



CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2017, 8-10 November 2017, Barcelona, Spain

Knowledge Discovery from Surgical Waiting lists

Cristiana Neto^a, Hugo Peixoto^b, Vasco Abelha^b, António Abelha^b and José Machado^{b*}

^aUniversity of Minho, Campus Gualtar, Braga 4710, Portugal

^bAlgoritmi Research Center, University of Minho, Campus Gualtar, Braga 4710, Portugal

Abstract

Methods for knowledge discovery in data bases (KDD) have been studied for more than a decade. New methods are required owing to the size and complexity of data collections in administration, business and science. They include procedures for data query and extraction, for data cleaning, data analysis, and methods of knowledge representation. The part of KDD dealing with the analysis of the data has been termed data mining. Data mining is the process of pattern discovery and extraction where huge amount of data is involved. Data mining has been used intensively and extensively by many organizations. In healthcare, data mining is becoming increasingly popular, if not increasingly essential. Data mining applications can greatly benefit all parties involved in the healthcare industry. In this work is presented an approach for the use of data mining in the context of waiting lists for surgery, namely for predicting the type of surgery (programmed or additional) for a record in the list.

© 2017 The Authors. Published by Elsevier B.V.

Peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies.

Keywords: Knowledge Discovery in Databases, Data mining, Surgery waiting list, Decision Support Systems, Classification

* Corresponding author. Tel.: +351 253 604 430; fax: +351 253 604 471.

E-mail address: jmac@di.uminho.pt

1. Introduction

Among the huge and varied number of fields that currently exist, data are being collected and stored at a tremendous rate¹. More and more organizations place the needs that information provides on other entities but often these organizations, such as healthcare institutions, find themselves with a huge amount of data but without truly valuable information².

So, in order to be able to assist humans in extracting useful information (knowledge) from the growing amount of digital data, there is an urgent need for a new generation of computational theories and new tools. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD)³. The KDD is the term used to refer a multistep process of discovering useful knowledge from data and one particular step in this process is the application of data mining algorithms for extracting patterns (models) from data. Recently, data mining has invoked significant interest in its applications to healthcare.

The healthcare environment is usually information rich but knowledge poor. However, data mining techniques can be applied to create a knowledge rich healthcare environment. The specific part of healthcare that this work focuses on is the analysis of information about waiting lists for surgery in a hospital.

Waiting lists have a strong influence on access to health care and may mean that there is an inability to satisfy basic human health needs. The waiting lists increased in Portugal in the last decades and one of the main complaints of patients is the high waiting times for surgery, which can often be years. It's important to emphasize that waiting lists for surgery tend to be more pronounced in countries that combine health insurance⁴.

In 2004, an information system was implemented (SIGIC – Integrated System for the Management of Enrolled Patients for Surgery) within Portuguese health care institutions and led to a better planning and programming of the institution's activity, reducing waiting times through management.

One of the measures taken with the implementation of the SIGIC in order to promote the recovery of surgical waiting lists was the creation of additional production. So nowadays, two types of surgery can be distinguished: programmed/base and additional. The type of surgery that a patient is receiving is very important in terms of hospital management, especially with regard to its financial management.

Due to this importance, it would be interesting to predict the type of surgery of a record in the waiting list given certain attributes. This prediction can be done using data mining algorithms (classification or regression). In our case, as the objective is to predict the type of surgery, the most adequate algorithms are the classification ones. For this process, several tools can be used and the chosen was the Weka tool. This paper presents a contextualization on the subject of knowledge extraction, particularly the data mining technique and the results obtained, as well as its discussion.

Regarding the structure of this paper, it is divided in six major sections. After the current introduction, the second section presents related work and the background concepts of the given problem which is waiting lists. The methodologies, data and methods used are then described in next section. Section four describes the steps of the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology that was followed during the DM process. In section five the obtained results are discussed. Finally, section six includes conclusions and future work.

2. Background and Related Work

As defined by Fayyad, Piatetsky-Shapiro, and Smyth⁵, “*Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.*” A KDD is a multistep process that automatically examines data to determine the types of relationships, using a variety of techniques from statistical, artificial-intelligence, or machine-learning algorithms⁵. The quality of health services can be improved using clinical systems with a built-in decision support system, where KDD is very relevant, reducing error and costs.

Portela et.al.⁶ presented a way how INTCare, an Intelligent Decision Support Systems (IDSS) developed in the intensive care unit of the *Centro Hospitalar do Porto*, could accommodate new functionalities. This paper presented an approach to the KDD procedure in order to enable a pervasive, online and real-time processing of data in ICU,

bringing improvements in the information availability and consequently a more proactive attitude by the doctors is facilitated.

Although data mining and KDD are often treated as equivalent, in essence, data mining is an important step in the KDD process. Data mining is the process of using artificial intelligence techniques and statistical and mathematical functions to extract knowledge from a dataset stored in a database⁷. Data mining methods can be divided into two categories: supervised and unsupervised⁸. The supervised methods are used to predict a value and require the specification of a target attribute and unsupervised methods are applied to discover the intrinsic structure, patterns, or affinities between the data.

Braga et.al. present an approach to improve the decision making process using induced data mining models to predict if a patient would need to take a vasopressor¹⁰. This paper focus in the decision making process in intensive care units (ICU), where there's a need of quick and accurate decisions. The data mining model were induced using data from vital sign monitors, laboratory analysis and information about the patient's Electronic Health Record. These models can reduce the need of vasopressor drugs by helping intensivists to act and take accurate decision before the vasopressor be need by the patient¹⁰.

Portela et.al. also presented an approach to improve the decision making in ICU¹¹. In this paper, the authors used data mining techniques to predict the probability of a patient be discharged in the next hour, inducing models with a very good sensitivity. The results they achieved also allow for predicting the bed occupancy rate in ICU for the next hour. The work done provides new knowledge in real time, which represents an improvement in the decision making¹¹.

2.1. *Waiting Lists*

Waiting lists have a strong influence on access to health care and may mean that there is an inability to satisfy basic human health needs. This problem raises both efficiency and equity concerns⁴. It's considered impossible the absence of waiting lists, what can be done is reduce waiting lists' average duration to minimum values¹².

The increase in the waiting lists for surgery suffered in Portugal in the last decades, can be explained by the aging of the population, the introduction of new technologies, leading to an increased demand for surgical interventions^{13,14}. Over the last few years, information systems (IS) have been implemented with the aim of combating waiting lists for surgery and they intend to help in patients' management. SIGIC – Integrated System for the Management of Enrolled Patients for Surgery was implemented within Portuguese health care institutions in 2004 and led to a better planning and programming of the institution's activity, reducing waiting times through management^{15,16}.

Nowadays, two types of surgery can be distinguished: programmed/base and additional. The “base production” is the production contracted at the beginning of the year. “Additional production” reeferes the production that exceeds the base production contracted with the hospitals of the SNS as well as that effected by the entities contracted under the SIGIC¹⁶. Additional production is a way of promoting the recovery of surgical waiting lists¹⁷. So, the type of surgery that a patient is receiving is very important in terms of hospital management, since the EPE (Public Business Entities) hospitals contract with the ARS (Regional Health Administration) their production, stipulating, among others, targets for the base production and for the additional production. At the end of each year (or contractual period) it is verified which part of the production was carried out beyond the contracted base value. Failure to perform the additional production represents a penalty for the hospital and the realization of additional production beyond established only favors the hospital up to a certain value¹⁷.

3. **Methodology, Material and Methods**

In this section, it is presented the data, tools and techniques that were used and the methodology used during the elaboration of this work.

3.1. Methodology

During the Data Mining Process, the CRISP-DM Methodology was followed. This is a six phase hierarchical process, divided in the following steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment¹⁸. As seen in other studies this methodology increases the success of application of the DM process and allows the implementation of the DM models in real environment¹⁹.

3.2. Materials

In a first phase of the project, it was necessary to choose a dataset from a large set of data stored in a data warehouse that was in the SQL format. This data was previously extracted, transformed and loaded (ETL process) using mainly the Oracle SQL Developer tool. The data used in this project is a subset of some clinical data extracted from a hospital in the north of Portugal. This data contained information about the waiting list for surgery, such as information about the patient, information about the surgery itself (type of surgery, service where it was performed, etc.) as well as waiting time from the time the surgery is scheduled to perform. Thus, the final dataset contained 7 attributes presented further in this work and 21526 cases, which is a large amount of data. Each case represented an entry in the surgery waiting list.

3.3. Methods

In this section, it is presented the data mining techniques were used to induce de Data Mining models, namely: Logistic Regression (LR); Naive Bayes (NB); Decision Tree (DT) and k-Nearest Neighbors.

LR is a binary classification technique. This technique, assumes that the input variables are numeric and have a Gaussian (bell curve) distribution. The algorithms learn a coefficient for each input value, which are linearly combined into a regression function and transformed using a logistic (s-shaped) function. This technique is a fast and simple technique, but can be very effective on some problems²⁰. The NB is a classifier based on conditional probabilities. It uses formula that calculates a probability by counting the frequency of values and combinations of values in the historical data, named Bayes' Theorem. Given the probability of an event that has already occurred, the Bayes' Theorem finds the probability of another event to occur. So, NB calculates the posterior probability for each class and makes a prediction for the class with the highest probability²⁰. The DT technique, like NB, is based on conditional probabilities. They work by creating a tree to evaluate an instance of data, start at the root of the tree and moving to the leaves (roots) until a prediction can be made. The process of creating a decision tree works by greedily selecting the best split point in order to make predictions and repeating the process until the tree is a fixed depth. After the tree is constructed, it is pruned in order to improve the model's ability to generalize to new data. DT can support classification and regression problems²⁰. The k-nearest neighbor's technique, also called kNN for short, also supports both classification and regression. Its operation is based on storing the entire training dataset and querying it when making a prediction to locate the k most similar training patterns. As such, there is no model other than the raw training dataset and the only computation performed is the querying of the training dataset when a prediction is requested²⁰.

4. Data Mining Process

4.1. Business understanding

The goal of this work is to predict the type of surgery that will be performed considering several given attributes of the entry in the waiting list. The final results should be very accurate regarding the sensitivity of the context. In addition, predicting the type of surgery that will have to be performed will improve the management of the healthcare institution in question, more specifically financial and resource management, since the existence of additional surgeries means higher remuneration for physicians and greater need for resources.

4.2. Data understanding

Table 1 presents the ranges of maximum and minimum values that they take in the dataset.

Table 1. Dataset attributes with respective minimum and maximum values.

Attribute	Min	Max
Service	1	8
Surgical Intervention	43	8689
Diagnosis	32	99678
Sex	1	2
Age	0	99
Waiting Time	0	397
Type of Surgery	1	2

The last attribute presented is the type of surgery and represents what we want to predict. This attribute can take two values, namely 1 for base/programmed surgery and 2 for additional surgery. The initial distribution of cases for each type of surgery is shown in Figure 1.

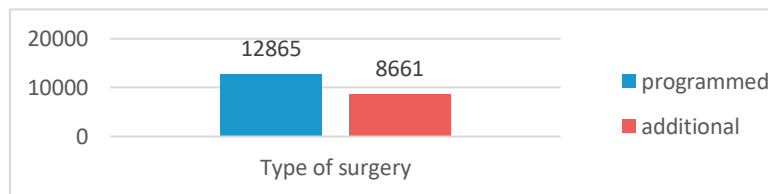


Figure 1 - Distribution of cases by type of surgery where blue stands for base/programmed and red stands for additional.

4.3. Data preparation

After selecting the dataset, it was necessary to carry out a normalization process, that is, to transform all data into values from 0 to 1. This normalization was performed using the equation (1):

$$\text{normalized value} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

Where $X_{\max}$ and $X_{\min}$ correspond to the maximum and minimum that the values of this attribute (X) take, presented previously. For example, the type of surgical intervention had values between 1 and 8, so $X_{\min}=1$ e $X_{\max}=8$.

4.4. Modelling

This phase consisted of inducing the Data Mining Models (DMM) in Weka using the prepared data. Since Classification was the chosen Approach (A), there were 4 different DM techniques (DMT) that were used: LR, NB, DT and KNN. These techniques were implemented by the use of the following classification algorithms: functions.Logistic, bayes.NaiveBayes, lazy.IBk and trees.J48. These algorithms were used with the standard configurations. For each DM technique, only one sampling method (SM) was tested. Cross Validation, using 10 folds and where all data is used for testing. In addition, there were two data approaches (DA) tested: Without Under or Oversampling; With Undersampling or With Oversampling. The target variable was the Type of surgery, and the considered scenarios, which consist of different combinations of attributes, were the following:

- S1: {All attributes}
- S2: {All except Sex}
- S3: {All except Waiting Time}
- S4: {All except Sex and Waiting Time}

These different scenarios were composed in order to evaluate which attributes were the most relevant to predict the need for performing neonatal resuscitation. The first scenario considers all attributes; the second scenario removes the patient gender. The third scenario maintains the patient gender and removes the waiting time. Finally, the fourth scenario removes both patient gender and waiting time. Each DMM can be described as belonging to an approach (A), being composed by a scenario (S), a data mining technique (DMT), a sampling method (SM), a data approach (DA) and a target (T), in total, 48 models were induced:

$$\text{DMM} = \{A, S, \text{DMT}, \text{SM}, \text{DA}, T\}$$

The final DMM induced were:

$$\text{DMM} = \{1 \text{ Approach}, 4 \text{ Scenarios}, 4 \text{ DM Techniques}, 1 \text{ Sampling Method}, 3 \text{ Data Approaches}, 1 \text{ Target}\}$$

4.5. Evaluation

All four types of algorithms presented theoretically previously were tested and the comparison between the accuracy (Correctly Classified Instances) of these results is in Table 2. In this phase it was used the Data Approach without under or oversampling and only the first scenario (S1).

Table 2. Comparison of the results between different types of algorithms for S1 and without under or oversampling.

Algorithm	Scenario	Data Approach	Correctly classified instances
functions.Logistic	S1	Without under or oversampling	87.58%
naive.NaiveBayes	S1	Without under or oversampling	91.52%
trees.J48	S1	Without under or oversampling	92.89%
lazy.IBk	S1	Without under or oversampling	88.65%

Observing the results obtained initially with the application of the four selected algorithms (Table 2) we can conclude that the algorithm with better performance was trees.J48. Although the result obtained with J48 was good, the next approach, was trying to improve that result even more.

First, it was verified which attributes were more and least important, for that, the "Select attributes" tab of the Weka tool was used. Several "Attribute Evaluator" options were tested and the majority indicated that the Gender attribute to be the least important. The attribute Waiting Time was also indicated as one of the least important. This discovery led to the inclusion of scenarios 2, 3 and 4 (S2, S3 and S4). Therefore, the same algorithms were tested again but now without these two attributes, first without sex (S2), then without waiting time (S3) and then without both (S4). It was observed that the results didn't improve in all this scenarios, in contrary, they got worst as can be seen in Table 3, where is presented the results obtained removing the sex attribute (S2).

Table 3. Comparison of the results between different types of algorithms for S2 and without under or oversampling.

Algorithm	Scenario	Data Approach	Correctly classified instances
functions.Logistic	S2	Without under or oversampling	87.33%
naive.NaiveBayes	S2	Without under or oversampling	91.52%
trees.J48	S2	Without under or oversampling	92.87%
lazy.IBk	S2	Without under or oversampling	89.45%

As the attribute removal technique did not improve the results, other DA were tested. Changing the dataset used to build the predictive model to have more balanced data was another hypothesis. Undersampling and Oversampling was used and tested. In this case the under-represented class is type 1 of surgery, i.e. additional surgery. Thus, in the oversampling technique the number of cases of additional surgery was increased to a number closer to the number of cases of programmed surgery (from 8661 to 12468, maintaining the 12865 cases of programmed surgery – 25333 cases total). The results improved to 93.33%. In Table 4 are presented the results for the S1 using over sampling.

Table 4. Comparison of the results between different types of algorithms for S1 and with oversampling.

Algorithm	Scenario	Data Approach	Correctly classified instances
functions.Logistic	S1	With oversampling	90.51%
naive.NaiveBayes	S1	With oversampling	91.77%
trees.J48	S1	With oversampling	93.33%
lazy.IBk	S1	With oversampling	91.65%

5. Discussion

After careful analyzing the achieved results, the algorithm that fits the given data the best is trees.J48. This is achieved mainly because of the great amount of data available and because of the accurate result the J48 algorithm can produce. The DMM achieved 93.33% of accuracy.

One of the attempts made to increase the accuracy of the results was to delete some attributes that were considered to be the least important. The removal of attributes lead to a simpler model (more transparent and easier to interpret) and to faster model induction. But in terms of accuracy, this removal may not be advantageous. In our case, the removal of sex and waiting time attributes led to a decrease in the accuracy, what can be explain by the fact that being the least important attributes doesn't mean they are expendable. So, we can conclude that sex and waiting time are related to the type of surgery and are essential to its prediction.

The next attempt was to use under and over sampling. Using oversampling improved the results since it corrected for a bias in the original dataset. In our case, oversampling led to a balanced dataset, since programmed surgery had a lot more cases than additional, what improved the accuracy. The reason because oversampling was better than undersampling is that with over sampling we keep all the information in the training dataset. With undersampling a lot of information is dropped even if this dropped information belongs to the majority class, it is useful information for a modeling algorithm. Despite all these modifications that were made to improve the results the improvement was not very remarkable, increasing only 0,43%. Thus, it was a whole set of factors that contributed to the success of the results obtained almost from the start, among them: the normalization of data, the absence of empty fields in the data, the elimination of irrelevant attributes from the initial dataset, among others.

6. Conclusion and Future Work

Healthcare data mining provides countless possibilities for hidden pattern investigation from the large datasets that health organizations present. These patterns can be used by physicians to determine diagnoses, prognoses and treatments for patients in healthcare organizations. In this study case, the data mining was used to predict the type of surgery in the waiting lists for surgery in Portugal (programmed or additional) of each record in the dataset. This prediction can be very useful in terms of managing the health organization in question because the type of surgery involves mainly monetary issues since additional surgeries are paid from the usual table of operations. The main incentive for additional surgeries is that even having to pay half the additional value to the professionals of the teams who work outside working hours is more advantageous than sending the patients to the private and having to pay for the entire operation there. Thus, a prediction of the type of surgery can help predict costs in order to improve the planning of surgical activity in hospitals.

To perform this process several Data Mining Models were tested and despite all attempts to improve the result only over-sampling increased slightly the results. The best result was achieved using all the attributes (S1), Sample Method of cross-validation, Data Approach with over sampling and trees.J48 algorithm. The DMM reached 93,33% of accuracy. This good initial results may indicate potential use in a real environment and applicability to a future Decision Support System. In conclusion, this work fulfilled all the objectives outlined, mainly raising our knowledge about the data mining process, its algorithms and its application in the health area. In the future bringing new attributes and collecting new data may help explore the full potential of this kind of system and its application in the Healthcare domain.

Acknowledgments

This work has been supported by Compete: POCI-01-0145-FEDER-007043 and FCT within the Project Scope UID/CEC/00319/2013.

References

1. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3).
2. DeGruy, K. B. (2000). Healthcare applications of knowledge discovery in databases. *Journal of Healthcare Information Management*, 14(2), 59-69. PubMed ID: 11066649
3. Fayyad, U., & Uthurusamy, R. (1996). Data mining and knowledge discovery in databases. *Communications of the ACM*, 39(11), 24-26. doi:10.1145/240455.240463
4. Barros, P. P. (2008). As listas de espera para intervenção cirúrgica em Portugal. *Iprisverbis*.
5. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11), 27-34. doi:10.1145/240455.240464
6. Portela, F., Gago, P., Santos, M. F., Silva, A., Rua, F., Machado, J., Neves, J. (2011). Knowledge Discovery for Pervasive And Real-Time Intelligent Knowledge Discovery for Pervasive and Real-time Intelligent Decision Support in Intensive Care Medicine. *Proceedings of KMIS 2011 – International Conference on Knowledge Management and Information Sharing*, Paris, France.
7. Oliveira, S., Portela, F., Santos, M. F., Machado, J., Abelha, A., Silva, Á., & Rua, F. (2016). Clustering Data Mining models to identify patterns in weaning patient failures. *International Journal Of Biology And Biomedical Engineering*, 10.
8. Han, Jiawei & Kamber, Micheline. (2006). *Data Mining: Concepts and Techniques*. Illinois: Morgan Kaufmann, Second edition.
9. Hand, D.J. (1981), *Discrimination and Classification*, New York: John Wiley & Sons.
10. Braga, A., Portela, F., Santos, M. F., Machado, J., Abelha, A., Silva, A. & Rua, F.. (2016). Data Mining to predict the use of Vasopressors in Intensive Medicine Patients. *Jurnal Teknologi*. Volume 78, No 6-7, pp 1-6. ISSN: 0127-9696. Penerbit UTM Press.
11. Portela, F., Veloso, R., Oliveira, S., Santos, M. F., Abelha, A., Machado, J., Silva, A. & Rua, F. (2015). Predict hourly patient discharge probability in Intensive Care Units using Data Mining. *Indian Journal of Science and Technology*. Volume 8, Issue 32. ISSN: 0974-6846. Indian Society for Educat.
12. Siciliani, L., & Hurst, J. (2005). Tackling excessive waiting times for elective surgery: a comparative analysis of policies in 12 OECD countries. *Health Policy*, 72(2), 201-215.
13. Fernandes, A., Perelman, J., & Mateus, C. (2010). *Health and Healthcare in Portugal: Does Gender Matter?* Instituto Nacional Ricardo Jorge, Lisboa, Portugal.
14. OPSS Observatório Português dos Sistemas de Saúde (2003). *Saúde: que rupturas? Relatório de Primavera*, 1 – 124. Lisboa, Portugal.
15. Oliveira, O. (2012). *Extração de Conhecimento nas Listas de Espera para Consulta e Cirurgia*. Dissertação de Mestrado, Universidade do Minho, Braga, Portugal.
16. Portaria n.º 852/2009 de 7 de Agosto, *Diário da República*, 1.ª série — N.º 152.
17. Portugal, Tribunal de Contas. (2013). *Auditoria às remunerações mais elevadas pagas pelas unidades hospitalares que integram o Serviço Nacional de Saúde (Vol. 1), Relatório n.º 12/2013 – 2.ª Secção, Processo n.º 31/2011 - AUDIT*.
18. Silva E., Cardoso, L., Portela, F., Abelha, A., Santos, M. F. and Machado, J. (2015) Predicting nosocomial infection by using data mining technologies. In *New Contributions in Information Systems and Technologies*, pages 189–198. Springer.
19. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wir, R.th. (200). *Crisp-dm 1.0 step-by-step data mining guide*.
20. Brownlee, J. (2016, July 22). *How To Use Classification Machine Learning Algorithms in Weka*. Retrieved May 31, 2017, from <http://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/>