Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 577–584

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Healthcare Interoperability and Pervasive Intelligent Systems (HiPIS 2017)

"Predicting Resurgery in Intensive Care - A data Mining Approach"

Ricardo Peixoto[a], Lisete Ribeiro[a], Filipe Portela*[a], Manuel Filipe Santos[a] and Fernando Rua[b]

[a]*Algoritmi Research Center, University of Minho, Guimarães, Portugal*
[b]*Intensive Care Unit, Centro Hospitalar do Porto, Portugal*

Abstract

Every day the surgical interventions are associated with medicine, and the area of critical care medicine is no exception. The goal of this work is to assist health professionals in predicting these interventions. Thus, when the Data Mining techniques are well applied it is possible, with the help of medical knowledge, to predict whether a particular patient should or not should be re-operated upon the same problem. In this study, some aspects, such as heart disease and age, and some data classes were built to improve the models created. In addition, several scenarios were created, with the objective can predict the resurgery patients. According the primary objective, the resurgery patients' prediction, the metric used was the sensitivity, obtaining an approximate result of 90%.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Data Mining, Classification, Interventions, Reinterventions, INTCare.

1. Introduction

This study aims to use classification approaches in order to predict the patients who are resurgeried together with the medical knowledge in a view to helping the health professionals. It is expected that with the development of this study, it is possible to improve the performance of Intensive Care Units (ICUs) and assist their healthcare professionals in making decisions about their patients. The dataset used in this project was provided by Hospital Santo António in Porto, however, to improve the quality of the results, these have been modified. The strategies used was standardization of data to create the models, but without changing the accuracy of the results. The standardization of data is a set of

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: cfp@dsi.uminho.pt

1877-0509
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.08.291

rules that aims to reduce data redundancy and increase data integrity. Reinterventions have only recently been identified as a problem, which means that there is not a very extensive work in this area. Sometimes, health professionals have not a standard to understand if a particular patient should or should not be resurgery, so this project aims to help the reader understand that situation.

The objective of this study is to identify the health problems and the characteristics of resurgery patients in order to prevent them from being again intervened. This work was conducted by following the CRISP-DM methodology.

This work is divided into five sections. The first section is the introduction where the main ideas of this work are presented, the second is the background where the problem is defined as well as the theory behind the work, the third section is the description of the study methods, where the tools used are described. In the fourth chapter of this paper is the discussion where some views on the results are presented. Finally, the last section presents some conclusions and basic ideas about the work to be done in the future.

2. Background

This section aims allow collecting information about the work that has already been developed in practical and research terms. In practical terms, the works presented are similar to the one developed in this project. In terms of research, the concepts that are linked to this project were presented. First, important concepts for this work are presented, such as intensive care, INTCare, and Surgery and resurgery. After this, the data mining technique used are explained. Finally, the works that exist related to this project are presented.

2.1. Intensive Care and INTCare

In 90's experts realized that the knowledge gained so far was not enough to solve the complex problems that appeared in real life [1]. After the 90's, a change was made, having gained more interest because of the vast complexity of the data collected [2]. Intensive care is defined as a multidisciplinary field of medical science that deals with the diagnosis, prevention, and treatment of potentially reversible disease conditions in patients with imminent failure of vital functions [3]. In intensive care, risk forecast has always had an important role, being one of the areas of medicine with greater severity and which deserve more attention and research [3]. According to Silva [3], intensive medicine increased the ability to save lives at risk. For this, it is necessary to make a correct diagnosis and develop treatment plans to improve the conditions of patients and save their lives [4]. Information Systems recently provided a lot of information for intensive care units that are well crafted, allowing health professionals in this area to make the best decisions for their patients. These information systems allow clinicians perceived the complexity of disease, improve monitoring of patients and increase the size of the resource. New technologies and progressive computerization units of intensive care have played a significant role AT this point [3]

The INTCare aims to develop an intelligent system with the ability to develop clinical events [2]. The INTCare is a system that was designed for use in ICUs, supporting the decisions of doctors, using Data Mining [1].

The system can be accessed anywhere and anytime to obtain information such as the patient data, monitoring of clinical events, doctors scores and the organ failure probability. This system was developed based on an automated process paradigm and knowledge discovery and agents. The agents are social computing entities that are predefined by system creators whose activity contributes to the goal of the global system [1]. The main characteristics of these agents are that they have intelligent behavior, accuracy, robustness, flexibility, and efficiency [1]. These agents have important roles as the input of clinical data, pre-processing, data mining, performance, model initialization, data recovery, forecasting, assessing the scene and interface [1]. An essential requirement in achieving good results in knowledge discovery is the quality of data [5].

2.2. Surgery and resurgery

Due to the evolution of science, surgical interventions came to be seen as a treatment. With the development of new techniques and the growth of knowledge, surgery methods have become crucial driven through the advances in anesthesia, antisepsis, radiology, blood transfusion as well as the use of the bone or the prosthesis [6]. However, in the

past, surgical interventions were seen only as a last resort for some diseases. In 2004, in a survey conducted in 56 countries, the World Health Organization (WHO) stated that there were, on average, one intervention every five people per year [7]. Surgical intervention wasn't routinely considered part of the traditional model of public health since it is regarded as the last alternative to curing a patient. Surgical interventions are always present in a significant proportion of the population, especially in developing countries, where conventional treatment sometimes is not available and where there is a huge reserve of diseases not treated surgically [8]. Currently, it is not possible to obtain any statistics for interventions in intensive care once this area is still under investigation. However, it is possible to conclude that there are countries where surgical operations are normal and others where there aren't ethical conduct in the interventions. A resurgery happens when a patient needs to be again operated to something that has already been before.

2.3. Classification

According to Freitas 9, "The classification task can be considered an ill-defined, in deterministic function, which is inevitable in that involve forecasting." The classification is designed to identify the class of determined record. In this task, the model analyses the set of provided records. Each record already contains an indication of what class it belongs and it is able to 'learn' how to classify a new record [10]. This technique is used to predict values of the categorical type variables, is that the model created aims to classify which category a record belongs [10]. This task aims to recognize, together with the data, the observations that have the same characteristics. The goal is to predictr the class of an item from the database. If a data record contains the Region field, then some of the typical values of the field, such as North, South, East, West, can define the class [11]. In the classification task, the most common techniques are decision trees, support vector machines, neural networks, classifiers Bayes and genetic algorithms [12]. Turban, Sharda and Delen [12] argue that the most important factors in the evaluation of a classification model are the accuracy of prediction, speed, robustness, scalability, and interpretability.

2.4. Related Work

As mentioned before, the reintervention only has been recently recognized as a problem, which means that there is little work in this area. In practical terms, some data mining work has been developed in the area of intensive care, to predict interventions. However, in the case of reintervention, nothing has yet been developed, which means that this is a new project in cases where the patient was resurgeried and requires further intervention.

Some work that has already been developed in the ICUs are presented in the references [13, 14, 15].

The first study aims to use data mining techniques to multi-prediction of organ failure. This is based on patient's characteristics (such as age, Admission Type, Admission Origin, Diagnostic and Gravity Indexes) as well as Clinical Adverse Events occurred during internment in the UCI. It's presented a framework that allows the management and characterization of the group of models generated [13]. In the second study, a model was developed to support the decision of the doctors providing the level of sepsis and the best treatment for patients with microbiological problems. For this, it used the prediction task, the classification model, the supervised learning method and the algorithms: Decision Trees, Support Vector Machines, and the Naïve Bayes classifier. The results have shown that, in general, there is a weak correlation between the level of sepsis and the therapeutic plan, considering the drugs group [14].

The last study aims to predict readmissions in intensive care units, enabling health professionals to make decisions for more efficient planning. In this study, was used Naïve Bayes technique. The models obtained allow health professionals to know the probability of a patient being readmitted in a UCI [15].

3. Description of the study

This section aims to understand the methodologies and tools used in the execution of this project. Subsequently, the results obtained in each of the phases of CRISP-DM are presented and explained.

3.1. Methods and tools

Data mining is a complex process that requires many different tools, and people. Which leads to the success of a project of this area depends on the right combination of good tools and expert analysts (people).

For a work of this complexity, it is necessary to follow guidelines and methodologies to ensure that the entire process is performed correctly. That way, the practice methodology to be followed is the CRISP-DM. The methodology Design Science Research has an important role in assisting the research project since together with the CRISP-DM Methodology, provided an essential guide in developing the project. The design phase enables the development of artifacts to solve problems. According to Wirth and Hipp [16], the CRISP-DM methodology provides an overview of the life cycle of a data mining project. A Data Mining project can be divided into six tasks: Business understanding (understand the purpose of the project through a business perspective); Data understanding (collection and analysis of data and identification of problems therein); Data preparation (construction of the final data taking into account the initial data); Modelling (selection and application of various data modeling techniques); Evaluation (construction of a model that should achieve the objectives initially defined.) and Implementation (development of a model that the client can use). This methodology provides users with a structured approach to project planning and can serve as a common reference point in the use of Data Mining [16]. In the developing phase, the tool used for data exploration, preparation and creation of scenarios and collation of data was Oracle SQL Developer.

The scenarios used were designed according to the results obtained previously. Thus, together with the medical knowledge, assist health professionals in decision making.

3.2. Business Understand

This work presents the reintervention prediction of patients in intensive care. The idea of this project is to provide information for professionals in order to understand what the characteristics of patients who need to be resurgeried.

3.3. Data Understanding and Data preparation

Once the dataset used was provided by a hospital, the data presented showed several errors and inconsistencies. In developing this project, several changes were necessary at the end dataset in order to standardize the data and to enable the best possible information regarding the models created. First, the data errors and null values were deleted. There are errors in the data when they are badly written or they present a value that does not make sense, considering the attribute. Classes were defined for some attributes, as shown in Table 1 and 2.

A data change was made in attribute "MainDiagnosis" by creating a "SurgeryCategory" attribute, where the value is assigned as a surgery that this patient has been exposed that is, each operation is related to a medical field.

To facilitate interpretation was established, in a simple form, a category of surgery which allows realizing what type of surgery the patient was subject. This, through the principal patients' diagnosis. To understand which category of each surgery, an intensive study of these surgeries and medical areas was carried out, as well as a study by health professionals. These studies helped to dispel some doubts and certainties about certain areas. Then all types of surgery created by describing the surgery that the patient was submitted can be: General surgery; Maxillofacial surgery; Vascular surgery; Dermatology; Neurology; Orthopedics; Otorhinolaryngology; pneumology; Urology; Diagnosis unspecified. Once the attribute "MainDiagnosis" is a VARCHAR, a patient with the same diagnosis can be described differently. With this change, does not only reduce the number of different names, such as there is a standard in the information, i.e., it allows all patients with the same disease, even if written differently, to be recognized. This attribute was used in an initial phase of the project when other data mining techniques were applied, however, with the use of oversampling, the patients with reinterv = 0 did not possess this information, the number of NULL cases was high, and the results were not as expected, so this variable was not used in the final forecasts. In general, this process was run for sufficient attributes. Oversampling was used only in the best scenarios. When these attributes have values that can be more easily grouped by “classes” we, in fact, proceeded to this change but always ensuring that the accuracy of the models was not affected.

Table 1 illustrates the classes created with the days of hospitalization of each patient, calculated through the current date and the day they were hospitalized. The same table presents the similar process that was used for the patients' age, where in both cases, classes were created consisting in 5 in 5 values, where the first was built from 0 to 5, the second 6 to 10 and so on. Each class has many associated records. The percentage of each created class of records is presented in Table 1. The creation of these categories facilitates the creation and understanding of the models since it does not alter their accuracy, but simplifies the existing information. One of the attributes that were important in this dataset and was not present was the age of patients. Therefore, there was the need to go through the date of birth and the current date, to calculate the age of each patient. After calculated the age of each patient, was then created the attribute "CLASS OF AGES" where, through the age of the patient estimated previously, classes were created, in order to improve the data and thus facilitate the reading of the data in the creation of various scenarios. This class was pooled with the registration 5 to 5 ages, as can be seen in Table 2 as well as the percentage of records of each class.

Table 1. Class of hospitalization

Days	Percentage	Days	Percentage
0-5	72.06%	41-45	0.09%
6-10	14.44%	46-50	0.30%
11-15	6.41%	51-55	0.09%
16-20	3.12%	56-60	0%
21-25	2.05%	61-65	0.04%
26-30	0.64%	66-70	0%
31-35	0.56%	71-75	0%
36-40	0.17%	76-80	4.27%

Table 2. Class of ages

Ages	Percentage	Ages	Percentage
16-20	0.17%	61-65	21.82%
21-25	1.03%	66-70	11.92%
26-30	1.54%	71-75	9.31%
31-35	2.09%	76-80	12.39%
36-40	3.76%	81-85	8.12%
41-45	6.36%	86-88	4.96%
46-50	9.43%	91-95	2.18%
51-55	12.90%	96-100	1.03%
56-60	9.61%		

In addition to the creation of the classes mentioned above, others have been developed with the same goal.
In the next Table 3 shown, it's possible to see the remaining changes to the data.

Table 3. Changed attributes

Original Value	Changed Value	Original Value	Changed Value
Physical disability	Physical_or_Psychic_Disability	Operating Room	Note Situation
Psychic disability		Ward	
Cardiac Insufficiency	Card_Hepa_Cron_or_Resp_Insuf	Emergency Room	
Chronic Hepatic Insufficiency		Another Situation	
Chronic Renal Insufficiency		Intermedia Unit	
Chronic Respiratory Insufficiency		Urgency	
Blind	Blind_or_Deaf_or_Dumb		
Deaf			
Dumb			

In addition to the data listed above, then the illustrated Table 4 shows other attributes present in the preparation of scenarios and their respective possible values.

Table 4. Other attributes

Variable	Distinct Value	Variable	Distinct Value
Alcoholism	2 (0 or 1)	Radiotherapy	2 (0 or 1)
Diabetes Insulin-treated	2 (0 or 1)	Sex	2 (0 or 1)
Diabetes Non-Insulin treated	2 (0 or 1)	Drug addict IV	2 (0 or 1)

DPOC	2 (0 or 1)	Grafted	2 (0 or 1)
HTA in Treatment	2 (0 or 1)	By air	2 (0 or 1)
Other immunosuppressant	2 (0 or 1)	Reinterv	2 (0 or 1)
Pacemaker	2 (0 or 1)	Long-term corticosteroid therapy	2 (0 or 1)
Chemotherapy	2 (0 or 1)		

Once all the mentioned changes were made, it was possible to use a final DataSet with all the existing information organised with a total of 22 different attributes and 4682 patients.

3.4. Modeling

This layer was developed with the aim of being able to translate business goals by classification techniques. The modelling was performed using the Oracle SQL Developer that allows having an integrated environment for data mining. In order to achieve the expected results were created 17 scenarios that allow physicians of the ICUs to understand the characteristics of resurgeried patients. The first scenario was created using all the attributes of the final dataset. The remaining sets were created using some selection criteria, such as physically or mentally handicapped patients or patients with heart, hepatic, chronic or respiratory insufficiency. In total, 136 models were generated. These models can be represented by:

DMM1 = {17 Scenarios, 4 techniques, 2 Sampling Method, 1 Representation Method, 1 Target}

The scenarios score was measured by combining multiple attributes. With the combination of attributes has been revealed that some attributes had more influence than others. In the Table 5 are presents the 17 scenarios and the different attributes. In the columns are presented the 22 attributes of the final Data set. In the lines are presented the 7 attributes, where each X being an attribute of this scenario.

Table 5 - Multiple scenarios

Scenario	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
S1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
S2	X															X	X	X				X		
S3												X				X	X	X						
S4																X	X	X		X	X	X		
S5																X			X		X	X		
S6																X	X	X		X	X	X		
S7											X	X				X	X	X	X	X	X	X		
S8												X				X	X	X	X					
S9		X												X		X	X				X			
S10														X		X		X	X					
S11		X														X	X	X						
S12		X														X		X	X					
S13												X				X	X						X	
S14														X		X			X		X			
S15									X	X						X			X					
S16	X				X		X	X	X	X				X		X	X		X					
S17						X								X			X	X						X

A – Alcoholism
B – Long-term corticosteroid therapy
C – Diabetes Insulin Treated
D – Diabetes Non-Insulin Treated
E – DPOC
F – HTA In Treatment
G – Other immunosuppressants
H – Pacemaker
I – Chemotherapy
J – Radiotherapy
M – Drug addict IV
N – Grafted
O – By air
P – Reinterv
Q – Class of Days of Hospitalization
R – Class of Ages
S – Note situation
T – Insuf_Card_Hepa_Cron_Ou_Resp
U – Physical_or_Psychic_disability
V – Blind_or_Deaf_or_Dumb

K – Sex L – Type of admission	W – ObservationStat X – SOFA

3.5. Evaluation

All scenarios were studied in detail in order to understand which models had the best results. In order to define which the best scenarios are, some calculations were performed according to the classification technique.

TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives
ACC = Accuracy
Sens = Specificity
Espec = Specificity

$$(TP+TN)/(TP+TN+FP+FN) = ACC. \quad (1)$$
$$TP/(TP+FN) = Sens. \quad (2)$$
$$TN/(TN+FP) = Espec. \quad (3)$$
$$TP/(TP+FP) = Precision\ 0. \quad (4)$$
$$TN/(FN+TN) = Precision\ 1. \quad (5)$$

The values shown in Table 6 are the best one for each scenario and metric.

Table 6. Best Scenarios

Scenario	ACC	Sens	Espec	Precision 0	Precision 1
5	34.69%	89.03%	20.89%	22.22%	88.24%
14	34.00%	89.69%	20.90%	21.06%	89.59%
15	35.03%	88.57%	20.85%	22.85%	87.59%

4. Discussion

After the assessment of all scenarios, it is possible to see that some of the models hit the forecast of resurgeried patients in about 90%. These models with high probability have attributes that can be seen as significant in predicting resurgery. The knowledge that these models translate into along with the knowledge and experience of health professionals can be crucial to the ICUs. Regarding the scenarios, scenarios 5, 14 and 15 have satisfactory results in terms of sensitivity, however, the accuracy and specificity not have very significant results, which leads to the conclusion that the data provided are only good at predicting the resurgeried patients.

As can be seen in Table 7 if the clinical goal is predicting resurgery patients, the oversampling technique should not be used. The models created after using oversampling are more suitable not predict not resurgery patients. The values of accuracy and specificity are higher.

Table 7. Best scenarios, with and without oversampling

	Without oversampling			With oversampling		
Scenario	ACC	Sens	Espec	ACC	Sens	Espec
5	34.69%	89.03%	20.89%	54.05%	66.98%	51.37%
14	34.00%	89.69%	20.90%	54.10%	68.49%	51.38%
15	35.03%	88.57%	20.85%	53.94%	65.96%	51.31%

The created models can't guarantee by 100 percent that a patient will be resurgeried, however, these models combined with the medical knowledge will indeed allow to make that prediction.

5. Conclusions and Future Work

Finalized this study, it can be observed that the size information of each patient resurgery is quite high. As such, this work allows conclusions that through human observation would be quite difficult to obtain. The created models cannot guarantee that resurgery patients are the patients with their characteristics presents in the built models. However, by adding this information to the useful knowledge, can be vital in the treatment of the patient because this

helps the professional health in their decision making. By analyzing the work done, it is possible seeing that the only practical work done to date is the first part of this project, whose goal is to characterize the resurgery patients in intensive care. This work is important in clinical terms since it can help the physicians have new knowledge and experiences. This fact is essential for the health of patients in ICUs. In scientific terms, since resurgery only recently were considered a problem, this study allows realizing that through a given dataset the patient's illness can be predicted.

Acknowledgments

"This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013." This work is also supported by the Deus ex Machina (DEM): Symbiotic technology for societal efficiency gains - NORTE-01-0145-FEDER-000026

References

1. P. Gago, M.F. Santos, Á. Silva, P. Cortez, J. Neves, L. Gomes, INTCare: a Knowledge Discovery Based Intelligent Decision Support System for Intensive Care Medicine, (2006).
2. F. Portela, J. Machado, A. Abelha, Álvaro Silva, F. Rua, M.F. Santos, Pervasive and Intelligent Decision Supportin Intensive Medicine – The Complete Picture, (2014) DOI: 10.1007/978-3-319-10265-8_9.
3. Á.J.B.M. da Silva, Modelos de Inteligência Artificial na análise da monitorização de eventosclínicos adversos, Disfunção/Falência de órgãos e prognóstico do doente critico., (2007).
4. C.W. Hanson, B.E. Marshall, Artificial intelligence applications in the intensive care unit, 29 (2001) 427–435.
5. F. Portela, M. Vilas-Boas, M.F. Santos, Improvements in Data Quality for Decision Support in Intensive Care, in: M. Szomszor, P. Kostkova (Eds.), Electron. Healthc. Third Int. Conf. eHealth 2010, Casablanca, Morocco, December 13-15, 2010, Revis. Sel. Pap., Springer Berlin Heidelberg, Berlin, Heidelberg, 2012: pp. 86–94. doi:10.1007/978-3-642-23635-8_11.
6.. E.C.S. Maziero, Avaliação da implantação do programa de cirurgia segura num hospital de ensino, (2012).
7. T.G. Weiser, S.E. Regenbogen, K.D. Thompson, A.B. Haynes, S.R. Lipsitz, W.R. Berry, A.A. Gawande, An estimation of the global volume of surgery: a modelling strategy based on available data., Lancet. 372 (2008) 139–44. doi:10.1016/S0140-6736(08)60878-8.
8. H.T. Debas, R. Gosselin, C. Mccord, Chapter 67 Surgery, (2004).
9. A.A. Freitas, Understanding the Crucial Differences Between Classification and Discovery of Association Rules – A Position Paper, 2 (2000) 65–69.
10. C. Camilo, J. Silva, Mineração de Dados: Conceitos , Tarefas , Métodos e Ferramentas, (2009).
11. V. Devedzic, Knowledge Discovery and Data Mining in Databases, (2001) 1–24.
12. E. Turban, R. Sharda, D. Delen, Decision Suport and Business Intelligence.pdf, (2010).
13. J.J.R. Pereira, Modelos de Data Mining para multi-previsão : aplicação à medicina intensiva, (2005).
14. João M. C. Gonçalves, Filipe Portela, Manuel F. Santos, Álvaro Silva, José Machado, António Abelha, Fernando Rua. Real-time Predictive Analytics for Sepsis Level and Therapeutic Plans in Intensive Care Medicine. IJHISI - International Journal of Healthcare Information Systems and Informatics. Volume 9, Issue 3, pp 36-54. ISSN: 1555-3396. IGI Global. (2014) DOI: 10.4018/ijhisi.2014070103.
15. P. Braga, F. Portela, M.F. Santos, F. Rua, Data mining models to predict patient's readmission in intensive care units, 6th Int. Conf. Agents Artif. Intell. ICAART 2014. 1 (2014) 604–610..
16. R. Wirth, J. Hipp, CRISP-DM: Towards a Standard Process Model for Data Mining, (2000).