise. In order to keep accuracy, the classification model must be updated, ideally, automatically and without external interference. Novelty detection is a machine learning task that aims to identify new classes, however this task cannot always be done by a supervised learning technique, because the instances can appear without labels. A solution to that issue is to apply clustering techniques to train and update the classification model. MINAS is a multi-class novelty detection algorithm for data stream that is able to update it classification model using a partitional clustering algorithm. This paper proposes two different clustering approach for training and updating on MINAS: one is a density-based clustering and the other is a hierarchical approach. The empirical experiments shows that by changing the clustering approach it is possible to improve the classification accuracy over time.

Keywords: novelty detection, data stream, clustering, multi-class

1 Introduction

In real world applications, data may arrive in a stream continuously with a distribution that can change over time. This scenario is known as data stream [10] and represents new challenges to machine learning task. Especially because traditional machine learning techniques consider that the data is always available for consultation, however, this does not happen in data streams, because the data is discarded as new data emerge, making impossible further consultation [4].

Due to changes in data distribution, new classes may appear, existing classes may disappear or evolve. In that context the classification model needs to be updated constantly to detect these changes and keep the accuracy over time.

Novelty detection is a classification technique which aims to identify when a change in the data may be considered novelty [6].The literature has treated the novelty detection problem as a only one class problem: the "Normal" class and the "Not Normal" class. However, as in [4], this paper considers that in a data stream more than one class may arise as a novelty class. Thus, the novelty detection problem should be treated as a multi-class problem.

Most algorithms presented in literature also consider that eventually, amount all data, some labeled instances will arrive in the stream. However, in real world applications, this assumptions cannot be mandatory to re-train the classification model. Due to the impossibility of using a supervised technique in the absence of labels, in [4], it is proposed the algorithm MINAS (MultI-class learNing Algorithm for data Streams) for novelty detection in multi-class scenario with two phases: an offline supervised phase and an online unsupervised phase with novelty detection. In online phase, when the classification model cannot classify an arriving instance, the instance is placed in a temporary memory. The algorithm CluStream [1] is used to find micro-clusters representatives of novel classes or extensions of known classes in the temporary memory, when it reaches a preset size. Clustering is a interesting choice in data streams scenario because of the lack of information about data distribution, which includes the appearance of novel classes.

The micro-cluster is a structure containing statistical summary of the data. Hence, it is based on K-means algorithm, which requires fixed parameters. But this information may not be available during the execution. Moreover, this technique is not good to find clusters with arbitrary shape. These characteristics can result in a low quality clustering, which implies on a low classification accuracy.

The main objetive of this paper is analyze the impact in MINAS accuracy when different clustering technique is used for updating the classification model. For that, it was used two known and well established clustering algorithms for data streams: DenStream [3] and ClusTree [8] algorithms. Unlike CluStream, the DenStream find clusters of arbitrary shapes and have a strategy to find and treat outliers. The ClusTree also finds clusters with arbitrary shapes, but stands out because is parameter-free. That way, ClusTree makes no assumption on the size of the clustering model.

To enable a analysis of the algorithms in terms of suitability for changes in data streams the experiments were elaborated with artificial benchmarks and real data stream. The evaluation demonstrates the gain of the proposed algorithms in comparison to current state of MINAS [4]. It exclusively achieves highly competitive results throughout all experiments, demonstrating its robustness and the capability of handling changes in data streams.

This paper is organized as follows. In section 2, it will be discussed the main related work about novelty detection in data stream. The micro-cluster structure used on CluStream algorithm is show in section **??**. The algorithm MINAS is detailed in section 3, it is also described the characteristics of the DenStream and ClusTree algorithms and how they were incorporated in MINAS. In section 4, the modifications was compared to the original MINAS. Finally, the conclusions are presented in section 5.

2 Related Work

In the literature, there are some novelty detection techniques that uses unsupervised techniques to update the classification model in online phase. However, in this cases, when a novel class is detected, this algorithms wait until the instances classified as novelty receive a true label, which in a real world scenario may no happen. In that case, the algorithm can wait for a long period of time to be updated to changes in the data stream, compromising accuracy.

The DETECTNOD algorithm is proposed in [7] and treats the novelty detection problem as a clustering task, using the K-means algorithm. In this approach, the final clusters are subjected to the discrete cosine transform (DCT) in order to generate the most compact representatives of classes. These clusters are used to detect novelty or concept drift in a data stream. However, DETECTNOD has the limitations present in K-means algorithm. Also, does not have the cluster validation to check the quality of the obtained clusters.

The OLINDDA algorithm (Online Novelty and Drift Detection Algorithm) [11] uses the final clusters of the K-means clustering algorithm to detect novelty and update the classification model. OLINDDA is divided into two phases: online and offline. In the offline phase is generated a labeled model of the "Normal" class with labeled examples. In the online phase, the model "Normal" can evolve or a new class can arise. When this occurs is generated a model "Extension" or "Novelty". Unlike DETECTNOD, the OLINDDA algorithm uses clustering validation to check the quality of the clusters and determine whether in fact the clusters can be considered novelty or extensions of the known class. However, this algorithm assumes that the novelty detection is a one-class problem, not considering that in data streams more than one novel class can appear at the same time.

In [9] is presented ECSMiner, a classification technique that allows automatic novelty detection. This algorithm assumes that all objects are labeled. In the algorithm, a ensemble of classifiers are trained by equal-size chunks. The ensemble is continuously with the novelty find by all classifiers. The ECSMiner algorithm considers that the label of an object is always available for consultation after a certain time. Unlike that, the MINAS algorithm updates the classification model using a clustering task, eliminating the need to get the label of the object.

All this approaches use K-means or a variant of it to update their classifier. Because of that, this approaches are subject to the same problems that K-means are: they only find hyperspherical clusters, it is necessary to set a fixed number of K clusters and the initial prototype may not be representative. Hence, we propose to adapt MINAS algorithm with different clustering approaches. That way, we intent to increase the final classification accuracy by obtaining clusters that better represents the data distribution on online updating phase.

3 MINAS With Different Clustering Approaches For Novelty Detection

MINAS is a multi-class novelty detection algorithm to data streams [4]. In this algorithm, the classification model is updated by CluStream algorithm, using this unsupervised approach, MINAS is capable to detect unknown classes and update the classification model automatically without the necessity of waiting for true labels. When an unknown class is found, it is validated by a cluster validation technique. If the pattern is considered valid, the classification model is updated with representative micro-clusters, otherwise, the known class is considered an outlier and is discarded. The micro-clusters are used as a summarize representation of the distribution state of the data stream.

The offline phase in the original MINAS, Figure 1, is the first task and it is performed only once in the begging of execution. The classification model is initially trained by a labeled data set. The labeled data are separated by their labels in sub-sets, each one is submitted to a clustering process with CluStream to generate micro-clusters representative to each class. The micro-clusters have relevant information from classes, by them it is possible to calculate the centroid (average representative of the clusters) and the radius of the micro-clusters. These micro-clusters are used as representatives of the classes for the classification model training.

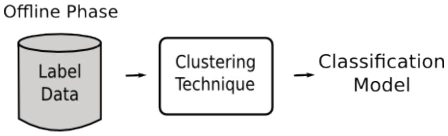


Fig. 1: MINAS - Offline Phase

In online phase, Figure 2, the classification model receive unlabeled instances coming from the data stream. If the instance can be identified by the model, it is labeled. Otherwise, it is stored in a temporary memory for future analysis. When the temporary memory has certain quantity of unlabeled instances, starts a clustering process with the unlabeled instances. Valid micro-clusters are marked as novelty or extension of a known class and then incorporated into the classification model. The invalid micro-clusters are considered outliers and discarded. To decide what is a valid micro-cluster and what is not, MINAS use silhouette measure.

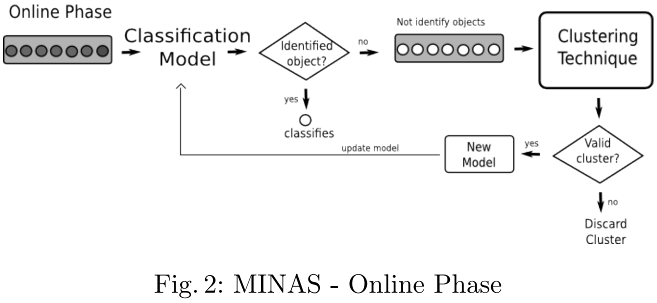


Fig. 2: MINAS - Online Phase

The CluStream [1] algorithm have limitations that prejudice the final clustering, because only finds a fixed number of spherical micro-clusters, is sensitive to order that which instance arrives in the streaming and needs initial representative prototype to generate a final clustering with good quality, according to some validation criteria. Because of these limitations, the MINAS algorithm may have low accuracy depending on the data distribution. Other problem involving CluStream is that the algorithm do not treats outliers. Only the most recent micro-clusters are stored in a fixed size temporary memory. That way, the relevance of the existing micro-clusters are decreased if it does not receive a new instance for a certain time. When outliers form new micro-clusters, older micro-clusters representing real novel classes can be discarded.

3.1 MINAS With DenStream

DenStream is a clustering algorithm for data streams capable of finding micro-clusters with arbitrary shapes and automatically detect outliers. In addition, does not require as parameter a fixed number of micro-clusters, which is ideal for data streams in which the number of classes is undetermined and may change over time. This algorithm is based on density approach where the dense regions are separated from non-dense regions.

In online phase of MINAS, DenStream is used to obtain p-micro-clusters, each one indicating a dense region. The concept is similar to the micro-clusters in [1]. For this algorithm, p-micro-clusters contains: the weighted square sum of the examples, the weighted sum of the instances within the micro-cluster and the weight (resulting from the fading function). With that information, it is possible to calculate: centroid and the radius of the micro-clusters.

If the p-micro-cluster has a lower density, it is marked as an outlier and is reserved in a secondary memory (making it a o-micro-cluster). If over time the density of the o-micro-cluster is equal to or greater than the specified per parameter, then the o-micro-cluster becomes a representative p-micro-cluster, otherwise it is discarded [3].

At the end of each execution of DenStream, the MINAS classification model is updated with the found p-micro-clusters. The classification model performs the classification of new instances from the data stream comparing the centroids of the micro-clusters with the instances arriving in the streaming. If the instances are identified, they receives a label that refers to it nearest centroid.

3.2 MINAS With ClusTree

Another clustering approach used for novelty detection in MINAS algorithm was ClusTree algorithm. ClusTree is a parameter-free algorithm that adapts by itself and is capable of organizing the found micro-clusters in a hierarchical tree, which makes the manipulation of the new micro-clusters faster [8]. To maintain only the newest micro-clusters, it is used the same decay function used on [3]. Another advantage of ClusTree is that if there is no time to determine the closest micro-cluster from a instance, the instance is stored in the temporary memory until it is time to be incorporate on the closest micro-cluster [8]. Because of this properties, MINAS show better results with this implementation, as we can see in section 4.

In MINAS online phase, when the temporary memory reaches a determined size, each instance inside it become a micro-cluster. After that, each new instance is compared with the centroids of the existent micro-clusters. If the instance is too far from the micro-clusters, it become a new micro-cluster. The concept to aggregate a new example to a micro-cluster is the same used in the previous approachs. It is done by a distance measure comparison with each instance and all micro-clustres centroids. The results are stored in a R-Tree to fast recover.

All the presented clustering data streams approaches, [1], [3], [8], use micro-clusters similar to the concept presented in [12]. The formal description of a micro-cluster is a set of d-dimensional examples $X = x_{i_1}...x_{i_n}$ which one with a specific time. It is defined as three components: $CF(n, C\bar{F}1, C\bar{F}2)$, n is the number of examples. $C\bar{F}1$, is the sum of examples and $C\bar{F}2$ the sum squared of the examples. With that information, it is possible to obtain the centroid e radius of a micro-cluster.

The main difference between this approaches is how they use the micro-cluster concept. In [1], besides the concept presented in [12], the micro-cluster also have the sum and sum squared of the time that each point was inserted on the micro-cluster. That way it is possible to know the micro-clusters that composed any time requested by the user. In [3] the centroid and radius weight are used to determine if a certain micro-cluster is a real micro-cluster or an outlier. This weight define which micro-cluster will be excluded. In [8], it is maintain the same micro-cluster in [12], but they are structured in R-tree family hierarchy. The micro-clusters are also submitted to a weight function with the same process as in DenStream algorithm.

4 Results and Discussion

In this section, we present the experimental evaluation for original MINAS from [4] and the proposed MINAS modifications presented in this paper.

4.1 Data sets e Settings

All algorithms were implemented in Java, using MOA Project Libraries [2] to simulate data streams. The experiments were executed on a computer with CORE I7 processor, 16 GB of RAM memory. The experiments was done to analyze the performance of MINAS algorithm with the CluStream, DenStream and ClusTree clustering approaches.

It was used synthetic data set also cited in [4]. The data set MOA has concept drift, appearance and disappearance of classes. The SynD and SynEDC data sets were also used in [9]. The first has only concept drift and the second has concept drift and appearance of new classes. However, it is important to consider how the algorithms behave in a real scenario, because of that, the real Forest Cover data set [5].

The MINAS version with DenStream has fixed initialization parameters, so it always produces the same micro-clusters. It was used as parameters: $\lambda = 0.006$, $\epsilon = 0.01$, $\beta = 0.2$, $\mu = 1$. The parameters was set by empirical experiments.

The CluStream algorithm, despite the fixed number of micro-clusters, it generates different micro-clusters due to the initial prototypes selected for the initialization of K-means algorithm. It was used as parameters: K-micro-clusters = 100 e maximum amount of not identify instances in temporary memory = 2000.

The ClusTree algorithm is a parameter-free algorithm, so does not need parameter initialization.

4.2 Results Analysis

In order to evaluate the results this paper adopted the evaluation measures for the multi-class problems presented in [4]. The results were analyzed by their CER, Combined Error Rate, which is calculated by the average weights of false positive and false negative per class. The accuracy measure F-measure, defined by the weighted harmonic average between precision and recall. And Unk, a rate of instances that were not explained by the classification model.

The Table 1 shows the performance of the algorithms based on the presented evaluation measures. The MINAS with ClusTree has higher accuracy in three of the data sets. The ClusTree algorithm maintain only the most recent micro-clusters, besides never let an instance of the clustering process, that way, when a micro-cluster appear as novel class MINAS is automatically updated. However the DenStream algorithm has low accuracy values and high CER, because parameters that defines density areas are not updated over time, which also indicate that the approach is not suitable to deal with changes in the stream over time, because of that MINAS loses it accuracy.

Table 1: Performance of the algorithms

Data Set / algorithm	MOA			CoverType			SynD			SynED		
	Unk	FMacro	CER	Unk	FMacro	CER	Unk	FMacro	CER	Unk	FMacro	CER
Minas_CluStream	7,92	0,96	0,13	6,75	0,38	11,7	0,45	0,65	0,35	5,24	0,73	0,45
Minas_ClusTree	1,04	0,99	0,09	0	0,37	13,84	0	0,74	0,25	0	0,76	0,42
Minas_DenStream	3,96	0,98	0,64	0	0,69	0	99,56	0,33	0,69	99,75	0,30	0,67

In Figure 3 the CER (green line) is low, but indicates that the MINAS with CluStream make some mistakes. Which is acceptable, given the fact that the classification model updates over time. However it is tricky to establish the parameters that generate the best result, because is necessary to know the right number of clusters and find good initial prototypes.

However in Figure 4 it is possible to see that the MINAS with DenStream starts with a CER close to zero, but in the middle of execution increases the

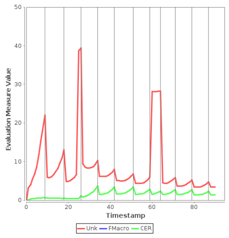


Fig. 3: Clustream

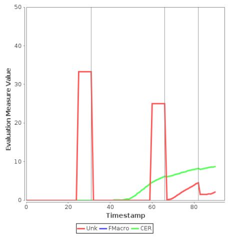


Fig. 4: DenStream

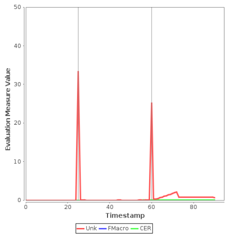


Fig. 5: ClusTree

mistakes. That happens because DenStream algorithm has fixed parameters, especially the elipsion that determines the size of the radius that a micro-cluster must have. As the data distribution changes over time, this parameters can not be the same. It has to evolve with the data to avoid this inconvenient. That also justifies the Unk, red line, that only has two peaks, indicating that the algorithm had two periods of time with more unclassified instances, which were the time when more new classes were found.

The MINAS with ClusTree algorithm presents betters results than the others, first Unk, Figure 5, shows only two peaks of unclassified instances. AlsoCER is practically zero, showing that the model barely misses. That happened because the algorithm structure provides that only the most relevant micro-clusters remain. That way, with few micro-clusters it is possible to have a classification model with good accuracy.

5 Conclusion and Future Work

This paper presents two new versions of the algorithm MINAS using DenStream e ClusTree clustering algorithm. This approaches shows that is possible to get better classification accuracy by changing the clustering technique for novelty detection in MINAS algorithm. ClusTree algoritm shows better results because of it memory structure that guarantees that only the most relevant micro-clusters remain. That way, with few micro-clusters it is possible to have a classification model with good accuracy.

As future work, we propose to examine the implementations in others data streams.

References

1. Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.: A framework for clustering evolving data streams. In: Proceedings of the 29th international conference on Very large data bases-Volume 29. pp. 81–92. VLDB Endowment (2003)
2. Bifet, A., Holmes, G., Pfahringer, B., Kranen, P., Kremer, H., Jansen, T., Seidl, T.: Moa: Massive online analysis, a framework for stream classification and clustering. (2010)

3. Cao, F., Ester, M., Qian, W., Zhou, A.: Density-based clustering over an evolving data stream with noise. In: SDM. vol. 6, pp. 328–339. SIAM (2006)
4. Faria, E.R., Gama, J., Carvalho, A.C.: Novelty detection algorithm for data streams multi-class problems. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing. pp. 795–800. ACM (2013)
5. Frank, A., Asuncion, A.: UCI machine learning repository (2010), `https://archive.ics.uci.edu/ml/datasets/Covertype`
6. Gama, J.: Knowledge discovery from data streams. Chapman & Hall/CRC Boca Raton (2010)
7. Hayat, M.Z., Hashemi, M.R.: A dct based approach for detecting novelty and concept drift in data streams. In: Soft Computing and Pattern Recognition (SoCPaR), 2010 International Conference of. pp. 373–378. IEEE (2010)
8. Kranen, P., Assent, I., Baldauf, C., Seidl, T.: The clustree: indexing micro-clusters for anytime stream mining. Knowledge and information systems 29(2), 249–272 (2011)
9. Masud, M.M., Gao, J., Khan, L., Han, J., Thuraisingham, B.: Classification and novel class detection in concept-drifting data streams under time constraints. Knowledge and Data Engineering, IEEE Transactions on 23(6), 859–874 (2011)
10. Silva, J.A., Faria, E.R., Barros, R.C., Hruschka, E.R., de Carvalho, A.C., Gama, J.: Data stream clustering: A survey. ACM Computing Surveys (CSUR) 46(1), 13 (2013)
11. Spinosa, E.J., de Leon F de Carvalho, A.P., Gama, J.: Novelty detection with application to data streams. Intelligent Data Analysis 13(3), 405–422 (2009)
12. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: an efficient data clustering method for very large databases. In: ACM SIGMOD Record. vol. 25, pp. 103–114. ACM (1996)

Predicting winning teams for regular season of the NBA

Saulo Ruiz[1]

MAP-i Doctoral Programme in Computer Sciences, Porto, Portugal
up201502083@fep.up.pt

Abstract. The sports industry generate a huge amount of data. It's not only the box score at the end of a game. With new technologies, every sport has translated each sequence of play into a instance on a database. Teams are using machine learning techniques to improve their game and also to understand their rivals in the field or court. This papers present two different approaches to use the basketball data to predict the teams that end on top of the National Basketball Association (NBA). We use Decision Tree (DT), Bayesian Network (BN) and K-Nearest Neighbors (KNN) to predict the true performance of a team within the league. We also compare how this algorithms perform when applying different transformations to the initial dataset while avoiding overfitting at the same time. We achieved near 90% accuracy when predicting the true value of the top tier teams using only three seasons of data. This papers shows that the data from the sports industry can be useful to anticipate teams performance and to optimize the value of the team with a limited budget.

Keywords: Machine Learning, Predictive Modeling, Bayesian Network

1 Introduction

Data science is a growing field, not only on techniques but also in applications. Every industry is on the pursuit of better predictions, better classifications, better evaluations of each element within their activities. Retailers want to know what products offer on a given period. Telecommunication companies want to target which customers are more likely to leave. These have become daily challenges solvable with the use of machine learning.

As every other industry, sports has it's own challenges. But was not until 2001 when baseball, one of the major sports in North America showed the real potential of statistical techniques in the sports industry. These techniques quickly turned into formal use of data science for different challenges in the industry [3]. At that time, the main challenge was more of an optimization problem, on which the management of the Oakland's Athletics, a team in the Major Baseball League (MLB) had to maximize the quality of the team with a limited budget. However, the main problem is the subjectivity on this industry, what is the true value of a player despite the media attention or fan base. Then we can see how two problems arise. First, the business problem which is to maximize performance of the team with limited money and second the machine learning problem which focuses in assigning a real and objective value to a given player.

However, statistical analysis was used before in other sports going unnoticed. Since 1996, the NBA was using the IBM's Advanced Scout System [2]. This was a decision support system with the goal to help the coach and the rest of the management team to pick the best players based on expected performance.

The remainder of this paper is organized as follows. In the next section we will revise related work on applications of machine learning to the sports industry. Section 3 presents the dataset used in the experiments. Section 4 describes the methodology used in the experiments. Section 5 present the results of the prediction model experiments. Conclusions and the outlook of this approach are highlighted in the Section 6.

2 Related Work

Since the mid 90's, data mining has been used in the sport industry. The appearance of more accurate predictive algorithms, the sports industry has turned more and more into a data-driven industry. Since 1995, there was already data driven decision for scouting future talents [2]. But some years later, the use of statistical methods was used even to create plays and style of play for some teams [4]. NBA teams have created Data Science departments in order to analyze not only their plays but also try to find patterns and weakness in the teams they have to face.

The analysis does not limit to analyze teams but also players as individuals. Our work focuses in proposing an analysis similar to the one presented in [1]. However, we intent to predict the team that wins the league instead the season Most Valuable Player (MVP). As implied on the name, this is a player's award and is based on individual performance rather than team performance. The season MVP does not necessarily belongs to the league winning team.

Similar tasks have been performed but in different sports. [7] show how data is used to predict winner in football competitions. Using a mix of previous performance and then matching with the teams within the competition. In our case we consider only the one league which is the NBA. The case presented in [7] compares teams from different leagues as it analyzes the Union of European Football Associations (UEFA). The UEFA by it's nature, unite 32 teams from at least 8 different leagues each year. Therefore, this league it's dynamic as the teams vary each year.

To create a proper rank for the UEFA implies to normalize the leagues to obtain the true performance of each team on a standard way. We propose two different transformation methods to standardize the teams in our dataset and assign a performance relative to their league. There is evidence that the transformation method has an important role to achieve better predictions [6], [5].

In next section we present the dataset used for the experiments we performed. Presenting a description of features and statistical analysis that allow to identify noise in the dataset.

3 Dataset

The dataset used in this papers consist of team stats of three full seasons of the NBA. The period in the dataset goes from the 2014 season to the 2016 season. It covers a total of 30 teams. The teams are divided into conferences (East Conference and West Conference) under the NBA schema. Each conference holds 15 teams. However, as the winner of the league can be form either conference and they are geographically split we considered it irrelevant for our study. Bellow we present the description of features in the initial dataset and the short name of the stat that will be used throughout the rest of the paper.

- **GP:** The number of games played.
- **W:** The number of games won by a player or team.
- **L:** The number of games lost by a player or team.
- **WIN%:** The percentage of games played that a player or team has won.
- **MIN:** The number of minutes played by a player or team.
- **PTS:** The number of points scored.
- **FGM:** The number of field goals that a player or team has made. This includes both 2 pointers and 3 pointers.
- **FGA:** The number of field goals that a player or team has attempted. This includes both 2 pointers and 3 pointers.
- **FG%:** The percentage of field goal attempts that a player makes.
- **3PM:** The number of 3 point field goals that a player or team has made.
- **3PA:** The number of 3 point field goals that a player or team has attempted.
- **3P%:** The percentage of 3 point field goal attempts that a team makes.
- **FTM:** The number of free throws that a player or team has made.
- **FTA:** The number of free throws that a player or team has attempted.
- **FT%:** The percentage of free throw attempts that a player or team has made.
- **OREB:** The number of rebounds a player or team has collected while they were on offense.
- **DREB:** The number of rebounds a player or team has collected while they were on defense.
- **REB:** A rebound occurs when a player recovers the ball after a missed shot. This statistic is the number of total rebounds a player or team has collected on either offense or defense.
- **AST:** The number of assists – passes that lead directly to a made basket – by a player.
- **TOV:** A turnover occurs when the player or team on offense loses the ball to the defense.
- **STL:** Number of times a defensive player or team takes the ball from a player on offense, causing a turnover.

- **BLK:** A block occurs when an offensive team attempts a shot, and the defense team tips the ball, blocking their chance to score.
- **BLKA:** Number of times an opponent has registered a block on the shot of a player or team.
- **PF:** The number of personal fouls a player or team committed.
- **PFD:** The number of personal fouls that are drawn by a player or team.
- **+/-:** The point differential when a player or team is on the floor.

By definition we can identify variables that do not have any use for a machine learning model. Take the example of GP or W. For the case of GP, this feature is equal for all teams in all seasons since every team must play 82 games each season. As for the W, we know it's correlated with WIN% since WIN% it's calculated as W/GP. As GP is constant for each team each season then we know there is a direct relation between W and WIN%.

The only not intuitive stat is the +/- Box Plus Minus (BPM). BPM is a box score-based metric for evaluating basketball players' quality and contribution to the team. It is the latest version of a stat previously called Advanced Statistical Plus/Minus; it is not a version of Adjusted Plus/Minus, which is a play-by-play regression metric. BPM relies on a player's box score information and the team's overall performance to estimate a player's performance relative to league average. BPM is a per-100-possession stat, the same scale as Adjusted Plus/Minus: 0.0 is league average, +5 means the player is 5 points better than an average player over 100 possessions (which is about All-NBA level), -2 is replacement level, and -5 is really bad. To get this indicator to the team level is performed a weighted average of the team weighting the possessions that included a player on his given team. In Table 1, we show some statistical measures from the dataset. These are average (Avg), standard deviation (StdDev), max, min and median. This will help not only to have an overall idea of the dataset but also to validate it.

Other insight we can get it's the magnitude of the different stats, this is related to the nature of the stat itself. We notice that blocks are somewhere between 3.6 to 6.8. However, if we compare this stat to points and we know the nature of the game, we know that it's impossible for a game to end with 7 points for a team. We deal with this scale discrepancy using 2 types of transformations that will be detailed in following sections.

In order to select the features used in our models we created a correlation matrix. This will help us to clearly identify what variables can bring noise through co-linearity to the models. We did not consider GP for this matrix since the nature of this feature is not determined by the performance of the team. In Table 2, we can observe some expected results.

A inverse relation between W and L. This is expected because they W can be obtained from L mixed with GP, as GP is constant a inverse relation is remaining. Can be seen more easily in Equation 1 and Equation 2.

$$W = GP - L \tag{1}$$

$$L = GP - W \tag{2}$$

Table 1. Features from the NBA season stats.

Stat	Avg	StdDev	Max	Min	Median
GP	82	0	82	82	82
W	41	12.68	73	10	41
L	41	12.68	72	9	41
WIN%	0.50	0.15	0.89	0.12	0.50
MIN	48.4	0.2	48.8	48.1	48.4
PTS	102.8	4.6	115.9	91.9	102.8
FGM	38.3	1.6	43.1	33.7	38.3
FGA	84.5	2.3	89.2	77.2	84.4
FG%	45.3	1.5	49.5	40.8	45.2
3PM	8.7	1.8	14.4	5.0	8.6
3PA	24.5	4.6	40.3	14.9	24.7
3P%	35.3	1.8	41.6	31.7	35.1
FTM	17.6	1.6	22.3	13.9	17.4
FTA	23.1	2.1	29.4	18.5	22.9
FT%	76.0	3.1	81.5	66.8	76.3
OREB	10.5	1.1	13.1	7.9	10.5
DREB	33.0	1.4	36.2	29.3	33.1
REB	43.5	1.7	48.6	38.6	43.6
AST	22.3	2.1	30.4	18.0	22.1
TOV	14.2	1.2	17.7	11.5	14.0
STL	7.8	0.9	10.0	5.7	7.8
BLK	4.8	0.7	6.8	3.6	4.8
BLKA	4.8	0.7	6.3	3.0	4.9
PF	20.1	1.4	24.8	16.6	20.3
PFD	20.1	1.2	23.9	17.5	20.1
+/-	0.0	4.7	11.6	-10.2	0.2

Table 2. Correlation Matrix of the initial dataset.

Stat	W	L	WIN%	PTS	FGM	FGA	FTA	FT%	TOV	BLKA	PFD	+/-
W	1.00											
L	-1.00	1.00										
WIN%	1.00	-1.00	1.00									
PTS	0.56	-0.56	0.56	1.00								
FGM	0.53	-0.53	0.53	0.86	1.00							
FGA	-0.01	0.01	-0.01	0.52	0.61	1.00						
FTA	-0.00	0.00	-0.00	0.27	-0.09	-0.13	1.00					
FT%	0.17	-0.17	0.17	0.24	0.21	0.10	-0.19	1.00				
TOV	-0.27	0.27	-0.27	-0.02	-0.06	-0.07	0.23	-0.28	1.00			
BLKA	-0.48	0.48	-0.48	-0.22	-0.32	0.10	0.26	0.06	0.33	1.00		
PFD	0.10	-0.10	0.10	0.16	-0.14	-0.24	0.80	-0.16	0.15	0.05	1.00	
+/-	0.97	-0.97	0.97	0.59	0.56	0.01	0.00	0.17	-0.27	-0.52	0.09	1

Other interesting relation can be seen in PTS against FGM. As FGM contribute directly to PTS, it's not the only way to increase PTS. As in Table 1 we can see an average of 17.6 FTM which is around 17 PTS per game. But teams average nearly 103 PTS per game, therefore, presenting a 0.86 relation between PTS and FGM can be deduced.

Most of these indicator are positive indicators, this means that the higher the indicator the better the performance of the team. There are three negative indicators. These are BLKA, PF and TOV. They are considered negative indicators since they do not contribute to the PTS of the teams.

We can also see how the BPM is related to W. This is expected also as the BPM is an indicator related to the gain of the teams each possession. IF a team is having gains each possession it's expected to perform well, therefore winning more games.

Other relation found among the features is the FTM and FTA to PFD. As for the nature of the sport, we know that a PFD occurs on the motion of shooting then the shooting team is granted 1 to 3 FT attempts.

After analyzing all this relations we proceeded to drop some features. We dropped W and L to keep only WIN%. Even after finding relations in other features, we decided to keep them since they might be useful after transforming in next section.

4 Comparing two different representations for the data transformation

In this section we describe the two different ways of transforming the dataset in order to create a team based dataset. As for now, we have only seen summarized data of the dataset. But this data is different for each team. Therefore, the way to identify winning teams among all teams is to compare the different metrics between themselves.

For this task we propose two different approaches. First, a ranking system which will re-code the existing value of the feature to the rank from 1st to 30th. Second, a distributions system. For this transformation we propose to use the normalization process using Equation 3. In this equation, *i* represents the feature under evaluation of the ones listed before. While *j* represents the instance under evaluation.

$$Z = \frac{x_{ij} + \mu_i}{\sigma_i} \tag{3}$$

The idea under this transformation is to avoid scaling issues. As presented in the previous sections, the scale of some features dominate others. Therefore, some models will tend to assign the weights on this high-scale features since they do not take into account the nature of the feature. Important features like BPM will get irrelevant when compared to features like PTS or FGA. The normalization procedure keeps the order. This is, the team with highest PTS will remain as highest but just in a smaller scale.

4.1 First Experiment: Ranking System

The idea of this system is simple. It bases on the logic that a team that a winning team is a complete team. This is that the winning team has the highest average rank when compared to the rest on the league. After applying the transformation to a set like Table 3, we have a new dataset as presented in Table 4.

Table 3. Example set by team.

Team	PTS	TOV	+/-
CLE Cavaliers	110.3	13.7	3.2
GSW Warriors	115.9	14.4	11.6
DET Pistons	101.3	11.9	-1.1

As seen by the examples, we considered the nature of the indicator. This means that TOV, even that is presented as a positive number we know that it does not contribute to a good performance of the team. Therefore, the highest TOV will be translated into the last in the rank. This is contrary if we see PTS which is a direct contributor to a good performance of a team and the higher the indicator the smaller the rank.

We also created an average rank to help us to have a general idea of the performance of the team overall. Since some teams focus on defense, they might

Table 4. Ranking transformation example set by team.

Team	PTS	TOV	+/-	AVG-Rank
CLE Cavaliers	2	2	2	2.00
GSW Warriors	1	3	1	1.67
DET Pistons	3	1	3	2.33

not have flashy numbers in stats like PTS and AST but have high BKL and DREB. So the AVG-Rank will allow us to balance high offensive teams with high defensive teams. Hereinafter, the dataset obtained throught the rank transformation will be know as DB-A.

4.2 Second Experiment: Normalizing Data

For the second experiment we transformed the dataset using the Equation 3. Using the same example on Table 3, we can then reconstruct and create a a dataset following Table 5.

Table 5. Normalize transformation example set by team.

Team	PTS	TOV	+/-
CLE Cavaliers	0.19	0.35	-0.26
GSW Warriors	1.12	1.01	1.33
DET Pistons	-1.31	1.36	-1.07

The idea of this transformation is to split the teams by performance. Take for example PTS which is a positive indicator. Then, teams with performance above average will be with positive number while, on the contrary, if the team is performing below average of the league then will have a negative indicator. This also re-scale all the features. The dataset obtained after using the normalization transformation will be called DB-B from now on.

4.3 Training of algorithm

After having the two datasets we trained three algorithms, DT, KNN and a BN based on the nature of the problem. The goal of this algorithms is to provide a probability of winning the league for each team. Notice that the probability split among the 30 teams. For example, if Team A has 99% of winning then the sum of all the other 29 teams must be equal to 1%.

We added the label of the previous positions of each team in the last 3 seasons. This label is the target fot the models and then a distance between winning team and the predicted position is calculated using Euclidean Distance.

We used 10-fold cross validation for each dataset. Then we compared the prediction with the current performance of the team. This is, we have the position of each team at mid-season for the 2017 season which is out of the initial dataset. We created the average rank for each team based on their current performance to evaluate the models.

5 Analysis of Results

For measuring the performance of the models we used the Mean absolute percentage error (MAPE). MAPE is a measure of prediction accuracy of a forecasting method in statistics. As we used the average ranked we turned the prediction into a continuous variable instead of an ordinal variable. We applied the models to both DB-A and DB-B. DB-A provides a better perspective since it takes into account the rank of the stats and can then balance a defensive team with a offensive team as stated before. In Table 6, we can observe that with DB-A create algorithms more accurate than the algorithms trained with DB-B.

Table 6. MAPE of the algorithms for DB-A and DB-B.

Algorithm	DB-A	DB-B
DT	13.5%	23.2%
KNN	15.9%	21.6%
BN	10.3%	18.1%

The main insight from these results is the difference of the performance between DB-A and DB-B. Therefore, we can affirm that the ranking transformation provides a better representation of the performance of the teams and allow the algorithms to better produce better predictions.

6 Conclusions

As we have shown in the previous section, the use of machine learning techniques can be useful for the sports industry. Even though the DB-B did no provide accurate results, the models trained with DB-A proved to be near 90% accurate using our metrics of evaluation. Only traditional data was considered for the dataset. Our dataset can be expanded adding other features related to the performance of the teams which are more specific, such as, pace, average time offense, passes made, among others. Also, only three seasons were considered.

These seasons are considered imbalanced by the experts since the same two teams have reached the finals in all three seasons. Moreover, one of the seasons considered have the case of the most wins for a team in a season in the NBA history. Even if the seasons have the same number of teams and the same amount of games each year, the presence of extremely dominant teams can bias a season.

For future work, we will focus in using more seasons to avoid dominance of teams. We will also expand the dataset in terms of features, we believe that adding team-performance related features can help to identify the good teams. Moreover, we intend also to expand the algorithm selection and try the Weight of Evidence (WoE) transormation as a dataset transformation.

References

1. Chen, M.: Predict nba 2016-2017 regular season mvp winner. Fuzzy Systems and Data Mining III: Proceedings of FSDM 2017 299, 15 (2017)
2. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine 17(3), 37 (1996)
3. Lewis, M.: Moneyball: The art of winning an unfair game. WW Norton & Company (2004)
4. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. In: ACM Sigmod Record. vol. 29, pp. 427–438. ACM (2000)
5. Ruiz, S., Gomes, P., Rodrigues, L., Gama, J.: Credit scoring in microfinance using non-traditional data. In: Portuguese Conference on Artificial Intelligence. pp. 447–458. Springer (2017)
6. Van Gool, J., Verbeke, W., Sercu, P., Baesens, B.: Credit scoring for microfinance: is it worth it? International Journal of Finance & Economics 17(2), 103–123 (2012)
7. Van Haaren, J., Zimmermann, A., Davis, J.: Mlsa15-proceedings of"machine learning and data mining for sports analytics", workshop@ ecml/pkdd 2015 (2016)

Tuberculosis Classification and Drug Resistance Detection in Medical Images with Deep Learning

João Figueira Silva, Jorge Miguel Silva, Eduardo Pinho, and Carlos Costa

DETI - Institute of Electronics and Informatics Engineering of Aveiro
University of Aveiro, Portugal
{joaofsilva,jorge.miguel.ferreira.silva,eduardopinho,carlos.costa}@ua.pt

Abstract. Tuberculosis (TB) is one of the top 10 causes of death worldwide. Its correct treatment depends on the correct diagnosis of TB type, which is extremely important to avoid the development of multidrug resistance (MDR) as the MDR form is considered a health security threat by WHO. Current methods to detect tuberculosis are either costly and time-consuming - mainly involving blood tests - or inaccurate - mostly based on medical imaging. Therefore, there is a need for a quick, cheap and efficient diagnosis solution. With recent developments in technology, medical imaging-based solutions have regained much interest. We propose an improved version of the solution developed for an international medical imaging challenge, ImageCLEFtuberculosis 2017, whose tasks were not known in forehand to be solvable with medical imaging approaches. Our solution uses 3D Convolutional Neural Networks (CNNs) and Computed Tomography (CT) volumes from TB patients to classify TB in one of five types, and to detect multidrug resistance. Obtained results still have much margin for improvement, proving the difficulty of the challenge, but hold significant promise for the field of TB diagnosis.

Keywords: Deep Learning, Medical Imaging, Computed Tomography, Tuberculosis

1 Introduction

Tuberculosis (TB) is an infectious disease, spread through the air, that presents grave concern to human health as it is one of the top 10 causes of death worldwide. TB can be divided in two main categories, extrapulmonary and pulmonary TB, with the latter being more prevalent as TB generally affects the lungs [1]. Each of these categories can be further divided, for instance, pulmonary TB can be classified in five distinct types. Successful treatment is reliant on a correct diagnosis of the TB type [2].

The most critical aspect of TB is its capability of acquiring drug resistance, with the Multidrug-Resistant (MDR) form being considered a health security threat [1]. Drug resistance can arise from an incorrect drug usage during treatment, and the resulting MDR-TB form is much more expensive and difficult to treat, with results being generally poor [3]. It is thus of utmost importance to

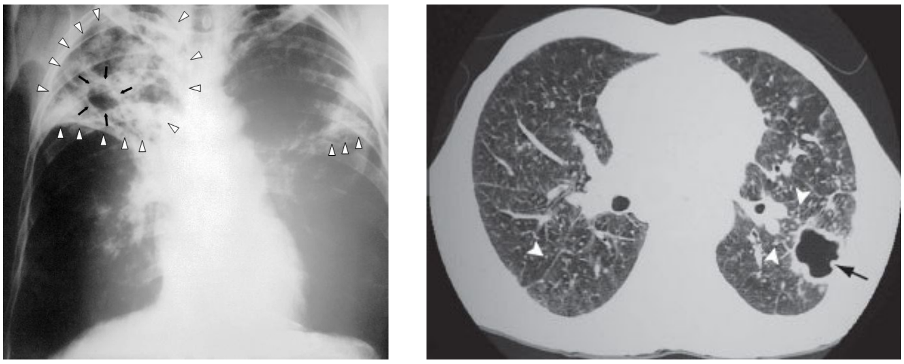


Fig. 1. Chest X-ray (left) and CT (right) scan obtained from TB patients. Images retrieved from [8] and [9], respectively.

have a quick, cheap and efficient diagnosis solution to control the disease. However, such solution is currently an impossible due to the existing methods being either costly and time consuming, or inaccurate [4].

While most solutions for TB detection involve blood tests [5], medical imaging can be used as a tool for diagnosis, namely, using chest X-ray and Computed Tomography (CT) scans (Fig. 1), but is usually associated with inconclusive results [6, 4]. Despite recent progress in medical imaging-based solutions, which spurred interest once again in this field [7], the development of these solutions is an extremely hard task as working with medical imaging datasets usually encompasses distinct challenges such as limited access to data, its reduced size and unbalanced class distributions.

Since deep learning (DL) has had a significant appearance in the field of medical imaging with promising results in recent years [10], we decided to create a deep learning-based solution for the detection of multidrug-resistance and classification of TB type using chest CT scans. It is important to mention that current medical image-based solutions focus on the detection of tuberculosis, not on TB type classification nor on drug resistance which are harder tasks *per se*.

This paper presents an improved version of the system described in [11], which was tested in the ImageCLEFtuberculosis task [2] from ImageCLEF 2017 [12], an international challenge centered on medical imaging. Regarding paper structure, developed methodology is presented in Section 2, results are presented and discussed in Section 3, and finally Section 4 draws some conclusions and lines for future work.

2 Methodology

The problem to address consisted in detecting multidrug resistance and classifying the TB type from CT scans. Since datasets for MDR detection and TB type classification were not interchangeable, each subproblem was tackled sepa-

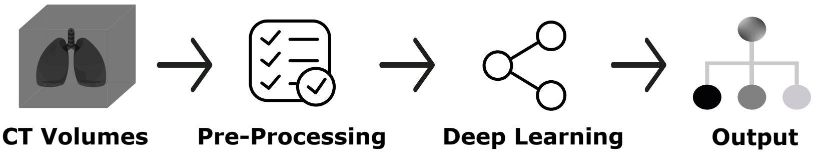


Fig. 2. Pipeline Overview.

rately. We propose a two-stage pipeline consisting of data pre-processing and a DL model (Fig 2). The same pre-processing was applied for both subproblems whereas the DL model was fine-tuned for each subproblem.

Pre-processing consisted of using the CT images to segment the lung region, and resizing data to be used as input in the DL model. The DL model then used batches of pre-processed data and classified it. Each of the stages will be further explained in more detail.

An extremely important aspect for the selection of our DL model comes from the observation that a CT volume is composed by several images and the analysis of a single slice might provide poor classification results. Thus, we opted to feed the DL models with volumes composed of stacks of CT slices, leading us to the use of a 3D-CNN model instead of a conventional CNN model. This option naturally brought implications regarding input tensor shape. Such implications were solved with pre-processing.

2.1 Data Pre-processing

As previously mentioned, different datasets were provided by ImageCLEF for each subproblem. For multidrug resistance detection, a train dataset with 230 CT volumes and a test dataset with 214 CT volumes were provided, each with two possible classes - TB and MDR-TB. For TB classification, a train dataset with 500 CT volumes and a test dataset with 300 CT volumes were provided, each with five possible classes - 5 different pulmonary TB types. The number of slices varied across CT volumes, and slice had a size of 512x512 pixels [2].

In the training datasets, lung segmentation was performed using masks created with the following method: a threshold was applied to the images where intensities below -300 Hounsfield units were set as background, pixel values were normalized to have an intensity range between 0 and 255, and resulting images were passed through a binary thresholding process with a threshold value of 20. Using scikit-image[1], small holes and small objects were removed, using methods with the same name and parameterized with minimum size of 100 and connectivity of 4. Next, the two methods were reapplied but with a minimum size of 1000, resulting in the final masks.

Obtained masks were highly similar compared to those provided to the participants [13]. Dice's coefficient, which varies between 0 (no similarity) and 1 (full similarity), was computed to assess the similarity between created and provided

[1] http://scikit-image.org

masks, with a global average value of 0.9755 being obtained. Regarding the test dataset, provided masks were used to ensure that test data was not tampered.

Resulting volumes, with the lungs segmented, were reshaped to comply with the NHWC channel ordering (number of samples x height x width x channels) used in CNNs. Here, the number of samples corresponds to the number of CT slices. Each CT slice was then resized to dimensions of 256x256 pixels.

Obtained volumes were resized, in respect to the number of slices, so that all volumes had the same number of slices. This was achieved by padding the top and bottom of each volume, resulting in a final volume with fixed size (real data in the center, and padding in the extremities). Finally, data was normalized to have zero mean. Pre-processed datasets were saved in HDF5 files, resulting in two HDF5 files per problem (MDR detection and TB type classification) containing train and test data.

2.2 Model Architecture

As previously mentioned in this section, it was our decision to use a 3D-CNN model due to the nature of the problem. The model was implemented with TensorFlow [14] version 1.0.0 with support for GPU, which massively increases the speed and efficiency of training and developing models such as neural networks. Moreover, TensorBoard was used during the development of the 3D-CNN model for debugging and optimization purposes. Some additional functions needed for the models' development were imported from TFLearn[2] (v0.3), a DL library that provides a higher-level API to TensorFlow. The 3D-CNN training scripts ran on an Ubuntu server machine equipped with an NVIDIA Tesla K80 GPU accelerator.

An overview of the built model with the respective composition of each layer is presented in Fig 3. The decision to use a 3D-CNN model with seven convolutional layers and two fully connected layers was empirical.

Existing literature supports that deeper models can be more powerful than shallow ones, as the former can learn how to represent high-level abstractions, presenting particular interest for the fields of vision, language and other AI-level tasks [15]. However, it is also known that deeper models are more difficult to train due to problems such as the vanishing gradient problem, where initial layers learn at slower speeds than final layers. Naturally, deeper networks are more prone to the vanishing gradient problem [16], and demand bigger compute power, which is a very significant overhead. Thus, considering the implications of deep neural networks, and the existing limited compute power, it was decided to build a small network.

As it is possible to observe in Fig 3, the network's first six layers share the same structure (but not the hyperparameters). In these six layers, the incoming tensor is passed through a sequence of 3D convolution, batch normalization, non-linear activation function and 3D max pooling. This sequence reflects a change in the previous proposal [11], as pooling was previously being executed prior to

[2] TFLearn: http://tflearn.org

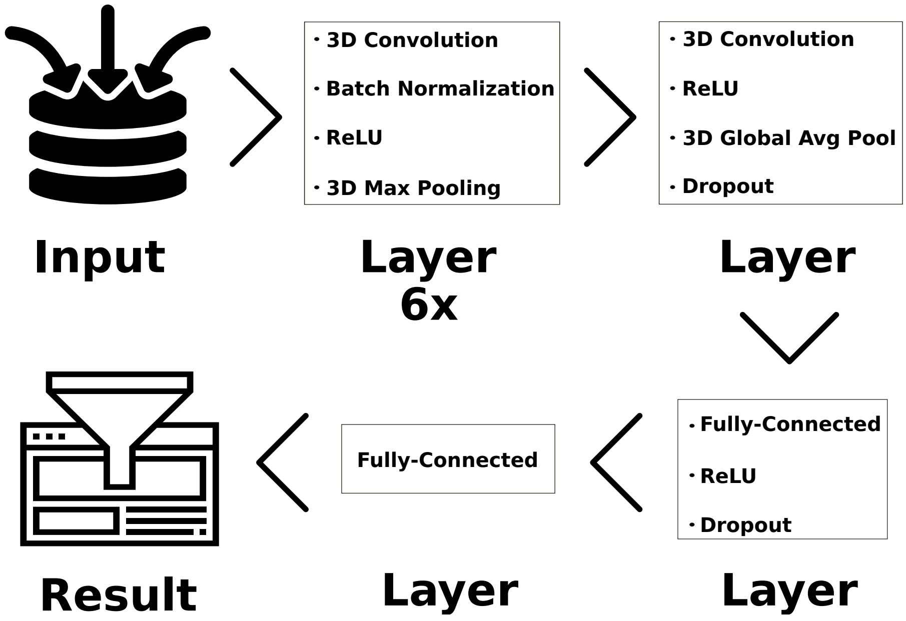


Fig. 3. Diagram of the neural network model used.

batch normalization and activation function and could be resulting in the loss of important information.

Batch normalization is very important as it addresses a phenomenon called internal covariate shift, which slows down the training of neural networks [17]. Concerning the activation function, since the sigmoid activation function can cause problems when training deep neural networks [18], the rectified linear unit (ReLU) was used. The solution presented in [11] erroneously mentions the use of the leaky variant of ReLU, since TFLearns implementation of leaky ReLU contained a bug which made it work as a regular ReLU. That aspect has been reflected in the architecture presented in Fig 3.

Overfitting is another serious concern in neural networks, specially when working with medical imaging datasets which frequently consist of small amounts of data, with unbalanced distributions. For that reason, dropout [19], a regularizer used to reduce overfitting in neural networks, was used in our model. However, it was only applied to the fully-connected part of the network as convolutional layers have considerable inbuilt resistance to overfitting [20]. Also, L2 regularization was used in each convolutional layer to reduce model overfitting, and the last Fully-Connected layer has a softmax activation function.

The same model was used both for MDR detection and TB type classification, though with different hyperparameters due to the fine-tuning procedure performed for each subproblem. Dropout was used with a drop probability of 0.5 for both subproblems. A detailed description of the remaining hyperparameters for models' layers is presented in [11]. It should be noted that the hyperparameters were defined with compute power constraints in mind. All weights were initialized as described in [21].

2.3 Model Training

Before training the model with pre-processed data, the training dataset was split into 80/20 parcels, for training and validation splits respectively. Data distribution had moderately balanced classes for the MDR detection problem, whereas for the TB type classification a less balanced dataset was provided. For each subproblem, data was split taking into account class distributions, in order to ensure the same class distribution in training and validation splits. Even though the network was prepared to work with K-fold cross validation, due to time constraints and the inherent nature of the training process of a neural network, the network was validated offline using a single combination of the 80/20 split.

Both 3D-CNNs were fed with mini-batches of data containing complete CT scans, where each sample corresponds to one of the CT volumes being forwarded through the net. Since these networks are demanding in terms of memory and computational cost, in order to enable the use of bigger batch sizes, each preprocessed volume was cropped into a fixed smaller number of slices corresponding to the size of the smallest volume in the original dataset, with the cropping method extracting data from the center of each CT volume.

In order to prevent the networks from learning a given data sequence/order, data splits were shuffled in each epoch, prior to being fed to the models. Four metrics were used to assess model performance, namely: cross entropy, accuracy, precision, and recall.

Moreover, an Adam optimizer [22] was used for stochastic optimization, with the following settings being used: $\alpha = LearningRate$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. Table 1 summarizes some of the hyperparameters used in order to train the neural networks. All hyperparameters were defined through empirical testing.

Table 1. Best hyperparameters used in the neural network for each subproblem.

Hyperparameter	**MDR detection**	**TB type**
Batch Size	30	30
Learning Rate (LR)	3 x 10^{-5}	8 x 10^{-6}
LR Decay	5%	5%
Decay Interval	10 *epochs*	15 *epochs*

Learning rate was reduced by a fraction of 5 percent of its value after 10 and 15 epochs and validation was performed in intervals of 3 and 2 epochs for the MDR detection and TB type classification models, respectively.

3 Results and Discussion

Since ImageCLEFtuberculosis organization provided contestants with unlabelled test datasets, we faced the nuance of not being able to assess test performance for the newly improved version of the models. With the lack of test labels, a comparison between test performances for the old and new versions cannot be made, thus to perform a direct comparison between the two versions we resort to performances obtained with the validation dataset.

The best validation results for the previous models, presented in [11], and for the improved versions herein described are shown in Table 2. As it is possible to observe, the new versions outperform previous ones in every evaluation metric, though with just slight improvements in some metrics.

Table 2. Performance metrics obtained by the models in the validation phase.

Metrics	**MDR Detection**		**TB Type**	
	Old	**New**	**Old**	**New**
Accuracy	0.5501	0.6023	0.1744	0.2367
Precision	0.5470	0.5624	0.5223	0.5371
Recall	0.3440	0.3592	0.4413	0.4639

Regarding MDR detection, a two class problem, the model shows the same behaviour as before, favoring the retrieval of the most frequent class but significantly struggling to detect the less frequent and more relevant class, resulting in a substantially lower recall comparatively to obtained accuracy and precision.

For the TB type task which is a multi-class problem (five classes), as expected, the new model still achieves lower accuracy than the MDR detection model, while maintaining slightly similar precision and improved recall. The higher number of classes, combined with a less balanced dataset, greatly impacts on the validation accuracy which was significantly lower than in the MDR detection task. Such accuracy value demonstrates that the neural network had difficulties in identifying the classes in data, explaining why some classes had no occurrence registered in the validation dataset.

Despite observing improved performances for the new models during validation, the most important part of the analysis is missed due to the impossibility of measuring test performances. Since all entries in ImageCLEFtuberculosis were tested with the same untampered test dataset, the challenges final list of results could serve as a benchmark for solution comparison provided test labels were known.

Since the observed behaviour for our previous solutions consisted in a decrease in performance during the test phase, we can speculate that the same behaviour should occur for the new version of the models, thus anticipating lower performances than those seen in Table 2. This means that in spite of the better performance during validation, test metrics are expected to be significantly lower.

The list of test results for the two subtasks comprised in ImageCLEF's tuberculosis task [2] clearly demonstrates the high difficulty associated with the task at hands. For MDR detection, the top ranking model had an accuracy just slightly over 50% whilst our original model's test accuracy was nearly 47%. It is expected that the new model's test performance should be close to that of the top ranking model. Concerning TB classification, test results were in general worse comparatively to those of MDR detection. Here, the top ranking entry had an accuracy of 40% compared to our model's 24.3%. Despite the improved performance of our new model during validation, test results are expected to maintain significantly worse than those of the top ranking entry.

In spite of our models' seemingly poor performance, the final ImageCLEFtuberculosis result list [2] shows that other entries are not very far, having comparable performance. Since in our approach the search for the best hyperparameters was empirical and not extensive enough due to limitations in terms of available time, it is our belief that there exists margin for progress and improvement in our work, provided there is more time to better train the models, and correctly fine-tune them.

4 Conclusion and Future Work

Tuberculosis presents great concern to the health community as it is one of the top ten causes of death worldwide. While its correct diagnosis is paramount, existing solutions still have to progress towards a cheap, quick and reliable solution. Medical imaging, which was once the go to solution, is showing once again great promise with the research community making great efforts in that direction.

The ImageCLEFtuberculosis challenge is an example of such effort, focusing on the creation of solutions to detect drug resistance and classify the type of pulmonary TB (from a pool of five possible types) in CT scans from TB patients.

Herein, we presented an improved version of the previously created solutions for the ImageCLEFtuberculosis 2017 challenge. The improved version uses the same pre-processing but comprises modifications in the neural networks architecture, resulting in better performances both for MDR detection and for TB type classification during validation.

The absence of labels for the test dataset provided by ImageCLEFs organization severely limits our ability to assess our new models performance, since without performance metrics obtained with the test dataset it is not possible to directly compare our models with those in ImageCLEFtuberculosis result list, a list which serves as a solution benchmark. Nonetheless, we expect the improved MDR detection model to be closer to the best MDR detection solution in ImageCLEFtuberculosis 2017. Regarding TB type classification, in spite of the improvements, we expect the new solution to maintain far from the best ranking model.

It is interesting to notice that various solutions submitted by other research groups to the ImageCLEFtuberculosis challenge used DL approaches, which shows that DL is an area that holds great promise and that we are following the right path.

As future work, and since overfitting was an effective reality during the development of the neural network models, in the future we hope to evaluate the impact of techniques such as data augmentation and weight normalization on our models' results. Furthermore, running the model with K-fold cross-validation and performing an ensemble of the resulting networks could further improve our results.

Acknowledgments

This work is financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme, and by National Funds through the FCT Fundação para a Ciência e a Tecnologia. João Figueira Silva is funded by the research grant of PTDC/EEI-ESS/6815/2014 project and Jorge Miguel Silva is funded by the research grant of CMUP-ERI/ICT/0028/2014-SCREEN-DR project. Eduardo Pinho also was funded by the FCT under the grant PD/BD/105806/2014.

References

1. WHO: Tuberculosis Fact Sheet. Available in: http://bit.ly/1iT0dPm (October 2017)
2. Dicente Cid, Y., Kalinovsky, A., Liauchuk, V., Kovalev, V., , Müller, H.: Overview of ImageCLEFtuberculosis 2017 - predicting tuberculosis type and drug resistances. In: CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, Dublin, Ireland, CEUR-WS.org <http://ceur-ws.org> (September 11-14 2017)
3. Ahuja, S.D., Ashkin, D., Avendano, M., Banerjee, R., Bauer, M., Bayona, J.N., et al: Multidrug resistant pulmonary tuberculosis treatment regimens and patient outcomes: An individual patient data meta-analysis of 9,153 patients. PLOS Medicine **9**(8) (08 2012) 1–16
4. Al-Zamel, F.A.: Detection and diagnosis of mycobacterium tuberculosis. Expert Review of Anti-infective Therapy **7**(9) (2009) 1099–1108
5. Anochie, P.I., Onyeneke, E.C., Ogu, A.C., Onyeozirila, A.C., Aluru, S., Onyejepu, N., Zhang, J., Efere, L., Adetunji, M.A., Sánchez, J.G.B.: Recent advances in the diagnosis of mycobacterium tuberculosis. Germs **2**(3) (2012) 110
6. Ryu, Y.J.: Diagnosis of pulmonary tuberculosis: recent advances and diagnostic algorithms. Tuberculosis and respiratory diseases **78**(2) (2015) 64–71
7. Miller, C., Lonnroth, K., Sotgiu, G., Migliori, G.B.: The long and winding road of chest radiography for tuberculosis detection. European Respiratory Journal **49**(5) (2017)
8. CDC: Public Health Image Library. Available in: http://bit.ly/2Aq6UFZ (October 2017)

9. Barboza, C.E.G.A., Winter, D.H., Seiscento, M.A., Santos, U.d.P., Terra Filho, M.A.: Tuberculosis and silicosis: epidemiology, diagnosis and chemoprophylaxis. Jornal Brasileiro de Pneumologia **34** (11 2008) 959 – 966
10. Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang, G.Z.: Deep Learning for Health Informatics. Biomedical and Health Informatics, IEEE Journal of **21**(1) (jan 2017) 4–21
11. Silva, J.a.F., Silva, J.M., Pinho, E., Costa, C.: 3D-CNN in Drug Resistance Detection and Tuberculosis Classification. In: CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, Dublin, Ireland, CEUR-WS.org <http://ceur-ws.org> (September 11-14 2017)
12. Ionescu, B., Müller, H., Villegas, M., Arenas, H., Boato, G., Dang-Nguyen, D.T., Dicente Cid, Y., Eickhoff, C., Garcia Seco de Herrera, A., Gurrin, C., Islam, B., Kovalev, V., Liauchuk, V., Mothe, J., Piras, L., Riegler, M., Schwall, I.: Overview of ImageCLEF 2017: Information extraction from images. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference of the CLEF Association, CLEF 2017. Volume 10456 of Lecture Notes in Computer Science., Dublin, Ireland, Springer (September 11-14 2017)
13. Dicente Cid, Y., del Toro, O.A., Depeursinge, A., Müller, H.: Efficient and fully automatic segmentation of the lungs in CT volumes. In: Proceedings of the VISCERAL Anatomy Grand Challenge at the 2015 IEEE ISBI. CEUR Workshop Proceedings, CEUR-WS (2015) 31–35
14. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al: TensorFlow: Large-scale machine learning on heterogeneous distributed systems. (2016)
15. Bengio, Y., Yoshua: Learning Deep Architectures for AI. Foundations and Trends® in Machine Learning **2**(1) (2009) 1–127
16. Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J.: Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies. In: Field Guide to Dynamical Recurrent Networks. IEEE Press (2001)
17. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. CoRR **abs/1502.03167** (2015)
18. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10). Society for Artificial Intelligence and Statistics. (2010)
19. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research **15** (2014) 1929–1958
20. Nielsen, M.A.: Neural Networks and Deep Learning (2015)
21. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1026–1034
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014)

SESSION 4

Bioinformatics, Education and Gamification

Upper Limbs Movement analysis for Medical classification of Breast Cancer Patients

António Guerra, Hélder P. Oliveira, João P. Monteiro, André T. Magalhães,
Sérgio Magalhães, Edgar Costa

FEUP/INESC-TEC, Porto, Portugal,
bio10060@fe.up.pt

Abstract. Breast cancer is the leading cause of cancer in women, with a high survival rate. However, common treatment procedures, such as radiation therapy or surgical removal of the axillary lymphatic nodes, lead to a decrease in patients quality of life due to impairments in upper limb function. A premature detection of these types of problems is vital to decrease the impact in Upper Body Function (UBF) and further complications. Nevertheless, detection is currently performed using subjective methods, either using volume and angle measurements or inquiries. The present work aims to create an objective method to perform this evaluation and provide a more accurate analysis of upper limb impairments. For that, a number of exercises were selected and RGB-D and skeleton tracking data were acquired during these. Features extracted from this data allows to characterized patient condition, evaluating the presence of pain, stiffness, weakness, lymphedema and functionality. Results obtained in this work were very promising for breast cancer patients' classification, and are better and more complete than those presented in the literature.

1 Introduction

Breast cancer is the leading cause of cancer in women, with 1.383.500 new cases in 2010 worldwide [1] with only a mortality rate is 27%, which is even lower in developed regions [2]. This is due to the development of effective therapies and medical procedures, as well as prompt diagnosis. Due to low mortality rate, concerns with the Quality of Life (QOL) of breast cancer patients have been growing, usually associated with problems of Upper Body Function (UBF) on women after breast cancer surgery, and the impacts of each treatment in it. Usually, this problems are related with surgical procedures and post-surgical treatment, with no defined healing process. Several studies have been trying to address UBF and psychological problems such as depression, anxiety and problems with body image [3] in this group.

Other problem associated with post-surgery diagnosis is the difficulty to perform it in early stages using traditional methods. Studies showed that subjective self-reports are more sensitive and can lead to a more premature detection of reduced UBF [4] than traditional methods. However, this can lead to incorrect results. Therefore, there is currently a demand for an objective method that can assess UBF and QOL. This method needs to be accurate, low-cost, able to produce results in quickly, reproducible and capable of performing early diagnosis of upper-limb impairments.

The purpose of this study is to improved the existing methodology, initially developed by Moreira et al. [5], to assess breast cancer patients condition using Microsoft Kinect. For this, we will use a more diverse and complete set of exercises. After that, several methodologies of signal and image processing were tested to understand which will lead to a more robust approach. Lastly, a classification step will be perform to evaluate the patients.

This work was performed with the support of medical staff, more specifically Dr. André Magalhães and nurse Sérgio Magalhães from Hospital S. João.

The remainder of this paper is structured as follows: section II provides a brief overview of two main concepts; in section III the methodology implemented is explained; the experimental results obtained is explained section IV and, finally, section V contains the concluding remarks regarding the developed work.

2 Background

2.1 Methods for UBF Assessment

There is evidence that suggests that subjective assessment of UBF can detect lymphedema more prematurely and is more sensitive to the development of lymphedema [4] than any other method. Also subjective assessment through patient self-report allows functional but also psychosocial aspects to be taken into account [6], like pain or stiffness, for instance.

From all the inquiry present in the literature, the most used in breast cancer patients is Disability of Arm, Shoulder and Hand (DASH) [7], in about 46% of the times [8]. However, this inquiry was not designed for breast cancer patients and no validation for this population could be found in the literature.

This report consists of 30 questions that assess pain, weakness, numbness and functionality, and, possibly, an optional module that is related with work and sports. Studies found that DASH shows better results in breast cancer population, even when compared with other inquiries, some of then design to this condition [8].

Other self-report that is interest to this group is Kwan's Arm Problem Scale (KAPS). This report was created to assess the condition of breast cancer patients in upper limb related to function, pain, stiffness and swelling. This inquiry also has a subscale that rates impairment in activities of daily living. This scale has been shown to be convergent and exhibits discriminant validity in breast cancer survivors [9]. However, this report does not take into account psychometric properties. Other self-report include Upper Limb Disability Questionnaires (ULDQ) and 10 Questions by Wingate.

2.2 Upper Limb Assessment using Microsoft Kinect

Microsoft Kinect can be used to perceive human pose, behaviour or posture. Regarding upper limb movement, Pastor *et al.* [10] uses the sensor in the rehabilitation process of stroke patients. For this they use the skeleton tracking algorithm applied to a serious game that scores the exercises performed by the patients when compared with a expected pattern.

Kurillo *et al.* [11] performed a study to evaluate the ROM of upper limb extremity, an important factor in the rehabilitation of breast cancer patients. As well as Pastor [10], this author uses the skeleton tracking algorithm and obtained an accurate and robust to perform this type of evaluation in medical environment.

Lu *et al.* [12] performs Volume Estimations in patients with lymphedema using the RGB-D sensor. A problem with this system is that requires a manual selection of the ROI. Furthermore, there are other requirements (the device needs to be approximately at 80 cm of distance, for instance) that affects its practicality. Results obtained showed that the method is robust.

Lastly, Moreira *et al.* [5] also to study Upper Limb Volume, ROM and other features to evaluate aesthetics and lymphedema presence. To validate the data, the author uses self-reports. Problems associated with this work is that only one exercise is used when performing ROM evaluation. Also, the work was limited to a not-diversified database, which can harm the results obtained. Precision on the detection of lymphedema was also a problem.

2.3 Methods for Volume Assessment

Detection of lymphedema can be detected by tonometry (applying pressure for one minute on a region where edema fluid is known to accumulate) [8], for instance. However, this technique only works in advanced stages of the disease, a problem since an early detection of lymphedema allows to minor the problems associated with it.

In ideal conditions, and using the most accurate methods, the diagnosis should be performed with comparison between pre and post surgery limb volume. However, since the volume of the limb before surgery is not usually measured, is usually impossible to perform this technique. To solve that a bilateral comparison is performed to assess the presence of edema [13]. Nevertheless, some authors defend that, since difference between arms occur in healthy women, this measure is imprecise [13]. In this subchapter we will focus in some techniques used in the detection of lymphedema.

Subjective Methods Some criteria are associated with the detection of lymphedema: the method must be efficient, easy to use, non-invasive, inexpensive, hygienic and reliable. Although there are no golden standard defined in literature [14], Water Displacement is usually considered the most complete method when considered the previous mention criteria [13]).

However, this technique has several problems. First, it is hard to be precise in the performance of this procedure, since patients usually have difficulties due to low mobility of the arm, a problem usually associated with lymphedema. In addition, the definition of the exact point to perform comparison is very hard to do [15]. Additionally, Water Displacement is time consuming, not portable, unhygienic and messy [16, 13]. Lastly there are some problems with related to standard deviation (up to 25 mL) [17] of the method.

Objective Methods As can be seen, even the golden standard in subjective methods have a lot of problems associated. So an objective and fast method to measure lymphedema have been perused. This can be possible with 3D technology, being a example the Computer Aided Measurement Laser (CAML), proposed by Trombetta *et al.* [18] in 2012. In this work the authors used 3D IR laser scanning for limb analysis and data collection of its shape, size and appearance, and computer aided design (CAD) system, to create a model through the data collected, as circumferential and limb volume measurements. The IR laser scanning used was Polhemus FastSCAN^{TM}, that was tested in 2007 by McKinnon *et al.* [19] for volume measurement.

For comparison and results validation, the previously mentioned water displacement method was used chiving similar. The authors concluded that the errors between the system and circumferential measurements are minimal and acceptable.

Other systems can be found in literature, with Insignia^{TM} laser scanning system [20] being the most relevant, showing to be suitable for volume assessment.

3 Methodology

The first task was to select a RGB-D sensor to UBF assessment. For that purpose, initially, it was research what was the best sensor in the market to achieve the purpose of this work[21]. From this search, it was found that Microsoft Kinect was the more appropriate sensor, due to its properties such as the skeleton tracking system.

After that, a group of exercises to be performed by the patients was selected with the help of medical staff. For it some criteria were defined: for instance, the exercises needed to be performed by the patients in medical environment. An interesting group of exercises that meet the requirements are rehabilitation exercises, being selected the 5 more relevant to the medical professionals since would allow to extract different evaluation features. Exercises were adapted from [22] and can be seen in Figure 1.

Also, and since lymphedema is characterized by arm swelling, is important to access the volume of both limbs, to understand if there is or not an increased size of the affected arm. To perform this acquisition, the patient maintain a position with arms in a T shape for at least 3 seconds.

Data acquisitions were performed at Centro da Mama in Hospital S. João.

3.1 Database Characterization

The database is composed of 63 patients, from whom were acquired skeleton tracking system data, colour images and depth maps. Also for each one of it, upper-body functionality was computed by performing DASH questionnaires to all of them. Patients were divided in two groups, 0-39 showed excellent or good

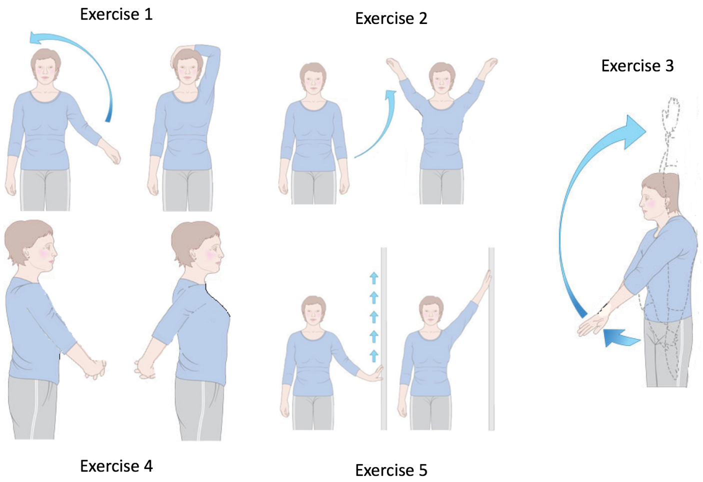


Fig. 1: Exercises selected for patient evaluation

upper body functionality (class 0), 40-100 showed regular or bad upper body functionality (class 1) [23]. Since DASH also allows the possibility to compute if the patient has pain, weakness or stiffness, the patients were also divided in two classes, where 1 represents the presence of this condition and 0 the absence. Lastly, it was assessed if patients had lymphedema by the medical physician. Table 1 represents the distribution of each class for each condition. From this table we can see a quite balanced distribution in each metric evaluated.

3.2 Data Acquisition and Processing

Since each exercise have different characteristics and goals, the data acquisition is defined by its purposes: skeleton data was extracted for all exercises and RGB-D data was only acquired for Volume Measurements.

Data based on Skeleton Tracking System Data Using the skeleton tracking system from Microsoft Kinect is possible to retrieve coordinates of joints and other important points of the human body over the exercise, allowing the characterization of the patients movement during this period of time with high precision [24]. Nevertheless, ambient conditions as luminosity can have impact in precision [25].

Table 1: Physical condition of the patients

	Pain	Weakness	Stiffness	Lymphedema	Functionality
1	39	36	31	27	28
0	24	27	32	36	35

Table 2: Summary of features extracted using Skeleton Tracking System Data

	Exercise 1	Exercise 2	Exercise 3	Exercise 4	Exercise 5
Elbow Height	1	8	15		37
Wrist Height	2	9	16		38
Elbow Width				24 & 25	
Wrist Width				26 & 27	
Range of Motion	3	10	17		39
Elbow Flexion	4		18	30 & 31	40
Duration of the Exercise	5	11	19	32 & 33	41
Velocity	6	12	20	34 & 35	42
Acceleration	7	13	21	36	43
Compensatory Movement Hip-Shoulder		14	22		44
Inclination				23	
Angle Hip-Shoulder-Wrist				28 & 29	

One way that we minor this problem is using signal processing. In this work a selection of filters based on literature was made, with five filters being selected: Median Filter,Moving Average Filter, Gaussian Filter, Low-Pass Filter and Kalman Filter.

Filter selection was performed comparing a filtered signal with a signal obtained from color image that were manually marked. Raw signal was obtained using the skeleton tracking system algorithm of Microsoft Kinect in the most relevant joints for each exercise. These signals were then filtered and compared with the signal acquired by manual tracking (ground truth). This comparison was done using a covariance matrix.

After filter selection, the data obtained from each one of the previous mentioned exercises allowed to extracted unique 44 features. Table 2 summarizes feature acquisition for each exercise and numbers that are attribute. Important to notice that Exercise 4 has two phases, so two different feature extractions can be performed.

Data based on RGB-D Data Volume can be an important feature to detect lymphedema. Therefore, in addition to all the previously mention features extracted, we asked the patients to perform an exercise to estimate volume of both arms. Three methods were tested to retrieve the volume: method 1 is purely based in depth map data, method 2 constructs a segmentation mask using Skeleton information and then segments the arm using colour images and, lastly, method 3 constructs a segmentation mask using the depth information and then segments the arm using colour images. Segmentation in colour images were performed using Graph Cut.

For all this methods, it is important to define the region of interest - the upper-arm - defined superiorly by the shoulder, inferiorly by the elbow and communicates medially with the axilla [26]. Therefore, there are five points that

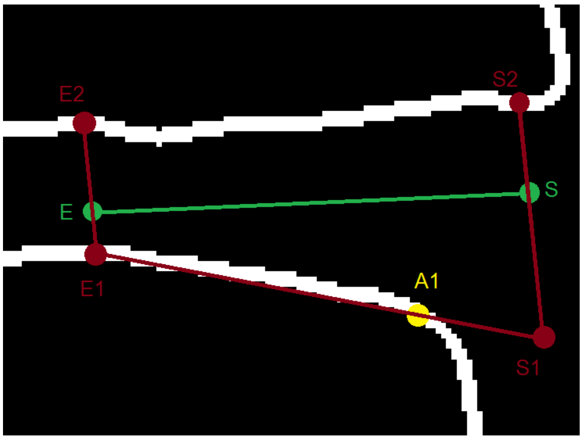


Fig. 2: Arm segmentation

define the region: the armpit (A1), the medial (E1) and lateral (E2) elbow point, and the medial (S1) and lateral (S2) shoulder point. The detection of this points are performed for both arms and using the skeleton information (Shoulder (S) and Elbow (E) - points in green in the Figure). The final mask, as well the relevant points, are present in Figure 2. Important to mention that, to compute it, a silhouette of the patient is obtained by segmentation explained before.

After that, the obtained masks are compared with the marked mask to select the more correct approach using to metrics: Dice coefficient and Jaccard Index.

After method selection is necessary to perform a volume estimation of the arm. For this approach colour and depth data are necessary to compute the point cloud of each arm. Then, a convex hull algorithm is used on the set of computed points. To validate this step two objects with known volumes were used, achieving with good results.

To volume estimation, 40 images of each patient were selected (20 of depth data and 20 of colour data). Only obtained values within a range of 40% of the median of each patient were considered, in order to discard possible outliers. Then, for each arm the mean value was computed. Ratio between the two arms are computed and was the 45th feature extracted.

3.3 Classification

In this work three supervised learning classifiers were used: Fisher Linear Discriminant Analysis (LDA), Nave Bayes Classifier and Support Vector Machines (SVM). Furthermore, due to the high number of extracted features, three feature selection methods were tested: Mutual information [27], Sequential Floating Forward Selection (SFFS) [28] and Forward Selection, that consists in the following: Firstly, the error associated with the classification of one feature is computed. This feature (Feature A) is then kept and a new subset of features is formed by the combination of Feature A with the remaining ones. The error of all these combinations is computed, and the best combination is fixed. Important to mention that features were normalized to a scale of 0-1.

A grid search of the best parameters for each classifier was performed, and a Leave One-Out scheme was used during this tests. For each type of classifier that was described, the number of features that led to the lowest misclassification in each output was selected. In the cases where more than one subset of features obtained minimum misclassification, the model with less complexity was chosen. Metrics as Precision and Recall were also computed.

4 Results and Discussion

Signal Processing - Filter Selection In these results, Kalman filter presents a higher variance in 5 of the 10 values. Furthermore, in those Exercises where Kalman Filter does not have the best variance, this filter still presents similar values to those of the best filter. So, and after other tests that validate this selection, the Kalman Filter was chosen to perform signal processing.

Segmentation method selection Results from each method and respective metrics (Dice Coefficient and Jaccard Index) can be found in Table 3.

Considering the results presented in Table 3, the best segmentation process is the one which uses only depth data, showing a good overlapping between ground truth and the detected arm. Also, the lower results for segmentation when using Colour data are due to problems in segmentation by using GrabCut, something that can seen in Figure 3.

Classification results

First, and respecting feature selection, Forward Selection achieve the best results for all classifiers in all classifications. The results obtained are present in the following sections.

Pain Classification The best classifier associated with pain and its characteristics are presented in Table 4.

As expected, results obtained for SVMs were very good, since it proved to be a good option when there is a limited amount of data available (only 66 patients. Also, it is relevant to notice that the feature set is composed by features from Exercise 3 (Shoulder Height and Compensatory Movement Hip-Shoulder) and Exercise 4 (Shoulder Width and Duration of the Exercise). Furthermore, pain classification showed a high recall result due to the low error in classification of

Table 3: Dice coefficient and Jaccard Index results of arm segmentation for three methods

	Method 1	Method 2	Method 3
Dice coefficient (D)	0.794	0.721	0.724
Jaccard Index (J)	0.672	0.635	0.642

Table 4: Parameters of best classifier for pain classification

Classifier	Kernel	C	y	MER	Recall	Precision	Feature Set
SVM	rbf	1	0.25	0.19	0.974	0.776	[15 22 24 33]

patients with pain. Precision results were lower. However, these results, as well as the associated error, are very promising for the performance of pain classification in breast cancer patients.

Stiffness Classification The best classifier and its associated characteristics are presented in Table 5.

In stiffness classification, the more relevant features are obtained from Exercise 3, Exercise 4, Exercise 5 and Volume. Furthermore, stiffness classification shows high recall and precision, and low error in classification. These results are very promising and allow us to conclude that this methodology is a suitable solution for stiffness classification. Also, this explains the importance of the volume extraction, which can be relevant to more than lymphedema detection.

Weakness Classification The best classifier and its associated characteristics are presented in Table 6.

Weakness classification as something interesting: the feature set is composed by features obtained from all exercises. Furthermore, weakness classification shows high recall and precision, and also a low error in classification. These results are very promising and allow us to conclude that this methodology is

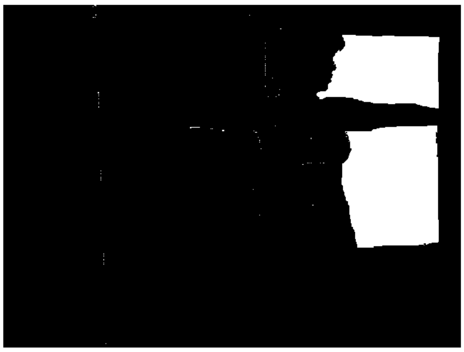


Fig. 3: Problems in patient segmentation using GrabCut method for Patient 47

Table 5: Parameters of best classifier for stiffness classification

Classifier	Kernel	C	y	MER	Recall	Precision	Feature Set
SVM	rbf	8	0.75	0.127	0.833	0.939	[21 28 36 38 39 42 45]

Table 6: Parameters of best classifier for weakness classification

Classifier	Kernel	C	y	MER	Recall	Precision	Feature Set
SVM	rbf	8	1	0.111	0.839	0.929	[4 10 15 16 28 35 43]

a suitable solution for weakness classification. Also, this shows the importance of different exercises, and that a more extensive number could revealed to be important.

Lymphedema Classification The best classifier and its associated characteristics are presented in Table 7. The feature set form lymphedema classification is composed by features obtained from Exercise 3, Exercise 4 and Exercise 5. It is important to compare these results with those obtained by Moreira [5], since in this work lymphedema is also classified. Results obtained in this work shows better precision and error (0.87 vs. 0.86 obtained by Moreira and 0.159 vs. 0.19 obtained by Moreira, respectively), with similar results for recall (0.741 vs. 0.75 obtained by Moreira). This can be explained by the higher number of exercises used and also by the higher number of patients from whom data was acquired.

Table 7: Parameters of best classifier for lymphedema classification

Classifier	Kernel	C	y	MER	Recall	Precision	Feature Set
SVM	rbf	2	0.75	0.159	0.741	0.87	[16 21 24 26 28 29 38 43]

Furthermore, these results are very promising and allow us to conclude that this methodology is a suitable solution for lymphedema classification. However, since the volume is not present in this classification can revealed some problem in its computation.

Functionality Classification The best classifier and its associated characteristics are presented in Table 8.

The feature set is composed by features obtained from Exercise 1, Exercise 2, Exercise 3 and Exercise 4. Moreira [5] also performed a functionality evaluation, but using UEFI inquiries instead of the DASH inquiries. Although it is possible to compare results, with our approach achieving better results in the two evaluated metrics (error and recall), this comparison needs to be performed conservatively due to the different methodology used. Even with this in mind, the results obtained are very promising.

Table 8: Parameters of best classifier for functionality classification

Classifier	Kernel	C	y	MER	Recall	Precision	Feature Set
SVM	rbf	8	0.55	0.127	0.857	0.857	[2 3 8 11 15 27 33 36]

5 Conclusion

In this work, the best classification results were obtained in the SVMs classifiers with RBF kernels, with a associated misclassification error of 0.19 for pain, 0.127 for stiffness, 0.111 for weakness, 0.159 for lymphedema and 0.127 for functionality classification. These results are very promising, and are better when compared to results from existing literature. In the future it is important to create a global model, instead of using five separate models for each state, and resolve some problems associated with volume estimation.

We conclude that the proposed methodology appears to be suitable for the evaluation of upper-body functional status in breast cancer patients and can be used to help medical physicians during diagnostic the phase.

References

1. Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D.: Global cancer statistics. CA: a cancer journal for clinicians **61**(2) (2011) 69–90
2. Jemal, A., Center, M.M., DeSantis, C., Ward, E.M.: Global patterns of cancer incidence and mortality rates and trends. Cancer Epidemiology Biomarkers & Prevention **19**(8) (2010) 1893–1907
3. Vahdaninia, M., Omidvari, S., Montazeri, A.: What do predict anxiety and depression in breast cancer patients? a follow-up study. Social psychiatry and psychiatric epidemiology **45**(3) (2010) 355–361
4. Gärtner, R., Jensen, M.B., Kronborg, L., Ewertz, M., Kehlet, H., Kroman, N.: Self-reported arm-lymphedema and functional impairment after breast cancer treatment–a nationwide study of prevalence and associated factors. The Breast **19**(6) (2010) 506–515
5. Moreira, R.: Dynamic analysis of upper limbs movements after breast cancer surgery. FEUP (2014)
6. Armer, J.M., Radina, M.E., Porock, D., Culbertson, S.D.: Predicting breast cancer-related lymphedema using self-reported symptoms. Nursing research **52**(6) (2003) 370–379
7. Solway, S., of Orthopaedic Surgeons, A.A., et al.: The DASH outcome measure user's manual. Institute for Work & Health (2002)
8. Harrington, S., Michener, L.A., Kendig, T., Miale, S., George, S.Z.: Patient-reported upper extremity outcome measures used in breast cancer survivors: A systematic review. Archives of Physical Medicine and Rehabilitation **95**(1) (2014) 153 – 162
9. Nesvold, I.L., Fosså, S.D., Naume, B., Dahl, A.A.: Kwans arm problem scale: psychometric examination in a sample of stage ii breast cancer survivors. Breast cancer research and treatment **117**(2) (2009) 281–288
10. Pastor, I., Hayes, H.A., Bamberg, S.J.: A feasibility study of an upper limb rehabilitation system using kinect and computer games. In: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, IEEE (2012) 1286–1289
11. Kurillo, G., Han, J.J., Obdrzálek, S., Yan, P., Abresch, R.T., Nicorici, A., Bajcsy, R.: Upper extremity reachable workspace evaluation with kinect. In: MMVR. (2013) 247–253

12. Lu, G., DeSouza, G., Armer, J., Anderson, B., Shyu, C.R.: A system for limb-volume measurement using 3d models from an infrared depth sensor. In: Computational Intelligence in Healthcare and e-health (CICARE), 2013 IEEE Symposium on, IEEE (2013) 64–69
13. Stanton, A., Badger, C., Sitzia, J., et al.: Non-invasive assessment of the lymphedematous limb. Lymphology **33**(3) (2000) 122–135
14. Ridner, S., Montgomery, L., Hepworth, J., Stewart, B., Armer, J.: Comparison of upper limb volume measurement techniques and arm symptoms between healthy volunteers and individuals with known lymphedema. Lymphology **40**(1) (2007) 35–46
15. Taylor, R., Jayasinghe, U.W., Koelmeyer, L., Ung, O., Boyages, J.: Reliability and validity of arm volume measurements for assessment of lymphedema. Physical Therapy **86**(2) (2006) 205–214
16. Megens, A.M., Harris, S.R., Kim-Sing, C., McKenzie, D.C.: Measurement of upper extremity volume in women after axillary dissection for breast cancer. Archives of physical medicine and rehabilitation **82**(12) (2001) 1639–1644
17. Swedborg, I.: Voluminometric estimation of the degree of lymphedema and its therapy by pneumatic compression. Scandinavian journal of rehabilitation medicine **9**(3) (1976) 131–135
18. Trombetta, C., Abundo, P., Felici, A., Ljoka, C., Di Cori, S., Rosato, N., Foti, C.: Computer aided measurement laser (caml): technique to quantify post-mastectomy lymphoedema. In: Journal of Physics: Conference Series. Volume 383., IOP Publishing (2012) 012018
19. McKinnon, J.G., Wong, V., Temple, W.J., Galbraith, C., Ferry, P., Clynch, G.S., Clynch, C.: Measurement of limb volume: laser scanning versus volume displacement. Journal of surgical oncology **96**(5) (2007) 381–388
20. Vukotich, C., Geyer, M., Erdeljac, F.: Use of a laser scanning system to measure limb volume in chronic edema. ehabilitation Engineering & Assistive Technology Society of North America (RESNA) (2011)
21. Litomisky, K.: Consumer rgb-d cameras and their applications. University of California, Riverside. Ano (2012)
22. exercises after breast reconstruction surgery using muscle from your back — Cancer Research UK: http://www.cancerresearchuk.org/about-cancer/type/breast-cancer/treatment/surgery/reconstruction/exercises-after-breast-reconstruction-using-back-muscle
23. Clé, P.G.V., Tasso, L.E., Barbosa, R.I., Fonseca, M.d.C.R., Elui, V.M.C., Roncaglia, F.B., Mazzer, N., Barbieri, C.H.: Estudo retrospectivo do estado funcional de pacientes com fratura do rádio distal submetidos à osteossíntese com placa lcp. Acta fisiátrica **18**(4) (2011)
24. Galna, B., Barry, G., Jackson, D., Mhiripiri, D., Olivier, P., Rochester, L.: Accuracy of the microsoft kinect sensor for measuring movement in people with parkinson's disease. Gait & posture **39**(4) (2014) 1062–1068
25. Huber, M., Seitz, A., Leeser, M., Sternad, D.: Validity and reliability of kinect skeleton for measuring shoulder joint angles: a feasibility study. Physiotherapy (2015)
26. Seeley, R., Tate, P., Stephens, T.: Anatomy & Physiology. McGraw-Hill (2007)
27. Perez, A., Larranaga, P., Inza, I.: Supervised classification with conditional gaussian networks: Increasing the structure complexity from naive bayes. International Journal of Approximate Reasoning **43**(1) (2006) 1–25
28. Somol, P., Pudil, P., Novovičová, J., Paclık, P.: Adaptive floating search methods in feature selection. Pattern recognition letters **20**(11) (1999) 1157–1163

Neurocognitive stimulation game:

Serious game with adaptive difficulty for stimulation and assessment

João Costa[1,2,3], Jorge Neto[2,3], Ricardo Alves[3]

[1]Faculdade de Engenharia da Universidade Porto, Portugal
[2]Instituto Superior de Engenharia do Porto, Portugal
[3]Games Interaction and Learning Technologies - GILT, Porto, Portugal,
jncosta@gmail.com

Abstract. The ageing process is naturally accompanied by impairment in people's cognitive processes. The European population ageing is a challenging task for the European social policy and for the mental health specialists. Emerging technologies can play an important role in the neurocognitive stimulation area as they possess characteristics that might reduce the anxiety levels of patients while participating in neurocognitive stimulation or assessment therapies. In particular, serious games provide a setup that can be explored to improve the easy access to neurocognitive stimulation and assessment environments, regardless of place and time, at a lower cost could reach more people than traditional methodologies. This paper presents a serious game aiming to analyse neurocognitive deficits and stimulate the players' deficitary neurocognitive processes according to their problems. This game is based on traditional neurocognitive psychotherapy for adults, mainly addressing the cognitive processes of attention and memory. The game will simulate real world scenarios, allowing a better generalization process due to ecological validity.

Keywords: Active Ageing, Neurocognitive Deficits, Neurocognitive Stimulation, Assessment, Serious Games, Unity, DDA.

1 Introduction

With the increasing number of the elderly in our population, it is possible to verify the consequent increase of the cognitive decline occurrence. Neurocognitive stimulation has been a highly approached research area for the past years, as it offers new opportunities for people with cognitive impairments. Several neurocognitive stimulation programs are implemented in medicinal context, with the aim of decelerating cognitive decline [1] and, consequently, improving the lifetime quality of the patients. Although, these programs have some limitations that might compromise the desired impact on the person's quotidian. These boundaries involve, for example, the lack of ecological validity and patients' low motivation, due to the high emotional pressure they feel on following these neurocognitive stimulation programs and assessments [3]. Ecological validity can be seen as a transitional phenomenon, which analyses the current behavior, within specific environments related to the real world, by using discrete and reliable research methods [2].

The use of serious games is turning into a remarkable resource, as it offers computerized alternatives to neurocognitive stimulation and assessment programs. By focusing our efforts on reducing the impact of the above-mentioned limitations, i.e., reinforcing the need for ecological validity and adapt the tasks' difficulty levels, there is a high probability that the results obtained from a serious game's neurocognitive stimulation program can show positive results [4, 5].

Although the few projects that use serious games try to solve these problems, none of them covers efficiently the core aspects, such as ecological validity, that we address. Current approaches usually focus on only computerizing the assessments and do not always consider the content of the program itself. In addition, there is no special attention given to the users' interaction.

Our methodology consists on developing a serious game, called SynapseToLife, which focus on the neurocognitive stimulation of the players, by making them perform several tasks, immersed in well-known scenarios, thus strengthening ecological validity. More importantly, the serious game will create a group of familiar daily life scenarios (e.g. kitchen) to the user, in order to allow an easier transfer of the stimulated cognitive abilities into the users' quotidian, given the ecological validity variable that we aim to approach [3, 6].

2 Problem description

With the increasing number of the elderly and the consequent incidence of cognitive decline associated, it becomes important to invest in mental health, to minimize the social and economic impacts of this phenomenon, promoting active ageing.

Since the neurocognitive deficits may be present about 20 years before the clinical diagnosis on dementias, such as the Alzheimer disease [7], it is necessary to develop more effective and motivating strategies of monitoring and stimulating people's cognitive abilities, allowing them to follow an healthier life style [8]. SynapseToLife will be able to perform an early intervention, which is of major importance in order to slow down possible pre-clinical manifestations of neurocognitive [9] deficits, which, consequently, will contribute for public health's cost reduction [10].

2.1 State of the Art

Serious games allow the monitoring and presentation of stimulus, capable of motivating the user [11] and which show a greater accessibility [12], presenting positive results [13].

The use of serious games has been increasingly referenced as an important resource for psychological assessment and intervention [14]. Showing positive results in multiple domains, such as prevention [15], rehabilitation [14], neurocognitive stimulation [16], assessment [17] and monitoring neurocognitive changes [18], leading to beneficial changes, when it comes to brain plasticity [19], changing the brain's structure [20] and facilitating the impact of neurocognitive stimulation on everyday functioning [6]. However, the use of serious games, with older people, is still in an early phase [21] and there is little information when it comes to the impact of these programs on the users' quotidian activities and on their quality life [22].

There are several available games in the market, which aim neurocognitive stimulation [23, 24]. These, however, are not specifically developed to target a certain population and, in most cases, are not supported by robust studies on ecological validity. Although there are empirical evidences of neurocognitive improvements, several games do not evaluate the impact of a serious game in patients' daily life and do not offer content, which benefits the generalization process of, trained tasks, to their daily reality.

3 Proposed Solution

In this work, we propose to develop the ACT-Age platform, which includes the serious game SynapseToLife, aiming to promote neurocognitive stimulation and assessment. The serious game will enable an easier transfer of the neurocognitive stimulation results to users' quotidian activities, by simulating real life scenarios and users' interactions with them. This will be supported by the ecological validity concept, previously outlined, and which will play a significant role when it comes to reduce the users' anxiety levels, consequently enhancing their motivation, while being cognitively stimulated, and increasing the efficiency of the neurocognitive stimulation's results.

Throughout the game, tasks, adapted from the neurocognitive stimulation programs traditionally implemented, will be presented to the users along with a calculated difficulty level. The purpose of the game's tasks, is to simulate real life situations where users need to evoke their cognitive processes. A dynamic difficulty adjustment (DDA) component will also be an essential tool to develop, as it is a powerful expert control system, capable of studying and interpreting users' performances, throughout the game, and adapting (controlling) the game and tasks' difficulty according to the users' cognitive capabilities, directly assisting each one of them.

3.1 Main Components

SynapseToLife is organized in four different scenarios, where each will simulate real world situations, as the user will need to go through them, while performing the intended stimulation tasks. By structuring the game flow in mini games, adapted to real life situations or problems, users will hardly percept they are performing stimulation tasks and will only worry about having fun by completing these random tasks.

Another important component to be developed is the DDA system, as it is crucial when it comes to adapt the mini games' difficulty to the user's cognitive ability. Its job involves analyzing the player's performance and assure that he or she keeps relaxed and in a concentrated state of mind, by constantly presenting challenges and rewarding them accordingly (game flow). Shortly, the DDA allows the automatic readjustment of the game's difficulty, based on the player's performance results. These results, along with all users' interactions while playing the game, will be recorded in a centralized server.

Later on, the expert will analyze these same results, by accessing this server. These actions are transformed into useful and careful information and, more specifically, in data that the expert needs in order to study the player's performance, such as the number of right answers, the number of attempts, the response times or the cognitive processes approached (e.g. memory).
Lastly, and before the game starts, a diagnostic test will be presented to the user with the aim of setting a baseline, enabling posterior comparisons and analyzing the evolution of the players' cognitive status. This way, it is possible to understand if neurocognitive stimulation occurred and if the game itself presents all the tools needed to perform this stimulation.

3.2 Game Structure

Figure 1 presents the relationship between the business concept (cognitive stimulation) and the project development. With this scheme, it becomes easier to visualize the interaction between the user and the expert.

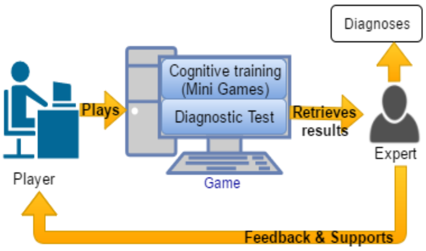


Fig. 1. Architectural Model.

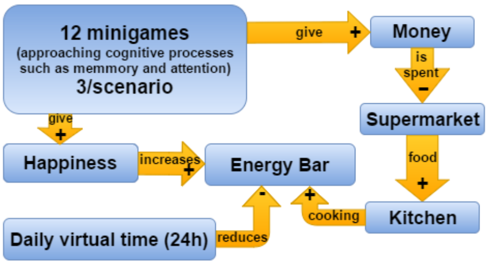


Fig. 2. Game Flow

Players will need to complete the mini games and a diagnostic test, which will be better approached further on. The final results are reported to the expert, to perform the diagnosis. Both the IT staff and the expert will provide assistance to the players, in case they have doubts related to the activities they will need to perform. In figure 2, it is possible to understand, more specifically, the game flow, which consists on the game's life cycle.

3.3 Game Scenarios

All scenarios, presented in figure 3, will be developed in 3D, by using the Unity3D platform. The tridimensional model was chosen due to the ecological validity factor. The more detail the game presents, the more the players become concentrated and the easier it is for them to understand the logic behind the tasks they perform throughout the game, which mainly focus on certain cognitive processes, like memory and attention. Each scenario is responsible for asking the completion of tasks related to that specific scenario. For example, in the Kitchen scenario, the player only performs kitchen related tasks, such as cooking, promoting stimuli on specific cognitive processes. The same applies to the other scenarios and this is the base of ecological validity, i.e., answer questions or performing tasks that are inserted in a given context and that can actually happen in a real life situation.

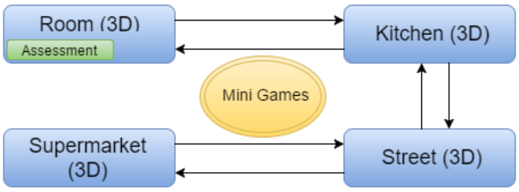


Fig. 3. Game scenarios.

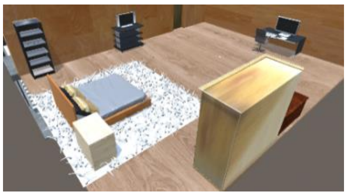


Fig. 4. Scenario example.

3.4 Dinamic Dificulty Adjustment component

It is proven that a person does not behave cognitively in the same way every day. In this way it is necessary to identify the person daily cognitive ability and try to make the adjustment as fast as possible so that the person does not lose interest in playing.

As the game progresses, the player will face more or less difficult situations (depending on his performance), so eventually the DDA component will find a balance, so that the player does not advance too much, to a greater difficulty when compared to the player's cognitive abilities, and get the best out of the stimulation performed.

The dynamic and automatic readjustment component of difficulty is based on the analysis of the player last results, so it is possible to assign the best possible difficulty to the mini game that the player is performing at a given moment. In order to improve accuracy, in assigning the difficulty, the last N results (hits and failures) of each mini game are analyzed, and the smaller the number of analyzed results (N), the more weight each result will have in the algorithm final decision.

When invoked, the component is responsible for analyzing these last results and assigns a difficulty to the game during runtime. The same procedure applies to the various phases of the mini games. Since each mini game has several stages, each stage invokes the DDA component and gets the best possible difficulty to the current phase of the mini game.

It is simple to demonstrate the process of changing the difficulty during the game. Exemplifying through a minigame, the difficulty increases or decreases depending on the last attempts of the player. The component is invoked before each stage of the mini game, so the difficulty changes throughout the mini game.

If the player is able to complete the current task without fail, he will find himself, in the next mini game, with a difficulty possible greater than the previous one applied. In this way, it is interpreted that the player does not have so many cognitive difficulties and, therefore, manages to surpass the more advanced levels of the following mini game. If it does not, the difficulty can decreases, establishing a favourable balance between the player cognitive abilities and the tasks presented to him.

The diagram illustrated in Figure 5, demonstrates the interaction process between the game and the DDA component. As described the communication always starts from the mini games ending in the DDA component.

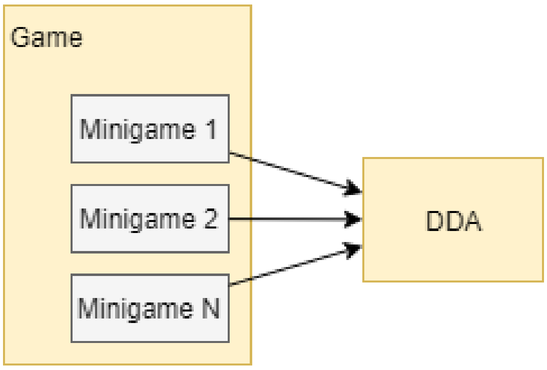


Fig. 5. Interaction between game and DDA.

The DDA is created only once (Singleton pattern), being passed by parameters, all the mini games that the current game contains. All the mini games implement an interface called *IGame*, in this way, the interaction between the DDA component and any of the mini games can be performed through the interface, allowing the decoupling between the DDA and the Game itself. The classes that represent the mini games implement the *IGame* interface and are responsible for storing the information in the database, promoting decoupling between the DDA component and the database. The DDA is therefore a component that presents only the coupling with an interface (low coupling) and is responsible for determining which next mini game and with what difficulty it must execute (high cohesion), and, in this way, can be reused for any game that works with mini games, missions or any task that cam implement the *IGame* interface. Another feature that facilitates this reuse of the DDA, is that this component is implemented according to a client / server logic, so communication always starts in the game (client), simplifying the whole process of interaction, like is shown on Figure 5 all mini games call the DDA component.~

The entire process is described by:

- The game makes a request to the DDA component, asking what is the next mini game to execute, and start it.
- Once the mini game is started, it invokes the DDA component, asking him the difficulty with which he must execute, the DDA decision is based on the data available through the interface.
- Finally, the mini game ends, sending its result to the class that represents the mini game, that is, the representative class that implements the mentioned interface, which is responsible for storing this same result in the internal database.

4 Expected Outcomes

After the project development phase, the users and the serious game will be both evaluated, in different ways. The game must be able to retrieve, from the users' interaction with the game, information needed for the expert, in order for him to draw his conclusions. The serious game will also need to be able to analyze the players' responses throughout the game, since the tasks' difficulty levels must be adapted according to the user needs and since it is essential for us to study these same responses and observe their interaction. During this assessment phase, users' responses will be assessed and carefully analyzed, i.e., the extracted and quantified results, from the neurocognitive stimulation program performed, will be studied.

4.1 Output Model

The players' actions are directly connected to their performance, i.e., there is a set of features that help us understand how the user performance during the game or how this same user is dealing with it. The variables (e.g. right answers) together, form this output model so that the expert can analyze the players' cognitive ability in the best way possible. The output model, in turn, will be allocated in the available server.

However, the game has an error-free structure, which means that the player will not advance to the next phase without understanding what he or she did wrong and without knowing how can the current problem must be solved. This way, the neurocognitive stimulation has more impact on the player, as the probability of him solving an equal problem, in the future, increases.

4.2 Assessment Plan

The assessment will be based on a study with a quasi-experimental design, where an independent variable (cognitive training) is manipulated, verifying its effect on dependent variables (e.g. attention). Taking this into account, and shortly, the assessment process consists in the following steps:

- Interview and select the most suitable participants for the experiment, using the previously defined restrictions.
- Compare the pre-assessment results with the post-assessment results, analyzing all data obtained from the game and the neurocognitive assessment;
- Post-Post assess the participants after a few weeks or months, in order to confirm if there was retention of the cognitive skills acquired.

After the project development, six weeks will be reserved for this final assessment process, conducted by the expert and supported by the IT staff, in order to give a scientific answer to the work accomplished.

4.3 Main Contributions

Overall, the outcomes expected are based on the cognitive assessment itself, i.e., the game should present better results when compared with classic neurocognitive stimulation programs. The project ought to also show improvements on the interaction between the user and the computer, as efforts will be made related to this situation. Furthermore, it is essential making the system able to be sensitive to the players' interaction with the game's environment (e.g. verify if the tasks are having the correct reactions) and adapt its difficulty accordingly. We also expect to confirm that our serious game has ecological validity, since it has a preponderant role in the project's main purpose.

5 Conclusions

The serious game proposed in this work aims to overcome the limitations of current solutions addressing active ageing. It is based on a set of scenarios simulating familiar environments and everyday activities of an individual's daily life. This allows stimulating users' cognitive abilities and transferring the stimulation results to the tasks normally performed during their quotidian. The user interaction with the game, while playing it for stimulation, will be monitored and recorded in a cognitive model to allow guiding the effective stimulation towards deficitary cognitive processes. This model will also allow the neuropsychological profile monitoring and an effective intervention with ecological impact. We expect this serious game to become a robust tool being able to study, interpret and stimulate users' neurocognitive processes.

We aim to provide our future players a welcoming environment, by carefully analyzing which content is necessary to insert inside the game, in order to promote ecological validity, which is our main focus during the development phase, and by developing a tutorial which will help the player to get acquainted to the gameplay. Moreover, the DDA component must be able to analyze every situation possible and every user interaction throughout the game. It is also necessary to consider the interaction itself.

Thus, we seek to develop a solid serious game, in a technological and scientific point of view, which will promote the development of the knowledge necessary for the implementation of this technology in new markets with potential growth, aiming several domains, such as health, well-being, ageing and social inclusion. This project also seeks to give answers to questions based on the implementation, experimentation challenges, quality control in application domains and impact on users' quality life. These scientific evidences will allow a safer investment from the digital games' industry professionals in developing the "serious" market of games. Another important activity intends to disclose the project and its results.

Acknowledgments

The serious game described in this article was developed by Jorge Neto and João Costa (authors) while working on their master's degrees. The work was developed under the supervision of Ricardo Alves, Nuno Escudeiro and Paula Escudeiro.
The authors would also like to acknowledge the support in terms of infrastructures and technologies from GILT laboratories located in ISEP.

6 References

1. Glisky, E. (2007). Changes in Cognitive Function in Human Aging. In D. R. Riddle (Ed.), Brain Aging: Models, Methods, and Mechanisms. Boca Raton: CRC Press.
2. Rizzo, A., & Kim, G. J. (2005). A SWOT Analysis of the field of VR rehabilitation and therapy. Presence Teleoper. Virtual Environ. 14, 119–146. DOI= 10.1162/1054746053967094
3. Segal, R., Bhatia, M. & Drapeau, M. (2011). Therapist´s Perception of Benefits and Costs of Using Virtual Reality Treatments. Cyberpsychology, Behavior and Social Networking, 14 (1-2), 29-34.
4. Fissler, P., Kolassa, I.T., & Schrader, C. (2015) Educational games for brain health: revealing their unexplored potential through a neurocognitive approach. Front. Psychol. 6:1056. doi: 10.3389/fpsyg.2015.01056
5. Powers, K., Brooks, P., Aldrich, N., Palladino, M., and Alfieri, L. (2013). Effects of video-game play on information processing: a meta-analytic investigation. Psychon. Bull. Rev. 20, 1055–1079. doi: 10.3758/s13423-013- 0418-z
6. Spector, A., Orrell, M., & Woods, B. (2010). Cognitive stimulation to improve cognitive functioning in people with dementia. International Journal of Geriatric Psychiatry, 25(12), 1253–1258.
7. Rajan, K., Wilson, R., Weuve, J. (2015). Cognitive impairment 18 years before clinical diagnosis of Alzheimer disease dementia. Neurology published online June 24, 2015 DOI: 10.1212/WNL.0000000000001774
8. Mcguire, A.M., Anderson, D.J., Fulbrook, P. (2013). Perceived barriers to healthy lifestyle activities in midlife and older Australian women with type 2 diabetes. Collegian. http://dx.doi.org/10.1016/j.colegn.2013.07.001
9. American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing.
10. Direção-Geral da Saúde (2013). Portugal Saúde Mental – 2013. Programa Nacional para a Saúde Mental. Obtido em 7 de 2 de 2015, de Direcção-Geral da Saúde: http://www.dgs.pt/publicacoes.
11. Manera, V., Petit, P., Derreumaux, A., Orvieto, E., et al., (2015). 'Kitchen and cooking,' a serious game for mild cognitive impairment and Alzheimer's disease: a pilot study. Frontiers in Aging Neuroscience, 7:24. DOI: 10.3389/fnagi.2015.00024
12. Krebs, P., Prochaska, J.O., Rossi, J.S., (2010). A meta-analysis of computer-tailored interventions for health behavior change. Preventive Medecine. 51: 214–221.
13. Connolly, T., Boyle, E., MacArthur, E., et al., (2012). A systematic literature review of empirical evidence on computer games and serious games, Computers & Education, 59 (2): 661-686. DOI: 10.1016/j.compedu.2012.03.004
14. Robert, P. H., König, A., Amieva, H., et al. (2014). Recommendations for the use of Serious Games in people with Alzheimer's Disease, related disorders and frailty. Frontiers in Aging Neuroscience, 6: 54. DOI= 10.3389/fnagi.2014.00054

15. Wiemeyer, J., & Kliem, A. (2012). Serious games in prevention and rehabilitation – a new panacea for elderly people? European Review of Aging Physical Activity, 9: 41–50. DOI: 10.1007/s11556-011-0093-x
16. Nouchi, R., Taki, Y., Takeuchi, H., Hashizume, H., Akitsuki, Y., et al. (2012) Brain Training Game Improves Executive Functions and Processing Speed in the Elderly: A Randomized Controlled Trial. PLoS ONE 7(1): e29676. doi:10.1371/journal.pone.0029676
17. Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, et al. (2013). Video game training enhances cognitive control in older adults. Nature 501, 97–102. DOI=10.1038/nature12486.
18. Tarnana, I., Papagiannopoulos, S., Kazis, D., Wiederhold, M., Widerhold, B., Tsolak, M. (2015). Reliability of a novel serious game using dual-task gait profiles to early characterize aMCI. Front. AgingNeurosci.7:50, 1-15. doi: 10.3389/fnagi.2015.00050
19. Lövdén, M., Bäckman, L., Lindenberger, U., Schaefer, S., and Schmiedek, F. (2010). A theoretical framework for the study of adult cognitive plasticity. Psychol. Bull. 136, 659–676. doi: 10.1037/a0020080
20. Kühn, S., and Gallinat, J. (2014). Amount of lifetime video gaming is positively associated with entorhinal, hippocampal and occipital volume. Mol. Psychiatry 19, 842–847. doi: 10.1038/mp.2013.100
21. Tarnanas, I., Tsolaki, M., Nef, T., et al. (2014). Can a novel computerized cognitive screening test provide additional information forearly detection of Alzheimer's disease? Alzheimers Dement. 10,790–798. DOI= 10.1016/j.jalz.2014.01.002.
22. Kazmi, S., Ugail, H., Valerie, L., Palmer, I. (2014). Interactive Digital Serious Games for the Assessment, Rehabilitation, and Prediction of Dementia. International Journal of Computer Games Technology Volume 2014, Article ID 701565, 1-11. http://dx.doi.org/10.1155/2014/701565
23. Simon, M., Costas, B. (2013). Dementia Games: A Literature Review of Dementia-Related Serious Games. Serious Games Development and Applications - Lecture Notes in Computer Science ; Volume 8101. p. 15-27. Springer Publishing. The final publication is available at: http://link.springer.com/chapter/10.1007%2F978-3-642-40790-1_2
24. Arambarri, J., Torre, I., Coronado, M., Álvarez, I. (2014). Investigating the Potential market of a Serious Game for Training of Alzheimer's Caregivers in a Northern Spain region. International Journal of Serious Games Volume 1, Issue 4, October 2014 ISSN: 2384-8766 Doi: http://dx.doi.org/10.17083/ijsg.v1i4.36

Student-Centered Learning Environments for Self-Regulated Project-Based Learning in Higher Education: Qualification and Selection Study

Mohamed Yassine Zarouk[1] and Mohamed Khaldi[2]

[1] Faculty of Engineering of the University of Porto, Portugal
[2] Faculty of Science of Abdelmalek Essaadi University, Tetouan, Morocco

[1] up201610550@fe.up.pt; [2] medkhaldi@yahoo.fr

Abstract. Taking into account the remaining baffling problem for many online pedagogical designers to choose such a Learning Management Systems (LMS) as a Student-Centered Learning Environment (SCLE), because of confusing large landscape of sophisticated educational tools and strategies offered by these systems as well. Hence, in this paper we present a Qualification and Selection study of LMS used as Student-Centered Learning Environments fostering Self-Regulated Learning in case of Project-Based Learning suitable for higher educational context. The aim of this study is helping online pedagogical designers to qualify, compare and select the convenient LMS used as SCLE according to their specific considerations, by adopting a flexible selection/comparison mode based on the rating and weighting of a set of preliminary defined generic and specific criteria. Although, the study does not cover all of the most popular LMS but it remains applicable as a general method for qualifying and selecting such a Learning Management System.

Keywords: Qualification and Selection Method, QS Method, Student-Centered Learning Environment, Learning Management System, Self-Regulated Learning, Project-Based Learning, Self-Regulated Project-Based Learning, Higher Education.

1 Introduction

Self-Regulated (SRL), Project-Based (PBL) and Inquiry-Based Learning among others are innovative pedagogical approaches fostering a multitude of critical strategies for success in the twenty-first century. Students drive their own learning through inquiry, as well as work collaboratively to research and create projects that reflect their knowledge [1]. Moreover, Student-Centered Learning Environments (SCLE) could have potential to serve as fun and inspiring workshop settings, where students would engage in exciting Project-Based activities that integrate required curriculum material, while also simulating some aspects of real world “epistemic” contexts, challenging students to gain a richer understanding of learning material and processes in a more situated and relatable way [2].

Nevertheless, the implementation of an effective SCLE is still a baffling problem for online pedagogical designers despite the large landscape of sophisticated educational tools and strategies offered by the most of the Learning Management Systems (LMS) [3], [4].

Accordingly, this paper presents an ongoing work of the implementation of a proposed integrated framework for Self-Regulated Project-Based Learning (SRPBL) suitable for SCLE in higher education. For this purpose, we propose a Qualification/Selection study (QS) of the usage of LMS as SCLE.

Hence, in this sense we posed the following main research hypothesis:

- What are the LMS features suitable for SCLE?
- What are the required criteria (generic and specific) for QS study?
- How could we flexibly compare and select a LMS according to environmental, pedagogical and institutional considerations?

Finally, for this purpose the paper is structured as following: "material and methods" section, in which general process, necessary materials and data of the QS study are presented. Then, the discussions of obtained results exposed in forms of radar diagrams.

2 Material and Methods

This study is inspired from Selection and Qualification of Open Source Software (QSOS) method [5]. In this case, our proposed QS method (Qualification/Selection) is applied for E-Learning LMS.

As a preparatory phase, we pre-select seven LMS for the study (**Table 1**), based on their popularity and compatibility with our required criteria: PBL and higher education support (see **Webography**).

Table 1. Pre-selected LMS "Learning Management Systems" for QS study.

LMS		Version	Licence	Website
moodle	Moodle	3.4+	Free	https://moodle.org/
canvas	Canvas	9	Free	https://www.canvaslms.com/
Sakai	Sakai	11	Free	https://www.sakaiproject.org/
Bb Blackboard	Blackboard	9.1	Paid	http://www.blackboard.com/
Desire2Learn	D2L	10.6+	Paid	https://www.d2l.com/
docebo	Docebo	7.3	Paid	https://www.docebo.com/
efront	e-Front Pro	5	Paid	https://www.efrontlearning.com/

2.1 The "Qualification and Selection" Method

Generally, the choice to opt for software as a component of its information system, whether this System is free or paid, is based on the analysis of needs and constraints (technical, functional and strategic) and the adequacy of the software to these needs and constraints [6]. In this study, they are represented in two main groups: generic and specific criteria (**Fig.2**).

However, since it is envisaged to study the adequacy of different LMS, it is necessary to have a method of qualification and selection adapted to the specificities of this type of environments.

The general QS process is composed of several interrelated steps.

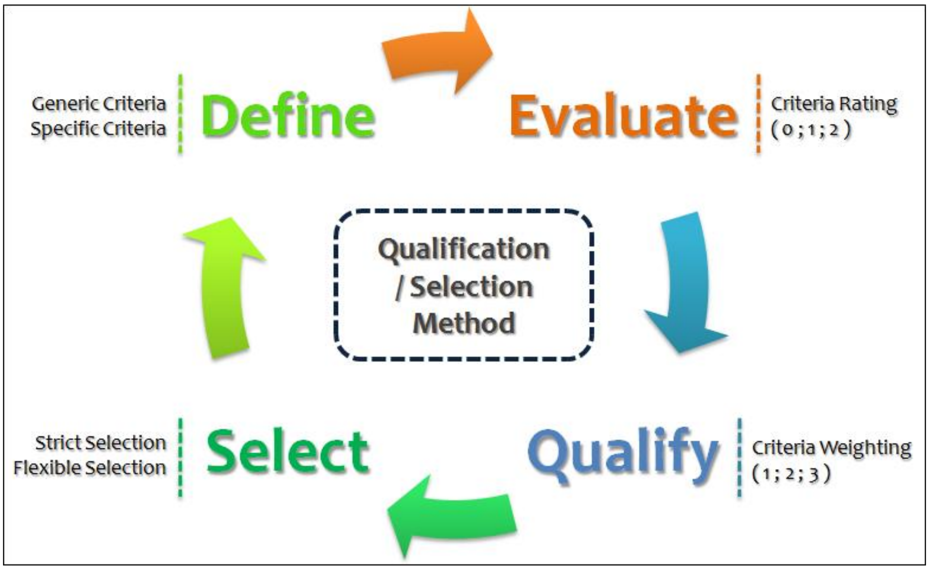


Fig.1. Processes of Qualification and Selecion (QS) method.

The general process presented can be applied with different granularities. This makes it possible to adapt to the level of detail desired in the qualification and selection process as well as to proceed by iterative loops for each of the four steps.

Briefly, the QS Method takes place in following four stages:

- Define LMS generic/specific criteria;
- Evaluate LMS criteria (rating);
- Qualify LMS criteria (weighting);
- Compare and Select a LMS (flexible selection).

2.2 Define

The objective of this step is to define different elements of typology reused by the three next steps in the overall process. First, we listed the different necessary LMS criteria for our study, and then we classified them into two main groups: generic and specific criteria (**Fig.2**).

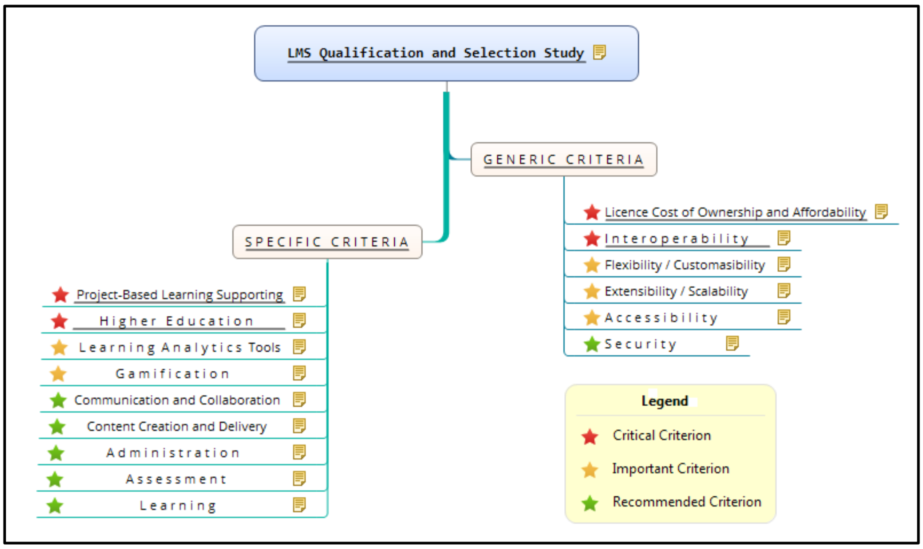

LMS Qualification and Selection Study
GENERIC CRITERIA
Licence Cost of Ownership and Affordability
Interoperability
Flexibility / Customasibility
Extensibility / Scalability
Accessibility
Security
SPECIFIC CRITERIA
Project-Based Learning Supporting
Higher Education
Learning Analytics Tools
Gamification
Communication and Collaboration
Content Creation and Delivery
Administration
Assessment
Learning
Legend
Critical Criterion
Important Criterion
Recommended Criterion

Fig.2. LMS generic and specific criteria.

Generic criteria

Generic criteria define the general functional aspect of usage of any E-Learning LMS such as "Affordability and Ownership", "Interoperability", "Flexibility and Customizability", "Extensibility and Scalability", "Accessibility and Security" (**Fig.2**). In addition, some criteria are defined by a set of different items and the rating of this kind of criteria is the rating average of the included items (**Appendices 1; 2**).

Specific criteria

These are the criteria defined by the specific pedagogical and technical of our case study. Consequently, nine criteria (**Fig.2**) have been defined, supporting Self-Regulated Based Learning principles [7], Project-Based Learning, Gamification mechanics and Learning Analytics tools among others.

2.3 Evaluate

The objective of this step is to proceed to the evaluation of the LMS by assigning the marks to each criterion previously defined.
For each criterion of the grid, the rating rule is as follows:

- 0: not covered;
- 1: partially covered;
- 2: completely covered;

NB. / In case of complex criteria such as “Communication and Collaboration”, “Interoperability” which is composed of a set of items; the criterion rating is the average value of its rated items.

In addition, there an exception for the criterion “Affordability and Ownership” rating, it rated only by 1 or 2. Because, the study proceed free LMS and paid LMS separately, then it doesn’t make sense to mention if the LMS is affordable in both cases but it reflect only the complexity of affordability.

2.4 Qualify:

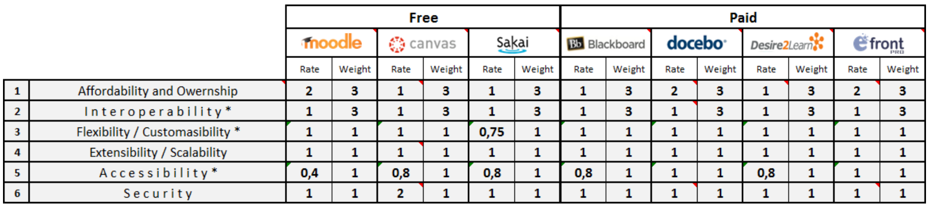
The objective of this step is to define a set of elements reflecting the needs and constraints related to the process of selecting an LMS. This is to be considered in the context of the use of the product, so as to obtain a "Select" of the general process.

Thus, each feature (Criterion) of the functional axis is assigned a requirement level, selected from the following:

- Critical functionality with weighting: 3;
- Important functionality with weighting: 2;
- Required functionality with weighting: 1;

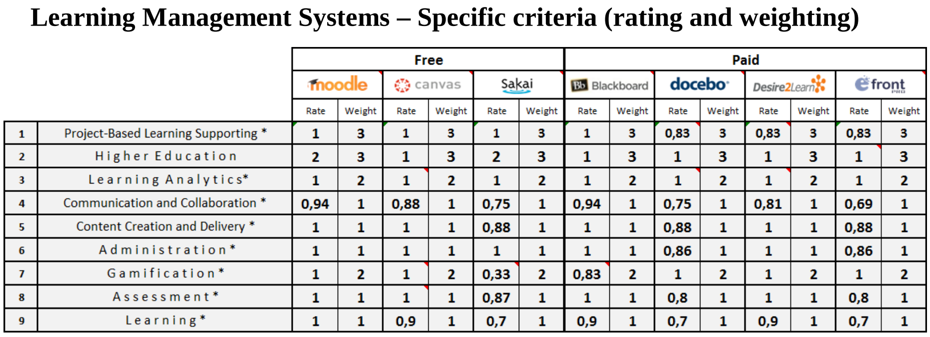
These requirements will be associated with weighting values in step “Select” depending on the chosen selection mode. Figures afterwards (**Fig.3** and **Fig.4**), the rating and weighting grids of the generic and specific LMS criteria:

Learning Management Systems – Generic criteria (rating and weighting)



		Free						Paid							
		moodle		canvas		Sakai		Blackboard		docebo		Desire2Learn		efront	
		Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight
1	Affordability and Owernship	2	3	1	3	1	3	1	3	2	3	1	3	2	3
2	Interoperability *	1	3	1	3	1	3	1	3	1	3	1	3	1	3
3	Flexibility / Customasibility *	1	1	1	1	0,75	1	1	1	1	1	1	1	1	1
4	Extensibility / Scalability	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	Accessibility *	0,4	1	0,8	1	0,8	1	0,8	1	1	1	0,8	1	1	1
6	Security	1	1	2	1	1	1	1	1	1	1	1	1	1	1

Fig.3. LMS generic criteria - rating and weighting.



Learning Management Systems – Specific criteria (rating and weighting)

		Free						Paid							
		moodle		canvas		Sakai		Blackboard		docebo		Desire2Learn		front	
		Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight
1	Project-Based Learning Supporting *	1	3	1	3	1	3	1	3	0,83	3	0,83	3	0,83	3
2	Higher Education	2	3	1	3	2	3	1	3	1	3	1	3	1	3
3	Learning Analytics*	1	2	1	2	1	2	1	2	1	2	1	2	1	2
4	Communication and Collaboration *	0,94	1	0,88	1	0,75	1	0,94	1	0,75	1	0,81	1	0,69	1
5	Content Creation and Delivery *	1	1	1	1	0,88	1	1	1	0,88	1	1	1	0,88	1
6	Administration*	1	1	1	1	1	1	1	1	0,86	1	1	1	0,86	1
7	Gamification*	1	2	1	2	0,33	2	0,83	2	1	2	1	2	1	2
8	Assessment*	1	1	1	1	0,87	1	1	1	0,8	1	1	1	0,8	1
9	Learning*	1	1	0,9	1	0,7	1	0,9	1	0,7	1	0,9	1	0,7	1

Fig.4. LMS specific criteria - rating and weighting.

2.5 Select:

The objective of this step is to select the LMS corresponding to the needs of the user taking into considerations the defined generic and specific criteria, or more generally to compare LMS of the same context as this case.

Two modes of selection are possible: strict selection and flexible selection.

Strict selection

Strict selection is based on a direct elimination process as soon as a software/LMS program does not meet the requirements of the step. This mode is very selective and can, depending on the level of requirement of the user, not return any eligible LMS.

A global score is then assigned to each LMS that has passed by the weighting, in the same way as in the flexible selection.

Flexible selection

This method is less strict than the previous one because instead of eliminating non-eligible software at the level of the generic or specific functional coverage, it simply classifies them while measuring the difference observed with respect to the filters defined above.

It is based on weighting values whose allocation rules are detailed in the previous step (Required, Important, and Critical).

Comparison

The software of the same domain can also be compared with each other according to the weighted scores obtained during the previous steps.

Thereafter, results are presented in the form of different radar diagrams, then we adopted a flexible selection and comparison mode which allows to select the convenient LMS according the pedagogical, environmental and institutional considerations without eliminating others systems.

3 Results and discussions

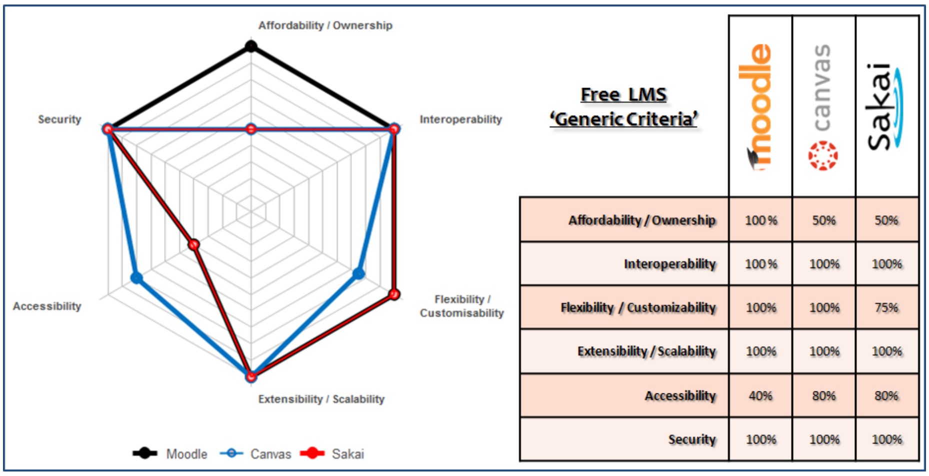
The results of this qualification and selection study are often presented by different radar diagrams for the purpose of offering a more global and holistic view of all the groups of the previously defined, evaluated and weighted criteria.

Hence, below, six different radar diagrams are presented:

- Free LMS generic criteria;
- Free LMS specific criteria;
- Free LMS global criteria;
- Paid LMS generic criteria;
- Paid LMS specific criteria;
- Paid LMS global Criteria;

In addition, all the criteria rating and weighting are formulated relatively in percentage values for the aim to well clarify and help to relatively compare different criteria produced in the diagrams.

3.1 Free Learning Management Systems - Free LMS generic criteria



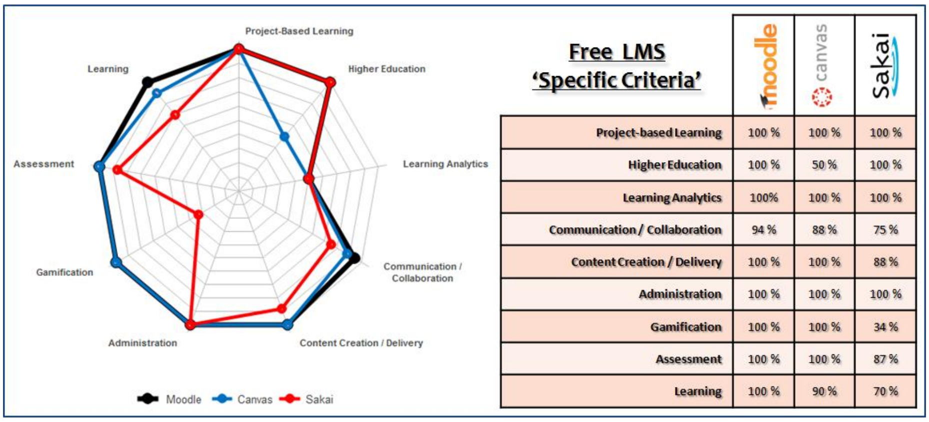
Free LMS 'Generic Criteria'	moodle	canvas	Sakai
Affordability / Ownership	100 %	50%	50%
Interoperability	100 %	100%	100%
Flexibility / Customizability	100%	100%	75%
Extensibility / Scalability	100%	100%	100%
Accessibility	40%	80%	80%
Security	100%	100%	100%

Fig.5. Radar diagram of free LMS generic criteria.

The radar diagram (**Fig.5**) shows a general equality regarding the general criteria of free LMS, with a benefit to the Moodle regarding "Affordability and Ownership", since it is completely free and open source and affordable to download directly from the official website without prerequisites (registration ...) as the case of Canvas and Sakai, but Moodle remains poorly accessible in terms of "Easy to use", "Easy to maintain" criteria relative to others.

In addition, we note that Sakai has the advantage of being more valuated in terms of accessibility taking into consideration its "Easy to use" and an active, widespread and effective support community.

3.2 Free LMS specific criteria



Free LMS 'Specific Criteria'	moodle	canvas	Sakai
Project-based Learning	100 %	100 %	100 %
Higher Education	100 %	50 %	100 %
Learning Analytics	100%	100 %	100 %
Communication / Collaboration	94 %	88 %	75 %
Content Creation / Delivery	100 %	100 %	88 %
Administration	100 %	100 %	100 %
Gamification	100 %	100 %	34 %
Assessment	100 %	100 %	87 %
Learning	100 %	90 %	70 %

Fig.6. Radar diagram of free LMS specific criteria.

Based on the specific criteria, Moodle remains a reference of Free LMS, thanks to a strong community of documentation and development while taking advantage a completely free and open source, which allows offering a very rich list of functionalities, basic features, plugins to download and a considerable available database of courses and case studies.

Furthermore, we could observe that Canvas LMS is rapidly getting forwards as a serious alternative to Moodle seeing it is already covering enough all important specific criteria such as "Gamification" and "Learning Analytics".

3.3 Free LMS global criteria

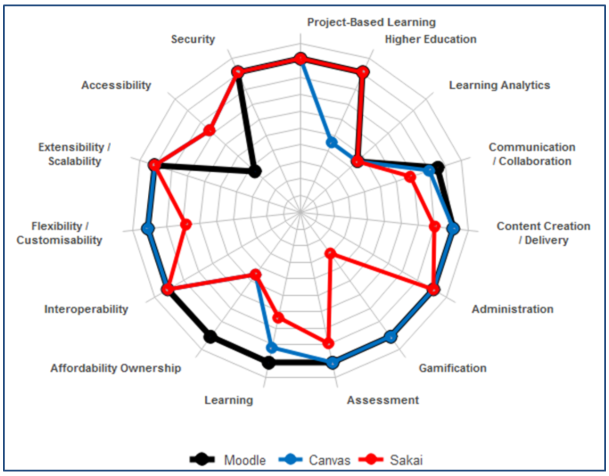
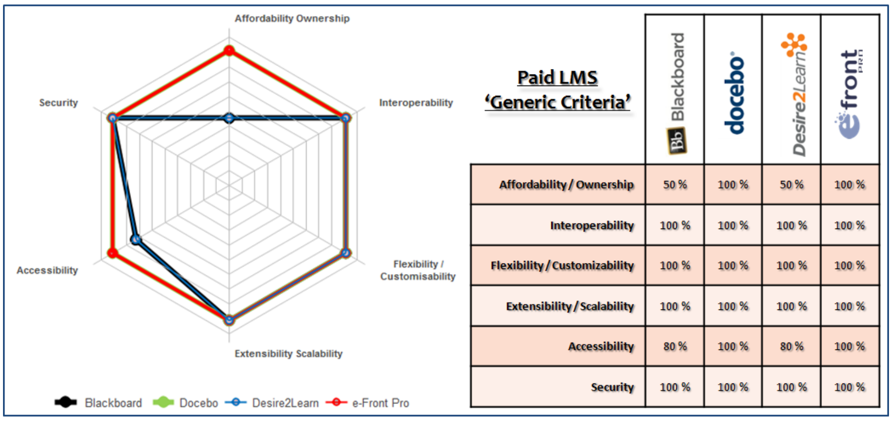


Fig.7. Radar diagram of free LMS global criteria.

Globally, Moodle, Sakai or Canvas remain very close in terms of functionalities and the selection differs only for some details, in terms of "Affordability and Ownership", “Accessibility” and “Gamification”. And in this case, the selection depends on the considerations defined by the user.

Accordingly, we could conclude that comparison between there three LMS in mainly determined by “Affordability and Ownership” criteria in addition to “Accessibility” and “Gamification” as important criteria.

3.4 Paid LMS generic criteria



Paid LMS 'Generic Criteria'	Blackboard	docebo	Desire2Learn	efront PRO
Affordability / Ownership	50 %	100 %	50 %	100 %
Interoperability	100 %	100 %	100 %	100 %
Flexibility / Customizability	100 %	100 %	100 %	100 %
Extensibility / Scalability	100 %	100 %	100 %	100 %
Accessibility	80 %	100 %	80 %	100 %
Security	100 %	100 %	100 %	100 %

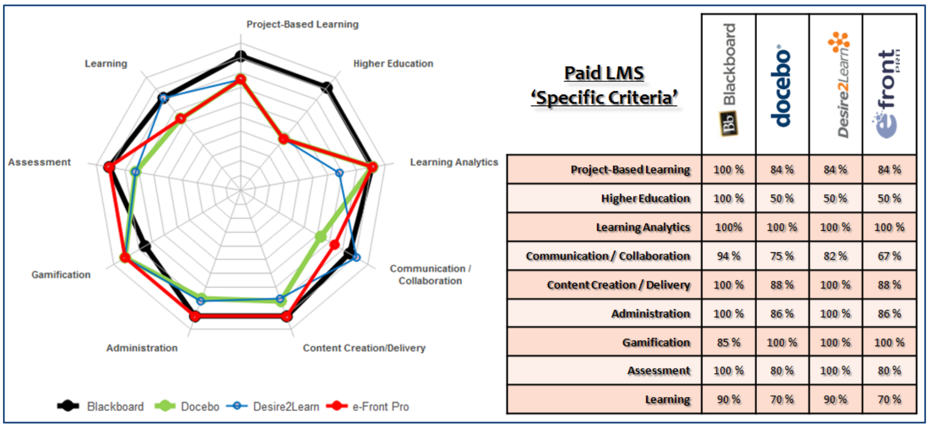
Fig.8. Radar diagram of paid LMS generic criteria.

At glance, we could notice that the given four LMS are divided in two groups taking in account “Affordability and Ownership” as remarkable criterion.

Regarding the paid LMS in terms of generic criteria, the difference remains in "Affordability and Ownership", and this criterion determines price to buy the license and also the procedure to obtain the necessary information to decide, and in this point Desire2Learn is the month discreet since it is difficult to know the prices before complicated required procedure.

We noted also that Docebo represents a serious competitor to Blackboard for its accessibility and affordability.

3.5 Paid LMS specific criteria



Paid LMS 'Specific Criteria'	Bb Blackboard	docebo	Desire2Learn	efront PRO
Project-Based Learning	100 %	84 %	84 %	84 %
Higher Education	100 %	50 %	50 %	50 %
Learning Analytics	100%	100 %	100 %	100 %
Communication / Collaboration	94 %	75 %	82 %	67 %
Content Creation / Delivery	100 %	88 %	100 %	88 %
Administration	100 %	86 %	100 %	86 %
Gamification	85 %	100 %	100 %	100 %
Assessment	100 %	80 %	100 %	80 %
Learning	90 %	70 %	90 %	70 %

Fig.9. Radar diagram of free LMS specific criteria.

Regarding the specific criteria, Blackboard is considered the dominant, in terms of features, documentation. It has a place similar to that of Moodle for free LMS. Despite such a costly affordability.

In the second, we remark e-Front in its Pro version has succeeded to achieve a drastic change in functionality, design compared to previous versions.

For Docebo, weaknesses are that it is more business oriented than Education despite their sophistical features in terms of Project Management and Learning Analytics.

3.6 Paid LMS global criteria

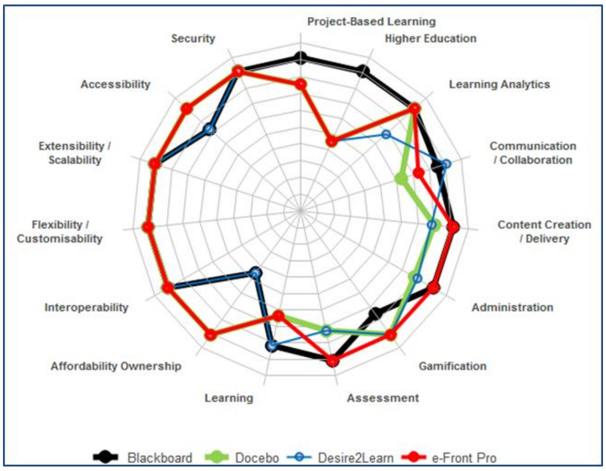


Fig.10. Radar diagram of free LMS global criteria.

Overall, the comparison between the four paid LMSs showed obvious performance and superiority for Blackboard's first and e-Front Pro as a very powerful LMS in terms of functionalities and a continuously evolving user community.

Finally, again despite the remarks quoted before, the choice is very flexible and personalized; because we think that the different LMS are more than enough for the implementation of the Student-Centered Learning Environment. Only that they have specific and tiny considerations to bear in mind.

4 Conclusions

The main objective of this study is to offer a handout for the qualification and flexible selection of LMS to different stakeholders contributing on the implementation of e-Learning case studies fostering Self-Regulated Learning in case of Project-Based Learning, using recent trends of e-Learning such as, Gamification, Portfolio, Learning Analytics… in the higher educational context.

Therefore, the aim of this paper is not to select the best LMS to use, but simply it helps to qualify, compare and flexibly select the convenient Learning Managements Systems taking into consideration different pedagogical, Environmental and institutional requirements based on two detailed groups of criteria (Appendices 1 and 2).

In addition, the study does not cover all of the most popular LMS but remains applicable as a general method for qualifying and selecting such a Learning Management System.

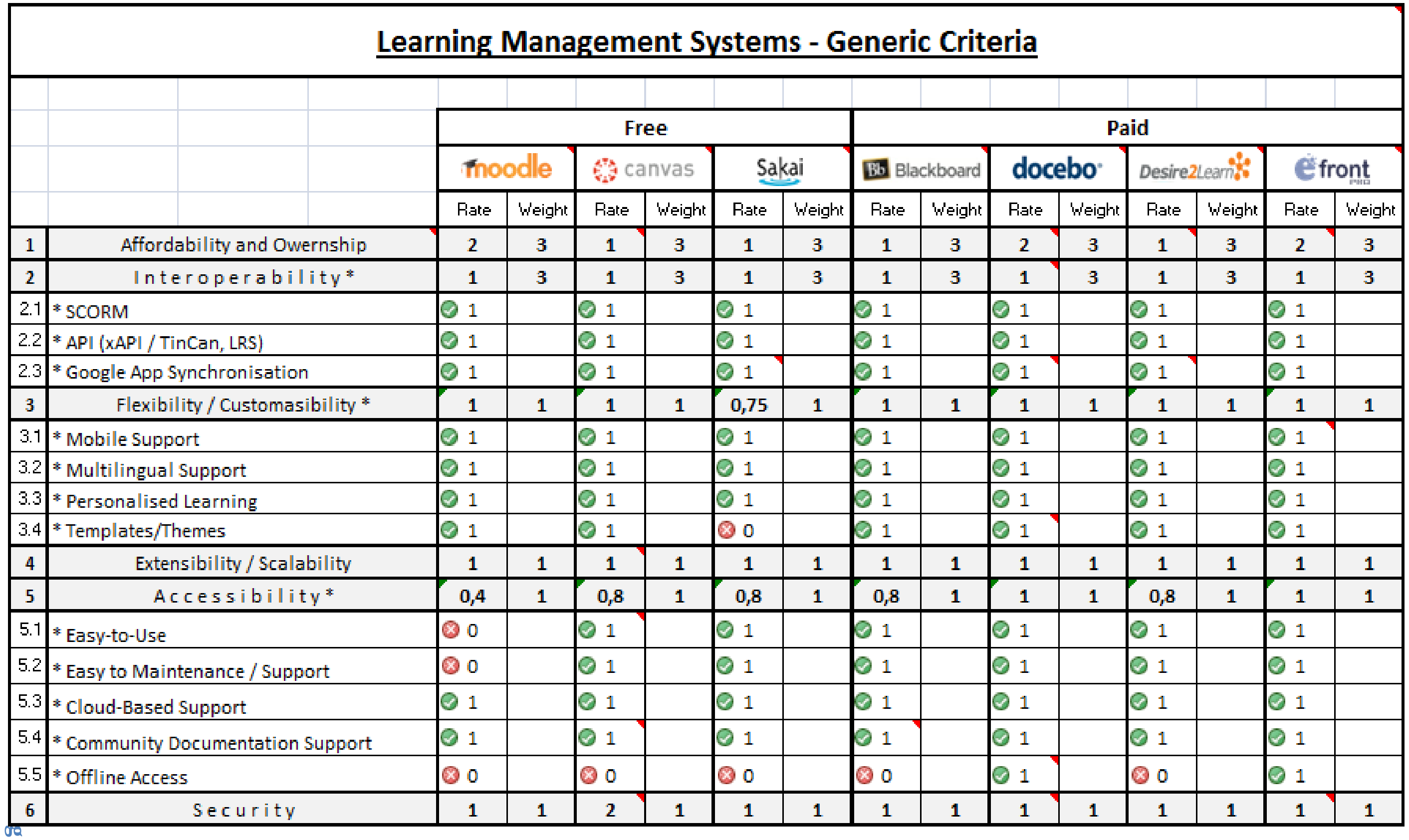
References

1. S. Bell, ‘Project-based learning for the 21st century: Skills for the future’, The Clearing House, vol. 83, no. 2, pp. 39–43, 2010.
2. S. K. W. Chu, R. B. Reynolds, N. J. Tavares, M. Notari, and C. W. Y. Lee, 21st Century Skills Development Through Inquiry-Based Learning: From Theory to Practice. Springer, 2016.
3. J. D. Basham, S. Stahl, T. Hall, and R. A. Carter Jr, ‘Establishing a Student-Centered Environment to Support All Learners’, in Handbook of Research on Classroom Diversity and Inclusive Education Practice, IGI Global, 2017, pp. 155–182.
4. M. J. Hannafin and S. M. Land, ‘The foundations and assumptions of technology-enhanced student-centered learning environments’, Instructional science, vol. 25, no. 3, pp. 167–202, 1997.
5. A. Origin, ‘Method for Qualification and Selection of Open Source Software (QSOS)’, Web published: http://www. qsos. org (Last visited: Jan., 2011), 2004.
6. J. P. Leal and R. Queirós, ‘A comparative study on LMS interoperability’, in Higher Education Institutions and Learning Management Systems: Adoption and Standardization, IGI Global, 2012, pp. 142–161.
7. N. Dabbagh and A. Kitsantas, ‘Using Learning Management Systems as Metacognitive Tools to Support Self-Regulation in Higher Education Contexts’, in International Handbook of Metacognition and Learning Technologies, R. Azevedo and V. Aleven, Eds. Springer New York, 2013, pp. 197–211.

Webography (accessed on 2nd January 2017)

https://blog.capterra.com/top-8-freeopen-source-lmss
https://www.capterra.com/learning-management-system-software/#infographic
https://corp-staging.raccoongang.com/
https://corp-staging.raccoongang.com/blog/lms-comparison-what-lms-suits-your-needs-best/
https://corp-staging.raccoongang.com/blog/important-lms-features-blended-learning/
https://campustechnology.com/articles/2017/10/16/2017-readers-choice-awards.aspx
https://elearningindustry.com/the-20-best-learning-management-systems
https://elearningindustry.com/directory/software-categories/learning-management-systems
https://elearningindustry.com/7-ways-integrate-technology-successful-project-based-learning
https://elearningindustry.com/elearning-trends-and-predictions-2017
https://elearningindustry.com/directory/software-categories/learning-management-systems
https://elearningindustry.com/the-20-best-learning-management-systems
https://er.educause.edu/~/media/files/articles/2014/4/selecting_lms.pdf?la=en
http://ethinkeducation.com/measure-moodle-data-xapi-lrs/
http://fraysse.eu/scorm-en-10-questions/
https://www.getapp.com/hr-employee-management-software/a/canvas-lms/?utm_medium=cpc&utm_source=network&utm_campaign=Ziff+Davis
https://www.joomlalms.com/learning-management-system-comparison.html
https://www.learndash.com/20-most-popular-learning-management-systems-infographic/
https://www.marketsandmarkets.com/Market-Reports/learning-management-systems-market-1266.html
https://mfeldstein.com/state-higher-ed-lms-market-us-canada-spring-2017-edition/
https://www.moodlenews.com/2017/moodle-the-lrs-getting-started-with-xapi-learning-record-stores/
https://www.moodlenews.com/2017/moodle-the-lrs-getting-started-with-xapi-learning-record-stores/
https://www.pcmag.com/search_redirect?qry=lms&searchSection=0&site=3
https://www.pcmag.com/article2/0,2817,2488347,00.asp
https://www.pcmag.com/business/directory/learning-management
https://www.trustradius.com/compare-products/canvas-vs-moodle
https://raccoongang.com/materials/
http://www.swiftelearningservices.com/how-to-choose-the-right-learning-management-system-lms-to-meet-the-training-requirements/
https://raccoongang.com/blog/lms-comparison-what-lms-suits-your-needs-best/

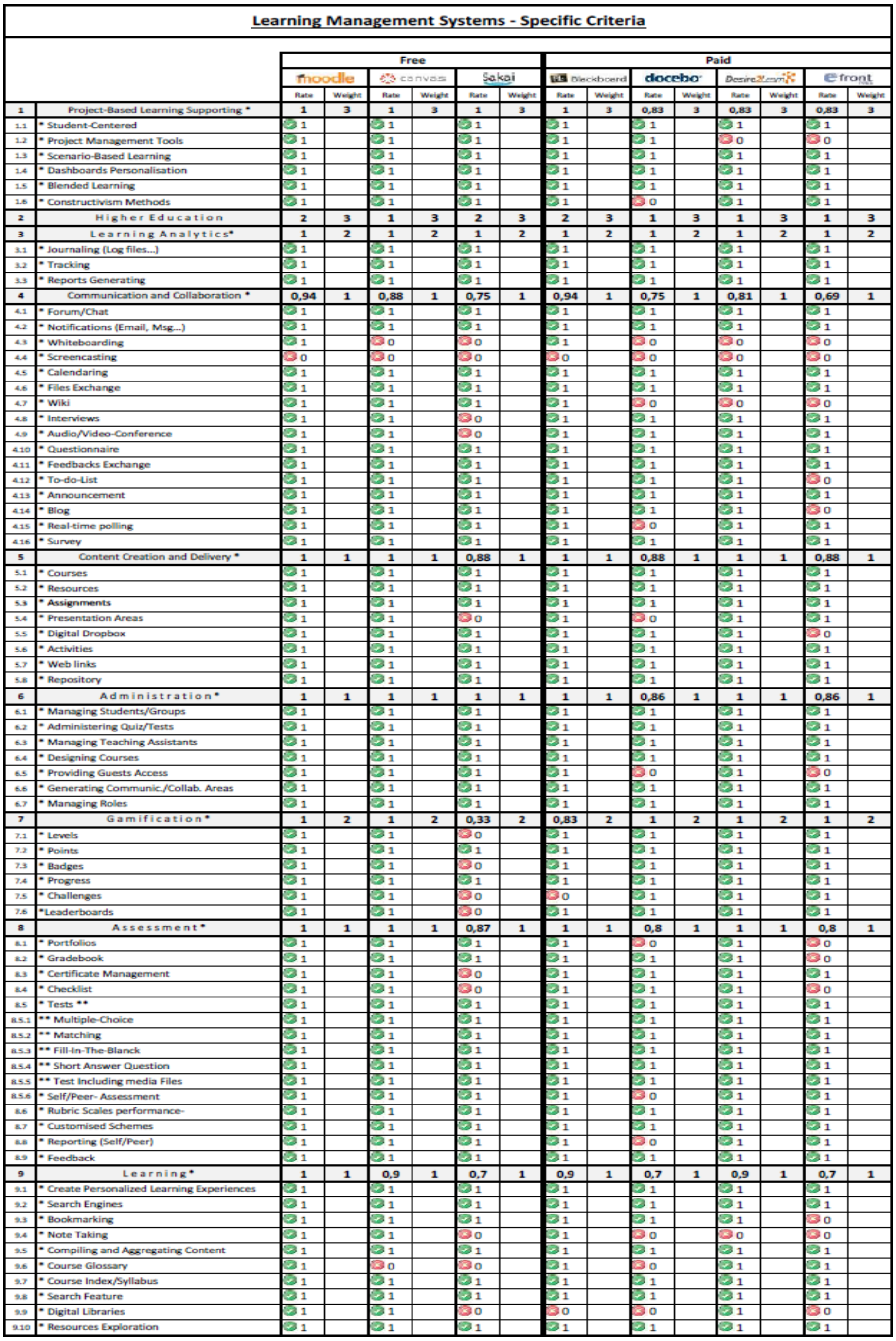
Appendix 1. LMS generic criteria - detailed rating and weighting.



Learning Management Systems - Generic Criteria

		Free						Paid							
		moodle		canvas		Sakai		Blackboard		docebo		Desire2Learn		efront	
		Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight
1	Affordability and Owernship	2	3	1	3	1	3	1	3	2	3	1	3	2	3
2	Interoperability*	1	3	1	3	1	3	1	3	1	3	1	3	1	3
2.1	* SCORM	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
2.2	* API (xAPI / TinCan, LRS)	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
2.3	* Google App Synchronisation	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
3	Flexibility / Customasibility *	1	1	1	1	0,75	1	1	1	1	1	1	1	1	1
3.1	* Mobile Support	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
3.2	* Multilingual Support	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
3.3	* Personalised Learning	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
3.4	* Templates/Themes	✓ 1		✓ 1		✗ 0		✓ 1		✓ 1		✓ 1		✓ 1	
4	Extensibility / Scalability	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	Accessibility*	0,4	1	0,8	1	0,8	1	0,8	1	1	1	0,8	1	1	1
5.1	* Easy-to-Use	✗ 0		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
5.2	* Easy to Maintenance / Support	✗ 0		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
5.3	* Cloud-Based Support	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
5.4	* Community Documentation Support	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
5.5	* Offline Access	✗ 0		✗ 0		✗ 0		✗ 0		✓ 1		✗ 0		✓ 1	
6	Security	1	1	2	1	1	1	1	1	1	1	1	1	1	1

Appendix 2. LMS specific criteria - detailed rating and weighting.



Learning Management Systems - Specific Criteria

		Free						Paid							
		moodle		canvas		Sakai		Blackboard		docebo		Desire2Learn		efront	
		Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight	Rate	Weight
1	Project-Based Learning Supporting *	1	3	1	3	1	3	1	3	0,83	3	0,83	3	0,83	3
1.1	* Student-Centered	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
1.2	* Project Management Tools	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✗ 0		✗ 0	
1.3	* Scenario-Based Learning	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
1.4	* Dashboards Personalisation	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
1.5	* Blended Learning	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
1.6	* Constructivism Methods	✓ 1		✓ 1		✓ 1		✓ 1		✗ 0		✓ 1		✓ 1	
2	Higher Education	2	3	1	3	2	3	2	3	1	3	1	3	1	3
3	Learning Analytics*	1	2	1	2	1	2	1	2	1	2	1	2	1	2
3.1	* Journaling (Log files...)	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
3.2	* Tracking	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
3.3	* Reports Generating	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
4	Communication and Collaboration *	0,94	1	0,88	1	0,75	1	0,94	1	0,75	1	0,81	1	0,69	1
4.1	* Forum/Chat	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
4.2	* Notifications (Email, Msg...)	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
4.3	* Whiteboarding	✓ 1		✗ 0		✗ 0		✓ 1		✗ 0		✗ 0		✗ 0	
4.4	* Screencasting	✗ 0		✗ 0		✗ 0		✗ 0		✗ 0		✗ 0		✗ 0	
4.5	* Calendaring	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
4.6	* Files Exchange	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
4.7	* Wiki	✓ 1		✓ 1		✓ 1		✓ 1		✗ 0		✗ 0		✗ 0	
4.8	* Interviews	✓ 1		✓ 1		✗ 0		✓ 1		✓ 1		✓ 1		✓ 1	
4.9	* Audio/Video-Conference	✓ 1		✓ 1		✗ 0		✓ 1		✓ 1		✓ 1		✓ 1	
4.10	* Questionnaire	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
4.11	* Feedbacks Exchange	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
4.12	* To-do-List	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✗ 0	
4.13	* Announcement	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
4.14	* Blog	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✗ 0	
4.15	* Real-time polling	✓ 1		✓ 1		✓ 1		✓ 1		✗ 0		✓ 1		✓ 1	
4.16	* Survey	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
5	Content Creation and Delivery *	1	1	1	1	0,88	1	1	1	0,88	1	1	1	0,88	1
5.1	* Courses	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
5.2	* Resources	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
5.3	* Assignments	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
5.4	* Presentation Areas	✓ 1		✓ 1		✗ 0		✓ 1		✗ 0		✓ 1		✓ 1	
5.5	* Digital Dropbox	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✗ 0	
5.6	* Activities	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
5.7	* Web links	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
5.8	* Repository	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
6	Administration*	1	1	1	1	1	1	1	1	0,86	1	1	1	0,86	1
6.1	* Managing Students/Groups	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
6.2	* Administering Quiz/Tests	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
6.3	* Managing Teaching Assistants	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
6.4	* Designing Courses	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
6.5	* Providing Guests Access	✓ 1		✓ 1		✓ 1		✓ 1		✗ 0		✓ 1		✗ 0	
6.6	* Generating Communic./Collab. Areas	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
6.7	* Managing Roles	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
7	Gamification*	1	2	1	2	0,33	2	0,83	2	1	2	1	2	1	2
7.1	* Levels	✓ 1		✓ 1		✗ 0		✓ 1		✓ 1		✓ 1		✓ 1	
7.2	* Points	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
7.3	* Badges	✓ 1		✓ 1		✗ 0		✓ 1		✓ 1		✓ 1		✓ 1	
7.4	* Progress	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
7.5	* Challenges	✓ 1		✓ 1		✗ 0		✗ 0		✓ 1		✓ 1		✓ 1	
7.6	*Leaderboards	✓ 1		✓ 1		✗ 0		✓ 1		✓ 1		✓ 1		✓ 1	
8	Assessment*	1	1	1	1	0,87	1	1	1	0,8	1	1	1	0,8	1
8.1	* Portfolios	✓ 1		✓ 1		✓ 1		✓ 1		✗ 0		✓ 1		✗ 0	
8.2	* Gradebook	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✗ 0	
8.3	* Certificate Management	✓ 1		✓ 1		✗ 0		✓ 1		✓ 1		✓ 1		✓ 1	
8.4	* Checklist	✓ 1		✓ 1		✗ 0		✓ 1		✓ 1		✓ 1		✗ 0	
8.5	* Tests **	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
8.5.1	** Multiple-Choice	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
8.5.2	** Matching	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
8.5.3	** Fill-In-The-Blanck	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
8.5.4	** Short Answer Question	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
8.5.5	** Test Including media Files	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
8.5.6	* Self/Peer- Assessment	✓ 1		✓ 1		✓ 1		✓ 1		✗ 0		✓ 1		✓ 1	
8.6	* Rubric Scales performance-	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
8.7	* Customised Schemes	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
8.8	* Reporting (Self/Peer)	✓ 1		✓ 1		✓ 1		✓ 1		✗ 0		✓ 1		✓ 1	
8.9	* Feedback	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
9	Learning*	1	1	0,9	1	0,7	1	0,9	1	0,7	1	0,9	1	0,7	1
9.1	* Create Personalized Learning Experiences	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
9.2	* Search Engines	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
9.3	* Bookmarking	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✗ 0	
9.4	* Note Taking	✓ 1		✓ 1		✗ 0		✓ 1		✗ 0		✗ 0		✗ 0	
9.5	* Compiling and Aggregating Content	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
9.6	* Course Glossary	✓ 1		✗ 0		✗ 0		✓ 1		✗ 0		✓ 1		✓ 1	
9.7	* Course Index/Syllabus	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
9.8	* Search Feature	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	
9.9	* Digital Libraries	✓ 1		✓ 1		✗ 0		✗ 0		✗ 0		✓ 1		✗ 0	
9.10	* Resources Exploration	✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1		✓ 1	

SESSION 5

Advanced Information Extraction and Machine Learning

Topic categorization in Portuguese news articles
André Santos)

NER: Supervised System to Recognize Participants and Location in Criminal News
Carla Abreu

Sentiment analysis techniques applied to regression tasks
José Ornelas

Topic categorization in Portuguese news articles

André F. Santos

CRACS & INESC-Porto LA
Faculty of Sciences, University of Porto
Porto, Portugal
afs@inesctec.pt

Abstract. Categorizing news articles according to their contents allows to decrease the information entropy in a world where the rate of publication of digital text documents is increasing fast. In this article we describe ongoing work which aims to evaluate the feasibility of implementing a classifier which is lightweight enough to be used in real time on the client side of a web application. More specifically, we gathered a corpus of Portuguese news and used it to train and evaluate several classification algorithms. We analyze the results obtained in terms of the classifiers error rate, training time and memory footprint.

Keywords: topic categorization, machine learning, text mining

1 Introduction

Online news articles first appeared as reprints from traditional newspapers; nowadays, however, they represent now the primary source of news for some segments of the population, both in developed and developing countries (whether consumed directly in the newspaper website, or indirectly e.g. through a social media application or a feed catcher) [2, 6, 8].

Unofficially known as *the fourth branch of government*, the press plays a vital role within our society, keeping us informed about the current state of affairs (at a local and global scale) and acting as a watchdog for the other three branches (legislative, executive and judicial). The (lack of) freedom of press and access to the news in a given country is even often considered an indicator of a lack of democracy [7, 10].

As such, improving the ways citizens can access the information (view it, query it and search for it) contained in news articles has the potential to contribute for a more informed and, ultimately, better, society [3].

On the other hand, the last decades have witnessed a fast increase on the rate of publication of digital text documents. Traditional document types, such as news articles, scientific papers or books are now published online along with new formats, such as blog posts or tweets, each having thousands or millions of new documents published each day [1, 9].

Publication is not the only step which has moved to the digital world; in fact, most often nowadays the whole document lifecycle happens digitally, with virtual tools available for researching, writing, styling, publishing and sharing [17].

Having the entire workflow happening within the digital world presents some opportunities when compared to the traditional process [12]. In particular, due to the current processing power commonly available, tasks related to the manipulation of the information contained within these documents (searching, compiling, annotating, sharing, . . .) can now be performed automatically and targeting a large amount of articles.

In addition to the document content (for example, in a news article, the *title*, *lead* and *body*), its metadata is also important: author(s), date of publication, source, topic, mentioned entities and their relations, etc [18,19]. Some of this metadata might be filled in and stored along with the document (e.g. *author* and *date of publication*); other is usually extracted from the document content (e.g. mentioned entities) [16].

An example of a feature which improves information access is the categorization of news articles by the topic (or topics) of its content [11]. The presence of such a categorization may influence the way the information is stored, organized, displayed and queried [15].

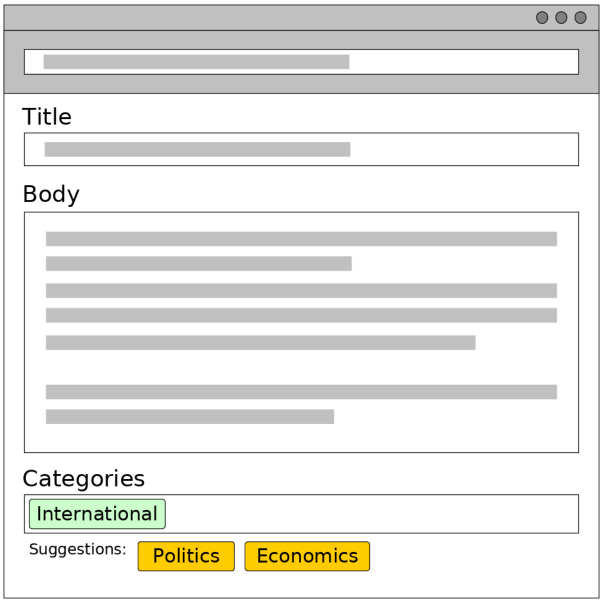


Fig. 1. Category classification and suggestion on the client side

The simplest way of achieving this categorization is to have the author of the article manually introducing it (e.g. the journalist typing it on the news article authoring framework); however, this solution presents some challenges:

- It increases the amount of work the author has to do.
- The author might not be sure which categories are available.
- The author might not be sure which category is the best (e.g. *Economics* vs *Finance*).
- It does not scale – e.g. if the goal is to categorize an existing (large) corpus.

Thus, an automated way of categorizing news articles could solve some of these problems and decrease the burden of this task. Additionally, a lightweight version of such a classifier could be implemented on the client side code of a web application, for example, allowing the categorization to happen in real time (i.e. as the author types in the article text). Figure 1 presents a suggestion of how this feature could look like if implemented on a web application.

The challenges of document classification have been well studied within the machine learning research field of study [4,13,14]. Given a corpus of already classified documents, several algorithms might be applied to train a classifier capable of determining the category of additional articles.

In this article, we describe the preliminary results obtained in developing a classifier to categorize news articles using a previously manually categorized corpus. Additionally, we evaluate the possibility of implementing such a classifier as lightweight as possible to allow it to run on the client side of a web application.

2 Methods

In order to train and evaluate classification algorithms, we first needed to choose and obtain a suitable dataset. Preferably, this dataset should contain documents which were previously categorized, allowing us to skip the time and effort-consuming task of manually categorizing the articles ourselves. Once this dataset was chosen and obtained, we would then clean and prepare it to be used to train the classifiers.

2.1 The dataset

We gathered a dataset of news articles published in *Observador*[1], one of the main Portuguese newspapers, which stands out from the others for being fairly young (it was created in May 2014) and for existing exclusively online. The initial dataset comprised 42.475 entries, the most recent ones dated from November 2016, from which we used only a subset, for reasons later described.

We gathered all the categories used by Observador, and ordered them from the most common to the least common. We selected the ones which had more than 1000 articles in our dataset, and reduced our original dataset to include

[1] `http://observador.pt`

articles from these categories only. Figure 2 presents an overview of the selected categories and the number of articles available for each one.

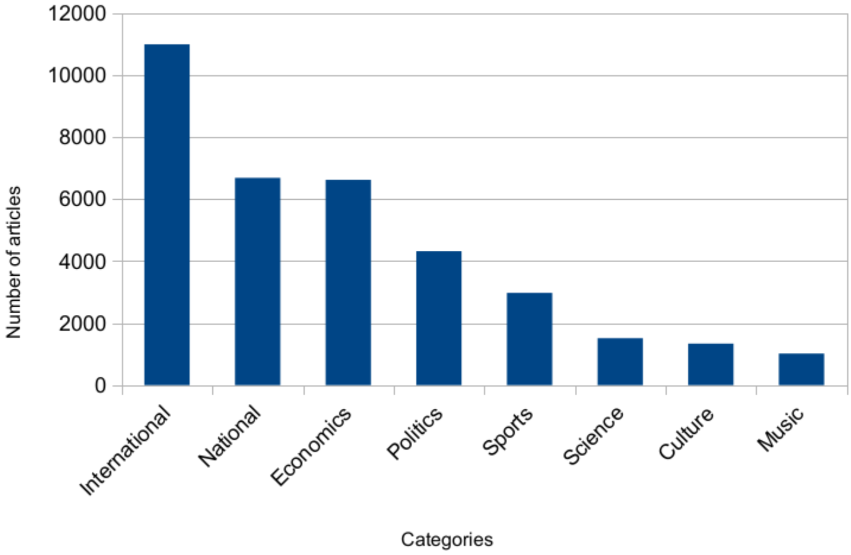


Fig. 2. Total number of articles retrieved for each category

We them randomly selected, from each category, 700 articles to be used to train the classifiers, and 200 to be used to evaluate their performance.

For each article, we had available its contents (title, lead, body) and several metadata fields (publication date, category, tags, etc). A truncated JSON representation of an article can be found in Listing 1.

2.2 Preprocessing the articles

Originally, the dataset was obtained as a large MongoDB collection (more than 2.5 million entries), containing articles from several Portuguese and international newspapers. The process needed to transform this collection into data our classifiers could process required querying the database, exporting the news articles and splitting them into a train and an evaluation datasets.

The database query selected articles from `Observador` where the body had a length greater than 100 characters (to discard some malformed articles which had an empty body or a body composed of only a few words), and the categories included at least one of the most common categories.

For each article returned by the query, the pretitle, title, subtitle and lead fields, if present, were simply copied to a plain text file, separated by blank lines. The body field, however, was stored in the database in HTML format. As such,

```
{
  Type: "sapo.obj.creativework.article",
  Source: {
      Name : "Observador"
  },
  Pretitle: "Benfica",
  Title: "Ruben Amorim com rotura total do ligamento cruzado",
  Author: {
      Name: "Observador"
  },
  Tags: [
      "benfica",
      "desporto",
      "futebol",
      "ruben amorim"
  ],
  PublishDate: ISODate("2014-08-25T18:33:00Z"),
  Lead: "Depois de Fejsa, mais uma baixa. O internacional português [...]",
  Body: "<p>O pior cenário confirmou-se. O Benfica informou esta segunda-feira [...]",
  URL: "http://observador.pt/2014/08/25/ruben-amorim-com-rotura-total-ligamento-cruzado/",
  CategoryPaths: ["Desporto"],
  Domain: "observador.pt",
  Language: "pt_PT",
  ...
}
```

List. 1. Example of JSON representation of an article the HTML tags had to be stripped, and then it was also added to the plain text file.

The files were stored in folders, separated by category. For each category, 700 articles were allocated to the train set, and 200 to the evaluation set. These preprocessing tasks were accomplished using Bash and Node.js scripts.

Both datasets were loaded into R using the `tm`[2] package. Each was then passed to a function responsible for preprocessing the text of the articles:

- The text was converted to lowercase characters.
- The Portuguese stopwords were removed.
- Diacritics were converted to their normalized form (e.g. $\boxed{\text{à}}$ → $\boxed{\text{a}}$).
- Punctuation signs were removed.
- Numbers were removed.
- Word were stemmed (e.g. $\boxed{\text{conseguiram}}$ → $\boxed{\text{consegu}}$).
- White space was removed and the text was tokenized.

A document-term matrix was calculated using the `DocumentTermMatrix` method from the `tm` package. For each of the 5600 documents present in the train set the matrix included all the terms which met the following requirements:

- The term length was between 3 and 30 characters (to discard things like URLs or badly tokenized sentences).
- The term appeared in the document at least twice.
- The term appeared at least in 10 documents.

The algorithm used to weigh the terms in the matrix was *tf-idf* [5]. The obtained matrix presented a sparsity of 99% and contained 3234 distinct terms.

The final step was to remove the sparse terms from the matrix. This allows to discard terms which might be too specific of the train set and which might negatively influence the performance of the classifiers by overfitting them to the train set. Additionally, a smaller list of terms reduces the execution time of both training and applying the classifier, and the memory footprint of the classifier. However, while discarding sparse terms we might end up removing relevant terms and increase our classifiers error rate.

We used the `removeSparseTerms` function from the `tm` package, and produced three distinct lists of terms:

- LT_{90} contained terms with 90% or less sparsity.
- LT_{95} contained terms whose sparsity was under 95%.
- LT_{99} contained the terms with a sparsity level below 99%.

2.3 Classification

To create the classifiers, we used 5 well known algorithms: decision tree (DT), k-nearest neighbors (KNN), naive Bayes (NB), neural network (NN) and support vector machine – with radial kernel ($\mathrm{SVM_{RK}}$) and linear kernel ($\mathrm{SVM_{LK}}$).

[2] `https://cran.r-project.org/web/packages/tm/`

We trained each of the classifiers three different times, one for each list of terms (LT_{99}, LT_{95} and LT_{90}). Then we evaluated each trained classifier using the test dataset.

The test dataset contained 1600 documents (200 belonging to each category) and was preprocessed in a way similar to the train set (described in the previous section, 2.2). In each iteration, however, the final list of terms in each test document was restricted to terms present in the corresponding list of terms (LT_{99}, LT_{95} and LT_{90}).

3 Results

A number of measurements and metrics were calculated regarding the lists of terms, the classifiers training process and their results in the evaluation process.

Table 1 presents the size of each list of terms after removing the terms whose sparsity was above the corresponding threshold.

Table 1. Lists size

	Sparsity threshold (%)	Number of terms left
LT_{90}	90	45
LT_{95}	95	139
LT_{99}	99	912

Table 2 presents the execution time for training the algorithm with the lowest error rate for each list of terms.

Table 2. Execution times for training

	Algorithm	Training time
LT_{90}	DT	12s
LT_{95}	KNN	3m
LT_{99}	KNN	57m

Table 3 presents the error rates obtained for each list of terms using each of the classifiers, with the value of the best classifier for each list highlighted in bold.

Table 3. Error rates

	DT	KNN	NB	NN	SVM$_{\text{RK}}$	SVM$_{\text{LK}}$
LT_{90}	**0.60**	0.71	0.85	0.61	0.88	0.88
LT_{95}	0.70	**0.56**	0.83	0.58	0.87	0.87
LT_{99}	0.70	**0.35**	0.87	0.76	0.87	0.87

Table 4 presents the confusion matrix generated using the k-nearest neighbors classifier with the LT_{99} list of therms, with the number of correct classifications for each category highlighted in bold.

Table 4. Categories confusion matrix (LT_{99} with KNN)

	science	culture	sports	economics	international	music	national	politics
science	**127**	19	1	6	12	24	11	0
culture	5	**102**	2	2	3	79	6	1
sports	1	3	**168**	7	3	14	4	0
economics	5	3	0	**146**	5	17	14	10
international	12	12	7	17	**90**	35	16	11
music	0	2	0	0	2	**184**	1	0
national	9	9	9	19	15	31	**88**	21
politics	1	3	0	30	8	17	14	**127**

4 Discussion

The analysis of the results obtained should take into account other metrics besides the error rates obtained for each classifier.

The size of the list of terms, for example, gives us an idea of the memory footprint of a classifier, a parameter which is of the utmost importance if the goal is to implement a classifier as lightweight as possible. The training time is also relevant, as shorter training times give the possibility of retraining the classifiers more often, allowing them to be updated as the corpus of articles grows in size.

Looking at the actual results obtained and represented in Tables 1 and 2 we can see that LT_{90} has simultaneously the smallest list of terms (45) and the shortest training time (12 seconds using the DT algorithm). However, LT_{90} also presents the worst error rates, even when looking at the algorithm which achieved its best results (0.60 using a DT classifier).

On the opposite side, LT_{99} presented the lowest error rates (0.35 using a k-nearest neighbors classifier), but it took almost one hour to train and used a list comprising 912 terms.

The results obtained confirm that there is a tradeoff between the size of the list of terms and the training time, on the one hand, and the classifier error rate, on the other. However, a list of 912 terms seems to be an acceptable memory footprint for a client side classifier; additionally, the higher training time would not present much problems in this scenario, as the classifier would be trained beforehand and thus not be visible on the client side.

Given the reasonable values for the training time and the size of the list of terms, and looking to the error rate obtained in each of tests performed, we can conclude that the best option was to use the largest list of terms (LT_{99}) and the KNN classifier.

It is worth noting that the categories of an article are not mutually exclusive. In fact, an article can be classified as belonging to more than one category. This might explain the greater error rates obtained in the categories *national* and *international*: one might argue that these categories correspond less to the topic covered by the article and are more related to the location of the news content.

5 Contributions and Future work

For copyright reasons, the corpus used to train and evaluate the classifiers described in this article cannot be shared. All the code used to process the documents, to implement the classifiers and evaluate them can be found at `http://github.com/andrefs/mapi-msr-categorization`.

The tasks and results previously described already provide useful insights into this matter. However, the lowest error rate obtained (0.35) might still be improved upon, either by leveraging new algorithms, fine tuning the ones already tested, or by increasing the train corpus.

We have established that it is in fact possible to develop a news article classifier which is lightweight enough to be used (in real time) on the client side of a web application. The following step will be the actual implementation of the classifier, probably in the form of a JavaScript library.

Acknowledgment

André Santos has a PhD scholarship from Fundação para a Ciência e Tecnologia.

References

1. Allan, S.: Online news: Journalism and the Internet. McGraw-Hill Education (UK) (2006)
2. Boczkowski, P.J.: Digitizing the news: Innovation in online newspapers. mit Press (2005)
3. Bollinger, L.C.: The tolerant society. Oxford University Press on Demand (1988)
4. Borko, H., Bernick, M.: Automatic document classification. Journal of the ACM (JACM) 10(2), 151–162 (1963)

5. Christopher, D.M., Prabhakar, R., Hinrich, S.: Introduction to information retrieval. An Introduction To Information Retrieval 151, 177 (2008)
6. Chyi, H.I., Lasorsa, D.: Access, use and preferences for online newspapers. Newspaper Research Journal 20(4), 2–13 (1999)
7. Goode, L.: Social news, citizen journalism and democracy. New media & society 11(8), 1287–1305 (2009)
8. Greer, J., Mensing, D.: The evolution of online newspapers: A longitudinal content analysis, 1997-2003. Internet newspapers: The making of a mainstream medium pp. 13–32 (2006)
9. Hilbert, M., López, P.: The world's technological capacity to store, communicate, and compute information. science 332(6025), 60–65 (2011)
10. House, F.: Freedom of the Press 2008: A global survey of media independence. Rowman & Littlefield Publishers (2009)
11. Kim, S.M., Hovy, E.: Extracting opinions, opinion holders, and topics expressed in online news media text. In: Proceedings of the Workshop on Sentiment and Subjectivity in Text. pp. 1–8. Association for Computational Linguistics (2006)
12. O'hara, K., Sellen, A.: A comparison of reading paper and on-line documents. In: Proceedings of the ACM SIGCHI Conference on Human factors in computing systems. pp. 335–342. ACM (1997)
13. Rubin, T.N., Chambers, A., Smyth, P., Steyvers, M.: Statistical topic models for multi-label document classification. Machine learning 88(1), 157–208 (2012)
14. Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) 34(1), 1–47 (2002)
15. Teitler, B.E., Lieberman, M.D., Panozzo, D., Sankaranarayanan, J., Samet, H., Sperling, J.: Newsstand: A new view on news. In: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems. p. 18. ACM (2008)
16. Vadrevu, S., Nagarajan, S., Gelgi, F., Davulcu, H.: Automated metadata and instance extraction from news web sites. In: Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on. pp. 38–41. IEEE (2005)
17. Williams, P., Leighton John, J., Rowland, I.: The personal curation of digital objects: A lifecycle approach. In: Aslib Proceedings. vol. 61, pp. 340–363. Emerald Group Publishing Limited (2009)
18. Yaginuma, T., Pereira, T., Baptista, A.A.: Design of metadata elements for digital news articles in the omnipaper project (2003)
19. Yaginuma, T., Pereira, T., Baptista, A.A.: Metadata elements for digital news resource description (2003)

NER: Supervised System to Recognize Participants and Location in Criminal News

Carla Abreu

Faculdade de Engenharia da Universidade do Porto
ei08165@fe.up.pt

Abstract. Information Extraction (IE) systems are designed to extract specific type of information from a given document. The sub-task that aims to recognize entities from the text, unit elements such as name of person, locations and organizations, is called Named Entity Recognition (NER). In this paper we investigate a supervised learning approach to recognize Locations and Participants in criminal news articles written in English, to address a common bottleneck on the NER the need of gazetteers to categorize an entity. We also study what are the relevant features to recognize Locations and Participants.

Keywords: Information Extraction, Named Entity Recognition, Text Mining, Supervised Learning

1 Introduction

Hundred of news articles are published, everyday, on several media sources. Being aware about all the events is a challenging task. For instance, if we would like to answer to questions like: How many people died on the shootings in Philippi on 30th September, 2017? How many people died last year on Birmingham? How many people were killed by John List? - an extensive research would be required. In general, the answers are found in one or a set of news articles. However, we can simplify the research by recognizing and extracting information from the news article and create a structured representation for all the extracted elements associated to it.

Named Entity Recognition (NER) is a task that aims to recognize entities on a given document. An important conference on the area, namely Message Understanding Conference (MUC) was launched by Defense Advanced Research Projects Agency (DARPA). It defines the entities as belonging to three categories[1]: Enamex - it includes names, such as Locations, Persons, Organization, and others; Timex - it includes Date and Time expressions; Numex it includes numerical elements, as, Numbers and Percentages. The documents used to recognize or extract information can also be categorized as mentioned on the NER

[1] (AFNER - Named Entity Recognition) - http://afner.sourceforge.net/what.html visited on 2017, November

Survey presented in [1]. There are two different elements that can be used to categorize a document: textual gender (i.e. journalistic, informal, scientific); and, domain (i.e. gardening, sports, business). The textual gender and the domain can have a significant impact on the NER tasks.

In this paper, we investigate a supervised learning approach to deal with the recognition of two different entities, Locations (i.e. Philippi, Birmingham) and Participants (i.e. John List). The documents used are classified as journalistic in a criminal domain.

By the mean of this work, we aim to participate in a SemEval-2018 competition Task 5[2]: "Counting Events and Participants in the Long Tail sent". The work pursued in this paper solves a part of the problem addressed on the aforementioned competition.

The remainder of the paper is organized as follows: Section 2 describes the related works on the NER, Section 3 describes the proposed approach, Section 4 presents the Data Resources, Section 5 describes the experiments, Section 6 presents the results obtained and Section 7 presents the conclusion.

2 Related Work

To deal with the NER problems, different approaches are developed. Early systems deal with this issue by making use of handcrafted rule-based algorithms. Moreover, today systems choose to handle with such problem using machine learning techniques (supervised learning (SL) and unsupervised learning). However, the major drawback of SL approach is its requirement of a large annotated corpus. For the case of unavailability of training examples, handcrafted rules remain the practical technique.

2.1 Handcrafted Rule-based Systems

The aim of NER task is to recognize an entity from a document. This task could be done using handcrafted rule-based systems. Normally, the handcrafted rule-based systems are composed by two components: (1) rule-based algorithms that aim to extract entities observed under some conditions; and (2) pre compiled list of entities, that can be used to verify if the extracted word belongs to a given category, this list could also be incremented with the entities extracted on (1). A well known system on this area was developed by Hearst [2] and is based on hyponymic and hyperonymic relations. When facing with NER tasks limitations, such as the absence of corpus, the rule-based systems continues to be used [3].

2.2 Supervised Learning Systems

Supervised learning techniques aim to infer a function from labeled training data in order to learn some patterns in which an entity appears. This technique

[2] (SemEval-2018 - Task 5) - https://competitions.codalab.org/competitions/17285 visited on 2017, November

requires a large corpus and also a definition of features. Several studies were pursued using this approach [4].

2.3 Semi-supervised and Unsupervised Learning Systems

In order to use machine learning approaches to avoid the large corpus requirement, semi-supervised learning and unsupervised learning approaches are applied to recognize entities. These approaches are a promise of fast deployment for many types of entities without the prerequisite of an annotated corpus. The most common method used on the unsupervised approach is the clustering. Some works were already pursued using this approach to solve the NER task, such as [5] and [6]. An interesting work done in this field, suggests one approach to solve a NER task and overcome the two major limitation on the area, existence of a gazetteers and an annotated corpus [6].

3 Supervised Learning System Approach

3.1 Natural Language Processing Tasks

Usually Natural Language Processing (NLP) tasks need to be done before the NER task. We started to remove journalistic patterns and expressions from a given news article. These patterns and expressions could be relevant for the reader but not for the entity recognition task. Table 1 shows three examples of journalistic patterns and two examples of journalistic expressions. In order to normalize the text, we create regular expressions to detect and remove these kind of occurrences on the given document.

Aditional NLP tasks were done on the text. For each token in a sentence was added supplementary information, such as its part of speech and stop words recognition and association.

Table 1. Journalistic Patterns and Expressions Examples

Type	Example
Pattern	Copyright 2017 by WJXT News4Jax - All rights reserved.
Pattern	Keep checking NBC4i.com for real-time updates on this story.
Pattern	To get alerts for breaking news, grab the free NBC4 News App for iPhone or Android.
Expression	contact kimber laux at klaux@reviewjournal.com or 702 - 383 - 0283.
Expression	Contact Jessica Terrones at jterrones@reviewjournal.com or at 702-383-0381.

3.2 Features

To work with machine learning techniques we need to categorize our data. In this work we have a set of news from where we want to extract information. It is common to label each word with a set of features. These features will be essential, for the SL approach, to recognize an entity in a given document.

The following features are selected as:

CAP Capitalized: indicates if a word has no characters capitalized, if it has the first or if it has all characters capitalized;
PT POSTagger[3] Association: identify parts of speech to each word, such as noun, verb, adjective, etc.
SWI Stop Words Identification: indicates if a word is or not a stop-word;
SWA Stop Words Association associate a corresponding stop-word;
NPA Paragraph: paragraph number where the word appears.

Table 2 presents an example how the features can be associated to a given word. For the sentence " ... shooting at a west Phoenix apartment that left one man dead... ", the word "Phoenix" is capitalized (1), corresponds to a noun (NNP), is not a stop-word (0).

Table 2. Categorizing each word on a sentence

	shooting	at	a	west	Phoenix	apartment	that	left	one	man	dead
CAP	0	0	0	0	1	0	0	0	0	0	0
PT	D	NN	IN	DT	NNP	NN	WDT	VBD	CD	NN	NN
SWI	0	1	1	0	0	0	1	0	0	0	0
SWA		at	a				that				

We believe that a simple association as demonstrated on Table 2 is not enough to categorize a word for a Named Entity Recognition task. For this latter reason, we also consider extra scenarios where the contextualization of the word in the text is taken into account. For these scenarios, we have the current word (C), previous word (P) and the next word (N). Our approach to represent the features, used in each scenario, is to indicate the word position following the feature abbreviation, such as C CAP, which indicates if a current word is or is not capitalized.

S1 C CAP, C SWI, C PT, P CAP, P SWI, P PT, P SWA, N CAP, N SWI, N PT, N SWA, C NPA;
S2 C CAP, C SWI, C PT, P CAP, P SWI, P PT, P SWA, N CAP, N SWI, N PT;
S3 C CAP, C SWI, C PT, P CAP, P SWI, P PT, P SWA;
S4 C CAP, C SWI, C PT, N CAP, N SWI, N PT, N SWA;

3.3 Data Cleaning and Transformation

Data quality is the main issue of quality information management. To guarantee the data quality, two processes were done: data cleaning and data transformation. Tables 4 and 5 present the data transformation for post tagger labels and stop words. When the word is not a stop word a value will miss on the SWA, so we replace an empty value by the character X and we will encode this value as demonstrated in Table 5.

[3] (POSTagger - All Tags) - http://www.nltk.org/book/ch05.html visited on 2017, November

Table 3. Data Cleaning and Transformation

Table 4. POSTtagger

PostTagger	Representation
DT	0
NN	1
NNP	2
VBD	3
...	...

Table 5. Stop Words

Stop Word	Representation
X	0
a	1
that	2
and	3
...	...

3.4 Classification Algorithms

Supervised learning aims to create a model that predicts the value of a target variable based on several input variables. In order to create a model, it was necessary to find a most appropriated algorithm, the classification algorithms choose to classify Locations and Participants as the following:

- Support Vector Classifier (SVC)
- Decision Tree Classifier (Tree)
- Random Forest Classifier (Random)
- Extra Trees Classifier (Extra)

These algorithms are provided by scikit-learn[4] and different configurations are taken in account as demonstrated on Table 6.

- Kernel: Specifies the kernel type to be used in the algorithm
- Criterion: The function to measure the quality of a split. Supported criteria are gini for the Gini impurity and entropy for the information gain.
- Splitter: The strategy used to choose the split at each node. Supported strategies are best to choose the best split and random to choose the best random split.
- Min samples split: The minimum number of samples required to split an internal node.
- Max features:The number of features to consider when looking for the best split
- N_estimators: The number of trees in the forest.

4 Data Resources

SemEval competition provides data for the propose of this paper. The data available by this competition is a set of criminal English written news articles. To extract locations and participants from criminal news, additional annotations were done. Were annotated a set of 9288 individual words in three categories: Locations, Participants and Others.

[4] (scikit-learn library in Python - Machine Learning in Python) - http://scikit-learn.org/stable/ visited on 2017, November

Table 6. Classification Algorithm Configurations

Alg/ID	Configuration
SVC 1	Default scikit learn configuration
SVC 2	kernel="linear"
SVC 3	kernel="sigmoid"
Tree 1	Default scikit-learn configuration
Tree 2	criterion="gini", splitter="best", min samples split=2
Tree 3	criterion="entropy", splitter="best", min samples split=2
Tree 4	criterion="entropy", splitter="random", min samples split=2
Tree 5	criterion="gini", splitter="random", min samples split=2
Tree 6	criterion="gini", splitter="best", min samples split=4
Tree 7	criterion="entropy", splitter="best", min samples split=4
Random 1	criterion="gini", n estimators=10
Random 2	criterion="gini" n estimators=5
Random 3	criterion="gini",n estimators=20
Random 4	criterion="entropy", n estimators=10
Random 5	criterion="entropy",n estimators=5
Random 6	criterion="entropy",n estimators=20
Extra 1	criterion="gini", max features="auto"
Extra 2	criterion="entropy", max features="auto"
Extra 3	criterion="gini", max features="sqrt"
Extra 4	criterion="entropy",max features="sqrt"
Extra 5	criterion="gini", max features="log2"
Extra 6	criterion="entropy" max features="log2"
Extra 7	criterion="gini" max features=None
Extra 8	criterion="entropy", max features=None

5 Experiments

5.1 Evaluation Metrics

In order to evaluate the proposed system three metrics were taken into account: Recall, precision and F-measure (F1). These metrics are generally selected ways of measuring system performance in this field.

Recall is the percentage of named entities present in the corpus and are found by the system.

Precision is the percentage of named entities found by the learning system and are correct.

F1 is the average of the precision and recall.

A named entity is correct only if the label assigned by the learning system is exactly the same that is present in the data file.

5.2 Experiments

A supervised learning system is needed to generate a model. The features selected in our experiments are grouped in scenarios (S1, S2, S3 and S4), as described on subsection 3.2. On these scenarios, distinct context elements are considered.

The classification algorithms used on our tests are all presented on subsection 3.4.

Our experiments are done based on the cross validation. We used the annotated data mentioned on subsection 4 partitioned on 4. The partitions represents a holdout of 75% of training data and 15% of testing data.

6 Results and Analysis

In order to be easy to retrieve a news article that answers a question, this work aims to create a simple news representation which consist on determining, where the news article will have attached all the locations and participants mentioned on it. For such, context, it is more important to capture all the information related with the news article instead of losing this information to boost the precision. For this latter reason, the most efficient evaluation metric is the recall. However, as we have three evaluation metrics we decided to use them all for an optimal decision.

As our evaluation was made with a consideration of the result of four partitions. The results presented are the mean and the standard deviation of the metrics results. The graph is read as follows: the best result should have a higher mean and a low standard deviation. When the standard deviation is low, it indicates a stable result.

The results obtained after classification of participants are presented on: Fig 1 precision mean, Fig 2 standard deviation precision, Fig 3 recall mean, Fig 4 recall standard deviation, Fig 5 F1 mean, Fig 6 F1 standard deviation. Location results are represented on: Fig 7 precision mean, Fig 8 precision standard deviation, Fig 9 recall mean, Fig 10 recall standard deviation, Fig 11 F1 mean, Fig 12 F1 standard deviation.

Based on the tested scenarios, the scenario S2 has the best performance in the classification of both Locations and Participants. This latter scenario has got the best precision combination (high precision mean and low standard deviation), did not reach an high mean of recall but a low standard deviation indicates a constant result on this evaluation metric. As for the F1 result, it also shows a best combination in extracting participants. However, the results of Locations extraction achieved by F1 metric are not encouraging.

We have analyzed a set of algorithms with a set of distinct configurations. The distinct configuration of Random Classifier and the Extra Tree Classifiers obtained exactly the same results. As a result, the distinct configurations have not affected the performance of these algorithms. The Decision Tree Classifier configurations show low variability with respect to the overall performance. A large overall performance was achieved by the SVC configurations, the second configuration for Decision Tree algorithm has the worst performance evaluation and the third configuration did not detect any positive case (for this reason is not on the graphs).

The analysis done to recognize locations are applied to recognize participants. For both the recognitions, the context is essential.

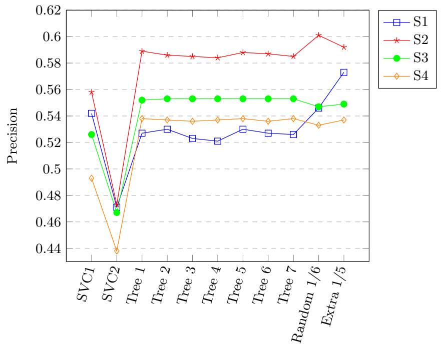


Fig. 1. Precision Average Extracting Participants

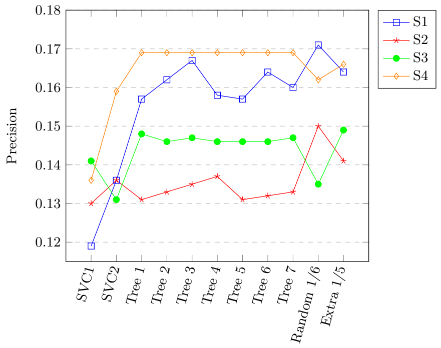


Fig. 2. Standard Deviation Of Precision on Extracting Participants

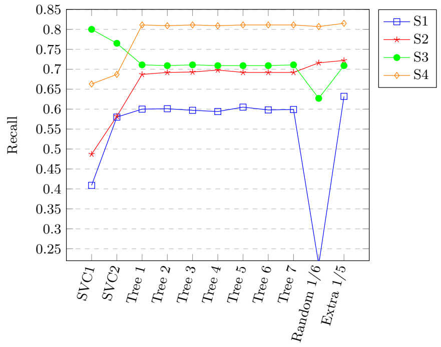


Fig. 3. Recall Average Extracting Participants

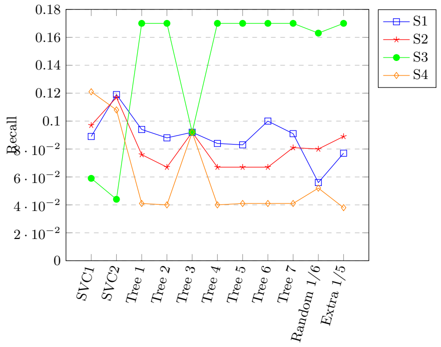


Fig. 4. Recall Standard Deviation on Extracting Participants

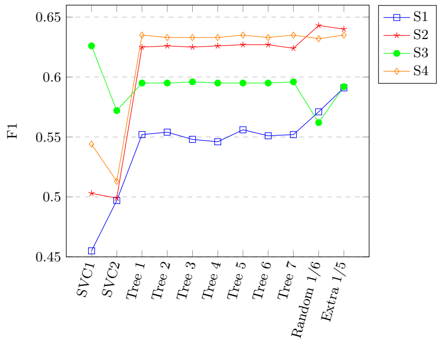


Fig. 5. F1 Average Extracting Participants

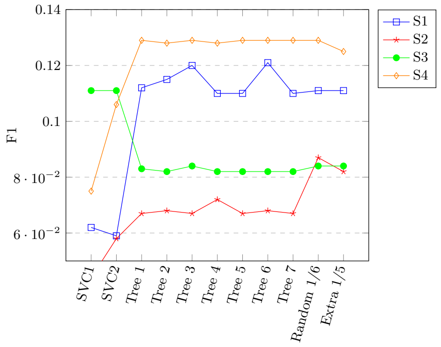


Fig. 6. Standard Deviation Of F1 on Extracting Participants

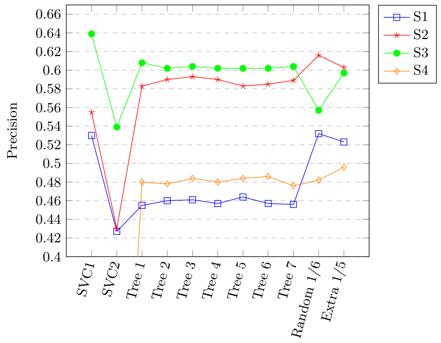


Fig. 7. Precision Average Extracting Locations

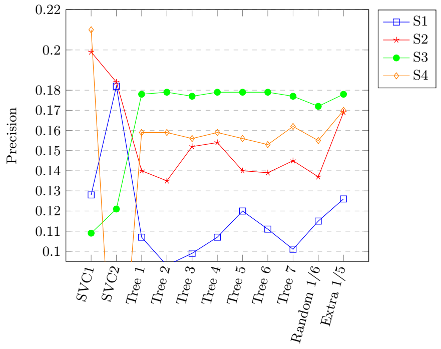


Fig. 8. Standard Deviation Of Precision on Extracting Locations

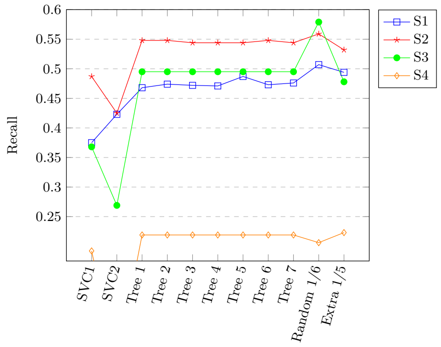


Fig. 9. Recall Average Extracting Locations

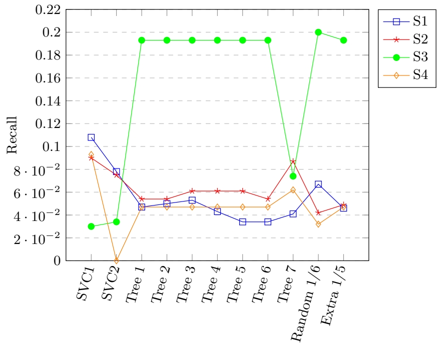


Fig. 10. Standard Deviation Of Precision on Extracting Locations

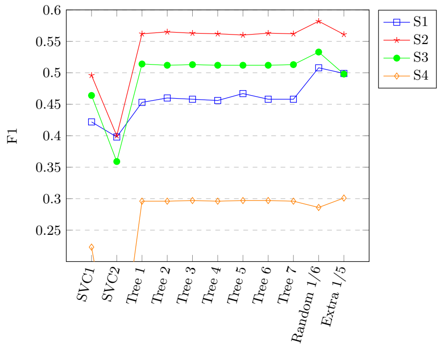


Fig. 11. F1 Average Extracting Locations

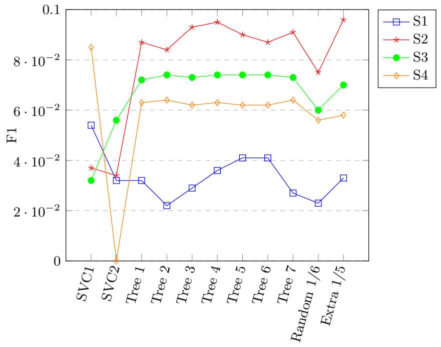


Fig. 12. Standard Deviation Of F1 on Extracting Locations

7 Conclusion and Feature Work

In the present paper, we have investigated an approach to recognize Locations and Participants in the context of criminal news articles by the support of a supervised learning system. We have also studied a set of features and their impact on the NER task. As a conclusion, the context of the word on the sentence has demonstrated a large impact in the recognition process. As future work, we will include another techniques for the recognition process taking into account the results achieved in the present work.

References

1. David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.
2. Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics, 1992.
3. Kashif Riaz. Rule-based named entity recognition in urdu. In *Proceedings of the 2010 named entities workshop*, pages 126–135. Association for Computational Linguistics, 2010.
4. Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 168–171. Association for Computational Linguistics, 2003.
5. Michael Collins and Yoram Singer. Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.
6. Andrei Mikheev, Marc Moens, and Claire Grover. Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics, 1999.

Sentiment analysis techniques applied to regression tasks

José Ornelas[1]

Faculty of Engineering of the University of Porto
jose.ornelas@fe.up.pt

Abstract. This paper evaluates how some of the literature based classification techniques applied to sentiment analysis over Twitter data behave in regression tasks. Our approach studied how the combination of bag-of-word models, in conjunction with part-of-speech tag and word polarity features, performed using random forest and support vector machine algorithms. We evaluated how these techniques improve the prediction of the arousal score for four sentiment classes: anger, fear, joy and sadness. The part-of-speech features presented the best results for most of the sentiment classes, although in some cases the polarity features presented similar results.

Keywords: data science, sentiment analysis, natural language processing

1 Introduction

Microblogs and social networks generate an enormous quantity of data. Extracting and understanding valuable information from the data has been a challenge for researchers since the rise of these platforms. Extracting and performing sentiment analysis on text has always, also, been a challenge. Microblogging increased the challenge through the specifities that each platform imposes to their users.

This paper focus on performing sentiment analysis over Twitter data. The users interact with the platform through texts and media, although we will only focus on the textual part. Users interaction with the platform has some specifities. The messages inserted in the platform are called tweets and most of the time include textual features such as emoticons, hashtags, links, slang and abreviations.

Sentiment analysis over Twitter data can help brands to understand their consumers, celebrities to understand their followers, politicians to understand their impact, etc. Most of the past analysis over this kind of data focused on understanding the binary (positive and negative) or 3-way polarity (positive, negative and neutral) of the messages. The past approaches addressed the problem as a classification task.

In this paper, our goal is to study how the approaches applied in classification tasks perform when applied in conjunction with regression techniques. In order to do this, we will use a dataset that is divided in four different sentiment classes: anger, fear, joy and sadness. For each class, we will try to predict the arousal score that displays the strength for each class. Each of the classes will be addressed separately. This work will be based on an existing dataset that contains an approximately total of 7000 messages annotated with the arousal score.

The result of this work will be submitted to the SemEval-2018 competition Task 1: "Affect in tweets"[1]. The work presented in this paper reveals some preliminary results that provide valuable insight that will be used during the participation in the competition.

The remainder of this paper is organized as follows. Section 2, contextualizes about the work made on the area of sentiment analysis over Twitter data. Section 3, describes the dataset organization and displays an analysis over the data. Section 4, describes the techniques applied during the preprocessing phase. Section 5, presents the techniques that were used to address the problem and how the impact will be measured. In Section 6 we present the results of our work. Finally, section 7 presents the conclusion and future work.

2 Related work

With the increase of the Web 2.0, blogs, microblogs and social networks started to be a popular field for Sentiment Analysis. Microblogs and social networks like Twitter became the place where people discuss and make opinions about everything, including brands, products or celebrities. These platforms generate an enormous quantity of data which contain relevant value, for example, brands are able to determine their users sentiments based on product reviews[4, 3].

Naturally, researchers started applying their effort on performing natural language processing techniques on these platform's data. Researching on Twitter data presented researchers with challenges mainly because of the specificity of the data generated by the platform. Tweets are short text messages, with a maximum of 140 characters long and are characterized by casual, compact and slang language. In order to fill a message in a Twitter post users resort on abbreviations, acronyms and emoticons. Also, these messages also contain hashtags, links to other websites or user references.

Research on sentiment analysis can be categorized in two main techniques. The first is an unsupervised technique and applies a sentiment lexicon with binary fashion terms, positive or negative, to evaluate the text[11]. The second is a supervised technique that uses textual feature representations coupled with

[1] http://alt.qcri.org/semeval2018/index.php?id=tasks

machine learning algorithms in order to extract the relationship between the opinion and text features [8, 7].

Regarding the sentiment analysis of Twitter data, most of the research derives from the literature[7]. Twitter provides an API[2] that helps collecting data from the platform. Some researchers collect and classify their own data[6, 9, 3], others use well-know datasets[5].

Pak and Paroubek[6], presented a method for automatically collect a corpus that is suitable to train a sentiment classifier. Their Naive-Bayes classifier resorts on n-gram and POS[3] tags, generated by TreeTagger[4], as features to determine positive, negative or neutral sentiments. Barbosa and Feng [2], presented a robust 2-step sentiment detection framework that focus on subjectivity detection and polarity detection. This framework abstracts the representation of sentences as features in order to categorize each sentence in positive, negative or neutral. Kouloumpis, Wilson and Moore[5], investigate the impact of linguistic features on sentiment analysis of twitter data. Their work evaluated the usefulness of existing lexical resources as well as features that capture information about the informal language specifities used in Twitter. Their models are trained with features from the HASH and EMOT datasets. The HASH dataset is a subset of Edinburgh dataset that helps determining the messages polarity based on the hashtags contained on the message. The EMOT dataset follows the same logic but resorts on emoticon to determine the message polarity. Their work evaluates the accuracy of using the combination of n-grams, lexical features and POS features both in HASH and HASH plus EMOT datasets. Although the authors determine that further analysis is necessary, their analysis reflects that POS features may not be useful comparing with the use of sentiment lexicon in conjunction with the microblogging features. Agarwal and Xie [1] work also focus on the use of POS polarity tags as features. Their results present a 4% gain in two similar classification tasks: binary (positive vs negative) and 3-way (positive, negative and neutral). The author's present tree kernel and feature based models that outperform the unigram baseline. Saif, He and Alani[9], focus on defining two sets of features that help to alleviate the data sparsity problem in Twitter sentiment classification, semantics features and sentiment-topic features. The author's works compare the use of semantic features vs semantic-topic features. Both methods outperformed the Naive Bayes based on unigram features only baseline, while using semantic-topic features returns better results than using semantic features with less features. Spencer and Uchyigit[10] work focus on Sentimentor platform that uses a Naive Bayes classifier to live classify Twitter data. Their work is aligned with Pak[6].

[2] Application Programming Interface

[3] Part-of-speach

[4] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger

As seen above most of the research work on sentiment analysis focus on binary or 3-way classification tasks. This paper will explore how some regression techniques perform for 4 specific sentiment types: anger, fear, joy and sadness.

3 Data description and analysis

Twitter is a social platform that restricts the users messages to 140 characters long. Due to nature of the platform (quick and short), users make some adjustments to their messages. They rely on the use of abbreviations, acronyms, hashtags and emoticons. The platform includes some specifities that will be described below. Emoticons represent facial expressions in a cartoonish way. Hashtags are used to mark topics and to increase the visibility of a tweet. Target(@) is used to refer and alert other users.

In this study we used a dataset from the SemEval[5] competition. The dataset contains 7102 tweets divided in four categories: anger, fear, joy and sadness. Table 1 presents examples for each of the categories. Each of the datasets row contain an id, the message, the type and the arousal score. The arousal score, that ranges between 0 and 1, represents the strength of the message for the sentiment type.

Figures 1 to 4 display the score distribution for each of the sentiment types. As seen in the figures none of the sentiments is biased to a weak or strong score. The segment 0.3 to 0.7 prevails for the four types.

Table 1. Examples of tweets

id	message	type	score
10023	Tasers immobilize, if you taser someone why the fuck do you need to shoot them one second later?! This is really sick! #rage #wtf,#murder	anger	0.812
40000	Depression sucks! #depression	sadness	0.958
20251	@AmyMek this is so absurd I could laugh right now (if I also didn't feel like crying for the future of our country). #despair #wakeupcall	fear	0.655
30780	People always tell me that they don't expect me to have anxiety because,I'm generally cheerful and don't act the way they expect me,to.	joy	0.160

[5] https://competitions.codalab.org/competitions/17751

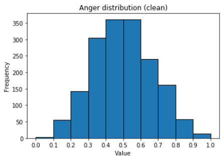


Fig. 1. Anger score distribution

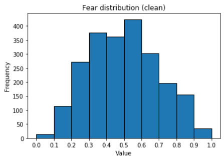


Fig. 2. Fear score distribution

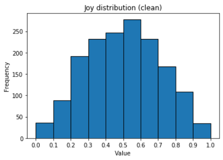


Fig. 3. Joy score distribution

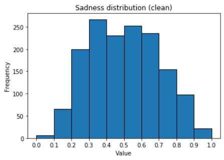


Fig. 4. Sadness score distribution

4 Pre-processing

The pre-processing phase encompasses 3 main steps. First, we will search and remove the most common words in English language that appear in each message. For the task, we will use the NLTK[6] Corpus Stopwords. Second, we will "translate" each of emoticons to their standard code based on the list provided by the unicode consortium[7]. The task will be supported by the Emoji for Python project[8]. Finally, we will group inflected forms of words. For this, we will use the WordNetLemmatizer algorithm provided in the NLTK package.

5 Approach

Our approach studies how some of the techniques seen in section 2 behave in regression tasks. Some of the features seen previously, such as POS tags and word's polarity, were studied in order to understand how they affect the overall score. These will be tested individually in combination with different models and algorithms. In order, to evaluate the performance of the classifiers we will use the Pearson Correlation Coefficient.

5.1 Models description

In order to understand how models affect the predictions we decided to use three variations of the bag of words (BoW) model. The first approach uses a

[6] http://www.nltk.org/
[7] https://www.unicode.org/emoji/charts/full-emoji-list.html
[8] https://github.com/carpedm20/emoji

simple BoW representation where each term in each tweet is represented by 0 and 1, depending if the word appears or not in the document. The second approach is a term frequency (TF) model, where each term will be measured by frequency of occurrence in all the tweets. The third and last method is a term frequency and inverse document frequency (TF-IDF) model, where each term will be represented by the overall importance in all the tweets.

5.2 Feature selection

The features selection process focused in studying how POS tags and word polarity features affect the tweet overall score. The POS tagger processes sequences of words and attaches a POS tag to each word. In this work we used the NLTK POS tagger. The word polarity was obtained through the TextBlob[9] library. TextBlob is a library that extends NLTK and pattern[10] libraries. The library includes a sentiment analyzer that provides words and sentences polarity between -1 and 1. Each word was classified with negative, neutral or positive tags as represented in Table 2.

Table 2. Word polarity classification

polarity value	polarity tag
less than 0	negative
more than 0	positive
exactly 0	neutral

5.3 Regression algorithms

The algorithms used in this work were the Random Forest Regressor (RF) and the Support Vector Regressor (SVR) from the Support Vector Machine (SVM) algorithms family. None, of the algorithms was enhanced, in both cases we have used the default options provided by the Sklearn[11] library.

5.4 Evaluation metrics

In order to evaluate how the models and algorithms performed we used the Pearson correlation coefficient, also known as Pearson's r. This metric provides a measure, between -1 and 1, of the linear correlation between the values. Where 1 represents a high positive correlation, -1 represents a high negative correlation and 0 represent no correlation between the values.

[9] https://textblob.readthedocs.io/en/dev/

[10] https://www.clips.uantwerpen.be/pages/pattern-en

[11] http://scikit-learn.org/stable/

6 Results and Analysis

This work aimed to help identifying which methods are best appropriated for each sentiment type. We used the Pearson correlation coefficient to determine the behaviour of six combinations of models and algorithms. The number results of the combination between the two algorithms used, random forest and support vector machine, and the three models presented in 5.1. For each of the six cases, we studied three different types of approach:

- baseline: composed by the cleaned dataset
- POS tags: the baseline plus the POS tags features
- Polarity: the baseline plus the word polarity features

Each type of sentiment was studied in separate, with different datasets for each type. Although the models and algorithms that were used were developed in the same way each type of sentiment has its own models.

6.1 Anger

In the anger results, presented in figure 5, its possible to see that the RF algorithms present better results. The SVM algorithms present more contradictory results. The polarity features provide better results in four of the six cases (RF-TFIDF and all three SVM variants). Although, the best overall score is represented by the RF-BoW using POS tag features.

6.2 Fear

Figure 6 presents the fear results. In this case the best overall score is again presented by the RF-BoW using POS tag features. In this case, the best score is followed closely by RF-TFIDF with polarity features. Once again, the polarity features outperform the baseline and the POS tag features in four of the cases (all the SVM variants and RF-TFIDF).

6.3 Joy

In figure 7 we present the joy results. The best overall score is obtained by the SVM-TFIDF with POS tag features. In most of the cases the baseline can was outperformed, and the performance was almos two times better. The SVM variants present contradictory results, the SVM-BOW presents the best baseline but including our features didn't present better results. Also, the SVM-TF with POS tags performs worse than most of the baseline cases. The RF cases present similar behaviours with all the three cases presenting better results with polarity features.

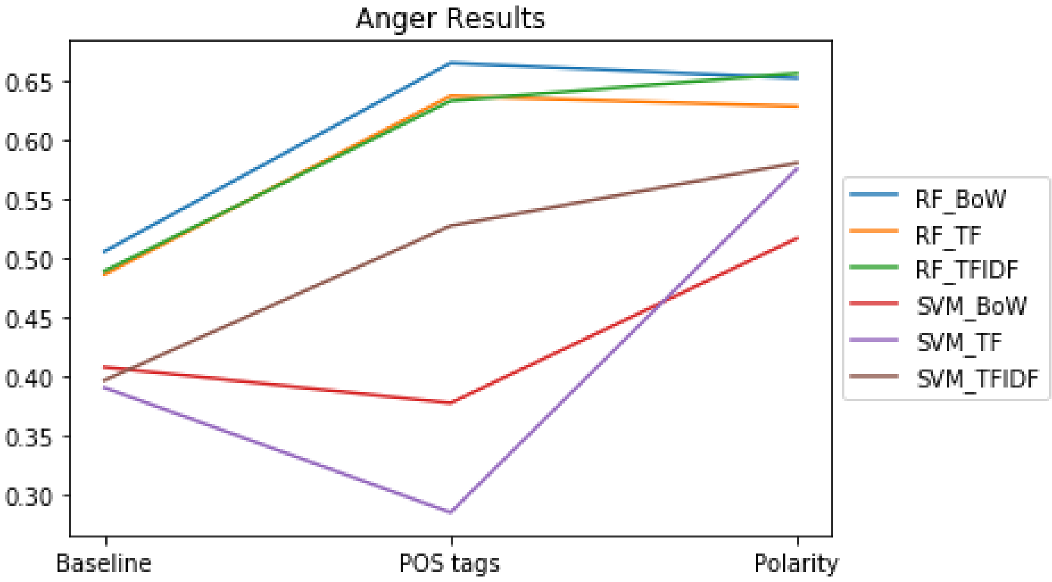


Fig. 5. Anger results

6.4 Sadness

The sadness results, figure 8, display that the two best overall scores for RF-TF and RF-BoW are accomplished with POS tags features. In all the cases the baseline is outperformed by both the POS tag and polarity features. The SVM-TFIDF present a behaviour similar two the two best scores. In the remaining cases display that the POS tag features are outperformed by the polarity features.

7 Conclusion and Future work

Sentiment analysis over twitter data is a complex field. The limited number of words in each tweet and the specificities of the platform (hashtags, emoticons, etc..) present several challenges when identifying different types of sentiments.

In this work we studied how some of the approaches used for sentiment analysis in classification tasks behave in regression tasks for four different types of sentiment. In our results we can see that for most of the sentiments the POS tags features present better results. Although, for some of the cases polarity features present close results. Also, the RF algorithms presented better results for three of the sentiments.

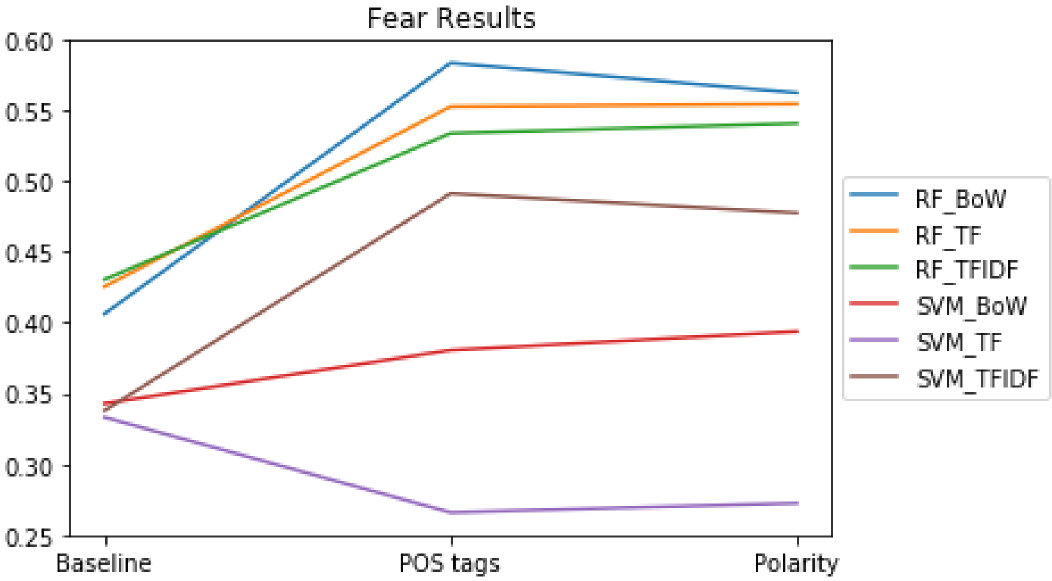


Fig. 6. Fear results

With all this information we will be able, in the future, to take better decisions on selecting the appropriate models, algorithms and features to perform sentiment analysis for these four types of sentiment.

As future work, we will add another two sets of features: emoji polarity and hashtag polarity. We think that both features can add valuable information that is likely to improve the overall score obtained in this work. Also, we will enhance both the algorithms to understand how we can improve the overall scores for each sentiment.

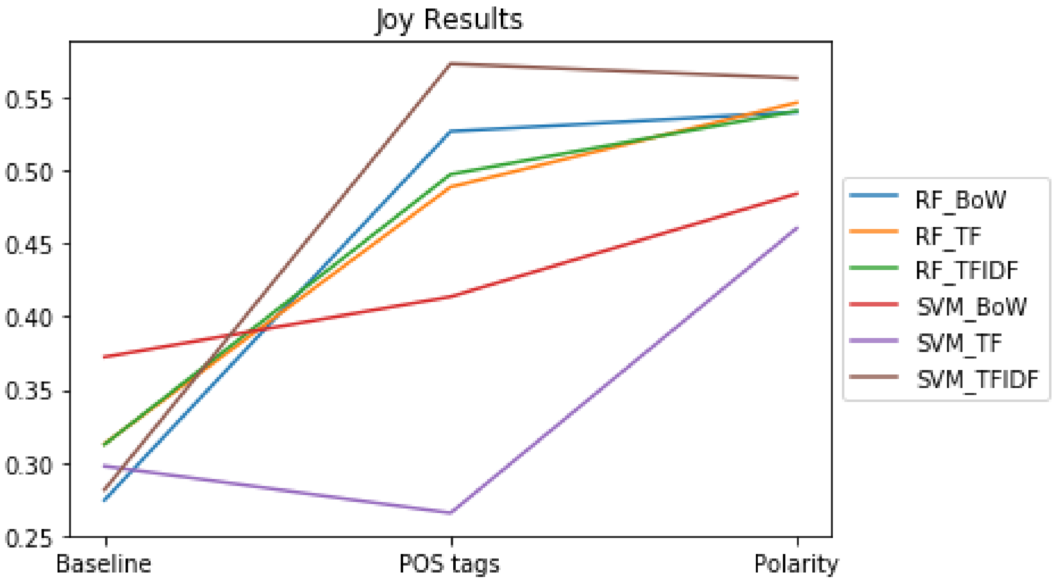


Fig. 7. Joy results

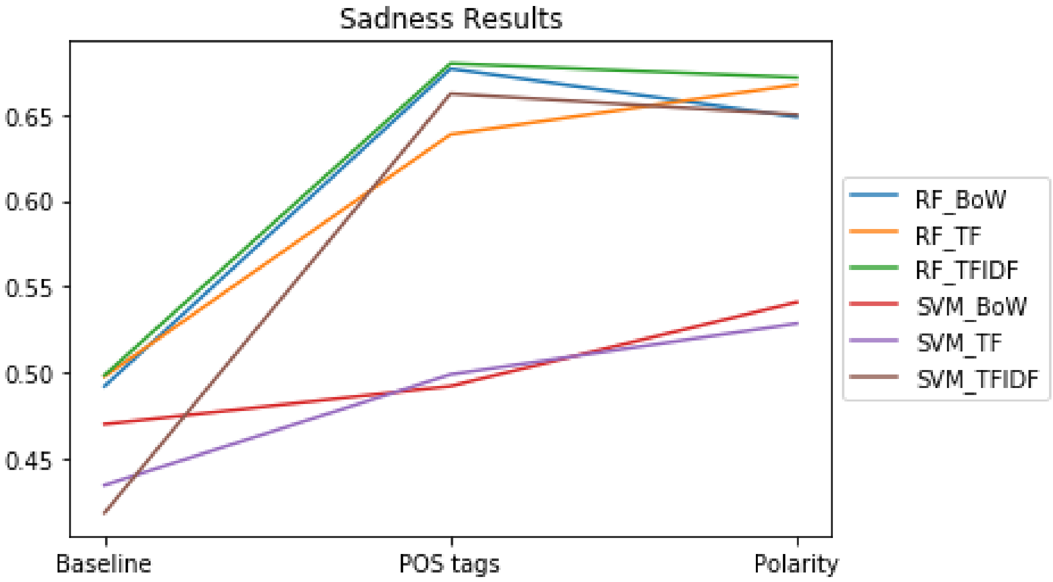


Fig. 8. Sadness results

References

1. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, 2011.
2. Luciano Barbosa and Junlan Feng. Robust Sentiment Detection on Twitter from Biased and Noisy Data. *Coling*, (August):36–44, 2010.
3. M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, 40(16):6266–6282, 2013.
4. Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, and Nick Bassiliades. Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, 40(10):4065–4074, 2013.
5. E Kouloumpis, T Wilson, and J Moore. Twitter sentiment analysis: The good the bad and the omg! *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 11)*, pages 538–541, 2011.
6. Alexander Pak and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *In Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pages 1320–1326, 2010.
7. Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in InformatioPang, B., & Lee, L. (2006). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 1(2), 91–231. doi:10.1561/1500000001n Retrieval*, 1(2):91–231, 2006.
8. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP '02*, volume 10, pages 79–86, 2002.
9. Hassan Saif, Yulan He, and Harith Alani. Alleviating data sparsity for twitter sentiment analysis. In *CEUR Workshop Proceedings*, volume 838, pages 2–9, 2012.
10. James Spencer and Gulden Uchyigit. Sentimentor: Sentiment analysis of twitter data. In *CEUR Workshop Proceedings*, volume 917, pages 56–66, 2012.
11. Peter D Turney. Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, (July):417–424, 2002.

SESSION 6

Computer Vision, Blockchain and Encryption

Computer-Vision-based surveillance of Intelligent Transportation Systems
João Neto and Rosaldo Rossetti

The contribution of blockchain technology for business innovation
Bruno Tavares

Towards an Access Control System for IoT on Blockchain
João Pedro Dias and Hugo Sereno Ferreira

Computer-Vision-based surveillance of Intelligent Transportation Systems

João Neto, Diogo Santos, and Rosaldo Rossetti

Faculdade de Engenharia, Universidade do Porto, Porto, Portugal,
jneto@fe.up.pt

Abstract. This project focuses on the development of a video analytics processor for detection and classification of vehicles. The motivation for this project comes from the need to have a plug-and-play solution to analyse traffic, as most of the existing solutions require training some sort of structure to recognize objects on the scene. This module can work as a data input for traffic management systems in addition to more traditional sensors such as the magnetic loop detectors. We also present a novel approach to the vehicle classification problem based on the use of a fuzzy set. To illustrate the proposed approach, the detection and classification implemented were tested with different cameras in different scenarios, showing promising results.

Keywords: computer vision, intelligent transportation systems, object tracking, object classification

1 Introduction

Traffic management in cities is a vast area as it studies the planning and control of the road network, with all the tasks associated with them. As the cities tend to evolve following the concept of Smart Cities, the management of how it is people move is a prime area where the information technologies can be applied. With the increase in the number of vehicles using the roads [1] so does the need to improve the methods of traffic management, and the development of multiple projects in multiple institutions around this issue comes as a confirmation of both its importance and the pertinence of the work under development in regards to this topic.

With the decreasing costs of cameras for video surveillance, the number of units installed around the world is rising, passing the 245 million mark in 2014 [2], over 65% of that number being from Asia. With this number of cameras placed globally the amount of data being collected every day is too large to be humanly processed, and thus the need to create autonomous processors arises. The market for automatic analysis of video is expected to be worth 11.17 Billion USD by 2022 [3], with the facial recognition area expected to have the highest CAGR (Compound Annual Growth Rate) due to the potential related to security applications.

This work reports on the development and implementation of a Computer-Vision-based framework to analyse moving-object scenes with multiple applications; the selected case study refers to road traffic networks, both in rural/motorway and urban settings. A number of supporting features were explored and practically implemented, namely object detection, object tracking, and object classification. Such features helped us implement analytics regarding traffic flow.

With the potential perspective of using computer-vision-based solutions in the specific domains of traffic and transportation engineering, the goal of this work is to implement appropriate mechanisms allowing for the identification and classification of moving objects on road networks. Providing traffic surveillance systems with the ability to automate the process of extracting information from video cameras is paramount to improve control and management of complex networks. Indeed, video streams capturing continuous images from selected spots can offer the appropriate means for incident detection, alarms, and complement data gathered from other traditional sensors, such as inductive loops installed beneath the pavement. Contrary to inductive loops, which are prone to reading failures more frequently than other analogous systems, video cameras can offer a wide range of other sensing capabilities. On the other hand, other challenges arise demanding robust and efficient algorithms. This work focuses on the specific problems of detecting and classifying moving objects in urban areas and motorway environments, with the objective of feeding more accurate and reliable information into simulation models underlying fully operational artificial transportation system platforms, such as the MAS-Ter Lab architecture [4].

The remaining of this manuscript is organised as follows. Section II briefly reviews the literature on topics related to this work, whereas Section III describe how the problem is approached, giving the reader details on the proposed approach. Section IV presents and discusses preliminary results, whereas Section V draws conclusions and suggests further steps in this research project.

2 Computer Vision in ITSs - A Review

Regarding intelligent transport systems (ITSs), the use of computer vision to aid in the analysis of traffic has been increasing in the last years due to the decreasing costs of hardware, both cameras, storage and processing power, as well as the growing knowledge to extract useful data from the video gathered. In contrast to the high installation and upkeep costs of other traffic control tools such as Inductive Loop Detectors and Microwave Vehicle Detectors, applying computer vision to handle these tasks is a profitable option for entities in charge of the analysis of this data.

One of the first works applying computer vision to analyse traffic, which was published in 1984 [5] and detected and measured movement in a sequence of frames. Since then, multiple fields of study have been created, not only for the measuring of traffic, as explored in [6], [7] and [8] where the authors evaluate methods to analyse traffic in urban environments, but also for the analysis of the environment around a self-driven vehicle, reading traffic signs using multiple approaches using convolutional neural networks [9] or using bag-of-visual-words [10], or to automatize the parking process as discussed in [11]. Recently some studies have surfaced where the authors discuss the analysis of passenger numbers and behaviour inside vehicles, in order to enforce traffic laws [12].

However our work upon this topic is focused on the aspect of traffic analysis. In order to understand what we need to accomplish, it is convenient to study what composes a typical traffic scene. Usually the scenes are viewed from a top-down perspective that places the cars against the road as background. The vehicles have a roughly regular rectangular shape when viewed from the top, with little variation when the camera is rotated, however their textures vary heavily, making it difficult for a detector to work based on the vehicles image representation [13]. However, depending on the camera position there can be object occlusion by other objects or scene components, which makes it difficult to detect and track vehicles relying solely on subtraction based segmentation techniques.

The scenes also have varying light according to the time of the day, and proposed solutions need to adapt to these changes as fast as possible, otherwise data might be lost. Other weather conditions can also interfere with the analysis, such as fog and rain, and need to be addressed when designing solutions.

3 Methodological Approach

3.1 Technology

In order to build this project we needed both a library of already implemented computer vision algorithms as well as a simple way to retrieve frames from both video streams and files. This section describes what were the chosen technologies including a brief description and why it was chosen.

OpenCV OpenCV is a library composed of implementations of useful computer vision algorithms implementation, widely used across the industry and academy. It has interfaces for multiple programming languages, like C++, Java and Python, but is natively written in C++ in order to take advantage of low level performance enhancements, as performance is an important factor in real time computer vision applications [14].

The library contains over 2500 algorithms ranging from the more basic image processing, such as filtering, morphology operators and geometric transformations, to more complex ones that are able to compare images, track features, follow camera movements and recognize faces, among others. Along with the image processing capabilities, OpenCV also ships with interfaces to stereo cameras such as Kinect that allow users to retrieve a cloud of 3D points and a depth map from the captured image. This was the chosen library as there existed already previous work at LIACC using it, which could be leveraged for this project.

JavaCV JavaCV is a wrapper for OpenCV written in Java that works on top of the JavaCPP Presets, a project that provides Java interfaces for commonly used C++ libraries, such as OpenCV and FFmpeg, the ones we are using, as well as CUDA, ARToolKitPlus and others. It provides access to all the functionalities of OpenCV inside a Java environment, and was the chosen solution as there was already experience inside LIACC working with this technology.

Even though the code is written in Java and runs inside a Java Virtual Machine, the code from OpenCV is compiled from C/C++ and the memory of the objects created there is allocated in a separate thread. This made it impossible to rely on the garbage collector to do the memory management, and necessary to manually delete the native objects. Failure to address this issue causes the system to run out of memory and a subsequent program crash.

3.2 Segmentation

This section describes how the segmentation of the received images is performed, and how it returns a foreground mask representing the moving areas of the scene.

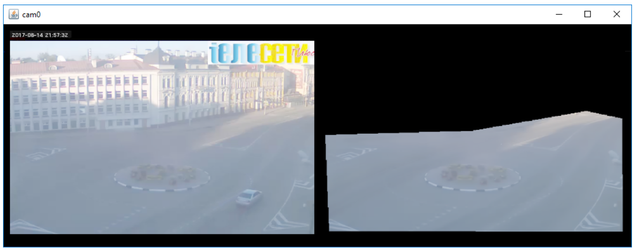


Fig. 1. Background Subtraction - Background Model

In figure 1 we can see the original frame on the left and the calculated background model of the scene on the right. To achieve this result, the first step is to mask the obtained frame in order to prevent uninteresting regions of the image from being processed by the Background Subtracter. The masking process consists in placing a binary image, called a mask, over the original image and removing all the information where the mask has false values.

The application of the mask solves an issue where moving or changing regions of the image outside our area of interest would create unwanted artifacts, for example moving trees due to the wind blowing, or the issue that occurred in this scene, where a car passing would be reflected in the windows of the building.

The next step of the Segmentation process is to feed the masked frames into a Background Subtraction algorithm that will use them to update its internal representation of the scene. This project uses the OpenCV implementation of the Mixture of Gaussians algorithm that allows the user to tune:

- The number of past frames considered on the background calculation
- The threshold value from which a pixel is considered to be foreground, compared to the difference between its current value and the one from the background model
- The learning rate of the algorithm, how much each new frame influences the model

After updating the background model we can retrieve from the Background Subtracter its foreground mask, a rough representation that is calculated by subtracting the background model from the current image and thresholding it, thus returning only the pixels where the difference is significant enough.

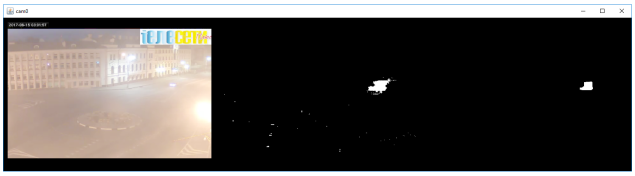


Fig. 2. Background Subtraction - Foreground Mask

As we can see in the middle image of figure 2 the foreground mask from the Background Subtracter can have a lot of noise due to lighting conditions. To solve this issue two morphology filters are applied to the image: a squared erosion filter to remove the small speckles that appear in the mask; followed by a larger circular dilation filter to consolidate the positive regions of the mask, as the wind shield and windows of the vehicles are usually detected as background due to their dark colour and/or reflection of the environment.

4 Object Classification

In order to classify the objects present in the scene into light or heavy vehicles a fuzzy set is used. This set is calculated at run time based on a mask drawn by the user via the web interface that roughly approximates the size of a light vehicle. The area of that mask ($\mathrm{A}_{\mathrm{Light}}$) is used as a base value to create the fuzzy set shown in figure 3. This set has 2 series, one for light vehicles, drawn in blue, and one for heavy vehicles, drawn in red.

The blue series peaks at $\mathrm{A}_{\mathrm{Light}}$, where we have 100% certainty that a matched object is a light vehicle. The point immediate points are at $2/3^*\mathrm{A}_{\mathrm{Light}}$, where the trust is 80% and at $4/3^*\mathrm{A}_{\mathrm{Light}}$ where it drops to 60%. Any object with area below $1/2^*\mathrm{A}_{\mathrm{Light}}$ or above $2^*\mathrm{A}_{\mathrm{Light}}$ are considered to have 0% chance to be light vehicles.

The red series peaks at $2^*\mathrm{A}_{\mathrm{Light}}$, and any object whose area is larger than this value is considered to be an heavy vehicle with 100% confidence. At the

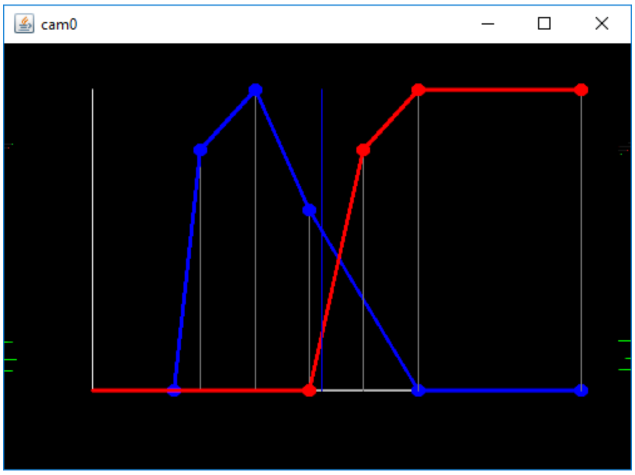


Fig. 3. Fuzzy Set used by the Object Classifier

same time, any object whose area is smaller than 4/3*A_{Light} is never considered to be an heavy vehicle, from where the certainty rises to 80% at 5/3*A_{Light}.

When an object is counted in one of the virtual sensors its area is passed to this classifier where an interpolation is calculated for each series, returning the certainty with which it belongs to each one of the classes. Using this output the process can distinguish classifications with certainty values below a user specified threshold to be treated as non-classified objects.

5 Results

In this section we will present the results achieved by our project on different settings, described below.

5.1 Video 1

- **Location** Via Cintura Interna, Porto
- **Description** Motorway scene
- **Traffic Intensity** Medium
- **Camera Position** Medium Height, Left side of lane

5.2 Video 2

- **Location** Via Cintura Interna, Porto
- **Description** Motorway scene
- **Traffic Intensity** Medium
- **Camera Position** Medium Height, Between Lanes

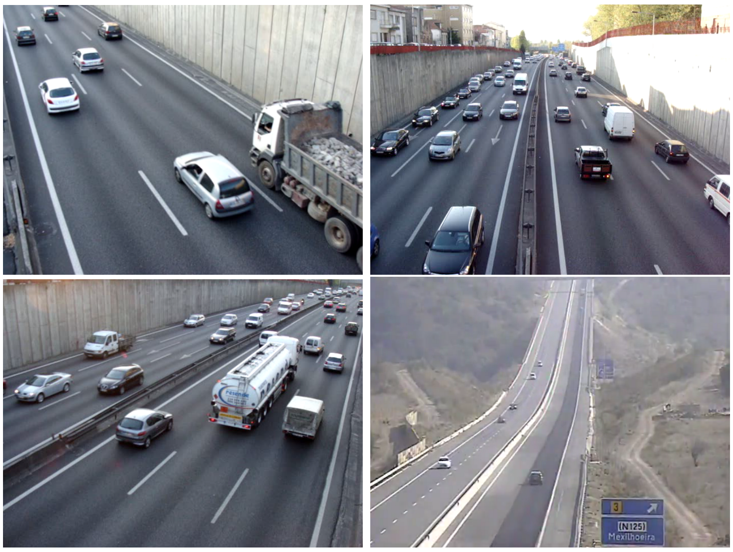


Fig. 4. Frames from Videos used

Table 1. Results from Video 1

	Light	Heavy
Observed	141	22
Detected	126	20
Ratio	89.3%	90.9%

Table 2. Results from Video 2

	Light	Heavy
Observed	155	30
Detected	131	28
Ratio	84.5%	93.3%

5.3 Video 3

- **Location** Via Cintura Interna, Porto
- **Description** Motorway scene
- **Traffic Intensity** High
- **Camera Position** Medium Height, Right side of Lane

5.4 Video 4

- **Location** A22, Mexilhoeira, Portugal
- **Description** Motorway scene
- **Traffic Intensity** Low

Table 3. Results from Video 3

	Light	Heavy
Observed	178	28
Detected	176	24
Ratio	98.9%	85.7%

Table 4. Results from Video 4 (Incoming and Outgoing Lanes)

	Light	Heavy	Light	Heavy
Observed	47	5	30	0
Detected	38	5	22	0
Ratio	80.9%	100%	73.3%	-

– **Camera Position** Elevated Height, Right side of Lane

Due to the fact that we use a fuzzy set to classify the vehicles based on their segmented region area, we can know with which degree of certainty the system classifies each vehicle as Light or Heavy. We can use this value to threshold weak classifications and, depending on the application, send the image to a more sophisticated classifier or even to a human operator.

The main drawback of the fuzzy set based classification is its dependency on the object identification process. As our segmentation process returns a binary foreground mask it is impossible to distinguish between multiple vehicles in a single foreground region, and only one object is detected. This object area is then taken into account by the classifier, and in cases where multiple Light vehicles occlude each other, they are classified as a single Heavy vehicle.

The main factors that contribute to the performance of our counting and classification processes are the camera position and the traffic intensity, as seen in the results. In videos one through three the medium to high density of traffic cause multiple occlusions which result in wrong classifications as explained before. In cases where the camera is placed in a high position overlooking the road, such as in video four, the occurrence of vehicle occlusion is reduced. When the viewing angle

6 Conclusion

This work reported on the development and implementation of a Computer-Vision-based framework to analyze moving-object scenes with multiple applications. Different case studies were initially selected with special emphasis on the identification and classification processes.

During the course of this project work an issue was found when trying to process a frame in multiple threads at the same time. The solution implemented creates a queue in each of the threads to where the frame collector sends the frames and from where the thread processing them reads them. This solution solves the waiting problem by making the threads totally independent from each other and ensures that the whole system performance is not capped if a slower thread is introduced.

One of the main contributions of the project is related to the tracking of stopped vehicles found in urban scenarios, as it was one of the gaps found during the literature review. To solve this issue an alteration is proposed in the segmentation process to keep track of the state of stopped vehicles. The first one is a stabilization step in the background subtraction initialization that waits for the background model to be steady. This steadiness is measured by the number of pixels updated in the last frame after the application of the background subtracter. This step ensures that the background is not initialized with stopped vehicles or other actors in the middle of the scene. All the work developed was made taken into account that the application was to be used by our collaborators for the purpose of vehicle counting and classification in high-way scenarios and as such this was the main focus of the use-cases in which the project was tested.

Throughout the course of this work a number of challenges appeared that would be interesting to address. Some of them are briefly presented below.

Improve the Segmentation Process: as of now the segmentation process cannot distinguish multiple objects occluded by one another or linked through a shadow, although there is written work about how to solve this. Implementing such a solution would improve the results of the counting process.

Time counting: The counting process can accurately count vehicles but it cannot detect when a vehicle was counted. To do so in videos, a starting time would be required as well as the frame count and rate. In streams the application needs to take into account both the starting time and the stream delay.

Intersection Analysis: It would be interesting to follow some of the work regarding intersection statistics and implement a module on top of the Analytics Module that would work specifically for such cases.

References

1. Navigant. Transportation Forecast: Light Duty Vehicles, 2017-04-18T20:30:25+00:00.
2. Niall Jenkins. 245 million video surveillance cameras installed globally in 2014 - IHS Technology, 2015.
3. ReportLinker. Global Video Analytics Market 2017-2021 - market research report, 2017.
4. Rosaldo JF Rossetti, Eugénio C Oliveira, and Ana LC Bazzan. Towards a specification of a framework for sustainable transportation analysis. In *13th Portuguese Conference on Artificial Intelligence, EPIA, Guimarães, Portugal*, pages 179–190. APPIA, 2007.
5. K. W. Dickinson and R. C. Waterfall. IMAGE PROCESSING APPLIED TO TRAFFIC, 1-A GENERAL REVIEW. *Traffic Engineering & Control*, 25(1), January 1984.
6. G. Lira, Z. Kokkinogenis, R. J. F. Rossetti, D. C. Moura, and T. Rúbio. A computer-vision approach to traffic analysis over intersections. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 47–53, November 2016.

7. N. Buch, S. A. Velastin, and J. Orwell. A Review of Computer Vision Techniques for the Analysis of Urban Traffic. *IEEE Transactions on Intelligent Transportation Systems*, 12(3):920–939, September 2011.
8. M. F. Hashmi and A. G. Keskar. Analysis and monitoring of a high density traffic flow at T-intersection using statistical computer vision based approach. In *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, pages 52–57, November 2012.
9. D. Soendoro and I. Supriana. Traffic sign recognition with Color-based Method, shape-arc estimation and SVM. In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, pages 1–6, July 2011.
10. C. Supriyanto, A. Luthfiarta, and J. Zeniarja. An unsupervised approach for traffic sign recognition based on bag-of-visual-words. In *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 1–4, October 2016.
11. K. Hammoudi, H. Benhabiles, A. Jandial, F. Dornaika, and J. Mouzna. Self-driven and direct spatio-temporal mechanisms for the vision-based parking slot surveillance. In *2016 SAI Computing Conference (SAI)*, pages 1327–1329, July 2016.
12. Robert P. Loce, Raja Bala, and Mohan Trivedi. Detection of Passenger Compartment Violations. In *Computer Vision and Imaging in Intelligent Transportation Systems*, pages 432–. Wiley-IEEE Press, 2017.
13. Jorge Badenas and Filiberto Pla. Applying computer vision techniques to traffic monitoring tasks. In *Methodology and Tools in Knowledge-Based Systems*, pages 776–785. Springer, Berlin, Heidelberg, June 1998.
14. OpenCV. About - OpenCV library, 2017.

The contribution of blockchain technology for business innovation

Bruno Tavares [0000-0001-8307-3923]

Faculty of Engineering University of Porto
Porto, Portugal
up201700372@fe.up.pt

Abstract. In many businesses, information is key, however the quality of the information is unclear, the blockchain technology promises to bring transparency and reliability and solve information quality problem for business. This work aims to find what business areas are looking into the technology and try to leverage the blockchain technology to innovate their business. Papers that related business integration with blockchain are part of this revision. This paper identifies the kind of applications have already been found for the technology. In this study we found out that for the technology be able to have massive acceptance in business, it is necessary that legal aspects of the technology progress and that the regulation is applied. It also becomes clear that in most systems, where information need to be certifiable and verifiable by a trusted third party, it is possible to leverage the blockchain technology. A considerable acceptance of the technology overtime is a great possibility and business must be ready to accommodate this change.

Keywords: Blockchain, business innovation, disruptive technology.

1 Introduction

The blockchain technology gathered a lot of traction in the past years, however academics work, that focus on the technology, only now started to look for new applications. Different kind of business are already trying to leverage this technology. This literature review aims to find what business areas have already integrated the technology, and how blockchain as contributed for the innovation in the business. For retaining literature, the criteria used was, papers related with blockchain, and business integration of the technology in preexisting areas. Papers that focused on the cryptocurrency capabilities of the blockchain, and blockchain frameworks are not the focus of this review.

Business innovation is a process for introducing new ideas, technologies, methodologies, services or products. Business innovation should improve products, services, processes, solve a problem, or it should reach new customers. The blockchain technology promises a high impact on old and new business. The blockchain offers decentralization, transactions, automation without the need for intermediaries. With the application of the blockchain technology, it is expected to reduce costs, processing time, risks, and create transparent ecosystems[7].

This paper first provides an overview of the blockchain technology. The third section exposes a literature review to identify the key topics and the technology roles in business integration. The fourth section analyzes and discuss the reason blockchain is selected. In the end the author presents his conclusions.

2 Background

A blockchain is a form of digital ledger consisting of 'blocks' of information. Each 'block' contains a record of the transactions that occur within a network[25].

Essentially, blockchain is a digital cryptographic ledge with a time-stamp, which works as a peer-to-peer database technology for managing and recording transactions[7]. Algorithms handle the verification and consensus between multiple entities, presumably creating a system that is immune to fraud, tampering, fraud, or central control. The Figure 1 shows a blockchain scheme and out the blocks connect to each other [23].

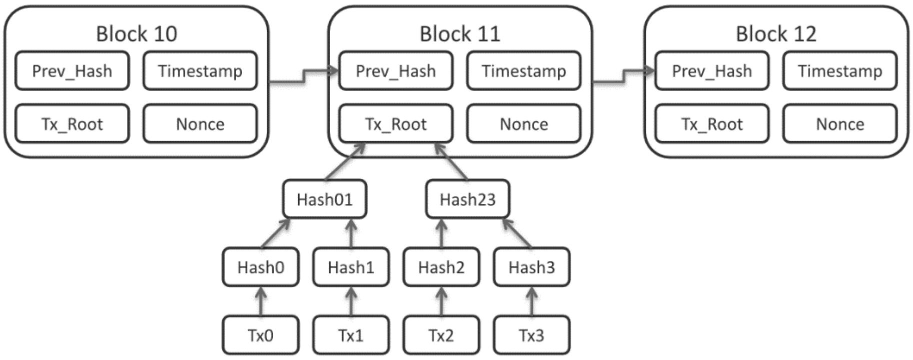


Fig. 1.

Blocks. each block holds information and after the others entities in the network have validated the block it can be added chronologically to the blockchain, together with a reference to the preceding block[9][26].

Transactions. The information in each block registers the transactions of properties or values in the blockchain. The technology ensures the security and correctness of the transactions.

Chain. The chain creation occurs because each block connects to the previous one, registering this way a chain of events. However there is one exception, the first block of the chain is common to all entities in a blockchain network and has no parent[5].

2.1 Blockchain Type

Public Blockchain. In a public blockchain, anyone can read and write information on the blockchain.
Private Blockchain. In a private blockchain, an organization controls the permissions to read and write information on the blockchain.
Permissioned Blockchain. This solution offers a hybrid between the public and private blockchain, for example, anyone can read the information, but only some entities can write information on the blockchain.

2.2 Sector

Govern. The public sector varies by country, but it usually includes services as the military, police, infrastructure (public roads, bridges, tunnels, water supply, sewers, etc.), public transportation, public education, health care and people working for the government. Govern and public agencies are largely funded by taxation and the decisions are made from o political point of view [4].
Commercial. Individuals or shareholders own private firms and the decisions are generally made with a commercial intent where the results constrains are imposed by the market force and the political system [4].

2.3 Proof of Work

A proof of work is a slice of data that is difficult to create but easy to verify and which satisfies certain requirements. If one person owns a distributed system of computers, he can assume that they will all cooperate because he controls their behavior. When this is not the case, there is a real need for different computers to prove that they are working toward the same goal.

3 Literature Review

The following Table 1 gives a summary of the main thoughts each paper presented. The x in the table means that the paper explicitly states the information. When the x' is used that means that the author inferred the information from reading the literature.

Table 1. Blockchain literature review

Authors	Type	Sector	Proof of Work	Notes

	Public	Private	Permissioned	Govern	Commercial	Yes	No	Issues Against (-) For Blockchain (+)	Integration (Role) Application (App)

Khaqqi et al. (2018) [12]			x'		x'	x'		Solves inefficient management (+)
								Solves fraud issues (+)
									Role: Reputation System;
									App: Regulation;

Sidhu J. (2017) [22]	x				x	x		Risk of failure with smart contracts (-)	Role: Set of hardened services; App: E-Commerce;
Yuan et al. (2017) [28]	x'				x'	x'		performance (-) privacy (+)	Role: new signature scheme and cryptographic technologies; App: Big data

Engelenburg et al. (2017) [6]			x	x			x	information sharing is:
								- reliable (+)
								- secure (+)
								- confidential (+)
									Role: business-to-government information sharing;
									App: E-Governance;

Shae et al. (2017) [21]			x'		x'	x'		trust transaction (+) data integrity (+) anonymity (+)	Role: assisting in medical decision-making research; App: Healthcare;
Li et al. (2017) [14]			x		x		x	administrative node (+) Mobile peers as field sensory agents (+)	Role: integrates a distributed event-based system with the traditional transaction based system; App: supply chain;

Zhang et al. (2017) [31]	x'				x	x'		Systematic (+)
								High efficiency (+)
								Flexible (+)
								Reasonable (+)
								Low cost (+)
									Role: distributed autonomous corporations;
									App: IoT E-business;

Wu et al. (2017) [26]			x		x		x'	Poor efficiency (-) Security (+) Intelligence (+)	Role: N+X hybrid blockchain storage: -manager block (private nodes) -storage block (public nodes); App: Energy Market;

Mengelkamp et al. (2017) [16]	x			x			x	Lack of regulation (-)
								Decentralized markets (+)
								Resolve conflicts of interest (+)
									Role: Market design framework;
									App: Energy Market;

Nijeholt et al. (2017) [15]		x'		x'			x	Prevents double-spending (+)	Role: Double-Financing prevention framework; App: Decentralized Registry / Regulation
Lee et al. (2017) [13]	x				x		x	Scalability issues (-) Security (+) Integrity (+)	Role: secure firmware update scheme; App: IoT;
Biswas et al. (2017) [2]	x				x	x		Improved reliability (+) Fault tolerance (+) Efficient operation (+) Scalability (+)	Role: blockchain based security framework to enable secure data communication; App: Smart City;

Ahram et al. (2017) [1]			x		x	x		Transparency (+)
								Privacy (+)
								Security (+)
									Role: HealthChain;
									App: Healthcare;

Gaetani et al. (2017) [8]			x	x		x		Performance (-) Compromising (+) Confidentiality (+) Cannot be reverted (+)	Role: SUNFISH proposal Federation-as-a-Service (FaaS) ; App: E-Governance;

Norta A. (2017) [18]			x		x		x	Smart contracts security flaws (-)
								Decentralized autonomous organizations (+)
									Role: distributed governance infrastructure
									
									application-layer smart-contract lifecycle;
									App: E-Governance;

Herbaut et al. (2017) [11]			x		x	x		Performance vs Scalability (-) Provisioning (+) Monitoring (+)	Role: Collaborative video delivery based on blockchain; App: E-Commerce;
Giancaspro M. (2017) [10]	x			x		x		Security concerns (-) Scalability (-) Workforce impact (-) Transparent (+) Anonymity (+) Distributed Ledge (+) Technology (+) Smart contract (+) Increased efficiency (+)	Role: The legal enforceability of smart contracts; it is uncertain whether they will easily adapt to current legal frameworks regulating 'conventional' contracts across jurisdictions; App: Regulation;

Geranio M. (2017) [9]		x			x	x		Nascent technology (-)
								Uncertain regulatory status (-)
								Large energy consumption (-)
									Role: Distributed ledger technology;
									App: Stock exchanges;

								Control, security, and privacy (-) Integration concerns (-) Cultural adoption (-) Cost (-) Irrevocable (+) Auditable (+)	

Mettler M. (2016) [17]	x			x		x		Decentralized database (+)	Role: public health management
									user-oriented medical research based on personal patient data
									drug counterfeiting;
									App: Healthcare;

Bogner et al. (2016) [3]	x				x	x		Don't need a Trusted Third Party (+)	Role: Decentralized Sharing App; App: E-Business sharing objects

Yue et al. (2016) [29]		x		x	x		x	Security (+)
								Don't need a Trusted Third Party (+)
								Decentralized platform (+)
									Role: Electronic Medical Record;
									App: Healthcare;

Zou et al. (2016) [32]	x				x	x		Monitoring (+)	Role: service contract management scheme dispute resolution protocol; App: E-Business;
Xu et al. (2016) [27]			x	x'		x		Security, privacy, scalability and sustainability immutable data storage (+)	Role: decentralized trading market secure data exchange and negotiation; App: E-Business;
Weber et al. (2016) [24]			x		x	x'		No central authority (+) Trust (+) Tamper-proof (+)	Role: technique to integrate blockchain into the choreography of processes in such a way that no central authority is needed, but trust maintained; App: supply chain;
Christidis et al. (2016) [5]		x			x		x	Enables trustless networks (+) Auditable manner (+)	Role: automate time-consuming workflows in the IoT ecosystems; App: IoT;

								Smart contracts allow us to automate complex multi-step processes (+) Scalability aspect of consensus mechanisms (-) Performance (-)	
Norta A. (2016) [19]			x'	x		x'		Agile business networking collaborations (+) Decentralized autonomous organizations (+)	Role: Establishing a decentralized governance infrastructure for enacting cross-organizational business-process aware collaborations; App: E-Governance;

Zhang et al. (2015) [30]								Share information (+)
								Distributed autonomous Corporations (+)
								Low cost (+)
								Flexible (+)
									Role: IoT E-business;
									App: IoT;

Norta A. (2015) [20]	x'				x'	x'		Decentralized autonomous organizations (+) Collaboration (+) Smart contracts (+)	Role: how to set up in a dependable way electronic communities of business collaborating; App: E-Governance;

3.1 Research Challenges

In this literature review, we can resume the future work to testing, implementation, further enhancements, lifecycle management, and regulation. For the development best practices emerge, it is necessary the encouragement of the blockchain research and development, this will also allow the creation of standards [8].

Integrating the blockchain technology to drive business innovation requires exhaustive validation and testing. The execution of testes must validate if the solution offers the performance, security, end user privacy, interoperability, and scalability, needed for the business. Some of the papers provide the academic solution. However, for the business be able to leverage the technology, is still necessary to implement the solution or at least a working prototype [2].

This technology is still emergent and new applications are being discovered, this creates a necessity for more features and further enhancements in terms of smart contracts, API design, ranking mechanism's, credit system's, incentive mechanism's, storage structures, and others. With the increased acceptance of the blockchain technology, regulation becomes necessary, industry experts and experts in the juridical domain must review legal enforceability of the technology. The technology lifecycle and management also must evolve to increase project's predictability and potentiate the growth of technology adoption [18].

4 Analysis and Discussion

4.1 Why blockchain?

Blockchain is a technology that replaces the need for personal trust, for trust in a system. It is a distributed database of records, where the verification of the records occurs through a consensus mechanism. With a blockchain solution, each interested party replicates, hosts, and maintain the ledger.

4.2 Role in Integration

Record keeping is a core function of any business. The records track past events and provide a view of the organization and relationships. Every business keeps its own records, sometimes in a master database, other times the information exists across different units. Unifying the information across different divisions takes a lot of time and is prone to error. In a blockchain solution when some entity enters new information to one copy, all the other copies are updated. The technology eliminates the need for third-party intermediaries to verify or transfer ownership. A decentralized solution was information sharing through transparency reduces the need for trust in a secure and verifiable manner [8] [14].

4.3 Applications

The blockchain technology it has being used to put proof of existence in: legal documents, health records, supply chain, IoT E-business, energy market, E-Governance, decentralized registry, stock exchanges and smart city. The economic, legal and political systems deal with contracts, transactions, and records. They protect assets and set organizational restrictions. They create and authenticate information. This information allows interactions among countries, societies, business, and individuals. Any system or organization that wants to share information with transparency, immutability safely between different parties, can leverage the capabilities offered by the blockchain technology [17].

5 Conclusions

In an information world, the bureaucracies that exist today need to evolve with the digital transformation. The way regulation and business interact need to advance. Blockchain is an open, distributed ledger that can record information between different parties efficiently and in a certifiable and perpetual way. It is also possible to program the solution to trigger operations automatically.

With the blockchain technology, it is possible to create a society where digital code register agreement and store the information in transparent, shared databases. The information is protected from deletion or tampering. In this society, every event (task, payment, agreement, and process) would have a signature that could identify the record, allowing validation and storage.

In this review, we conclude that any system where trust is necessary, for the information be able to retain value, is a good candidate to use the blockchain technology. There are still some technical issues that need to be addressed, for example, when it is not possible to delete or alter information in a public ledger, testing of the solution gains an additional weight in the development process. Another preoccupation that business have with the adoption of a new technology is the legal aspect, and how it can be enforced, regulation needs to evolve to take in consideration the capabilities that the technology offers. Blockchain is a technology that builds a foundation, and has the potential to turn into disruptive approach for business innovation[25].

References

1. Ahram, T. et al.: Blockchain technology innovations. 2017 IEEE Technol. Eng. Manag. Soc. Conf. TEMSCON 2017. 2016, 137–141 (2017).
2. Biswas, K., Technology, A.B.: Securing Smart Cities Using Blockchain Technology. 2016 IEEE 18th Int. Conf. High Perform. Comput. Commun. 5–6 (2016).
3. Bogner, A. et al.: A Decentralised Sharing App running a Smart Contract on the Ethereum Blockchain. Proc. 6th Int. Conf. Internet Things - IoT'16. 177–178 (2016).
4. Boyne, G.A.: Public and Private Management: What's the Difference? J. Manag. Stud. 39, 1, 97–122 (2002).
5. Christidis, K., Devetsikiotis, M.: Blockchains and Smart Contracts for the Internet of Things. IEEE Access. 4, 2292–2303 (2016).
6. Engelenburg, S. van et al.: Design of a software architecture supporting business-to-government information sharing to improve public safety and security. J. Intell. Inf. Syst. (2017).
7. Firica, O.: Blockchain Technology: Promises and Realities of the Year 2017. Calitatea. 18, S3, 51–58 (2017).
8. Gaetani, E. et al.: Blockchain-based database to ensure data integrity in cloud computing environments. CEUR Workshop Proc. 1816, 146–155 (2017).
9. Geranio, M.: Fintech in the exchange industry: Potential for disruption? Masaryk Univ. J. Law Technol. 11, 2, 245–266 (2017).
10. Giancaspro, M.: Is a "smart contract" really a smart idea? Insights from a legal perspective. Comput. Law Secur. Rev. 33, 6, 825–835 (2017).
11. Herbaut, N., Negru, N.: A Model for Collaborative Blockchain-Based Video Delivery Relying on Advanced Network Services Chains. IEEE Commun. Mag. 55, 9, 70–76 (2017).
12. Khaqqi, K.N. et al.: Incorporating seller/buyer reputation-based system in blockchain-enabled emission trading application. Appl. Energy. 209, October 2017, 8–19 (2018).
13. Lee, B., Lee, J.H.: Blockchain-based secure firmware update for embedded devices in an Internet of Things environment. J. Supercomput. 73, 3, 1152–1167 (2017).
14. Li, Z. et al.: On the Integration of Event-Based and Transaction-Based Architectures for Supply Chains. Proc. - IEEE 37th Int. Conf. Distrib. Comput. Syst. Work. ICDCSW

2017. 376–382 (2017).

15. Lycklama à Nijeholt, H. et al.: DecReg: A Framework for Preventing Double-Financing using Blockchain Technology. Proc. ACM Work. Blockchain, Cryptocurrencies Contract. - BCC '17. 29–34 (2017).
16. Mengelkamp, E. et al.: Designing microgrid energy markets. A case study: The Brooklyn Microgrid. Appl. Energy. (2017).
17. Mettler, M.: Blockchain technology in healthcare: The revolution starts here. 2016 IEEE 18th Int. Conf. e-Health Networking, Appl. Serv. Heal. 2016. 16–18 (2016).
18. Norta, A.: Designing a smart-contract application layer for transacting decentralized autonomous organizations. Commun. Comput. Inf. Sci. 721, November, 595–604 (2017).
19. Norta, A.: Service-Oriented Computing – ICSOC 2015 Workshops. 9586, July, (2016).
20. Norta, A., Norta, A.: Perspectives in Business Informatics Research. 229, August, (2015).
21. Shae, Z., Tsai, J.J.P.: On the Design of a Blockchain Platform for Clinical Trial and Precision Medicine. Proc. - Int. Conf. Distrib. Comput. Syst. 1972–1980 (2017).
22. Sidhu, J.: Syscoin: A peer-to-peer electronic cash system with blockchain-based services for E-business. 2017 26th Int. Conf. Comput. Commun. Networks, ICCCN 2017. (2017).
23. Wander, M.: Bitcoin Block Data, https://commons.wikimedia.org/wiki/File:Bitcoin_Block_Data.png.
24. Weber, I., Governatori, G.: Untrusted Business Process Monitoring and Execution. Int. Conf. Bus. Process Manag. 329--347 (2016).
25. White, G.R.T.: Future applications of blockchain in business and management: A Delphi study. Strateg. Chang. 26, 5, 439–451 (2017).
26. Wu, L. et al.: Democratic Centralism: A Hybrid Blockchain Architecture and Its Applications in Energy Internet. Proc. - 1st IEEE Int. Conf. Energy Internet, ICEI 2017. 176–181 (2017).
27. Xu, X. et al.: The blockchain as a software connector. Proc. - 2016 13th Work. IEEE/IFIP Conf. Softw. Archit. WICSA 2016. 182–191 (2016).
28. Yuan, C. et al.: SEC_Research on a New Signature Scheme on Blockchain. Secur. Commun. Networks. 2017, 10 (2017).
29. Yue, X. et al.: Healthcare Data Gateways: Found Healthcare Intelligence on Blockchain with Novel Privacy Risk Control. J. Med. Syst. 40, 10, (2016).
30. Zhang, Y., Wen, J.: An IoT electric business model based on the protocol of bitcoin. 2015 18th Int. Conf. Intell. Next Gener. Networks, ICIN 2015. 184–191 (2015).
31. Zhang, Y., Wen, J.: The IoT electric business model: Using blockchain technology for the internet of things. Peer-to-Peer Netw. Appl. 10, 4, 983–994 (2017).
32. Zou, J. et al.: A dispute arbitration protocol based on a peer-to-peer service contract management scheme. Proc. - 2016 IEEE Int. Conf. Web Serv. ICWS 2016. 41–48 (2016).

Towards an Access Control System for IoT on Blockchain

João Pedro Dias[1] and Hugo Sereno Ferreira[1,2]

[1] INESC TEC, FEUP Campus, Porto, Portugal
[2] Department of Informatics Engineering,
Faculty of Engineering, University of Porto, Portugal
{jpmdias,hugosf}@fe.up.pt

Abstract. Access control is a crucial part of the security of a system, restricting what actions users can perform on resources. Therefore, access control is a core component when dealing with data access policies and resources, discriminating which is available for a certain party. We consider that current systems that attempt to assure the share of policies between facilities are prone to system's and network's faults and don't assure the integrity of policies lifecycle. Such scenario is even more complex when dealing with the continuously growing Internet-of-Things. Distributed ledger, namely a fully-private blockchain, where the operations are stored as transactions, we can ensure that the different facilities have knowledge about all the parties that can act on the available resources. A developed proof-of-concept shows that this approach allows us to manage access control to systems and resources while maintaining integrity, auditability, and authenticity.

Keywords: Access Control, Blockchain, Internet-of-Things

1 Introduction

Lots of new smart objects are empowering the creation of cyber-physical smart pervasive systems, with application in a variety of domains[13]. These smart objects are under the umbrella of the Internet of Things paradigm, that foresees the advance towards new smart and inter-connected systems by the means of ubiquitous computing [14].

The arrive of IoT lead to an explosion of data being collected and, *a posteriori*, analyzed by different entities lead to the debut of data security and privacy issues. Especially when we take into consideration that smart devices may be connected to the Internet at some point for accessing its collected data anytime and anywhere [14].

As such, there is the need to control the accesses to these resources. Access control is concerned about determining the allowed activities of certain users, mediating every attempt by a user to access a resource in the system [7].

Blockchain was conceptualized by Satoshi Nakamoto and is used as a core component of the digital currency Bitcoin [11]. Data in a blockchain should

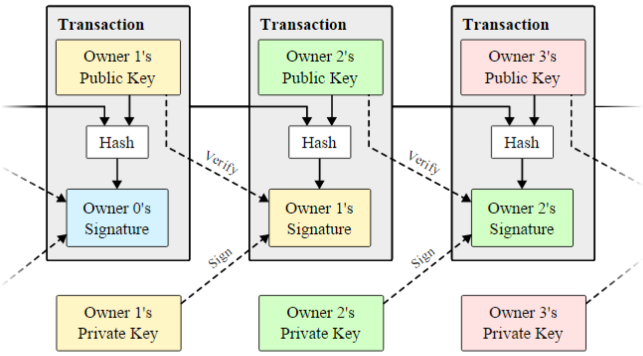


Fig. 1. Blockchain mechanism as proposed by Satoshi Nakamoto [11].

be tamper-proof, specifically accomplished by the use of cryptography, by the means of digital signatures and digital fingerprints (hashing), as shown in Figure 1 [11]. Also, consensus must be assured among peers considering scenarios where some of the peers are providing erroneous data, by partially or completed computer/network failures or, even, by malicious intent when some party tries to subvert the ledger [11].

A blockchain consists of a chain of blocks that contains information about transactions. Each of these transactions is digitally signed by the entity emitting them. Transactions are combined into a *block*, this *block* is committed to the chain, establishing the *blockchain*. Each block contains the hash of the previous block, being this propagated along the chain until the first block created when the blockchain was firstly created, designed *genesis* block [11].

We can then consider, from a technical viewpoint, that a blockchain works as a state transaction system, where there is a state that corresponds to the snapshot of the chain (the result of all transaction until now) and, after adding a new block of transactions to the chain, we got a new snapshot that corresponds to a new state of the system, as the result of the new transactions [5].

In order to a validate a block there is the need of a *proof-of-work*. This mechanism is used in order to get a consensus in the peer-to-peer network [11]. In Bitcoin, this process is called mining and consists of finding a *nonce* (by the means of *brute-force*) that satisfies the condition of generating a digest with the required number of leading zeroes. This *proof-of-work* guarantees consensus in a network following the principle that the nodes will always accept the longest available chain [11]. This also implies that older blocks - those further back in the blockchain - are more secure than newer ones.

There are alternatives to *proof-of-work*. In the *proof-of-stake*, as it is being considered to be used in Ethereum [5], the creator of the next block to be pushed in the chain is chosen in a deterministic way based on the wealth of the node. Another one, as used in the Sawtooth Lake [3], uses a *Proof of Elapsed Time*

(PoET), which is a lottery-based consensus protocol that takes advantage of the trusted execution environments provided by Intel's Software Guard Extensions.

Although the most common use of blockchain is for trade currencies, like Bitcoin, there exists an array of other applications for the technology. This is possible because, as blockchain is used to store *coin* transactions, it can be used to store any other domain transactions. Furthermore, it can be used as a general-purpose database distributed system, therefore making it useful in a large variety of situations [15].

Blockchains can be considered of two types, namely: public and unpermissioned, or private and permissioned. In the first type, anybody is allowed to use them (e.g. Bitcoin). In the other, there exists a closed group of known participants (e.g. a supply chain) [15].

In this paper, we suggest an approach to the problem of access control in large-scale and distributed systems, as it is observed in today IoT scenarios (e.g. wearables), where different entities and users should be able to access data with different permission levels and granularities. The owners should be able to manage the accesses to their data, by the means of adding, changing or revoking permissions. This system should be also capable of defining fine-grained permissions both at the user level and at the resource level. Notwithstanding, these users and resources are external to the system being only referenced by it.

Additionally, the system must be fault tolerant, which means that it must not be dependable on a centralized authority. Upon these considerations, such system should also assure consistency and integrity among nodes and operations along with the authenticity of the operations.

In our approach, we take into consideration the paradigm that is a Distributed Ledger Technology (DLT), more properly a specific type that is known as blockchain. A distributed ledger consists of a consensus of replicated, shared and synchronized digital data distributed along a set of nodes, working as a distributed database, generally geographically dispersed. It's important to note that, despite all blockchains being distributed ledgers, not all distributed ledgers are blockchains [4].

This paper is structured as follows: firstly is given an overview of related work in the scope of permission management in IoT systems. Afterwards, it is given a description of the purposed solution architecture. Then we address some core details of the *proof-of-concept* implementation. Finally, some final remarks are presented, summing up the contributions as well as pointing out further developments.

2 Related Work

The problem of access control has already been covered in the literature. We can observe different ways of controlling and managing accesses in different situations in our everyday technological systems. There exists a problem of defining permission rules, typically known as *policies*, alongside with the problems related with inconsistency between rules [9].

One of the more common and widespread approaches is the Access Control Lists (ACL), commonly used in modern operating systems. ACL consists of a list associated with an object that specifies all the subjects that can access the object, along with the access level (or rights) to the object [7].

Other systems use Access Control Matrix, in which each row represents a subject, each column an object and each entry is the set of access rights for that subject to that object [7].

There are also more configurable and extensive solutions like Role-Based Access Control (RBAC) and Attribute-Based Access Control (ABAC) [10]. RBAC define the user's access right basing itself on his/her roles and the role-specific privileges associated with them. The ABAC system extends the RBAC role-based features to attributes, such as properties of the resource, entities, and the environment [8].

Yet another approach is the Entity-Based Access Control (EBAC) system [1], which allows the definition of more expressive access control policies. This is accomplished by supporting both the comparison of attribute values, as well as traversing relationships along arbitrary entities. Moreover, Bogaerts et al. presents *Auctoritas* as an authorization system that specifies a practical policy language and evaluation engine for the EBAC system [1].

Some research has been done towards the use of blockchain in access control systems and IoT. Ouaddah et al. propose a new Access Control Model for IoT systems, called FairAccess, as a decentralized pseudonymous and privacy-preserving authorization management framework. Although the novelty of the solution, their solution introduces a new access model which implies the disregard of already widespread access control models [12].

3 Approach Overview

Our approach makes use of an Access Control Matrix, alongside with Blockchain technology, for access control management IoT-based environments. Such approach allows us to define fine-grained access control while maintaining highly-scalability, distributed with integrity, authenticity, auditability, and immutability.

A *proof-of-concept* of the approach hereby described and detailed was implemented in order to enable the concept to be tested and validated.

3.1 Access Control Model

In our solution, we use an approach similar to the Access Control Matrix, which allows the establishment of a correspondence between a subject, an object and a set of rights. In our approach we consider that each transaction of the blockchain corresponds to one entry of the type `<subject, permission, object>`, similar to Access Control Matrix, where:

- `subject`: entity or individual that have certain level of permission over an *object.*

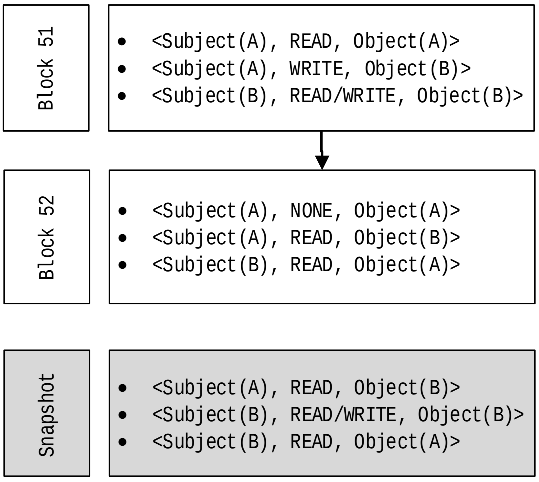

Block 51
• <Subject(A), READ, Object(A)>
• <Subject(A), WRITE, Object(B)>
• <Subject(B), READ/WRITE, Object(B)>
Block 52
• <Subject(A), NONE, Object(A)>
• <Subject(A), READ, Object(B)>
• <Subject(B), READ, Object(A)>
Snapshot
• <Subject(A), READ, Object(B)>
• <Subject(B), READ/WRITE, Object(B)>
• <Subject(B), READ, Object(A)>

Fig. 2. Example of two system blocks, transactions data and resulting snapshot.

- `object`: referenced object over which the *subject* have some access level permission. This object is stored outside the blockchain being only referenced by it.
- `permission`: permission level that a given *subject* has over some *object*. The value of this field can be one of four, namely:
  - `NONE`: Subject has no access to the object. Is used as a way to revoke previously given permissions.
  - `READ`: Subject has read access to the object.
  - `WRITE`: Subject has to write access to the object.
  - `READ/WRITE`: Subject has all the rights over the object.

There are a few core concepts in the definition of the logic of this access control model, as follows:

- A snapshot corresponds to the state of the blockchain at a given moment, reflecting the result of processing all the transactions in the timestamped and indexed order they appear in the chain, from the *genesis* block until the last block accepted in it.
- A new transaction establishes a relationship between a `subject` and a `object`, with a certain permission level.
- If there is a previous relationship between the same `subject` and `object` in the system it is overridden by the most recent one.
- From the moment that a permission is revoked (permission level `NONE`), the relationship `subject` and `object` does not appear in the blockchain snapshot.

These core concepts are demonstrated in Figure 2, where are shown the data of two blocks in the blockchain as well as a resulting snapshot of those two blocks of transactions in the order they appear.

3.2 Block Model

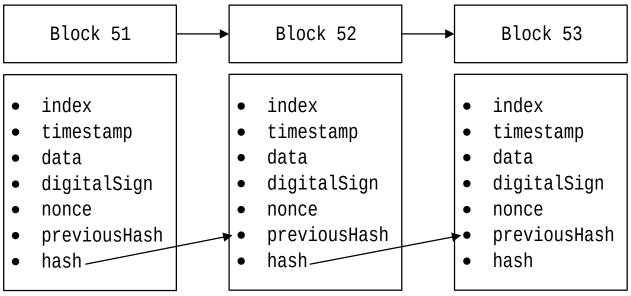


Fig. 3. Example of a excerpt of the blockchain with details of each block content.

We make use of a simple, implemented from scratch, blockchain. The model of each block is similar to the ones used in the Bitcoin system but with some particularities, one of which is the abstraction of the data present in the blockchain. The basic structure of a block in our chain is given in Figure 3, being each field detailed as follows:

- `index`: Corresponds to the index of the current block in the blockchain.
- `timestamp`: Timestamp corresponding to when the block was generated.
- `previousHash`: Hash of the previous block in the chain.
- `digitalSign`: Digital signature of the current block data.
- `data`: Content of the block. Corresponds to a set of transactions describing the access control policies. Further details on this are described in subsection 3.1.
- `nonce`: Value that is set so that the hash of the block will contain a run of leading zeros. This value is calculated iteratively until the required number of zero bits in the `hash` is found. This requires time and resources, making it so that the correct *nonce* value constitutes *proof-of-work*.
- `hash`: A SHA256 hash corresponding to the block data. This hash must have a leading *a priori* defined sequence being the size of the leading sequence what defines the effort of the *proof-of-work*.

The use of hashes allows us to maintain integrity along the immutable chain without a central authority, since any change in the data would result in a different hash, invalidating the blockchain. Authenticity is assured by the assign of a key-pair to each entity with access to the blockchain, identifying who write each block in the chain.

3.3 System Architecture

The blockchain is a technology distributed by default, working as a peer-to-peer network connecting the different nodes through a WebSockets interface. Nodes in the network synchronize between each other by following a set of rules:

- When a new block is generated by a node, this block is broadcast to the network;
- When a node connects to a new peer in the network it queries for the latest block;
- If a node finds a block that has a higher index than the last known block, it either adds the block to its current blockchain (in the case of the difference is equal to one node) or queries another node to for the full blockchain.

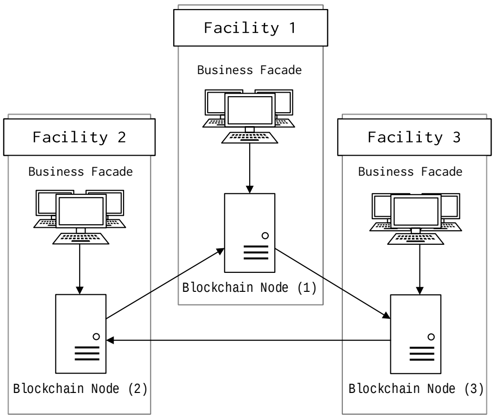


Fig. 4. Example of a distributed access control network integrating three distinct facilities, each one running one or more `Business Facade` over each `Blockchain Node`.

With the objective of building an architecture based on micro-services where we can easily adapt and rearrange logic components with minimal downtime and easily scale, we split the data model logic (access control model) from the blockchain logic, making two separated but interconnected micro-services. The network logic of this system can be observed in Figure 4, where the `Business Facade` have knowledge about the access control model. This `Business Facade` communicates with the corresponding node, `Blockchain Node`, in the blockchain network through an HTTP interface exposed by the node.

A diagram on how the different modules and entities are related is shown in Fig. 5. Any third-party with interest in the IoT data, including service or applications that relies on such, should first communicate with one of the `Business Facade` to certify that the entity trying to access the data, or using that service/application, is rightful to access the data exposed by it. Furthermore, if this is the first data access request by some `3`$^{\texttt{rd}}$ `party`, the request must be, in the first place, be approved by some administrative mechanism or entity on some `Edge Device/Application`. A sequence diagram exposing a more detailed representation of this interaction is given in Fig. 6.

It is relevant to point that the blockchain implemented is fully private. This means that the write permissions to the blockchain are kept centralized to one

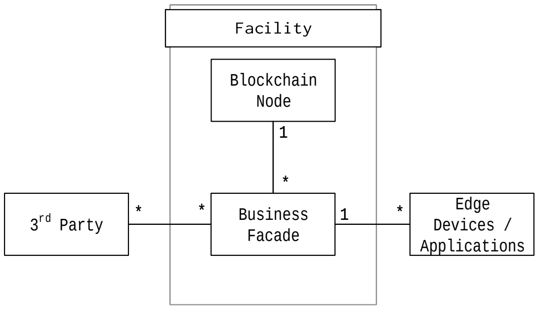


Fig. 5. Description of the different relationships on the architectural approach developed, focusing on how the different modules, namely, the `Business Facade` and `Blockchain Node`, interact with the external entities, namely, the `3rd Parties` and administration entities by the means of `Edge Devices / Applications`.

organization, generally, as pointed out before, the government, but read permissions to the chain can be public or restricted to an arbitrary extent.

Fully-private blockchain [2] are advantageous in the context of our approach, up front because they are cheaper. This is verified because transactions only need to be verified by a few nodes that can be trusted to have very high processing power (e.g. one node per facility), is this the blockchain miner nodes. Also, in this model, there is no need of incentives (e.g. in Bitcoin, miners are motivated by Bitcoin rewards) because it is of interest to the facilities to keep the blockchain running. Furthermore, if desired, the rules of the blockchain can be altered at some point by the owner institution and faults can be fixed by manual intervention if needed.

At the `Business Facade` level, a non-distributed basic blockchain was implemented as a way to store the different blockchain snapshots. This approach is used to reduce the need to execute all the transactions from the *genesis block* each time that we require a new snapshot in order to get the active policies. Yet, this concept is just a preliminary approach on how to store different snapshots along the lifecycle of a client, needing further developments.

4 Implementation Details

During the implementation of the *proof-of-concept*, some decisions were made; we describe these details here with the intent of helping the reader to re-implement a similar system.

The `Blockchain Node` and `Business Facade` were both implemented using `Node.js`, for no particular reason besides the simplicity of the language and availability of libraries. The built-in `crypto` module [6], for example, immediately provides us with a mechanism to digitally sign the blocks payload using public-key cryptography (using `RSA-SHA256`) and to calculate the hashes of each block using `SHA256` algorithm.

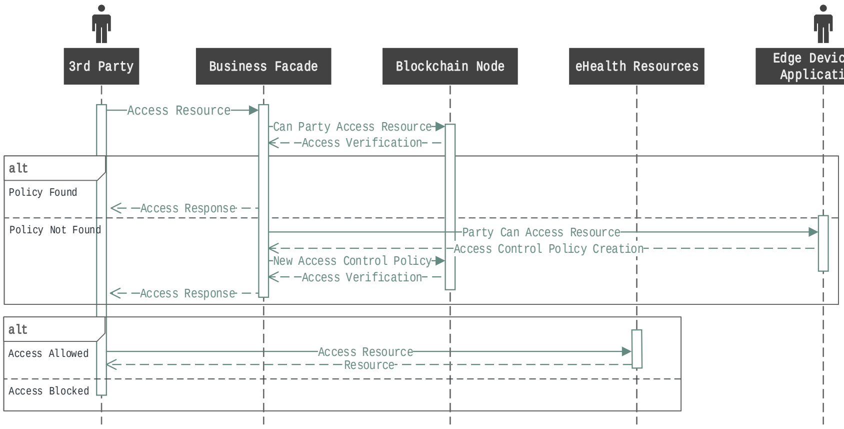


Fig. 6. Sequential view on how some 3rd `Party` can access or request access to an IoT resource or device data, detailing the communication between the inner modules of the architecture. Is it also visible the process of creation and/or renewal of access control policies.

This *proof-of-work* works on a *brute-force* basis. The *nonce* is iteratively incremented until the resulting SHA256 hash matches the *a priori* defined number of leading zeroes — this is similar to the Bitcoin system and establishes the "effort". We can easily tweak the "effort" to better suit our use-case.

5 Final Remarks

In this paper, we present an approach to solving the problem of managing access control in the IoT ecosystem. Access Control is a special complex task in IoT systems since resources and data are distributed among different entities. This approach consists of using blockchain alongside with the use of access control policies, stored as transactions.

For purposes of supporting the plausibility of our solution, a *proof-of-concept* was designed and implemented. This *proof-of-concept* allowed us to make some, even if preliminary, tests and validations over our approach, namely it's scalability, fault-tolerance, and correctness.

Overall, we determine that our approach is viable, giving diverse advantages when comparing to the in-place systems. This advantages, although not limited to, includes the integrity, auditability, and authenticity of the access control policies in the system, since this proprieties are inherent to a blockchain system.

Further research needs to be pursued in order to make such approach production-ready. In this context, further testing and validation are needed to assess the scalability proprieties of such system. This includes testing large-scale scenarios with different node dynamics, i.e., adding, removing ad invalidating nodes on-the-fly. Also, tests dealing with malicious attacks by third-parties or cases when one or more nodes of the blockchain are compromised should be pursued.

References

1. Bogaerts, J., Decat, M., Lagaisse, B., Joosen, W.: Entity-based access control: Supporting more expressive access control policies. In: Proceedings of the 31st Annual Computer Security Applications Conference. pp. 291–300. ACSAC 2015, ACM, New York, NY, USA (2015), http://doi.acm.org/10.1145/2818000.2818009
2. Buterin, V.: On public and private blockchains. https://blog.ethereum.org/2015/08/07/on-public-and-private-blockchains/ (August)
3. Corporation, I.: Sawtooth lake latest documentation. https://intelledger.github.io/ (2015), (Accessed on 06/02/2017)
4. Davidson, S., De Filippi, P., Potts, J.: Disrupting governance: The new institutional economics of distributed ledger technology (2016)
5. Ethereum), E.F.S.: Ethereum: A next-generation smart contract and decentralized application platform (2014), https://github.com/ethereum/wiki/wiki/White-Paper
6. Foundation, N.: Crypto module documentation. https://nodejs.org/api/crypto.html (2017), (Accessed on 06/07/2017)
7. Hu, V.C., Ferraiolo, D., Kuhn, D.R.: Assessment of access control systems. US Department of Commerce, National Institute of Standards and Technology (2006)
8. Hu, V.C., Ferraiolo, D., Kuhn, R., Friedman, A.R., Lang, A.J., Cogdell, M.M., Schnitzer, A., Sandlin, K., Miller, R., Scarfone, K., et al.: Guide to attribute based access control (abac) definition and considerations (draft). NIST special publication 800(162) (2013)
9. Krawiec, R., White, M.: Blockchain: Opportunities for health care (August 2016), (Accessed on 06/05/2017)
10. Li, M., Yu, S., Ren, K., Lou, W.: Securing personal health records in cloud computing: Patient-centric and fine-grained data access control in multi-owner settings. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering 50 LNICST, 89–106 (2010)
11. Nakamoto, S.: Bitcoin: A Peer-to-Peer Electronic Cash System p. 9 (2008), https://bitcoin.org/bitcoin.pdf
12. Ouaddah, A., Elkalam, A.A., Ouahman, A.A.: Towards a novel privacy-preserving access control model based on blockchain technology in iot pp. 523–533 (2017), https://doi.org/10.1007/978-3-319-46568-5_53
13. Rajkumar, R.R., Lee, I., Sha, L., Stankovic, J.: Cyber-physical systems: The next computing revolution. In: Proceedings of the 47th Design Automation Conference. pp. 731–736. DAC '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1837274.1837461
14. Tan, L., Wang, N.: Future internet: The internet of things 5, V5–376–V5–380 (Aug 2010)
15. Underwood, S.: Blockchain beyond bitcoin. Commun. ACM 59(11), 15–17 (Oct 2016)

SESSION 7

System Security and Networks

Heartbeat Biometrics: A Survey
Marcelo Santos and Luís Antunes

A framework to automate network packets construction
Cristiano Alves, Nuno Anacleto and Mário Antunes

A network packets sniffer with real-time graphical analytics
Daniel Pinto and Mário Antunes

Heartbeat Biometrics: A Survey

Marcelo Santos[1][0000-0001-8158-4682] and Luís Antunes[2][0000-0002-9988-594X]

[1] Center of Competence in Cyber Security and Privacy. Faculty of Science.
University of Porto
R. Campo Alegre, 1021, 4169-007 Porto, Portugal marcelo.santos@fc.up.pt
[2] Faculty of Science, University of Porto & CRACS/INESC-TEC
R. Campo Alegre, 1021, 4169-007 Porto, Portugal

Abstract. Security systems are a crucial part of our society in order to guarantee privacy and content protection. Despite their constant evolution they keep being exposed to new threats and consequently compromised. Biometrics were inserted in this context to try to provide additional security. In this paper we survey the current methodologies regarding heart-based biometrics, categorized in fiducial and non-fiducial based approaches, and exposed their achievements when considering Electrocardiogram (ECG) and Photoplethysmogram (PPG) signals. Moreover, several of the presented studies achieve 100% accuracy for individuals identification and negligible Equal Error Rate (EER) in authentication, lower than 5%.

Keywords: Security Systems, Biometrics, Electrocardiogram (ECG), Photoplethysmogram (PPG)

1 Introduction

Technology has become a significant part of our lives. Whether to be used for professional or recreational purposes, it is undeniable the presence of technological devices such as smartphones, personal computers or smart *wearables*. They provide fast and convenient access to several tools and services, such as online banking or fast money transactions, relying on the devices' security. However, when evaluating the current authentication systems it is evident the need for improvement due to security hazards.

Credential-based systems are yet the standards when protecting sensitive data in spite of being easily comprised, whether using social engineering, *phishing*, password guessing or even due to password hash databases leakage. Furthermore, passwords defined by people, tend to be easily cracked due to its simplicity. Nevertheless, when considering measures to enforce security or access control, different methods are used, such as *tokens*. Although, being a physical object, tokens are also exposed to vulnerabilities, from the most basic - misplacement or robbery - to more complex methods, such as cloning.
As a way to circumvent these vulnerabilities, a different concept emerged that resourced to physiological features to distinguish people - *Biometrics*. Through

out the years several modalities were presented, such as the usage of human fingerprint, face recognition, handwriting, the human eyes, specifically the iris and more recently the heartbeat [21,6,9,19]. However, and despite adding another layer of security to several systems, most of them were proved to be compromised, whether using high resolution photos to delude facial recognition, creating contact lets to replicate the iris or even by building physical models on latex to mimic fingerprints [24,18,20,12].
In this regard, systems that rely on the heartbeat are emerging due to its intrinsic properties desirable for the purpose of authentication and also identification, namely *universality*: since every person possesses this characteristic; *liveness detection*: as the signal derives from cardiac cycle, liveness analysis is trivial; *uniqueness*: meaning that, due to its physiological origin, and despite the general pattern being similar among individuals, there is a wide variability among them; *robustness*: since it relies on physiological events it naturally hamper its replication [1,26].

For the aforementioned reasons, this work will focus on evaluating this promising modality and its feasibility in the context of current authentication systems. We will start by exposing some background knowledge in Section 2, namely all the constituting stages of an authentication / identification system, specifically pre-processing, feature extraction and classification, and the waveform signals evaluated. Then, we will expose the two main approaches described in the literature, namely fiducial dependent and fiducial independent in Sections 3 and 4, respectively. Finally we will highlight the differences between methodologies and discuss their application and viability in Section 5.

2 Background Knowledge

Heart-based biometric systems can resource to several signals derived from the heartbeat, being the most common, Electrocardiogram (ECG), Photoplethysmogram (PPG), Phonocardiogram (PCG) and Electroencephalogram (EEG). However, for the purpose of this work we will focus of the first two, which will be subsequently described.

2.1 Waveform Signals

Over the past centuries, physiological signals have been used strictly for medical purposes. However, biometrics systems introduced a new paradigm in security systems. These waveform signals can contain a great amount of information while varying on the form of extraction.

Electrocardiographic (ECG) Wave

Caused by periodic depolarization and repolarization of atria and ventricle of the heart, ECG consists of a waveform signals, represented by several heartbeats, that allows measuring the variation of the electrical activity of the heart over

time.
Furthermore, a heartbeat consists of three major segments: P wave, QRS complex and T wave, illustrated in Figure 1, which corresponds to different stages of the cardiac cycle, *i.e.* atrial depolarization, ventricular depolarization and ventricular repolarization, respectively.

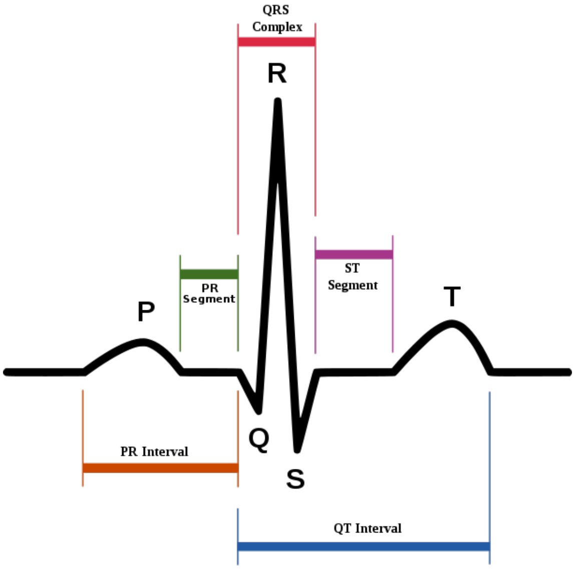


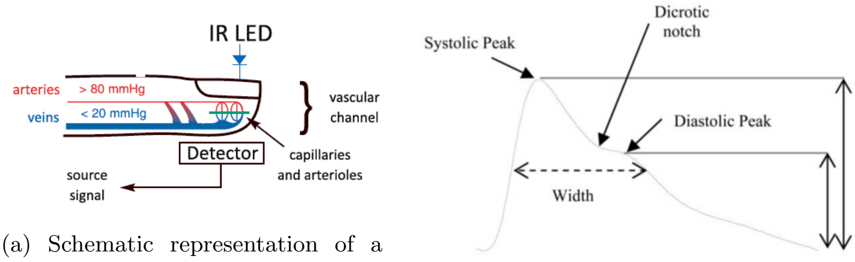
Fig. 1: Illustration of the ECG waveform traits.[29]

The PPG signal is obtained by capturing the variations of blood density in the capillaries. This variation is measured using an optical sensor composed by a light transmitter source (Light Emitting Diode (LED)) and a photo resistor (Light Dependent Resistor (LDR)). The sensor is placed directly over the skin to emit light that will penetrate into the blood vessels. The reflected light is then captured by the LDR creating a waveform signal as illustrated in Figure 2a.

The PPG waveform comprises three components, namely the *systolic peak*, the *dicrotic notch* and the *diastolic peak* (Figure 2b), related with the two specific occurrences. The systole, when the ventricles contract, ejecting the blood from the lower chambers to the organs, and the diastole, when the ventricles are relaxing and filling with blood.

2.2 Pre-processing

Data pre-processing in an essential step in almost every field that relies on extracting features from data. In this specific case, this stage consists of the disposal of unnecessary data that could interfere with the correct recognition of



(a) Schematic representation of a PPG sensor. [3]

(b) PPG waveform major traits. [10]

Fig. 2: Illustration of the PPG sensor and waveform signal.

individuals. Considering both ECG and PPG, this stage consists of removing noise artifacts from the waveforms. Generically, these noise artifacts are concentrated in two frequency ranges: *high-frequency*, cause by powerline interference and *low-frequency* caused by deep breathing or muscular movement [1].

2.3 Feature Extraction

After the pre-processing stage, features are extracted from the processed signal. It's in this stage that lies the main differences, consisting essentially on the usage, or not, of fiducials. Fiducial points consists of elements marked in the signal used for measurements. In the case of waveform signals, they can depend on time or amplitude.

2.4 Classification

Analogously to every security system, there is a stage at which the system decides based on the input. However, in biometric systems, the classification process does not depend exclusively on a direct match. Instead, the system will have to classify the sample considering that biologic data is changing over time. For this reason, these systems typically use machine learning algorithms, so it can learn based on the data acquired in the *enrollment* - where the user feeds his physiological data to the system, so it knows to whom the data belongs to.

3 Fiducial Based Approaches

Fiducial based, or fiducial dependent approaches, resources to the fiducial marks of the waveform signals. Regarding ECG, the data was obtained with different combinations of *leads* - the electrodes placed over the body to detect electrical activity. As for the PPG, signals were obtained using a one-piece device, thus we can consider it as one lead.

Biel *et al.* [4] evaluated the possibility of human identification using a 12-lead ECG analysis. They collected 50 samples from 20 individuals and extracted 360 features per subject. By evaluating feature's correlation between leads, they reduced the number of leads considered. by repeating the process they were able to narrow the number of features to 12. Using Soft Independent Modeling of Class Analogy (SIMCA) with a wide range of test sets, varying on the number of features, they were able to correctly classify 49 of the 50 samples in several tests, and obtained 100% accuracy in one of them.

Using an ECG one-lead extraction, Shen *et al.* attempted in [23] individuals identification by focusing on feature extraction from the QRS complex and the normalized QT interval. They achieved 95% and 80% accuracy by classifying samples using template matching, and Decision Based Neural-Network (DBNN), respectively. Later, in [22], and using a single ECG lead they added an extra step after the pre-processing, which they called *pre-screening*, consisting of filtering the samples to classify using template matching. By extracting 17 features, mostly from the QRS complex, they achieved 100% accuracy in small pre-determined groups, of 10, 20 and 50 people, and, 96% and 95.3% for groups of 100 and 168 subjects, respectively.

In [13], Gu *et al.* used PPG signals for human identification. They collected a 1 minute sample per subject with a sampling rate of 1KHz. The pre-processing consisted of a smoothing technique and, as features, they considered the number of peaks in each pulse, upward slope from the bottom to the top of the first peak, downward slope from the last peak to the bottom, and the time interval between the bottom point and the first peak point. For the classification they used Euclidean distance between the sample and the template obtaining 94% of accuracy.

Israel *et al.* evaluated in [15] the influence of the state of anxiety when using the heartbeat for identification. They started by extracting 15 features from the local maxima P, T and R, and from P', T' and L', being the prior and the second the endpoints of the P and T waves, respectively, and the latter the beginning of the P wave. Using stepwise correlation analysis they reduced the feature set to 12. With Linear Discriminant Analysis (LDA) and standard majority voting, they correctly classified 82% of the heartbeats and 100% of the individuals, for the data obtained from the neck, and 72% of the heartbeats and 100% of the individuals when obtaining the data from the chest.

Wübbeler *et al.* used *Einthoven triangle scheme* (three lead extraction) for human verification and identification with ECG, in [30], using Physikalisch-Technische Bundesanstalt (PTB) public dataset. They defined a bi-dimensional feature vector (H_x, H_y), where the prior consisted of the main lead data, and the latter the combination of the remaining leads. Features were extracted with a window interval of 100ms centered in the QRS complex. Using Nearest Neighbors (NN) and several thresholds, they correctly identified $98,1\%$ to 99% depending on the threshold, and an Equal Error Rate (EER) between 0.2% and 2.5%.

Yao *et al.* performed a pilot study in [31] resourcing to PPG derivatives to evaluate both, subject identification and discrimination. With a pulse oximeter, they collected 70 seconds samples from three subjects at a 300Hz rate and generated 9 datasets using a 9 step process. First they applied a Chebyshev low-pass filter, then randomly selected a pulse, which would be then fitted with a 10th order polynomial to extract the first 1st and 2nd derivatives. From the derivatives, they obtained respectively, the number of maximum and minimum points, and inflection points. The authors stated that high-order derivatives enable more discriminative attributes, but become more sensitive to noise. However, when considering low order derivatives, features become more robust and less sensitive, requiring weights assignment.

Using PPG signals, Wei *et al.* proposed a novel algorithm in [28] to eliminate baseline drift. The pre-processing stage consisted of outlier removal, Finite Impulse Response (FIR) low-pass filtering and baseline drift removal using the proposed algorithm. Their algorithm was based on the application of a series of low-pass and high-pass filters and recomposing the wave with wavelet reconstruction. They extracted the features using a cubic spline interpolation and enhanced differentiation. Using traditional differentiation they obtained an EER of 4.1% reduced to 0 when using their enhanced approach.

In [25], Spachos *et al.* explored the possibility of using PPG for identification testing their methodology with two PPG datasets, *OpenSignal* and *BioSec*, containing data from 14 and 15 healthy subjects, respectively. For the pre-processing stage, they started with peak detection, then segmented the signals and finally they normalized and performed time domain scaling. For the feature extraction, they used LDA and generated two datasets, for training and testing. For the classification they used NN and majority vote while using a threshold approach to accept or reject the input data. They achieved False Acceptance Rate (FAR) and False Rejection Rate (FRR) of 0.5% on *Opensignal* and and EER of 25% on *BioSec*.

Bonissi *et al.* presented in [5] a method based on correlation and template matching in a scenario of continuous authentication using PPG. Starting with the signal pre-processing, they used a third order high-pass *Butterworth* filter with a cut-off frequency of 0.1Hz. To extract the features, they generated several **T** templates containing a different amount of heartbeats each. Each template was obtained by performing the following iterations: 1) Segment the signal into a matrix **H**; 2) Apply Pan-Tompkins algorithm to detect peaks; 3) Compute the average number of hearbeats (**S'**) per entry in **H**; 4) Compute the maximum cross-correlation **C** between heartbeats from **H** and the average heartbeat **S'**; 5) Store **N** hearbeats corresponding to maximum number of **C** in **T**; 6) Discard the signal if the correlation was lower than the threshold. To classify the samples, they computed the similarity score of two templates by generating a maximum cross-correlation matrix (**M**) between every signal belonging to both templates. The matching score was obtained with the $mean(M)$, $median(M)$, $90\%(M)$, $95\%(M)$ and $max(M)$. From the signals, they generated three datasets, **DB20**, **DB30** and **DB40**, where each sample had the duration of 20, 30 and 40 seconds, respectively. For regular authentication, they achieved 5.29% EER with **DB40** while in a continuous authentication scenario the EER was 13.55% for an interval of 400 seconds.

Carreiras *et al.* exposed in [7] a biometric system model using ECG. The data was obtained from 12 leads, however the authentication was performed with a single lead, placed in the finger. They started by extracting the QRS complex, removed the outliers using *DMEAN* [17] and finally storing the generated templates. With an adaptive threshold approach, and using k-Nearest Neighbors (KNN) with $k = 3$ they classified the samples and discarded the ones exceeding the threshold. They tested their model with two datasets, **P618** and **Baseline**, obtaining an EER of 9.01% and Error of Identification (EID) of 15.64% for the prior, and 13.26% EER and 36.40% EID for the latter.

Akhter *et al.* evaluated in [2] the usage of Heart Rate Variability (HRV) for authentication with PPG signals. The dataset consisted of 2430 R-R Interval (RRI) sequences of 81 subjects, with 10 samples each. As features, they extracted: Root Mean Square of Successive Differences (RMSSD), Mean Heart Rate (MeanHR), Mean, Median, Standard Deviation of Heart Rate (SDHR) and the Maximum and Minimum interval duration in a particular RRI. Using Euclidean distance, with a threshold of 0.07, they achieved a recognition rate of 86.7% and an EER of 17%.

4 Fiducial Independent Approaches

Opposing to fiducial based approaches, fiducial independent approaches rather than relying on features obtained from time and frequency domains, are extracted from the morphology of the signal. Thus, enabling to bypass some constraints of the fiducial approach, such as signal synchronization.

Plataniotis *et al.* presented a novel approach in [14] for individuals identification using fiducial independent approach. Starting with an hybrid approach, they extracted appearance-based and analytic features. By presenting a novel method: *AC/DCT* which consists of computing the Auto Correlation (AC) coefficients and subsequently apply Discrete Cosine Transformation (DCT) for dimensionality reduction, they were able to discard the appearance-based features. Among several tests, they achieved 100% of individuals recognition on 14 subjects.

In [27] Wang *et al.* compared both approaches by extracting fiducial features and applying the AC/DCT method proposed in [14] on ECG signals. For the fiducial-based, they used P, Q, R, S, T and, P' and T', where the last two were the endpoints of the P and T waves, respectively. First they segmented the ECG samples into several windows containing multiple beats. Then, each window was submitted to the AC/DCT, discarding the least significant values. Using two datasets, PTB and MIT-BIH Arrhythmia Database (MIT-BIH), they obtained 100% and 94.88% accuracy with the fiducial approach, and 84.61% and 100% accuracy with the AC/DCT method, for PTB and MIT-BIH, respectively.

Using ECG and PCG, Fatemian *et al.* evaluated human identification with physiological signals in [11]. Their simple approach consisted of using Discrete Wavelet Transform (DWT) for noise reduction and a threshold decision for the classification. Settings the threshold to 0.85 and 0.98, they successfully identified respectively 95.37% and 90.6% of the samples.

Jung *et al.* presented in [16] a method to improve individual identification with ECG signals using DCT for window removal. For their experiment they used three datasets, Normal Sinus Rhythm Database (NSR), PTB and QT Database (QT) with a total of 672 subjects. The pipeline consisted of first, generate several windows of 5 seconds per signal, then pre-process the ECG signals using Daubechies 4 for noise removal, compute and scale AC coefficients and apply the DCT, and finally remove the unrecognizable windows. Regarding the classification, and comparing NN, Support Vector Machine (SVM) and LDA, they achieved the best results with the latter by correctly identifying 100% of the subjects of every dataset, while obtaining a window identification accuracy of 98.67%, 98.65% and 99.23% for NSR, PTB and QT, respectively.

5 Discussion and Conclusion

Among the studies presented, there are several particular aspects concerning each waveform signal. ECG signals contain more identifiable components, thus containing more information. However, the correct identification of each component require signal synchronization, specifically when considering fiducial based approaches. Moreover, as the signal is not periodic but highly repetitive [8], it is affected by HRV. Nevertheless, the great majority of the studies converge to this waveform due to its versatility in terms of the amount of information available.

Regarding PPG signals, it is composed essentially by three segments, thus providing less information than ECG. However, the PPG sensor is currently present in the majority of the smart *wearables* whereas ECG sensors are just starting to be introduced in this context.

Nevertheless, every study presents promising results leading to the conclusion that it is in fact possible to use heart-based signals for the purpose of human identification and consequently authentication. Notwithstanding, the signals duration may yet be a constraint when considering a heart-based biometric system, as the shortest duration observed was 5 seconds. Nonetheless, in a context of continuous authentication this may be a potentially optimal solution due to its intrinsic characteristic, specially liveness detection and robustness.

References

1. Foteini Agrafioti. *ECG in biometric recognition: Time dependency and application challenges*. PhD thesis, University of Toronto, 2011.
2. Nazneen Akhter, Hanumant Gite, Gulam Rabbiani, and Karbhari Kale. *Heart Rate Variability for Biometric Authentication Using Time-Domain Features*, volume 377. Springer, 2015.
3. AngioScan-Electronics. Schematic representation of an optical sensor mounted on the terminal phalanx of the finger. (Online; accessed march 15, 2016).
4. Lena Biel, Ola Pettersson, Lennart Philipson, and Peter Wide. Ecg analysis: a new approach in human identification. *Instrumentation and Measurement, IEEE Transactions on*, 50(3):808–812, 2001.
5. Angelo Bonissi, Ruggero Donida Labati, Luca Perico, Roberto Sassi, Fabio Scotti, and Luca Sparagino. A preliminary study on continuous authentication methods for photoplethysmographic biometrics. *2013 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications, BioMS 2013 - Proceedings*, (March 2016):28–33, 2013.
6. Roberto Brunelli and Tomaso Poggio. Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):1042–1052, 1993.
7. Carlos Carreiras, André Lourenço, Ana Fred, and Rui Ferreira. Ecg signals for biometric applications-are we there yet? In *Informatics in Control, Automation and Robotics (ICINCO), 2014 11th International Conference on*, volume 2, pages 765–772. IEEE, 2014.
8. Chuang-Chien Chiu, Chou-Min Chuang, and Chih-Yu Hsu. A novel personal identity verification approach using a discrete wavelet transform of the ecg signal. In *Multimedia and Ubiquitous Engineering, 2008. MUE 2008. International Conference on*, pages 201–206. IEEE, 2008.
9. D Dumn. Using a multi-layer perceptron neural for human voice identification. In *Proc. Fourth Int. Conf. Signal Process. Applicat. Technol*, 1993.
10. Mohamed Elgendi. On the analysis of fingertip photoplethysmogram signals. *Current cardiology reviews*, 8(1):14–25, 2012.
11. S Zahra Fatemian, Foteini Agrafioti, and Dimitrios Hatzinakos. Heartid: Cardiac biometric recognition. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1–5. IEEE, 2010.
12. Robert W Frischholz and Ulrich Dieckmann. Biold: a multimodal biometric identification system. *Computer*, 33(2):64–68, 2000.
13. YY Gu, Y Zhang, and YT Zhang. A novel biometric approach in human verification by photoplethysmographic signals. *Information Technology Applications in Biomedicine, 2003. 4th International IEEE EMBS Special Topic Conference on*, 2003.
14. IEEE. *ECG biometric recognition without fiducial detection*, 2006.
15. Steven A Israel, John M Irvine, Andrew Cheng, Mark D Wiederhold, and Brenda K Wiederhold. Ecg to identify individuals. *Pattern recognition*, 38(1):133–142, 2005.

16. Woo-Hyuk Jung and Sang-Goog Lee. Ecg identification based on non-fiducial feature extraction using window removal method. *Applied Sciences*, 7(11):1205, 2017.
17. André Lourenço, Carlos Carreiras, Hugo Silva, and Ana Fred. Ecg biometrics: A template selection approach. In *Medical Measurements and Applications (MeMeA), 2014 IEEE International Symposium on*, pages 1–6. IEEE, 2014.
18. Tsutomu Matsumoto, Hiroyuki Matsumoto, Koji Yamada, and Satoshi Hoshino. Impact of artificial gummy fingers on fingerprint systems. In *Electronic Imaging 2002*, pages 275–289. International Society for Optics and Photonics, 2002.
19. Michael Negin, M Salganicoff, and Grace G Zhang. An iris biometric system for public and personal use. *Computer*, 33(2):70–75, 2000.
20. P Jonathon Phillips, Alvin Martin, Charles L Wilson, and Mark Przybocki. An introduction evaluating biometric systems. *Computer*, 33(2):56–63, 2000.
21. Ashok Samal and Prasana A Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern recognition*, 25(1):65–77, 1992.
22. Tsu-Wang Shen, Willis J Tompkins, and Yu Hen Hu. Implementation of a one-lead ecg human identification system on a normal population. *Journal of Engineering and Computer Innovations*, 2(1):12–21, 2011.
23. Tsu-Wang Shen, WJ Tompkins, and YH Hu. One-lead ecg for identity verification. In *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*, volume 1, pages 62–63. IEEE, 2002.
24. Yogendra Narain Singh, Sanjay Kumar Singh, and Phalguni Gupta. Fusion of electrocardiogram with unobtrusive biometrics: An efficient individual authentication system. *Pattern Recognition Letters*, 33(14):1932–1941, 2012.
25. Petros Spachos, Jiexin Gao, and Dimitrios Hatzinakos. Feasibility study of photoplethysmographic signals for biometric identification. *17th DSP 2011 International Conference on Digital Signal Processing, Proceedings*, (March 2016), 2011.
26. Ya-Ting Tsao, Tsu-Wang Shen, Tung-Fu Ko, and Tsung-Hsing Lin. The morphology of the electrocardiogram for eevaluating ecg biometrics. In *e-Health Networking, Application and Services, 2007 9th International Conference on*, pages 233–235. IEEE, 2007.
27. Yongjin Wang, Foteini Agrafioti, Dimitrios Hatzinakos, and Konstantinos N Plataniotis. Analysis of human electrocardiogram for biometric recognition. *EURASIP journal on Advances in Signal Processing*, 2008:19, 2008.
28. Chen Wei, Lei Sheng, Guo Lihua, Chen Yuquan, and Pan Min. Study on conditioning and feature extraction algorithm of photoplethysmography signal for physiological parameters detection. *Proceedings - 4th International Congress on Image and Signal Processing, CISP 2011*, 4:2194–2197, 2011.
29. Psychology Wiki. Electrocardiography. (Online; accessed march 10, 2016).
30. Gerd Wübbeler, Manuel Stavridis, Dieter Kreiseler, Ralf-Dieter Bousseljot, and Clemens Elster. Verification of humans using the electrocardiogram. *Pattern Recognition Letters*, 28(10):1172–1175, 2007.
31. Jianchu Yao, Xiaodong Sun, and Yongbo Wan. A pilot study on using derivatives of photoplethysmographic signals as a biometric identifier. *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, 2007:4576–4579, jan 2007.

A framework to automate network packets construction

Cristiano Alves[1] and Nuno Anacleto[1] and Mário Antunes[1,2]

[1] Polytechnic Institute of Leiria, School of Technology and Management, Leiria, Portugal
{2170103,2170099}@my.ipleiria.pt, mario.antunes@ipleiria.pt

[2] INESC-TEC, CRACS, University of Porto, Portugal
mantunes@dcc.fc.up.pt

Abstract. Automatic network packet generation is widely used on several application domains, namely on producing network packet datasets, to launching malicious activity through the network or simply to understand and learn how computers networks behave. This paper aims to present a framework to automate the construction of IP network packets based on open source components. It is a work in progress where the main goals are twofold: to define an input format to identify the IP packets flows; and to automate IP packets construction for further injection in the network or offline processing. Results obtained so far are promising and prove the usefulness of such a framework in the context of network administration.

Keywords: Computers network, IP, dataset, TCP/IP, intrusion detection.

1 Introduction

IP network packets generation is widely used by networking practitioners, to test new network administration applications, like network intrusion detection systems, but also to understand how IP networks and protocols would behave in real scenarios. The overall idea is to automatically generate network packet datasets that could be further injected in the network interface. Metasploit (https://www.metasploit.com/) allows us to generate and launch anomalous and illegitimate network traffic. A wide set of network simulators can be used to generate traffic flows and to simulate their behavior in test scenarios, like the well-known NS2 (https://www.isi.edu/).

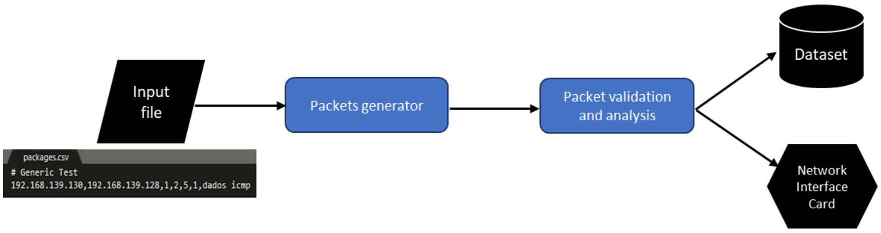
Existing network datasets, used to test classifiers or detectors, have some drawbacks, like those identified in the well-known KDDCUP 1999 DARPA dataset [1,2], mainly because they are based on synthetic and artificially generated flows.

In this paper we present a framework to automate the construction of network packets that can be further injected on the interface card or collected to a dataset for offline analysis and processing. The packets are generated based on an input specification, integrate an IP flow and may include both normal and anomalous network activity. The development was made in Perl and took advantage of the Net::RawIP Perl module [3]. In this first stage of development we have only used IP packets flows related with the most common TCP/IP protocols, namely TCP, UDP and ICMP.

The paper is organized as follows: section 2 describes the framework and its main components; section 3 presents the preliminary results obtained, and finally, on section 4 we discuss the conclusions and delineate the future work.

2 Framework

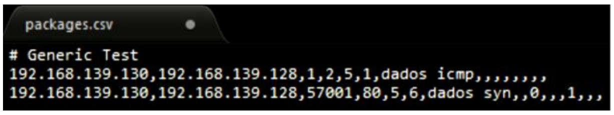
Figure 1 depicts the building blocks involved in the framework. There are four main blocks: the input text file that specifies the IP packets generation rules; the Perl script that implements the packet generator; a module for packets analysis and validation to verify the correctness of the packets produced, and finally, an output module that injects packets to the network interface or produce a file with the packets in PCAP format, for further offline usage.



```
# Generic Test
192.168.139.130,192.168.139.128,1,2,5,1,dados icmp
```

Fig. 1. General overview of packet production framework

Figure 2 depicts the format of the input file and an example of two entries. It is a part of a CSV file (one packet per line) in which the following IP, TCP and UDP header fields were considered: source and destination IP addresses; source and destination ports; TTL; upper layer protocol type; checksum; TCP flags (seq, ack, rst, syn,fin); packet and segment identification and application payload data. All the fields are separated by the symbol ",".

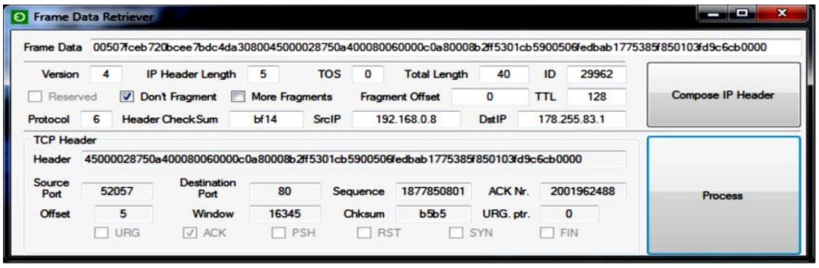


```
# Generic Test
192.168.139.130,192.168.139.128,1,2,5,1,dados icmp,,,,,,,,
192.168.139.130,192.168.139.128,57001,80,5,6,dados syn,,0,,,1,,,
```

Fig. 2. Input text file format

We have used Wireshark (http://www.wireshark.com/) to visualize and analyze what was being injected into the network, thus validating if the packets produced match the input parameters defined in the input text file.

We have also developed Frame Data Retriever (FDR) tool, a graphical application to extract and decipher the data collected by the network interface card, at both IP and transport TCP/IP layers. FDR, depicted in Figure 3, is also able to calculate IP, TCP and UDP headers checksum, useful to verify the integrity of IP packets constructed.



Frame Data Retriever

Frame Data: 00507fceb720bcee7bdc4da30800450000028750a400080060000c0a80008b2ff5301cb5900506fedbab1775385f850103fd9c6cb0000

Version: 4 IP Header Length: 5 TOS: 0 Total Length: 40 ID: 29962

□ Reserved ☑ Don't Fragment □ More Fragments Fragment Offset: 0 TTL: 128

Protocol: 6 Header CheckSum: bf14 SrcIP: 192.168.0.8 DstIP: 178.255.83.1

TCP Header

Header: 45000028750a400080060000c0a80008b2ff5301cb5900506fedbab1775385f850103fd9c6cb0000

Source Port: 52057 Destination Port: 80 Sequence: 1877850801 ACK Nr.: 2001962488

Offset: 5 Window: 16345 Chksum: b5b5 URG. ptr.: 0

□ URG ☑ ACK □ PSH □ RST □ SYN □ FIN

Fig. 3. Frame Data Retriever tool

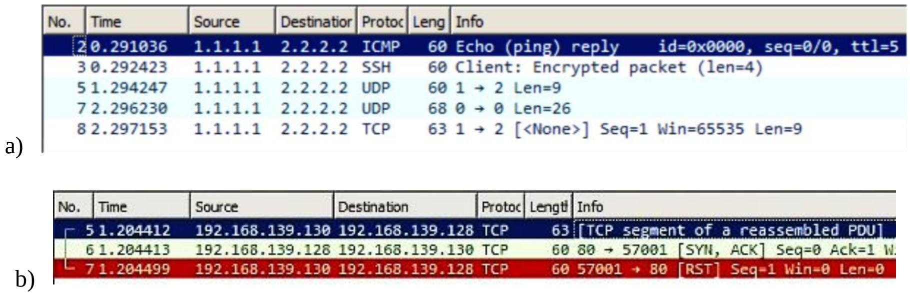
3 Results

A comprehensive set of experiments were carried out to evaluate the usefulness of the application and to validate the integrity of data produced.

Table 1. Description of tests.

Type of test	Description	Qty.
Generic tests	Synthetic and valid packets constructed to confirm the extension's correct functionality.	5
SSH tests	Specific to port 22.	2
Invalid packets tests	Erratic checksum fields, wrong source and destination ports, and oversized payloads.	7
Sender changes tests	IP addresses are public.	5

Table 1 describes the four types of tests considered and the number of tests executed. The types are the following: 1) in **generic tests** synthetic packets with no meaning were produced; 2) **SSH tests** involves network flows in which traffic is sent towards TCP port 22; 3) **invalid packets tests**, related to packets in which checksum fields of both the IP and upper layers headers were attempted to be altered. It was possible to identify incorrect and altered UDP header checksum that created invalid datagrams further discarded at the destination. Another invalid packet test consisted on increasing the payload length to frames above the maximum transmit unit (MTU) and see the influence on the injection and validation processes; 4) in **sender changes tests** the traffic injected into the network had public IP addresses as the Source IP address (Figure 4a).



a)

```
No. Time      Source   Destinatior Protoc Leng Info
  2 0.291036  1.1.1.1  2.2.2.2  ICMP   60 Echo (ping) reply    id=0x0000, seq=0/0, ttl=5
  3 0.292423  1.1.1.1  2.2.2.2  SSH    60 Client: Encrypted packet (len=4)
  5 1.294247  1.1.1.1  2.2.2.2  UDP    60 1 → 2 Len=9
  7 2.296230  1.1.1.1  2.2.2.2  UDP    68 0 → 0 Len=26
  8 2.297153  1.1.1.1  2.2.2.2  TCP    63 1 → 2 [<None>] Seq=1 Win=65535 Len=9
```

b)

```
No. Time      Source          Destination     Protoc Lengtł Info
  5 1.204412  192.168.139.130 192.168.139.128 TCP    63 [TCP segment of a reassembled PDU]
  6 1.204413  192.168.139.128 192.168.139.130 TCP    60 80 → 57001 [SYN, ACK] Seq=0 Ack=1 W.
  7 1.204499  192.168.139.130 192.168.139.128 TCP    60 57001 → 80 [RST] Seq=1 Win=0 Len=0
```

Fig. 4. a) Public IP addresses traffic; b) Response to SYN segments.

We have also executed some variations of these tests, to infer about the robustness of the application. Some examples are the creation of packets with invalid checksum, the injection of network traffic involving reserved TCP and UDP ports as origin and/or destination. In such case, as can be seen in Figure 4c, after sending the IP packet with SYN, the destination sent a SYN/ACK pair TCP segment and the operating system sent automatically a RST segment.

4 Conclusions and future Work

In this paper we have described a generic framework to automate the production and validate the integrity of IP packets. The packets produced can be injected in the interface card and/or stored in a dataset for offline processing. We have conducted a set of preliminary tests and obtained promising results. Despite the documented limitations of Net::RawIP Perl module, the automatization process is robust and the data integrity was validated. The framework deployed can help networking practitioners to produce their own datasets and to test offline applications and algorithms, such as intrusion detection systems.

Despite the promising results, this is an ongoing development and some issues are planned to be addressed. Enhancing network traffic injection and offline collection of packets is an eminent issue that is now under development and refinement. To refine the specification language of the input file is another enhancement to be produced. Basically, it should be possible to produce virtually any kind of packet, both synthetically or integrated in an IP packets flow. The use of a formal specification language for the input file, namely XML, is also in the list of developments to be made.

An interesting direction in the integration of this work in a network intrusion system, to test and evaluate intrusion detection systems. The flexibility of the input specification should give freedom to the network administrator to define the kind of normal and anomalous traffic to be produced. Again, in an intrusion detection perspective, the production of network traffic could be used to train and refine an intrusion detection sensor to the network profile being produced by the traffic generator deployed under this framework.

References

1. Lippmann, R., Haines, J. W., Fried, D. J., Korba, J., & Das, K. (2000). The 1999 DARPA off-line intrusion detection evaluation. Computer networks, 34(4), 579-595.
2. McHugh, J. (2000). Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. ACM Transactions on Information and System Security (TISSEC), 3(4), 262-294.
3. Sergey Kolychev, Gabor Szabo, "Net::RawIP - Perl extension to manipulate raw IP packets with interface to libpcap"; available at CPAN (http://search.cpan.org/)

A network packets sniffer with real-time graphical analytics

Daniel Pinto[1] and Mário Antunes[12]

[1] Polytechnic Institute of Leiria, School of Technology and Management, Portugal
2170104@my.ipleiria.pt,mario.antunes@ipleiria.pt
[2] INESC-TEC, CRACS, University of Porto, Portugal
mantunes@dcc.fc.up.pt

Abstract. Network traffic analysis is a key activity for networking practitioners, not only to understand the networking processes and behavior, but also to detect anomalous activities that could be related with intrusions or ongoing attacks. The existing graphical network sniffers are able to collect and display graphically the packets flows, but fail to identify real-time anomalous network activity. This paper aims to present a framework to collect network packets and to display in real-time a set of indicators that may help the network administrator to identify ongoing attacks or anomalous network activities. It is a work in progress that has, at this stage, the goal to develop the network packets collector and a front-end to analyse network flows indicators in real time.

1 Introduction

The network traffic analysis is a key responsibility of system administrators (SysAdmin) and other networking practitioners. Besides the understanding of network activity and behavior, the overall goal is to analyze packets flows and to further evaluate those that may be related to a network intrusion.

There is a wide set of applications devoted to collect, process, analyse and report network traffic activity. Network traffic sensors incorporated in intrusion detection systems (IDS [5]), like Snort-IDS (`www.snort.org`), are able to alert the SysAdmin about suspicious flows. Despite its abilities, it is a signature based IDS without learning and graphical features.

Wireshark [6] is a well-known network sniffer with good graphical and easy filtering abilities, but fails on the identification of suspicious online traffic. Other state-of-the-art applications, like `tcpdump`, allow SysAdmins to automate packets capture to offline analysis, but fail on graphical abilities and automatically collecting network packets flows indicators.

In the proposed framework we aim to sniff the network packets and also to analyse in real-time the flows indicators related with, for example, amount and type of flows and the amount of packets captured for each flow in a time frame, normally between five minutes to an half-hour. This paper is structured as follow. After introducing the fundamentals and nowadays solutions, we depict the framework and how the pilot application was implemented. Then we describe some preliminary tests and, finally, we present conclusions and delineate the future work.

2 Framework architecture

The framework architecture, depicted in figure 1, has two main modules: sniffer and web. In other words, the network packets sniffing and the Web data visualization. Modules are independent and the data collected is stored in text format for sharing between the modules by using JSON format for an easy understanding and processing.

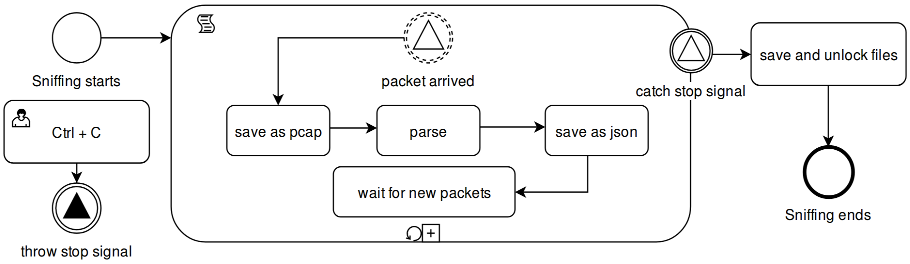
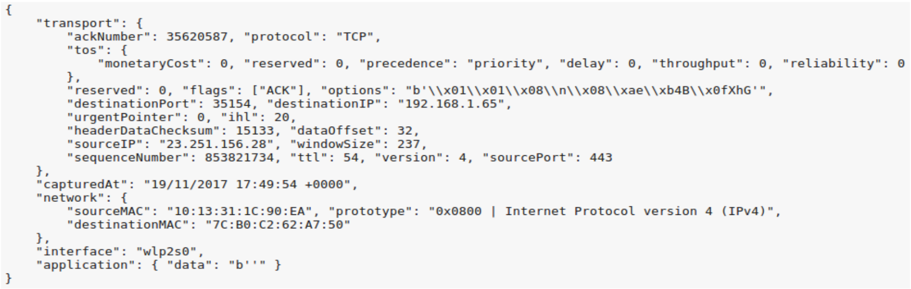


Fig. 1. General overview of the Sniffer Process (Business Process Model and Notation)

The process starts by collecting packets in the network interface card and further saving them in a Packet CAPture (PCAP) format file. Then all the binary information related with the TCP/IP protocols involved is parsed and stored in a JSON file, that can be further processed to be displayed in a web-based front-end application. Figure 2 illustrates the JSON representation of the parsing made to the packets, by identifying the header fields related with the TCP/IP layers [2].



```
{
    "transport": {
        "ackNumber": 35620587, "protocol": "TCP",
        "tos": {
            "monetaryCost": 0, "reserved": 0, "precedence": "priority", "delay": 0, "throughput": 0, "reliability": 0
        },
        "reserved": 0, "flags": ["ACK"], "options": "b'\\x01\\x01\\x08\\n\\x08\\xae\\xb4B\\x0fXhG'",
        "destinationPort": 35154, "destinationIP": "192.168.1.65",
        "urgentPointer": 0, "ihl": 20,
        "headerDataChecksum": 15133, "dataOffset": 32,
        "sourceIP": "23.251.156.28", "windowSize": 237,
        "sequenceNumber": 853821734, "ttl": 54, "version": 4, "sourcePort": 443
    },
    "capturedAt": "19/11/2017 17:49:54 +0000",
    "network": {
        "sourceMAC": "10:13:31:1C:90:EA", "prototype": "0x0800 | Internet Protocol version 4 (IPv4)",
        "destinationMAC": "7C:B0:C2:62:A7:50"
    },
    "interface": "wlp2s0",
    "application": { "data": "b''" }
}
```

Fig. 2. Packet captured to JSON

Web analytics module processes the data previously collected in JSON format into graphics. This module aims to present in real-time and using a web-based interface, the results obtained by the sniffing process. The visualization process is fast, intuitive and has useful information about the network packets flows, being possible to detect in real-time anomalous activity in the network.

Figure 3 illustrates one of charts produced by the application. It is related with the processing of 15118 packets collected during fifteen minutes on a home WiFi network. The IP packets carrying TCP segments were grouped by TCP flags. As shown, ACK is the most used flag by this protocol.

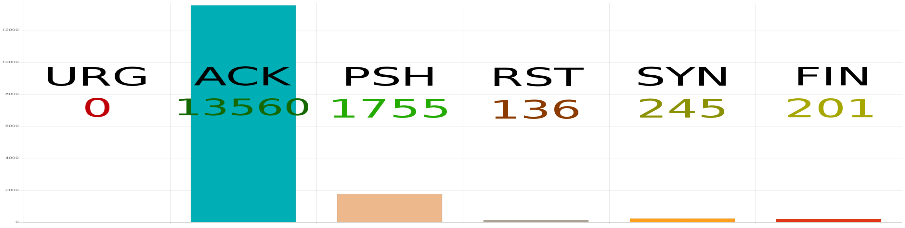


Fig. 3. 15897 grouped TCP flags from a total of 15118 packets

JSON[1] is lightweight format for data interchanging, making the information universal and readable for any human or tool, meaning that other programs can read the information because the schema is understandable, unlike binary data where we need additional knowledge and a specific application that knows the packet format.

A key for scalability is to keep things simple, dependent-free and easy to read. This was the driven to adopt JSON for interchanging data between the script module and the web module. With JSON the web module is not the only one that may be able to understand and create charts from the captured data, meaning that other external applications or other modules could also use it to exchange information. Also, with the advent of new web frameworks and technologies such as NoSQL[7] databases, it's possible to develop data-mining or business intelligence using technologies like Elastic Search[3].

3 Tests and Results

The obtained tests results were promising. Due the simplicity of the application and the little dependencies, we were able to accomplish almost instant results when loading the graphics using JSON files with more than ten thousand of entries of packets captured using real network traffic. This entries were then processed by the graphics to display the results by protocol (TCP, UDFP or ICMP), the TCP packets grouped by TCP flags and the packets grouped by source and destination IP. In this preliminary tests, the big slice is related with TCP protocol.

In another of those tests were captured 1552 packets in about 5 minutes using mainly web traffic and a video streaming. The results got the following distribution by protocol: TCP with 760 packets, UDP with 630 packets and ICMP with 162 packets.

To ensure results were correct, the parsed packets were printed out to a terminal while the script was running. The PCAP file was opened with Wireshark and compared with the JSON file to view the binary data and compare to check the integrity and truthfulness of the information in both files.

4 Conclusion and future work

In this paper we presented a work in progress network packets sniffer application to instantly allow the SysAdmin to get relevant information in a human readable way. The preliminary results were satisfactory, both the integrity of the packets collector and real-time graphical visualization. There are however some directions to explore in this work, namely: 1) to integrate a detection module to alert anomalous network activity observed by the abnormal amount of traffic in the flows, acting in some ways as an out-of-the-box intrusion prevention system (IPS); 2) to include a data mining and artificial intelligence module to better analyze the reposts produced and thus enhance the detection module; 3) to implement additional features in the collector, like the ability to save reports, to have multiple charts in real-time besides the flow analysis; 4) the development done so far is only for IPv4 networks, being an additional feature to implement the processing of IPv6 [4] packets.

In a short term the project intends to improve the actual modules by adding new features that allows the creation of new charts, advanced queries and filtering in the web module. For the script, more parameters should be added to better sniff the network. Also, the script must be able to translate a PCAP file as input, returning a JSON file as output and vice-versa.

In a long term development the framework should have an hybrid IPS discovering new attacks through the use of machine learning methods.

References

1. Tim Bray. The javascript object notation (json) data interchange format. 2017.
2. Walter Goralski. *The Illustrated network: How TCP/IP works in a modern network.* Morgan Kaufmann, 2017.
3. Clinton Gormley and Zachary Tong. *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine.* " O'Reilly Media, Inc.", 2015.
4. Robert Hinden. Internet protocol, version 6 (ipv6) specification. 2017.
5. Craig H Rowland. Intrusion detection system, June 11 2002. US Patent 6,405,318.
6. Chris Sanders. *Practical packet analysis: Using Wireshark to solve real-world network problems.* No Starch Press, 2017.
7. Kiran Fahd Sitalakshmi Venkatraman, Samuel Kaspi, and Ramanathan Venkatraman. Sql versus nosql movement with big data analytics. 2016.

SESSION 8

Distributed Computing

Utopia: A Generic Software-Defined Storage Architecture
Ricardo Macedo

State Synchronization by State Decomposition
Vitor Enes

Utopia: A Generic Software-Defined Storage Architecture

Ricardo Macedo[1]

HASLab, INESC TEC & U. Minho
ricardo.g.macedo@inesctec.pt

Abstract. The exponential growth of data produced is increasing both computational and storage demands of today's storage infrastructures, leading to the shift to the next-generation of storage solutions. However, current software-defined storage systems cannot attend these overwhelming requirements due to their non-scalable logically centralized controller. In this paper, we address this challenge by introducing Utopia, a generic software-defined storage architecture that provides high scalable and flexible control over multiple storage environments. By redesigning the control plane, we expect to deliver software-defined storage principles to centralized, hybrid and decentralized storage environments. Furthermore, this paper provides a vision of a software-defined storage architecture that we consider to be suitable for the next-generation of storage systems.

Keywords: Software-Defined Storage, Distributed Storage, Distributed Computing.

1 Introduction

With the adoption of the cloud computing paradigm, companies can deploy their solutions in a quick and flexible way, running them on an enterprise-grade IT infrastructure with a minimal up-front capital expenditure [1]. From a small network to a massive environment, enterprises can adjust both storage and processing capabilities with ease. However, in order to hide the system's complexity and at the same time, ensure quality of service (QoS), maintainability, flexibility and security to end users, several challenges have risen in the construction and management of these infrastructures. For instance, today's data centers comprise several layers along the IO path such as end user applications, virtual machines (VMs), several file systems and block devices [14]. Each one of these layers provide distinct interfaces and procedures to the IO, leveraging a strict and complex treatment of the IO flow, decreasing the overall system's performance. Additionally, since data forwarding, processing and management occurs at the same place, the ability to enforce policies in the system may be unfeasible. To address these challenges, *Software-Defined Storage* (SDS) has emerged recently to hide all complexities of management and improve control functionality in the traditional systems [9, 14, 15]. By decoupling the control and management tasks of the layers into two distinct components, the control plane and the data plane, it is easier to enforce policies due to the system's modularity and flexibility. The SDS control plane is logically centralized and tracks the status of each data plane layer in order to enforce policies across them. A policy defines a set of rules to be applied in the underlying storage devices to ensure QoS, for example, prioritize IO requests or define a maximum throughput limit to the system. As for the SDS data plane, it can be deployed across several storage components. By defining a modular and flexible data plane, besides expanding the policy domain, the controller has the ability to fine-tune each storage component to best meet the end user requirements. Current SDS systems, such as IOFlow, Retro and Crystal [9, 14, 15], have borrowed several ideas from the *Software-Defined Network* (SDN) paradigm such as the decoupling of the control and the data plane, the logically centralized controller and the data plane programmability [5, 7]. By adopting these ideas, current SDS approaches seems to be suitable for small-to-medium infrastructures [14]. However, produced data is growing at an unprecedented rate and latest projections predict a total of 44 *zettabytes* by 2020, a 10x increase from 2013 [16]. At the same time, data centers must increase their storage and computational power in order to keep the QoS delivered to the end user. As result, traditional SDS systems are unsuitable to attain these demands due to the overwhelming load on the centralized controller.

To address these challenges, this paper proposes Utopia, a generic software-defined storage architecture that provides a high scalable and flexible control over multiple distinct storage systems. Contrarily to the traditional logically centralized controller, we introduce a novel approach, by splitting the control plane into several controllers. Sharing the control and management actions over multiple nodes, improves the overall scalability of the system. In addition, the Utopia system is agnostic at two levels: *(i) model-agnostic*, while traditional SDS systems only can be applied over centralized models, Utopia is also suitable for hybrid and fully decentralized models [13]; and *(ii) resource-agnostic*, by analyzing the capacity and monitoring the underlying storage components, Utopia dynamically adapts and enforces policies over the system, thus being suitable for both homogeneous and heterogeneous environments. Additionally, since the previous works emphasize more on the data plane design and challenges, we focus on the control plane, providing a novel and detailed description of its components, roles and design challenges. As for the data plane, we present its general notions and we focus on how Utopia can leverage from existing solutions. With this paper we provide a vision of a software-defined storage architecture that we consider to be suitable for the next-generation of storage systems.

The remainder of this paper is organized as follows. We present the background of software-defined storage and current related work in Section 2. Section 3 details about the Utopia control plane design and depicts the general notions of the data plane and control applications. Section 4 describes the models that we consider Utopia is suitable for. The paper concludes in Section 5 with relevant observations and future work.

2 Background

The main goal of any software-defined system is to hide all the complexities of the management and control functionality of the system resources from the end users [8]. In this section we present a background of the general notions of software-defined storage and its challenges, alongside policy related topics.. Lastly, we present a survey of related SDS systems.

2.1 Software-Defined Storage

Recently, SDS solutions have emerged to facilitate the management and simplify the complexity of data storage systems, and at the same time, maintain an acceptable level of QoS. As in SDN, the separation of control and management tasks from the data layers into the control plane and the data plane, is the best-known principle in SDS. The control plane refers to the software component that manages and controls the storage resources through a policy-based mechanism. As for the data plane, it refers to the underlying infrastructure of the storage assets, enabling dynamic configuration of service and routing properties. Such isolation introduces novel design principles such as abstraction, flexibility and modularity, all key properties for supporting scalable solutions over heterogeneous applications. However, it also introduces new challenges that may be hard to solve. For instance, the control plane must provide dynamic configurable functionalities: on one hand to handle and enforce arbitrary policies; on the other hand, the system may need to self-adjust and apply orchestration policies to improve the overall system performance and at the same time, meet a certain QoS level. As for the data plane, it must be fast and cause minimal performance degradation. Any performance degradation will be noticed as the load on the system increases.

Another well-known design principle adopted by most of the SDS systems, is the logically centralized controller [9, 12, 14, 15]. With this centralized unit, yet physically distributed, the controller has global visibility over the system, being able to discover and interact with several storage components across the storage infrastructure. Typically, it maintains and exposes a topology graph to control applications. By consolidating the control operations to a single control unit, it eases both development of (centralized) control algorithms and management and enforcement of policies. As proved in current SDS solution, this traditional controller is suitable for small-to-medium data storage systems (*e.g.,* small data centers). However, shifting the current SDS paradigm to a larger storage environment with tens of thousands or even millions of nodes such as the Google's and Amazon's data centers, might be an unfeasible task, since the single point of control will be overloaded with massive amounts of logs and control requests. Furthermore, with the exponential growth of data produced, the adoption of the *Internet of Things* (IoT) paradigm as a storage and computational solution is becoming more realistic [2, 17]. Similarly, current solutions cannot deliver the SDS principles to this massive environment.

2.2 Policies

Policies are rules that define the behavior of a SDS solution, as well as the actions it must employ over the storage infrastructure. Policy representation can assume a simple tuple format, as presented in Table 1. The depicted policies are referred as high-level policies, which are sent from control applications to the controller. From a single high-level policy, a policy translator generates one or more low-level policies, corresponding to rules to be applied on the data plane.

The ability to control and enforce policies in storage systems, is one of the most important characteristics of SDS solutions. However, policy enforcement is difficult and raises many challenges. Some policies require *distributed enforcement*, since they may need to be enforced at more than one layer along the storage infrastructure. For example, policy P_2 entails that requests from components C_2 and C_3 should have higher priority, so it needs to be enforced at all layers along the IO path. Others require to be dynamically configurable (*dynamic enforcement*), since they may need to be adjusted to best meet the system requirements. For instance, policy P_4 requires that the bandwith limit for the component C_1 must be at least 2000 Mbit/s, but can be adjusted based on the spare system capacity. Additionally, policies are subjected to an *admission control* process, that decide their feasibility based on the capacity of the underlying resources. Finally, some policies may conflict with others due to causal dependencies or non-interoperable results, causing *policy overlapping.* For instance, consider policies P_1 and P_3. Policy P_1 entails that all requests from C_1 to the path */src/** must be encrypted with the AES-128 encryption scheme. As for P_3, it entails that all request from C_1 and C_4 to all paths, must be compressed with the snappy compression algorithm. Since data compression must be issued before the data encryption, the policy enforcement rules must act equally.

	Policy	**Policy Type**
P_1	{ C_1, Encryption, AES-128, /src/* }	Data-oriented
P_2	{ [C_2,C_3], Priority, High , * }	QoS-oriented
P_3	{ [C_1,C_4], Compression, Snappy, * }	Data-oriented
P_4	{ C_1, Bandwidth, $\geq$ 2000 Mbit/s, * }	QoS-oriented

Table 1. Example of data storage policies and the respective policy type. For ease of exposition, we represent policies of the form: {[*Storage components*], *Action*, [*Action properties*], [*Paths*]}. Data-oriented policies refers to direct actions over data. QoS-oriented policies refers to actions over the system behaviour (that could lead to undirect actions over data).

2.3 Related Work

IOFlow, now extended as sRoute, was the first complete SDS architecture [12,14]. Based on SDN principles, it decouples the control and the data plane and introduces programmable data plane queues that allow for flexible service and routing properties. IOFlow enables end-to-end policies to specify the treatment of IO flows from VMs to shared storage. It also introduces the notions of high-level policies translation into low-level data plane rules. Further, sRoute include IO routing abstraction, allowing tail latency control, replica set control and file cache isolation [12]. Retro is a framework for implementing resource management policies in multi-tenant distributed systems [9]. Similarly to IOFlow, it separates the control tasks from the data services into a system- and resource-agnostic data plane and a logically centralized control plane. Retro also introduces reactive policies that dynamically respond to the current resource usage of workflows in the system, instead of relying on static models of future resource requirements. In addition of the previous SDS solutions, Crystal is a SDS architecture that pursues an efficient use of multi-tenant object stores [15]. It provides new abstraction levels, allowing to add new functionalities at the data plane that can be immediately managed at the control plane. As for single data plane solutions, SafeFS is a modular FUSE-based architecture that allows system operators to simply stack building blocks, each whith a specific functionality (*e.g.,* compression, encryption, replication) implemented by plug-and-play drivers [11]. This modular and flexible design allows extending layers with novel algorithms in a straightforward fashion.

3 Design

Utopia is a generic software-defined storage architecture that provides a high scalable and flexible control over multiple storage environments. Its design aims to be model- and resource-agnostic, leveraging the SDS principles to centralized, hybrid (Figure 2) and decentralized (Figure 3) models, regardless the underlying storage devices. Similarly to previous SDS solutions, Utopia decouples the control and management tasks from the underlying data components into a control and a data plane. The data plane follows a queue-based design, delivering modularity and flexibility to both service and routing properties. It also exposes a simple control interface that allows the control unit to monitor, control, enforce low-level policies and dynamically configure the data plane queues to best meet both user and system requirements. Towards the control plane design, we introduce a novel principle by splitting the control plane into several controllers and distribute them across the storage infrastructure. This distribution depends on the underlying storage model. As result, the control plane is able to assume a logically centralized controller, a hierarchical controller or a decentralized controller (we analyze these assumptions with more detail in Section 4). Similarly to the major SDS solutions, the control plane exposes a simple interface to control applications that can be built on top. This interface allows users (*e.g.,* data center administrators, management software within the storage infrastructure) to issue high-level policies and deploy, control and configure the Utopia system.

3.1 Control Plane

The control plane is an essential part of any SDS solution. It is assigned to control and manage the storage infrastructure by enforcing policies. However, current SDS solutions have given more focus to the data plane design, causing the lack of a clear proposal about the architecture of this component. In this section, we introduce a novel proposal of the control plane architecture and detail about its internals. Figure 1 depicts the architecture of the Utopia control plane. The control plane comprises five major modules: *Handler*, *Watcher*, *Enforcer*, *Storage* and *Monitor* modules.

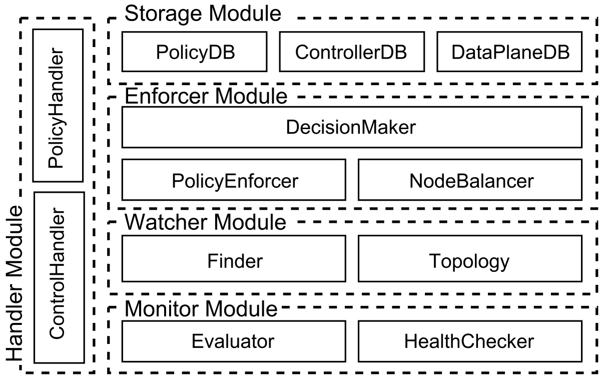


Fig. 1. Utopia control plane architecture.

Handler Module. This module handles all control and policy related communications. The *PolicyHandler* receives both high-level and low-level policies and converts the high-level ones into a set of low-level policies. It also sends all policies to the *PolicyDB* component to be stored. The *ControlHandler* sends and receives all control related requests such as data plane logs, control errors and control decisions. After digest the message, the *ControlHandler* sends it either to the *ControllerDB* or the *DataPlaneDB*.

Watcher Module. This module is assigned to discover new nodes and components in the system, in order to add and control new instances. This is done through the *Finder* component and it is particularly useful in decentralized models, where new nodes join the system without previous notice. The *Topology* component maintains a topology graph that contains the data planes structure and the underlying storage devices. This component provides global (or partial) visibility of the system and helps to improve the enforcement of policies by choosing where a policy should be enforced or which order a policy must follow to avoid conflicts.

Enforcer Module. Based on the *Topology* and the storage records (*Storage* module), the *DecisionMaker* performs decisions of control and management actions. A control decision may pass to help a data plane to choose the best configuration to its service and routing properties, or in a more advanced state, perform self-adjustments and reconfigurates them *on-the-fly*. Based on the overall system performance, the *DecisionMaker* regularly adjusts dynamic-state policies. Regarding to management decisions, this component decides and control background management tasks. For example, upon the failure of the primary controller in a replicated centralized environment, the remaining replicas can implement a leader election algorithm to choose the next primary controller. The *PolicyEnforcer* analyze and validate the feasibility and correctness of low-level policies. Depending on the hardness of a policy, the *PolicyEnforcer* can relax the policy action to provide a similar QoS level of the original policy. In addition, this component is in charge of policy enforcement and policy reconfiguration, accordingly the decisions of the *DecisionMaker* unit. Finally, the *NodeBalancer* component arranges storage nodes according to a specific parameter. This balancing task can be made in background, transparently to the user. This component is particularly helpful in hybrid and decentralized environments. For instance, consider the performance of the underlying storage devices as the balancing parameter. One group comprises the nodes that contain *Solid State Drive* (SSD) disks, as the other comprises the nodes that contain *Hard Disks* (HDD). Since the SSD group provides better performance, it will store the most frequently accessed data (*hot data*), as the HDD group will store the less frequently accessed data (*cold data*) [3, 4].

Storage Module. This module stores policies (*PolicyDB*), control logs (*ControllerDB*) and data plane logs (*DataPlaneDB*) so they can be used to perform control decision and policy validation.

Monitor Module. This module monitors the health of both control and data planes. The *HealthChecker* component periodically collects the health status of all storage components and sends them to the *ControlHandler*. By continuously checking the system status, the controller can adjust and manage the system behavior in order to ensure the stated QoS level. Additionally, writing the policies often requires intimate knowledge of the system profile. Hence, the *Evaluator* component collects a set of performance metrics as well as the storage nodes specifications and capabilities. Thus, the controller can deliberate if a given policy is able to be enforced in the storage infrastructure.

3.2 Data Plane

Contrarily to previous works, we introduce a formal proposal of the control plane architecture and we detail about its internals. Since current SDS solutions have emphasized on the data plane component, we focus on how Utopia can leverage from them and what design assumptions we should follow. Therefore, we resort to the designs of the IOFlow's and the SafeFS's data plane solutions [11, 14]. These solutions follow a queue-based design, offering a programmable data plane that allow for flexible service and routing properties. In addition, these solutions proved to be effective, compatible, flexible and modular. Regarding to their properties, service properties are related to the actions that one or more queues of the data plane must apply to the incoming IO. In detail, service actions can be oriented to the QoS such as ensuring a certain bandwidth value or give higher priority to the requests of a storage component, as presented in the policies P_2 and P_4 in the Table 1. Service actions can also be data-oriented such as data privacy, data granularity (*e.g.,* block, file or object storage), data consistency and storage efficiency (*e.g.,* compression, deduplication), as presented in the policies P_1 and P_3 in the Table 1. As for routing properties, we control where the requests should be routed. Since each data plane queue is associated with a default next-hop, we define which queue must handle the IO request. In order to specify both service and routing properties, data planes should expose a simple control interface that specifies low-level identifiers that can be used to direct requests to queues. This interface should allow the controller to create and remove queuing rules (low-level policies), define the hops of each queue, adjust the enforcement points and also collect logs, health status and performance metrics from all data plane queues. In the future, we plan to expose a simple and generic interface between the controller and the data plane, allowing flexible and modular control to the underlying storage components.

3.3 Control Applications

Similarly to the data plane design section, we focus on which design assumptions Utopia should follow to provide flexible control to the application built on top.

Control applications are the users' entry point of control and communication with the SDS system. As shown in IOFlow, sRoute and Retro, control applications can be used to many different ends such as performance control, malware scanning and latency evaluation [9, 12, 14]. However, this is only possible if the control plane exposes a simple control interface that specifies high-level policies to be used on the SDS system. This interface should allow control applications to create and remove (high-level) policies and also expose control and data plane logs as well the system status, so users can be aware of the system condition and manually control and adjust its behavior. In the future, we plan to expose a simple and generic interface between the controller and the control applications, comprising the described functionalities.

4 Models

In this paper, we propose a generic software-defined storage solution that delivers scalable and flexible control over multiple storage environments. In this section, we present the Utopia architecture in three distinct environments: centralized, hybrid and fully decentralized.

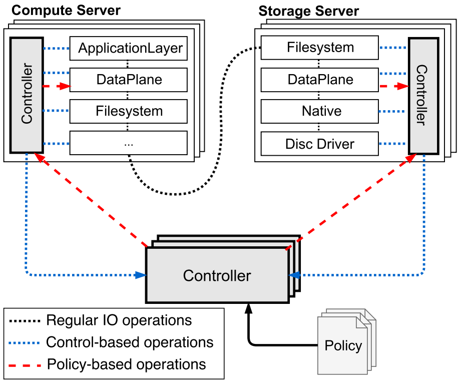


Fig. 2. UTOPIA architecture in a hybrid model. Storage components can either be a Compute Server or a Storage Server.

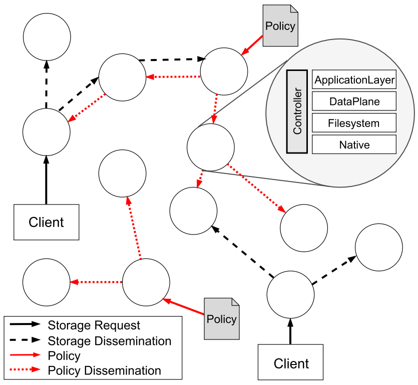


Fig. 3. UTOPIA architecture in a decentralized model. Each node comprises a data plane and a control plane.

4.1 Centralized Model

The centralized approach is the adopted model from current SDS solutions [9, 14, 15]. This model comprises the well-known logically centralized controller and several data planes distributed across the storage components. In this environment, UTOPIA can provide a similar approach to the current SDS solutions. As previously stated, this solution seems to be suitable for small-to-medium storage systems, but lacks scalability when we consider storage systems with higher computational demands.

4.2 Hybrid Model

Figure 2 depicts the architecture of UTOPIA in a hybrid environment. In this model we decoupled the logically centralized controller into two types of controllers. A subset of these controllers, form a central control unit. The remaining controllers are distributed across several storage components. Thus, this model can be seen as a control hierarchy. The central control unit (or major controller) has maximum control over the system (as the previous model). The distributed controllers (or minor controllers) however, only assume part of the control power. The control power is assigned by the control applications. The minor controllers are closer to their data planes and can control and manage minor tasks. For instance, this controller can receive logs and health status, analyze them and make decisions (red dashed lines) over the data plane if the action in question is in its control domain. Otherwise, the decision must be sent to the major controller so it can decide what to do. Moreover, in certain conditions, the minor controller can be seen as a *proxy* between the major controller and the data planes of its storage component. Regarding to management and monitorization tasks (blue dotted lines), the minor controller sends them periodically to the major controller. Additionally, note that the major controller continues to have global visibility. With this model we introduce a hierarchical control partioning that alleviates the major controller load, providing a more scalable solution that might be suitable for medium-to-large storage systems.

4.3 Decentralized Model

To address the incoming computational and storage demands, researchers have been working on very large scale systems (*e.g.,* DataFlasks [10]), adopting a fully decentralized environment, which many consider as the next-generation of storage solutions [2, 17]. Figure 3 depicts the architecture of Utopia in a fully decentralized environment. In this model, since it is not possible to centralize the control plane, we removed the central control unit and distributed the control power to each storage component. In this architecture, each node works for itself and its behavior is independent of other nodes. The same applies for the controller. Control tasks (*e.g.,* monitorization, policy enforcement) are done independently, making them easier to accomplish. However, each node comprise a set of neighbours, making it aware of the surrounding environment (partial view of the system). Therefore, nodes are organized into groups and can decide to store or discard data and policies according to the group they belong. Additionally, these groups can be arranged by a specific parameter (*e.g.,* storage space, hardware capabilities, proximity). Since we are dealing with thousands or even millions of connected devices, end-to-end communication with a major amount of nodes is impossible [10]. Therefore, node communication is handled through epidemic protocols such as gossip, since it can rapidly propagate information among a large collection of nodes using only local information [6]. This ensures that eventually, all nodes will receive the policies that are intended for them. With this model, we introduce a fully decentralized SDS solution, that might attain the high processing and storage demands of the next-generation of storage solutions.

5 Conclusion

This paper presents Utopia, a generic software-defined storage architecture that provides high scalable and flexible control over multiple storage environments. It does so by providing a modular control plane that is model- and resource-agnostic, leveraging the SDS principles to centralized, hybrid and fully decentralized storage environments. We also introduce a novel proposal of the architecture of the control plane and we detail about its internals. Utopia provides a vision of a SDS architecture that we believe to be suitable for the next-generation of storage systems. As future work, we envision to implement this architecture alongside the communication interfaces between the controller and the control applications to further extend the Utopia vision to real scenarios, and between the controller and the data plane so we could integrate our system with current (and novel) data plane solutions.

References

1. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I., Zaharia, M.: A view of cloud computing. Commun. ACM 53, 50–58 (2010)
2. Atzori, L., Iera, A., Morabito, G.: The internet of things: A survey. Information Systems Frontiers 17, 243–259 (2010)
3. Balakrishnan, S., Black, R., Donnelly, A., England, P., Glass, A., Harper, D., Legtchenko, S., Ogus, A., Peterson, E., Rowstron, A.I.: Pelican: A building block for exascale cold data storage. In: OSDI. pp. 351–365 (2014)
4. Eldawy, A., Levandoski, J., Larson, P.Å.: Trekking through siberia: Managing cold data in a memory-optimized database. Proceedings of the VLDB Endowment 7(11), 931–942 (2014)
5. Esch, J.: Software-defined networking: A comprehensive survey. Proceedings of the IEEE 103, 14–76 (2014)
6. Eugster, P.T., Guerraoui, R., Kermarrec, A.M., Massoulié, L.: Epidemic information dissemination in distributed systems. Computer 37, 60–67 (2004)
7. Jain, S., Kumar, A., Mandal, S., Ong, J., Poutievski, L., Singh, A., Venkata, S., Wanderer, J., Zhou, J., Zhu, M., et al.: B4: Experience with a globally-deployed software defined wan. In: ACM SIGCOMM Computer Communication Review. vol. 43, pp. 3–14. ACM (2013)
8. Jararweh, Y., Al-Ayyoub, M., Darabseh, A., Benkhelifa, E., Vouk, M.A., Rindos, A.: Software defined cloud: Survey, system and evaluation. Future Generation Comp. Syst. 58, 56–74 (2016)
9. Mace, J., Bodík, P., Fonseca, R., Musuvathi, M.: Retro: Targeted resource management in multi-tenant distributed systems. In: NSDI (2015)
10. Maia, F., Matos, M., Vilaça, R.M.P., Pereira, J.O., Oliveira, R., Riviere, E.: Dataflasks: Epidemic store for massive scale systems. 2014 IEEE 33rd International Symposium on Reliable Distributed Systems pp. 79–88 (2014)
11. Pontes, R., Burihabwa, D., Maia, F., Paulo, J., Schiavoni, V., Felber, P., Mercier, H., Oliveira, R.: Safefs: a modular architecture for secure user-space file systems: one fuse to rule them all. In: SYSTOR (2017)
12. Stefanovici, I.A., Schroeder, B., O'Shea, G., Thereska, E.: sroute: Treating the storage stack like a network. In: FAST (2016)
13. Tanenbaum, A.S., van Steen, M.: Distributed systems - principles and paradigms, 2nd edition (2007)
14. Thereska, E., Ballani, H., O'Shea, G., Karagiannis, T., Rowstron, A.I.T., Talpey, T., Black, R., Zhu, T.: Ioflow: a software-defined storage architecture. In: SOSP (2013)
15. Tinedo, R.G., Sampé, J., Zamora-Gómez, E., Artigas, M.S., López, P.G., Moatti, Y., Rom, E.: Crystal: Software-defined storage for multi-tenant object stores. In: FAST (2017)
16. Turner, V.: The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things (April 2009, accessed December 2017), `https://www.emc.com/leadership/digital-universe/2014iview/digital-universe-of-opportunities-vernon-turner.htm`
17. Zanella, A., Bui, N., Castellani, A.P., Vangelista, L., Zorzi, M.: Internet of things for smart cities. IEEE Internet of Things Journal 1, 22–32 (2014)

State Synchronization by State Decomposition

Vitor Enes

HASLab/INESC TEC and Universidade do Minho

Abstract. Data consistency often needs to be sacrificed in order to ensure high-availability in large scale distributed systems. *Conflict-free Replicated Data Types* (CRDTs) relax consistency by always allowing query and update operations at the local replica without remote synchronization. Consistency is then re-established by a background mechanism that synchronizes the replicas in the system. In state-based CRDTs replicas synchronize by periodically sending their local state to other replicas and by merging the received remote states with the local state. This synchronization can become very costly and unacceptable as the local state grows. Delta-state-based CRDTs solve this problem by producing smaller messages to be propagated. However, it requires each replica to store additional metadata with the messages not seen by its direct neighbors in the system. This metadata may not be available after a network partition, since a replica can be forced to garbage-collect it (due to storage limitations), or when the set of direct neighbors of a replica changes (due to dynamic memberships).
In this paper we further improve the synchronization of state-based CRDTs, by introducing the concept of join decomposition of a state-based CRDT and explaining how it can be used to reduce the synchronization cost of this variant of CRDTs.

Keywords: Eventual Consistency, CRDTs, Join Decomposition

Large-scale distributed systems resorting to replication techniques often need to sacrifice the consistency of the system in order to attain high-availability [1, 6]. One common approach is to allow replicas of some data type to temporarily diverge, making sure these replicas will eventually converge to the same state in a deterministic way. *Conflict-free Replicated Data Types* (CRDTs) [7, 8] can be used to achieve this.

CRDTs come mainly in two flavors: *operation-based* and *state-based*. In both, queries and updates are always immediate at the local replica, and this is why the system is available (as it never needs to coordinate beforehand with remote replicas to execute operations). In operation-based CRDTs [5, 7], operations are disseminated assuming a reliable dissemination layer, that ensures exactly-once delivery of operations. State-based CRDTs need fewer guarantees from the communication channel: messages can be dropped, duplicated and reordered. When an update operation occurs, the local state is updated through a mutator, and from time to time (since we can disseminate the state at a lower rate than the rate of the updates) the full state is propagated to the other replicas.

Although state-based CRDTs can be disseminated over unreliable communication channels [4], as the state grows, sending the full state can be very costly and become unacceptable. Delta-state-based CRDTs [2, 3] address this issue, by defining δ-mutators that return a delta (δ), typically much smaller than the full state of the replica, to be merged with the local state. The same δ is also added to an outbound δ-buffer, to be periodically propagated to remote replicas. This strategy requires keeping track of which updates (δ-groups) have been effectively received by other replicas of the system with which the local replica exchanges information directly, which leads to the maintenance of additional metadata that may have to be garbage collected (due to storage limitations) or non-existing (due to dynamic memberships). Delta-state-based CRDTs have been adopted in industry as part of Akka Distributed Data framework[1].

Current solutions perform bidirectional full state transmission when a replica joins the system (either for the first time or after a network partition) in order to receive the missed updates and to propagate the ones observed locally. After this initial synchronization, replicas can synchronize with a neighbor replica by sending groups of δs (all the δs that haven't been received by that neighbor), avoiding full state transmission at each synchronization step.

In this paper we present two novel algorithms, *state-driven* and *digest-driven*, used for efficient synchronization of state-based CRDTs when the metadata required for delta-based synchronization is not available. Given two replicas A and B, in the *state-driven* synchronization algorithm A starts by sending its local state to B, and the B replies with the missing updates. Convergence is achieved after two messages, since the replica B can merge the received state into its local state. In the *digest-driven* synchronization, instead of sending its local state, replica A starts by sending a digest of its local state (smaller than the local state) that still allows B to compute the missing updates on A. Besides sending the updates missed by A, B also sends a digest of its own local state, so that A can also compute the updates missed by B.

We then revisit the delta-state-based synchronization algorithm and propose two modifications that reduce the amount of state transmission necessary for synchronization: *avoid back-propagation of δ-groups* and *remove redundant state in received δ-groups.*

These contributions were only possible due to the concept of join decomposition of a state-based CRDT developed in the paper.

References

1. P. Ajoux, N. Bronson, S. Kumar, W. Lloyd, and K. Veeraraghavan. Challenges to Adopting Stronger Consistency at Scale. In *Proceedings of the 15th USENIX Conference on Hot Topics in Operating Systems*, HOTOS'15, pages 13–13, Berkeley, CA, USA, 2015. USENIX Association.
2. P. S. Almeida, A. Shoker, and C. Baquero. Efficient State-Based CRDTs by Delta-Mutation. In *Networked Systems - Third International Conference, NETYS 2015, Agadir, Morocco, May 13-15, 2015, Revised Selected Papers*, pages 62–76, 2015.

[1] https://doc.akka.io/docs/akka/2.4/scala/distributed-data.html

3. P. S. Almeida, A. Shoker, and C. Baquero. Delta State Replicated Data Types. *J. Parallel Distrib. Comput.*, 111:162–173, 2018.
4. P. Bailis and K. Kingsbury. The Network is Reliable. *Commun. ACM*, 57(9):48–55, Sept. 2014.
5. C. Baquero, P. S. Almeida, and A. Shoker. Pure Operation-Based Replicated Data Types. *CoRR*, abs/1710.04469, 2017.
6. H. Lu, K. Veeraraghavan, P. Ajoux, J. Hunt, Y. J. Song, W. Tobagus, S. Kumar, and W. Lloyd. Existential Consistency: Measuring and Understanding Consistency at Facebook. In *Proceedings of the 25th Symposium on Operating Systems Principles*, SOSP '15, pages 295–310, New York, NY, USA, 2015. ACM.
7. M. Shapiro, N. M. Preguiça, C. Baquero, and M. Zawirski. Conflict-Free Replicated Data Types. In *Stabilization, Safety, and Security of Distributed Systems - 13th International Symposium, SSS 2011, Grenoble, France, October 10-12, 2011. Proceedings*, pages 386–400, 2011.
8. M. Shapiro, N. M. Preguiça, C. Baquero, and M. Zawirski. Convergent and Commutative Replicated Data Types. *Bulletin of the EATCS*, 104:67–88, 2011.

PAPERS IN ALPHABETICAL ORDER

A framework to automate network packets construction

A network packets sniffer with real-time graphical analytics

A Study of Novelty Detection in Data Streams Using Different Unsupervised Approaches

Automated Fare Collection Data For Measuring Socio-economic Impacts On The Transport Supply*

Computer-Vision-based surveillance of Intelligent Transportation Systems

Ensemble Weighting for Quantile Regression

Heartbeat Biometrics: A Survey

Improving the efficiency of air traffic controllers in simulated environments

Metamorphic Virus Detection through Data Compression

Multidimensional Byzantine Approximate Agreement in Cyber-Physical Systems with Trust

NER: Supervised System to Recognize Participants and Location in Criminal News

Neurocognitive stimulation game: Serious game with adaptive difficulty for stimulation and assessment

Outlier Identification in Multivariate Time Series: Boilers Case Study

Predicting winning teams for regular season of the NBA

Sentiment analysis techniques applied to regression tasks

State Synchronization by State Decomposition

Student-Centered Learning Environments for Self-Regulated Project-Based Learning in Higher Education: Qualification and Selection Study

The contribution of blockchain technology for business innovation

Topic categorization in Portuguese news articles

Towards an Access Control System for IoT on Blockchain

Tuberculosis Classification and Drug Resistance Detection in Medical Images with Deep Learning

Upper Limbs Movement analysis for Medical classification of Breast Cancer Patients

AUTHORS IN ALPHABETICAL ORDER

André Santos

António Guerra and Hélder P. Oliveira

Arnaldo Pereira

Bruno Tavares

Carla Abreu

Cristiano Alves, Nuno Anacleto and Mário Antunes

Daniel Pinto and Mário Antunes

Joana Ribeiro

José Ornelas

João Costa

João Figueira Silva, Jorge Miguel Silva, Eduardo Pinho and Carlos Costa

João Neto and Rosaldo Rossetti

João Pedro Dias and Hugo Sereno Ferreira

Kemilly Dearo Garcia

Marcelo Santos and Luís Antunes

Marisa Reis

Mohamed Yassine Zarouk and Mohamed Khaldi

Ricardo Macedo

Saulo Carpio

Simão Reis

Tiago Neto

Vitor Enes

Yassine Baghoussi

9 789727 521739 >

DSIE'18

13th Doctoral Symposium
in Informatics Engineering

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

DEI DEPARTAMENTO DE
ENGENHARIA INFORMÁTICA

U.PORTO

aefeup

SUPPORTED BY:

IEEE
University of Porto - Student Branch

TALK<BIT
CONFERENCE

ISBN 978-972-752-230-9