

CENTERIS 2014 - Conference on ENTERprise Information Systems / ProjMAN 2014 -
International Conference on Project MANagement / HCIST 2014 - International Conference on
Health and Social Care Information Systems and Technologies

A clustering approach for predicting readmissions in Intensive Medicine

Rui Veloso^a, Filipe Portela^{a*}, Manuel Filipe Santos^a, Álvaro Silva^b, Fernando Rua^b,
António Abelha^c, José Machado^c

^a*Algoritmi Centre, University of Minho, Guimarães, Portugal*

^b*Serviço Cuidados Intensivos, Centro Hospitalar do Porto, Hospital Santo António, Porto, Portugal*

^c*CCTC, University of Minho, Braga, Portugal*

Abstract

Decision making assumes a critical role in the Intensive Medicine. Data Mining is emerging in the clinical area to provide processes and technologies for transforming data into useful knowledge to support clinical decision makers. Applying clustering techniques to the data available on the patients admitted into Intensive Care Units and knowing which ones correspond to readmissions, it is possible to create meaningful clusters that will represent the base characteristics of readmitted patients. Thus, exploring common characteristics it is possible to prevent discharges that will result into readmissions and then improve the patient outcome and reduce costs. Moreover, readmitted patients present greater difficulty to be recovered. In this work it was followed the Stability and Workload Index for Transfer (SWIFT). A subset of variables from SWIFT was combined with the results from laboratory exams, namely the Lactic Acid and the Leucocytes values, in order to create clusters to identify, in the moment of discharge, patients that probably will be readmitted.

© 2014 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/3.0/>).

Peer-review under responsibility of the Organizing Committee of CENTERIS 2014.

Keywords: Clustering, Data Mining, Intensive Care Units, Data Mining, SWIFT, Readmission, INTCare;

* Corresponding author. Tel.: +351-253510319; fax: 2+351-53510300.

E-mail address: cfp@dsi.uminho.pt.

1. Introduction

Currently, in large hospitals, there are at least twenty information systems in order to facilitate the whole process management and information that are needed by different departments [1]. The users of these systems need to access medical information in order to have the best clinical decisions and provide a cost-effective treatment to patients.

The large amount of data generated in the hospital and, more specifically, in intensive care units are quite complex to be analyzed using former methods. The application of Data Mining (DM) in these types of data provides processes and technologies for transforming such data into useful knowledge for clinical decision makers, resulting in a gain in the treatment of patients [2].

This work was based in a previous study about predicting readmission in Intensive Care Units (ICU) [3]. The results attained so far did not allow for an efficient characterization of the patients. This article aims to complement the previous work using clustering techniques to find common characteristics in ICU patients that can lead to a future unplanned readmission. All the work was done in the context of INTCare research project in the Centro Hospitalar do Porto, Porto, Portugal.

This article is structured in terms of the following chapters: Introduction – this chapter - presents the environment of the work developed. Background – the second chapter - defines the problem and the theoretical foundations of the work (Intensive Medicine, Readmissions, Clustering, INTCare and the Stability and Index for Load Transfer). The next chapter describes the study carried out (includes Methods and Tools, Business and Data Understanding, Data Preparation, Modeling and Evaluation). Chapter 4 discusses and interprets the results obtained in the work. Conclusions of the work and recommendations for future work in this subject close the article.

2. Background

2.1. Intensive Medicine

Intensive Medicine (IM) aims to diagnose and treat patients with serious illnesses and restore them to their previous state of health [4]. These ill patients are usually admitted to intensive care units (ICU) so that they can maintain their physiological functions through the various life-support devices. In these units the vital functions of patients are continuously monitored as well as the status of each of the organic systems: neurological, respiratory, hepatic, hematological, cardiovascular and renal. In order to ensure the life and condition of the patient these functions can be supported through drugs or by mechanical means until the patient has again its functions independently [5].

2.2. Readmissions

Bad decisions taken by the intensivist at the time of discharge is directly related to an unplanned readmission for the patient. Actually, the ability to predict a relapse in the patient after the discharge is limited [6]. It is considered a readmission when a patient is admitted to the same unit where he/she was earlier within thirty days after the discharge with the same diagnosis [7]. The number of readmitted patients in ICU are significant, having according to the literature review, in North America and Europe, an average rate around 7% [8].

2.3. Clustering

Cluster analysis divides the data into groups that have sense, are helpful or the booth. The groups are the objective and the clusters should capture the natural structure of the data. Unlike other techniques, the classification criteria are not defined by the analyst but are discovered along the process of clustering. The clusters are characterized by a great internal homogeneity and external heterogeneity [9].

The cluster analysis represent an important role in many areas like psychology and social sciences, biology, statistics, pattern recognition, information recovery, machine learning and data mining [10].

There are a large number of cluster algorithms and the choice of the methods to be adopted depends on the type of the data as well on the purpose and intended application. The majority of the clustering methods are included into five categories.

The Partition Methods build a set of partitions on the data, where each partition represents a cluster. The hierarchical methods execute a hierarchical decomposition of the data. These methods can be agglomerative or divisive. The agglomerative methods start with singular objects to form an isolated group. Then, successively the groups or objects are merged until only one group left. Divisive methods behave the other way. The density-based methods are useful to filter outliers or to discover with arbitrary form. Grid-based methods restrict the space of objects to a finite number of cells that form a grid structure. The Model-based methods formulate a model hypothesis for each cluster and find the best fit of the data to the model [11].

The evaluation of clustering results can be done laying on two factors: the compactness and separability. The compactness is a property that expresses how much the cluster elements are close. Lesser the variance value greater will be the compactness of the cluster. The calculation of the intra cluster distance is very useful to assess this characteristic. The separability evaluates how diverse the clusters are. This can be assessed by the inter cluster distance that will be the greater possible so the clusters are better [12]. To this work it was used the partition methods (k-means, k-medoids, x-means).

2.4. INTCare

This study is being developed under the research project called INTCare. INTCare is an Intelligent Decision Support System (IDSS) in constant development and testing. It is deployed in the ICU of the CHP. This IDSS is based on intelligent agents [13] and aims to support the decision making process and to predict clinical events as is patient organ failure (cardiovascular, respiratory, renal, hepatic, neurological and hematologic), patient outcome [14], readmissions, medical diseases and others. Regarding to the predictions made, the system is able to suggest procedures, treatments and therapies. This system is based in four autonomous subsystems (data acquisition, knowledge management, inference and interface) that use intelligent agents to perform actions [15, 16].

2.5. SWIFT

Several models or mathematical techniques can help predicting readmission probability of patients in ICU. A study was conducted to develop and validate a numerical index called Stability and Workload Index for Transfer (SWIFT) [6]. There are some variables that can be used to estimate the probability of unplanned ICU readmissions like the patient length of stay (LOS) in the ICU, measured in days, the source of patients admission, Glasgow Coma Scale (GCS), the ratio between partial pressure of oxygen in arterial blood (PaO₂) and fraction of inspired oxygen (FIO₂) and the evaluation of nursing care for respiratory problems (PCO₂). These variables are then scored taking into account the information available from the time of hospital discharge. Table 1 presents the SWIFT Variables and the scores to be assigned.

Table 1. SWIFT Variables.

Variables	Score
Original Source of ICU Admission	
Emergency Department	0
Transfer from a ward or outside hospital	8
Total ICU Length of Stay (in days)	
Lesser than 2	0
Between 2 and 10	1
Bigger than 10	14
Last measured PaO₂/FIO₂ ratio	
Bigger than 400	0
Lesser than 400 and bigger or equal to 150	5
Lesser than 150 and bigger or equal to 100	10
Lesser than 100	13

Glasgow Coma Scale at the time of ICU	
Greater than 14	0
Between 11-14	6
Between 8-10	14
Lesser then 8	24
Last arterial blood gas PaCO2	
Lesser than 45 mm Hg	0
Bigger than 45 mm Hg	5

2.6. Related Work

Included in the project INTCare a study has been made to understand ICU readmission phenomena. SWIFT method was used to create classification models to predict if a patient will be readmitted or not. The results were very satisfactory, obtaining 98.91% of accuracy. However, these results only have been possible due to the use of oversampling. More information about the study done and the results achieved using data mining models to predict patient's readmission in ICU is available in [3]. In order to make a deeper data exploration and improve the previous results, the current study will explore clusters to characterize possible readmitted patients.

3. Study Description

3.1. Methods and Tools

This study purposes the identification of readmitted and non-readmitted patients in an ICU using a set of clustering scenarios. For this work were used Oracle SQL Developer for data analysis, understanding and preparation and RapidMiner to build clustering scenarios. A benchmark analysis has been carried out comparing the following algorithms: k-means, k-means with kernels, k-means fast, k-medoids, x-means, expectation maximization clustering, top down clustering, DBSCAN, support vector clustering, random clustering and flatten clustering. From these techniques the ones that suited better (in terms of statistical and domain criteria) have been k-means, k-medoids and x-means, therefore considered in this work.

3.2. Business Understanding and Data Understanding

The main goal of this study is to decrease the number of readmissions through the use of Data Mining models. The Data Mining objective encompasses a characterization of groups of patients that are in risk of being readmitted in ICU. These models will support clinical decisions as well improve the quality of service. This study used real data acquired from the CHP databases. Data were collected from the patient clinical process and laboratory results. The data used is from April 23th, 2010 to February 10th, 2014 and corresponds to 1043 cases (patients).. The number of readmission cases verified throughout this period is 36 (about 3.5% episodes). 13 variables have been considered: age, sex, length of stay in the ICU (in days), emergency room (indicates if the patient came from the emergency room or if the patient was admitted from other hospital), PaO2/FIO2 (the ratio between the partial pressures of oxygen in blood and fraction of inspired oxygen), PaCO2 (partial pressures of carbon dioxide in blood), the scores relative to emergency room, length of stay, PaO2/FIO2 ratio, PaCO2 and the laboratory results of the quantity of lactic acid and leucocytes. The results from laboratory analysis considered the most recent results according to the date of patients discharge. Table 2 presents some statistics for each variable.

Table 2. Variables Considered.

Variable	Distinct Values	Average	Minimum Value	Maximum Value
Age	79	63.65	17.00	96.00
Sex	2 (1 or 2)	-	-	-
Length of Stay	38	6.39	0.00	62.00
Emergency Room	2 (True or False)	-	-	-
PaO2/FIO2 Ratio	754	287.83	37.80	6100
PaCO2	325	40.92	16.80	116.40
Lactic Acid	220	2.06	0.20	16.00
Leucocytes	766	36.79	0.00	8.77
PaO2/FIO2 Ratio Score	4 (0; 5; 10 or 13)	-	-	-
PaCO2 Score	2 (0 or 5)	-	-	-
Emergency Room Score	2 (0 or 8)	-	-	-
Length of Stay Score	3 (0; 1 or 14)	-	-	-
Readmission	2 (Yes or No)	-	-	-

3.3. Data Preparation

This process is iterative and was performed as many times as necessary to ensure the data quality. New data have been built from the existing one used in the first study [3]. Four derived attributes were created in order to meet SWIFT model. The birthing date was derived the age. The LOS was calculated by the difference between the date of discharge and the date of admission. The number of process, admission and discharge dates allow for the creation of a new attribute, called readmission, indicating if a patient corresponds to an unplanned readmission or not. Every patient admitted before 30 days from the last discharge are considered readmissions.

Using PaO2 and FIO2 it was possible to calculate the ratio PaO2/FIO2. Once this work is dealing with readmissions, were considered the results of PaO2, FIO2, Leucocytes and Lactic acid closest to the discharge date.

3.4. Modeling

The phase of modeling is focused on getting models to translate business goals through the application of data mining techniques. The modelling was done using Rapid Miner Studio 6. Rapid Miner is an integrated environment for data mining, machine learning, text mining, predictive and business analysis. The scenarios developed consider four main groups:

- Target Class = Readmission;
- Normal (N) = {emergencyroom, length_of_stay, PCO2, PaO2_FIO2_ratio};
- Scores (S) = {PCO2_Score, PaO2_FIO2_ratio_score, emergencyroom_score, length_of_stay_score};
- Case Mix (CM) = {sex, age};
- Lab Results (LR) = {lactic_acid, leucocytes}.

Considering these attributes the scenarios presented in table 3 were encoded.

Table 3.Scenarios.

Scenarios	Used Variables	Scenarios	Used Variables
S1	Normal	S8	Normal + Scores
S2	Normal + Case Mix	S9	Normal + Lab Results
S3	Normal + Case Mix + Scores	S10	Normal + Case Mix
S4	Normal + Case Mix + Scores + Lab Results	S11	Scores + Case Mix
S5	Scores	S12	Scores + Lab Results
S6	Case Mix	S13	Case Mix + Lab Results
S7	Lab Results		

These 13 scenarios originated 91 models, but only 39 models have been analyzed in this study – those representing points of interest from the clinical data results. To each one of the models is associated a scenario, a target and a clustering technique. As referred above were k-means, k-medoids and x-means demonstrated the best results. Table 4 permits to see the settings defined for these three algorithms.

Table 4.Algorithms Settings.

Algorithm	Setting	Value
k-means	K	2 to 11
	Max Runs	10
	Max Optimization Steps	100
	Measures Type	Numerical Measures
	Numerical Measure	Euclidean Distance
k-medoids	K	2 to 11
	Max Runs	10
	Max Optimization Steps	100
	Measure Types	Numerical Measures
	Numerical Measure	Euclidean Distance
x-means	K Min	2
	K Max	60
	Measure Types	Numerical Measures
	Numerical Measure	Euclidean Distance
	Clustering Algorithm	KMeans
	Max Runs	10
	Max Optimization Steps	100

X-Means algorithm calculates an optimal number of clusters (k) for the running but k-Means and k-medoids don't. For those techniques Davies-Bouldin Index was used to find the most correct number of clusters (the more lower the value is a better separation of the clusters and tightness inside clusters occur) and evaluated the Elbow method by observing the variations of the average within cluster distance (observe how the average within cluster distance varies from k to k and select the k where the natural progression of the measure dominates de structure). Table 5 demonstrates the optimal number of clusters considered for each model taking into account the Davies-Bouldin Index and the Elbow Method (returning the average within cluster distance for the elbow).

Table 5. Optimum number of clusters for k-means and k-medoids.

Model	Algorithm	Number of Clusters	Davies-Bouldin Index	Average within in cluster distance
M1	k-means	4	0.563	519.619
	k-medoids	3	0.733	43317.973
M2	k-means	4	0.466	9284.935
	k-medoids	3	0.768	43629.478
M3	k-means	3	0.468	9332.818
	k-medoids	3	0.775	43696.924
M4	k-means	5	0.541	9087.479
	k-medoids	3	0.827	46243.621
M5	k-means	6	0.528	8.857
	k-medoids	3	0.722	30.028
M6	k-means	7	0.513	8.530
	k-medoids	3	0.593	63.581
M7	k-means	2	0.388	519.652
	k-medoids	9	1.496	703.440
M8	k-means	4	0.454	9086.529
	k-medoids	3	0.740	43385.713
M9	k-means	6	0.503	6749.155
	k-medoids	5	1.297	35345.738
M10	k-means	4	0.466	9284.935
	k-medoids	3	0.768	43629.478
M11	k-means	3	0.882	89.845
	k-medoids	2	0.971	194.594
M12	k-means	2	0.393	571.905
	k-medoids	2	3.028	2222.068
M13	k-means	2	0.402	773.690
	k-medoids	5	3.098	2552.022

3.5. Evaluation

The evaluation phase was focused primarily on the assessment of the results provided by the use of k-means, x-means and k-medoids and then the results were compared against the initial goals of the project. Analyzing the results obtained considering the Davies-Bouldin index, is evident that the technique that achieves the worst results is k-medoids. Although the difference between the results obtained with k-means and x-means is not expressive, k-means got the best results. These results were expected since the x-means is based on k-means. Some models obtained interesting values for Davies-Bouldin index, unfortunately they don't achieve the lower limit imposed in this domain. Table 6 represents the three models that best suit both the domain and project goals and simultaneously present the best indexes.

Table 6. Models with best results.

Model	Algorithm + Number of Clusters	Davies-Bouldin Index	Clusters	Number of Readmission Cases
M9 (N+LR)	k-means with 6 clusters	0,503	C0	2
			C1 and C4	0
			C2	1
			C3	16
			C5	16
M4 (N+CM+S+LR)	k-means with 5 clusters	0,541	C0	16
			C1	2
			C2	1
			C3	0
			C4	16
M1 (N)	k-means with 6 clusters	0,563	C0	9
			C1	2
			C2, C4 and C5	0
			C3	24

Analyzing in more detail the model 9 it is possible figure out a great number of clusters of readmitted patients (C4 and C6 containing 16 cases).

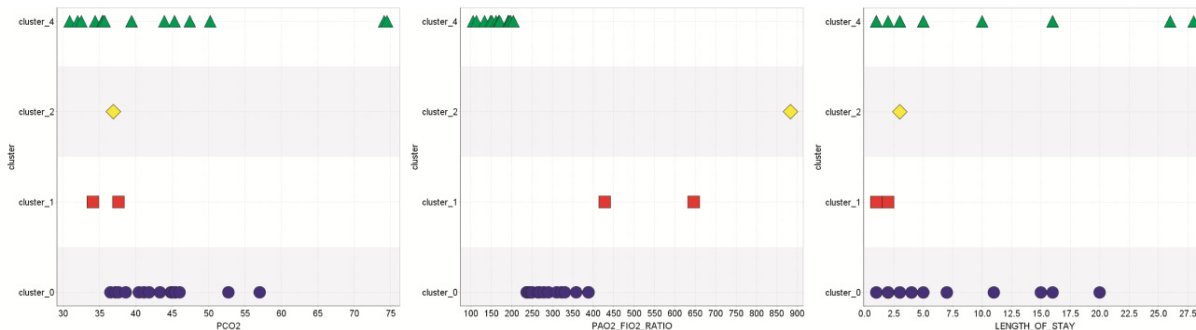


Fig. 1. Distribution of readmission cases by clusters in PaCO2, PaO2/FIO2 Ratio and Length of Stay

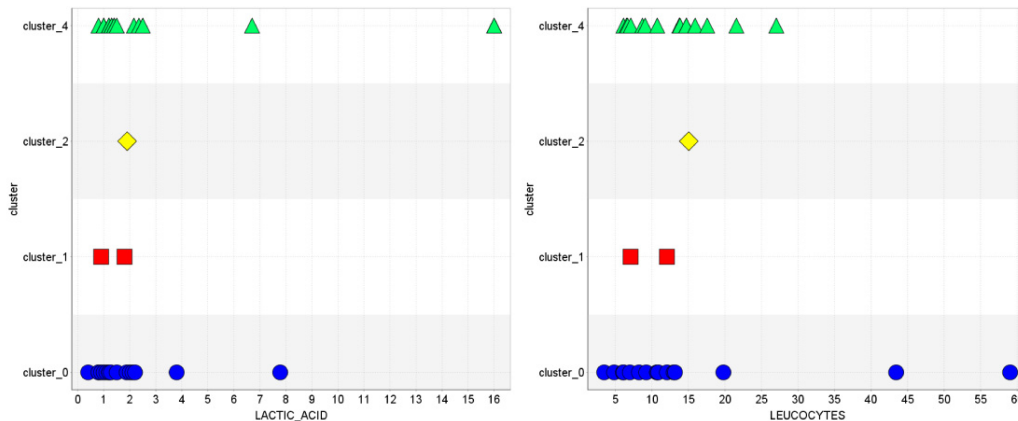


Fig. 2. Distribution of readmission cases by clusters in Lactic Acid and Leucocytes

Exploring the results obtained by model 9 (N+LR), and observing Fig. 1 and Fig. 2 we can identify the features that are influencing the clustering: PaO₂/FIO₂ ratio and the PaCO₂. PaO₂/FIO₂ in cluster 3 varies from 105 to 204 and in cluster 5 from 236 to 388. PaCO₂ varies from 31 to 75 in cluster 3 and from 36.5 to 57 in cluster 5. The length of stay do not presents a significant variation difference between the two main clusters since. For Lactic Acid cluster 3 varies from 0.8 to 16 and in the cluster 5 from 0.4 to 7.7. Leucocytes varies from 6 to 27 in cluster 3 and from 3 to 59 in the cluster 5. All the patients from cluster 3 came from a ward outside of the hospital and the cluster 5 contains patients that came from the other wards or from the emergency room of the hospital.

In model 4 (N+CM+S+LR) the results from PaCO₂, PaO₂/FIO₂ ratio, Leucocytes, Lactic Acid, Emergency Room and Length of Stay are the same with the difference that cluster 3 is now represented by cluster 4 and cluster 5 is represented by cluster 0. The remaining attributes, age, sex and the four scores are not preponderant on the result of the clusters since there are well distributed among the different clusters.

In model 1 (N), once more, there are two clusters containing the majority of the cases of readmission: cluster 3 and cluster 0. The most discriminative variable is PaO₂/FIO₂ Ratio, in cluster 3 varies from 105 to 268 and in cluster 0 from 275 to 430. The other variables present results with minor differences between them. It is possible, observing the other models, to conclude that the Scores and the Case Mix do not affect the clusters and the characterization of a readmitted patient.

4. Discussion

Based on the models 4 and 9 two global clusters can be created to better represent the characteristics of future readmitted patients. Table 7 presents these groups and the correspondent attributes and values.

Table 7. Relevant information about groups of patient readmitted.

Variable	Cluster 1	Cluster 2
Emergency Room	Patient's that came from an ward outside the hospital	Patient's that came from an ward outside the hospital and from the emergency room of the hospital
PaO ₂ /FIO ₂ Ratio	105 to 204	236 to 388
PaCO ₂	31 to 75	36,5 to 57
Length of Stay	1 to 28 days	1 to 20 days
Lactic Acid	0,8 to 16	0,4 to 7,7
Leucocytes	6 to 27	3 to 59
Age	24 to 88	50 to 93

The Scores and the Sex don't have impact on the groups. The Scores because they are a direct characterization of the Normal group values using SWIFT and sex because the original data is well balanced between man and woman.

5. Conclusions and Future Work

This work provided useful results which help to characterize the type of patients having a higher probability to be readmitted. The clusters developed cannot ensure which patients will be readmitted but they give information about which type of patients the physicians should consider in terms of clinical situation before to perform their discharge. For example, the Davies-Bouldin index tends to $+\infty$, however to a cluster be selected the value should be the nearest possible of 0, in the most cases the index is lower than 1 which represents a very good result to the clusters developed.

The attributes more related to readmissions are PaO₂/FIO₂ Ratio, PaCO₂, the Emergency Room and Length of Stay, as expected, because they are the major attributes of SWIFT. The results of analysis of lactic acid and leucocytes (suggested by the council of clinical experts) contributed for the segmentation of readmitted groups. In the case of the variables from the Case Mix group, only the age contributed for characterizing groups of readmitted patients.

Future work will include a more detailed study of the characteristics of not readmitted patients. Complementarily, explore the impact of the leucocytes and lactates influence in readmission of patients in the ICU (as suggested by clinical experts).

Acknowledgements

This work has been supported by FCT – Fundação para a Ciência e Tecnologia in the scope of the project: PEst-OE/EEI/UI0319/2014. The authors would like to thank FCT for the financial support through the contract PTDC/EEI-SII/1302/2012 (INTCare II).

References

- [1] Z. Chenhui, D. Huilong, and L. Xudong, "An integration approach of healthcare information system," 2008, pp. 606-609.
- [2] H. C. Koh and G. Tan, "Data mining applications in healthcare," *Journal of Healthcare Information Management—Vol*, vol. 19, p. 65, 2011.
- [3] Pedro Braga, F. Portela, and M. F. Santos, "Data Mining Models to Predict Patient's Readmission in Intensive Care Units," in *ICAART - International Conference on Agents and Artificial Intelligence*, Angers, France, 2014.
- [4] Á. Silva, P. Cortez, M. F. Santos, L. Gomes, and J. Neves, "Rating organ failure via adverse events using data mining in the intensive care unit," *Artificial Intelligence in Medicine*, vol. 43, pp. 179-193, 2008.
- [5] J. Ramon, D. Fierens, F. Güiza, G. Meyfroidt, H. Blockeel, M. Bruynooghe, et al., "Mining data from intensive care patients," *Advanced Engineering Informatics*, vol. 21, pp. 243-256, 2007.
- [6] O. Gajic, M. Malinchoc, T. B. Comfere, M. R. Harris, A. Achouiti, M. Yilmaz, et al., "The Stability and Workload Index for Transfer score predicts unplanned intensive care unit patient readmission: Initial development and validation*," *Critical care medicine*, vol. 36, pp. 676-682, 2008.
- [7] ACSS, "Administração Central do Sistema de Saúde, Circular Normativa nº 33/2012 ", ed, 2012.
- [8] A. L. Rosenberg and C. Watts, "Patients Readmitted to ICUs*," *Chest*, vol. 118, p. 492, 2000.
- [9] S. Tufféry, *Data mining and statistics for decision making*: John Wiley & Sons, 2011.
- [10] P.-N. Tan, Steinbach, M., & Kumar, V, *Introduction to Data Mining* 1ed.: Addison-Wesley Longman Publishing Co., Inc., 2005.
- [11] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*: Morgan kaufmann, 2006.
- [12] K. J. Cios, *Data mining: a knowledge discovery approach*: Springer, 2007.
- [13] M. F. Santos, F. Portela, M. Vilas-Boas, J. Machado, A. Abelha, and J. Neves, "INTCARE - Multi-agent approach for real-time Intelligent Decision Support in Intensive Medicine," in *3rd International Conference on Agents and Artificial Intelligence (ICAART)*, Rome, Italy, 2011.
- [14] F. Portela, M. F. Santos, J. Machado, A. Abelha, and Á. Silva, "Pervasive and Intelligent Decision Support in Critical Health Care Using Ensembles," in *Information Technology in Bio-and Medical Informatics*, ed: Springer Berlin Heidelberg, 2013, pp. 1-16.
- [15] F. Portela, M. F. Santos, Á. Silva, J. Machado, A. Abelha, and F. Rua, "Data mining for real-time intelligent decision support system in intensive care medicine," 2013.
- [16] Filipe Portela, Filipe Pinto, and M. F. Santos, "Data Mining Predictive Models For Pervasive Intelligent Decision Support In Intensive Care Medicine," presented at the *KMIS 2012 - International Conference on Knowledge Management and Information Sharing*, Barcelona, 2012.