Predictive Models for Hospital Bed Management using Data Mining Techniques

Sérgio Oliveira[1], Filipe Portela[1], Manuel F. Santos[1], José Machado[2], António Abelha[2]
[1]Algoritmi Centre, University of Minho, Guimarães, Portugal
sergiomdcoliveira@gmail.com; {cfp, mfs}@dsi.uminho.pt
[2] CCTC, University of Minho, Braga, Portugal
{jmac, abelha}@di.uminho.pt

Abstract. It is clear that the failures found in hospital management are usually related to the lack of information and insufficient resources management. The use of Data Mining (DM) can contribute to overcome these limitations in order to identify relevant data on patient's management and providing important information for managers to support their decisions.

Throughout this study, were induced DM models capable to make predictions in a real environment using real data. For this, was adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. Three distinct techniques were considered: Decision Trees (DT), Naïve Bayes (NB) and Support Vector Machine (SVM) to perform classification tasks. With this work it was explored and assessed the possibility to predict the number of patient discharges using only the number and the respective date. The models developed are able to predict the number of patient discharges per week with acuity values ranging from ≈82.69% to ≈94.23%. The use of this models can contribute to improve the hospital bed management because having the discharges number it is possible forecasting the beds available for the following weeks in a determinated service.

Keywords: Hospital Management, Management of Patients, Management of Beds and Data Mining.

1 Introduction

Organizations collect data from multiple business processes. Based on the idea that large volumes of data can be a source of implicit knowledge, new knowledge can be extracted when using appropriate tools (e.g., data mining) [1]. Hospitals operate within this principle, since their databases contain hidden knowledge that is important, for example, in a patient's clinical status and prognosis [2].

Data Mining (DM) implementation in the health sector can be an asset. In fact, DM can help health insurers, detect fraud or abuse in decision making over the relationship with their clients, doctors can identify more effective treatments, as well as best practices. As consequence patients can receive better health services. DM provides techniques to transform huge data volumes into useful knowledge for decision making [3]. Based in this principle it was determined whether it was possible to construct models of DM capable of supporting the process of hospital

management. The support is at the level of beds management using the number of patient discharges.

The beds management is currently a real problem for the hospitals because due the increase of the number of patients admitted, it is notorious for some services the lacking of resources. A system that can help to predict the number of patient's discharges can contribute to improve the hospital quality helping in the beds distributions process.

The main goal of this work is to predict the number of patient discharge by week using data mining techniques. To this problem were used classification techniques and a set of configuration applied to four services. The results were satisfactory with the accuracy being around 82.69 % and 94.23%. The results attained give confidence to use this models in order to help the decision process and enforces the idea to be possible forecast patient discharges using as input only the date and the number of patient discharged.

In the second chapter is presented the conceptual framework, DM techniques used, and metrics for evaluating the classification models. The third chapter presents the development of the practical component, guided by the CRISP-DM methodology. In the fourth chapter, the most relevant aspects of this work are analysed and discussed. The fifth chapter focuses on the work conclusions and, finally, possible directions for posterior work.

2 Background

The World Health Organization describes the hospital as being "... part of a medical and social organization whose function is to provide a comprehensive health service and care for the population, both curative and preventive, and whose outpatient services must reach families in the home environment, the hospital is also a centre for training of health workers and biosocial research. " [4].

To manage a hospital a deep understanding about the institution is mandatory. Managers need to have the perfect notion about the rules and routines of services. They have to be able of identifying the strengths and aspects that need improvement. From these aspects managers should outline a clear and organized plan to provide for an efficient and effective hospital management [5].Management consists of a variable set of technical tools and technologies used to ensure organizations' success [6]. In an organizational structure such as a hospital, where the core business is to optimize patients admission in the hospital, minimizing the length of stay and maximizing the quality of treatment should be sought [7]. In this sense, one of the most important features is the distribution of beds among the various services. The way of how the hospital is organized reflects the efficiency and quality of hospital management[8].

Introducing new technologies in hospitals has provided the opportunity for work improvement, e.g. some tasks can be performed faster, more consistently and with lower costs [9]. DM as well as some classic statistic methods have been used in hospital databases since 1990 [10], [11], [12]. Ever since, DM is becoming increasingly popular and essential in healthcare [3].

2.1 Hospital Bed Management

Hospital beds are one of the scarcest resources of hospitals. In most cases the beds are arranged according to hospital specialties in order to provide a better service for patients. The beds management reflects not only the services efficiency but also the quality of hospital management. In order to improve the hospital management with a focus on bed management, there are some studies using DM with special attention on patient admittion in the hospital. These studies presented as main goals decrease the length of stay of patients and performing a good beds management.

A project conducted at hospital Chiba University Hospital, studied which medical care tasks were directly correlated with patients' length of stay. This study revealed a strong correlation between some variables presenting coefficients between 0.837 and 0.867, in a range of [-1, 1]. The study concludes that the results obtained have shown a strong correlation between patient discharges and hospital management quality [11].

Another study was conducted at the National University Hospital of Singapore. This study had as main objective identify which was the key variable associated to mismatch allocation of beds by department. The models obtained had acuity values of 74.1% and 76.5% and allowed to find that the key variable was the medical speciality. Through this study was possible to determine strategies to the allocation of beds in the respective hospital [13].

2.2 Data Mining

Technological advances have provided new ways to create and store data. Organizations accumulate data related to their processes (billing, business transactions and accounting) based on the idea that the large volumes of data can be a source of knowledge [1].

From a technical standpoint, DM is a process that uses artificial intelligence techniques, statistics and mathematics to extract useful information and knowledge (or patterns) from large volumes of data. These patterns can be in the form of business rules, affinities, correlations, or terms of forecasting models [14]. The prediction methods goal is to automatically build a behavioural model, obtaining new samples and unknown samples, and being able to predict values of one or more variables related to the sample. The design patterns that enable knowledge discovery are easy to understand and can be easily used as a work base [15].

Classification models aim to identify a function that associates an event to a class within several discrete classification classes, i.e., classifiers map the input space into predefined classes. For example, the class patient has attributes that describe the patient; if a particular person meets the classification properties of the patient, then that person can be classified as patient [1].

For this work the implementation of DM was achieved through the statistical environment R. R presents itself as a programming language and an environment for statistical development [16]. The library e1071 [17] was used to implement the DM techniques Support Vector Machine (SVM), Decision Tree (DT), and Naïve Bayes (NB), and for troubleshoot forecasting - classification. To conduct the DM models evaluation the rminer library has been used [18].

3 Knowledge Discovering Process

The DM process is complex but when conducted in a methodological context becomes easier to understand, implement and develop. CRISP-DM methodology was followed to carry out the present study.

3.1 Business Understanding

Data used in this study were retrieved from Centro Hospitalar do Porto (CHP), Porto, Portugal. Models use as input values the number of patients discharged distributed across multiple services. The supplied sample comprised 62 302 records. CHP has no mechanism to predict the discharges flow. This evidence triggers an entire workflow process in accordance to the objective of this study, to predict the number of patients weekly discharged, primarily targeting the bed management improvement.

3.2 Data Understanding

The provided sample comprises the period between 01.01.2009 and 31.12.2012, referring to 1461 days. The years 2009, 2010 and 2011 have 365 days and the year 2012 has 366 days, which is the only bissextile year from the sample provided.
From the referenced timeline were collected 62,302 records, corresponding to discharges from ninety one hospital services. Each record consists of three fields:

- Date: corresponds to the date (day, month and year) of patient's discharged;
- Service: is the service associated to the patient in the date;
- Discharge number: contains the number of patients who were discharged. This field is directly related to the date and the hospital services, i.e., records were grouped by date and service.

As already mentioned, one of the goals is to predict the hospital discharges per week. From the ninety one existing services only four presented records for all week days during the period analysed. Based on this evidence, data was only exploited from four services: Orthopaedics, Obstetrics, Childbirth and Nursery.

The Table 1 presents a characterization of the data associated to each one of the services.

Table 1 – Data characterization

	Orthopaedics	Obstetrics	Childbirth	Nursery
Maximum	78	91	8	92
Minimum	19	40	0	34
Average	≈48.6	≈62.9	≈2.4	≈57.3
Coefficient of Variation	≈22.634%	≈17.011%	≈70.833%	≈19.197%
Standard Deviation	11	10.7	1.7	11
Total Discharges	10108	13080	495	11924

3.3 Data Preparation

In this study, as already mentioned, the objective it to make predictions of weekly discharges. So it was necessary to group the daily records into weekly records. By convention, one week begins on Sunday and ends in the next Saturday. Following this principle the discharges were grouped into 52 weeks and the respective years 2009, 2010, 2011 and 2012. Once these records were grouped in weeks it was necessary to use methods capable of determining intervals of values for hospital discharges, in order to make predictions using the classification approach.

The selection of the number of classes or intervals does not constitute a rigorous and scientific method, nor is there any selection method that can be considered appropriated [19]. Therefore, some methods were implemented to create classes: average, quartiles, average and standard deviation and sturges rule. Through the set of methods described, the data tables were created according to the number of weeks and the respective years, i.e., the rows are the number of weeks (52 in total), the columns (4 in total) correspond to the years, 2009 to 2012. This data representation is defined as being conventional. Having in consideration this data it was also explored a different approach using a sliding window data representation. However, this paper was focused in presenting the first approach – conventional.

3.4 Modelling

The techniques considered to induce classification models were: Support Vector Machine (SVM), Decision Trees (DT) and Naïve Bayes (NB). The data mining techniques selection was defined based in three aspects: easy understanding of the tecniques, engine efficient and the fact of being possible training the models with only a small dataset. Based in this three principles, SVM achieved the second and third goals and presents as being efficiently using a reduced number of data in the training phase. The DTs and NBs has in common the fact of being easy to understand and efficiently in the use. In the case of NBs, they also can present good results using a reduced number of data.

In order to implement a mechanism for model testing, two sampling methods have been selected: 10-folds Cross Validation (10-folds CV) and Leave-One-Out Cross Validation (LOOCV). The 10-folds CV was adopted due to the good results that it has demonstrated on multidisciplinary data [20]. LOOCV it was used because it is more suitable for databases with only a few dozens of tuplets [21] - since the data tables only have 52 or 205 records. All the techniques were submitted to the tune function. This function comes with the e1071 package. As main objective it executes network searches of hyper parameters intervals previously provided and consecutively identifies the best model and the respective hyper parameters.

For the SVM technique two distinct kernels were used, Radial-Basic Function (RBF) and Linear. Given the different kernels used it was necessary to perform different parameterizations because their hyper parameters differ from kernel to kernel. Depending on the kernel used by SVMs, a range of values for parameter C was defined. Its range has been defined by the values obtained by the power $2^{(1,..,4)} = [2, \dots, 16]$, in which $C > 0$. The cost parameter C introduces some

flexibility separating the categories in order to control the trade-off between errors in training or stiffness margins [22].
The hyper parameter Gamma (γ) was defined in the same way as *C*. The range was determined according to the values obtained by the power $2^{(-1,0,1)} = [0.5, 1, 2]$. Its parameterization was used in the RBF kernel. The γ value determines the curvature of the boundary decision [23].

The DTs were used to perform predictions through the classification approaches. The implementation of this technique was achieved through the CART algorithm. Two methods for attribute selection or splitting rules were used, Information Gain (IG) and the Gini Index (GI). The attribute selection measure IG determines the attribute with the highest information gain and uses it to make the division of a node [24]. The GI is determined by the difference between the original information requirement (i.e., based on only the ratio of classes) and the new requirement (i.e., obtained after partitioning *A*). This difference is expressed as: $Gain(A) = Info(D) - Info_A(D)$. The attribute A that has the highest information gain, Gain (A) is the division attribute of node *n* [24]. The objective of GI is to calculate the value for each attribute using the attribute for the node with the lowest impurity index [1]. The GI index measures the impurity of *D*, using a data partition or a training set of attributes $Gini(D) = 1 - \sum_{i=1}^{m} pi^2$, where pi corresponds to the probability of an attribute *D* of a class *Ci*. This value is estimated by $|Ci \cap D|/|D|$. The sum is calculated as a function of m classes [24].

Finally, the algorithm NB was used to perform predictions using the classification approach (was not carried out any manual parameterization for this algorithm). This algorithm used the tune function to identify the sampling method that should be used. These methods were previously determined.

The developed models can be represented by the following expression:

$M_n = A_f + S_i + C_x + MRD_z + TDM_y + MA_k$.

The model M_n belongs to the approach (A) classification and is composed by a service (S), a type of class (C) a method of data representation (MDR), a DM technique (TDM) and a sampling method (SM):

$A_f = \{Classification_1\}$
$S_i = \{Orthopedics_1,\ Obstetrics_2, Parturition_3, Nursery_4\}$
$C_x = \{Average_1,\ Quartils_2, Average\ and\ Standard\ Deviation_3,\ Sturges_4\}$
$MDR_z = \{conventional_1, Sliding\ Window_2\}$
$TDM_y = \{SVML_1, SVMRBF_2, DTGI_3, DTIG_5, NB_5\}$
$SM_k = \{10 - folds\ CV_1, LOOCV_2\}$

For instance, the model (M_1) uses the technique SVM with RBF kernel and the sampling method 10-folds CV and it is expressed by: $M_1 = A_2 + S_2 + C_5 + MDR_1 + TDM_2 + SM_1$.

3.5 Evaluation

This task is dedicated to models' evaluation. To evaluate the results achieved by the DM models it was used the accuracy or precision measure.

As the results presented by the DM models depend on the division of the mutually exclusive subsets, two procedures have been implemented: 10-folds CV and LOOCV. For the splitting procedures, ten executions were performed for each of them. Around 100 experiments were performed for each test configuration with models that use the 10-folds CV procedure. 520 experiments were performed to test models using LOOCV procedure. Table 2 depicts the obtained values of accuracy for the best classification models.

Table 2 – Evaluation of classification models

Model	Service	Technique	Sampling Method	Class	Accuracy
M_1	S_1	TDM_2	$SM_{1,2}$	C_4	≈82.69%
M_2	S_2	TDM_2	$SM_{1,2}$	C_4	≈82.69%
M_3	S_3	TDM_2	$SM_{1,2}$	C_4	≈94.23%
M_4	S_4	TDM_2	$SM_{1,2}$	C_4	≈90.38%

The DM technique that presented the best results was the SVM RBF, the two sampling methods used did not show to be crucial for the results, being equivalent. From table 1 it can be seen that rule sturges used in the creation of classes was the best method. Then, figure 1 presents the classes predicted by the models and their respective frequency for 2012 by week. Classes are represented by their maximum (green point) and at its minimum (blue point) value. If the expected values of the year 2012 (red dots) are between the maximum and minimum means that the class predicted was properly carried out, i.e., the actual values were classified in the predicted class limits. The presented classes were obtained using the sturges rule.

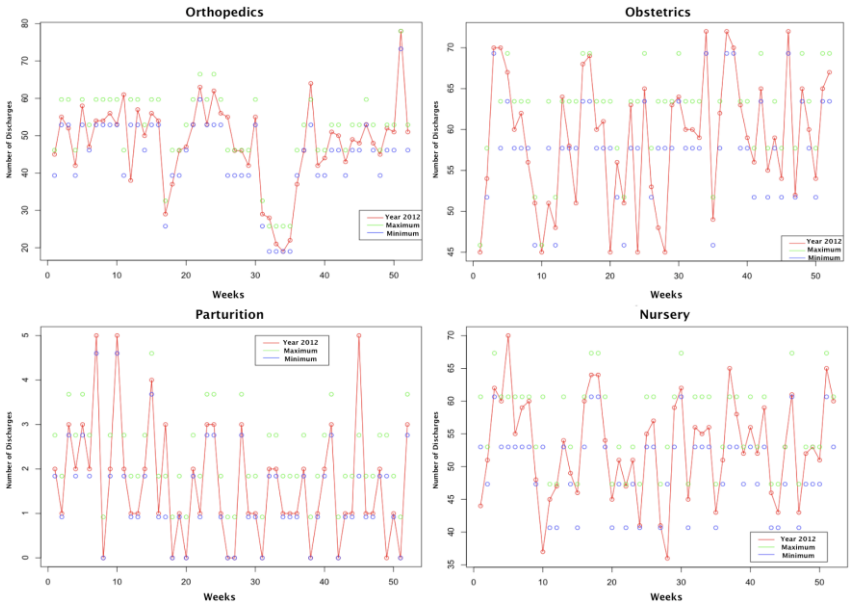

Figure 1 – Classes predicted

4 Discussion

The presented results are quite acceptable due to the prediction evaluations made. For the classification models, the best predictions resulted in acuities greater than ≈82%. Orthopaedics and obstetrics are the services with the worst results claiming ≈82.69% of acuity. The results obtained for the services of childbirth and nursery are ≈94.23% and ≈90.38% respectively. The predictions made for the childbirth and nursery service present results sufficiently satisfactory to support decision making. Table 3 presents the number of times DM models hit and missed.

Table 3 – Results of Predictions

Model	Wrong	Correct
M_1	9	43
M_2	9	43
M_3	3	49
M_4	5	47

From the two sampling methods used 10-folds Cross Validation (10-folds CV) and Leave-One-Out Cross Validation (LOOCV), 10-folds CV is the most suitable for this study, not because the results since the differences are very small and often are the same, but because the execution time of the models. Models that use the sampling method LOOCV take a long time to process.

Since the used data is real, the inclusion of these models in a Decision Support System (DSS) becomes expectable. Thus the knowledge generated from the usage of DM can be useful so that it can influence the operational efficiency, facilitating high-level decisions and service providing.

5 Conclusion

This work proved that is possible induce classification data mining models to predict weekly discharge of patients in a hospital using real data of daily discharge. A study was carried out considering data from CHP, corresponding to four years of activity. Good results were obtained in terms of precision for four services: ≈82.69% for orthopaedics and obstetrics; ≈94.23% and ≈90.38% for childbirth and nursery, respectively.

Two different methods of data representation were explored: conventional and sliding window. The second method is computationally more expensive and does not improve the results. Conventional representation in association with sturges rule, the sampling method 10-folds CV and the SVM technique, demonstrate to be the most suitable for this type of data.

Finally with this work it was possible assess the possibility to predict the number of patient discharges using only the number of discharges by day. The results attained using classification techniques can prove that fact.

6 Future Work

Future research will take into consideration the following aspects:

- To explore different types and configuration of Data Mining techniques;
- Incorporate new variables in the predictive models, such as gender and age of patients;
- Determine if the predictions could be more detailed. The introduction of the entry number of patients, bed occupation time and ratio should be taken into account to create a better bed management system;
- Repeat the experiments for more hospital services with new data.

Acknowledgements

This work is supported by FEDER through Operational Program for Competitiveness Factors – COMPETE and by national funds though FCT – Fundação para a Ciência e Tecnologia in the scope of the project: FCOMP-01-0124-FEDER-022674.

The authors would like to thank FCT (Foundation of Science and Technology, Portugal) for the financial support through the contract PTDC/EIA/72819/ 2006 (INTCare) and PTDC/EEI-SII/1302/2012 (INTCare II).

References

[1] M. Santos and C. Azevedo, *Data Mining Descoberta do conhecimento em base de dados*. FCA - Editora de Informática, Lda, 2005.

[2] M. Santos, M. Boa, F. Portela, Á. Silva, and F. Rua, "Real-time prediction of organ failure and outcome in intensive medicine," in *2010 5th Iberian Conference on Information Systems and Technologies (CISTI)*, 2010, pp. 1 –6.

[3] H. Koh and G. Tan, "Data mining applications in healthcare," *J Healthc Inf Manag*, vol. 19, no. 2, pp. 64–72, 2005.

[4] WHO, "Expert Committee on Health Statistics," 261, 1963.

[5] I. Santos and J. Arruda, "Análise do Perfil Profissional dos Gestores dos Hospitais Particulares da Cidade de Aracaju- SE," *Revista Eletronica da Faculdade José Augusto Vieira*, vol. Nª -7, 2012.

[6] J. Proença, A. Vaz, A. Escoval, F. Candoso, D. Ferro, C. Carapeto, R. Costa, and V. Roeslin, *O Hospital Português*. Vida Económica-Conferforum, 2000.

[7] M. Neves, "Os Médicos vão ter de ser os motores da reforma do sistema," *Revista Portuguesa de Gestão & Saúde*, no. 5, 2011.

[8] G. Yang, L. Sun, and X. Lin, "Six-stage Hospital Beds Arrangement Management System," presented at the Management and Service Science, 2010.

[9] A. Dwivedi, R. Bali, and R. Naguib, “Building New Healthcare Management Paradigms: A Case for Healthcare Knowledge Management,” presented at the Healthcare Knowledge Management Issues, Advances, and Successes, 2006.

[10] R. Bose, “Knowledge management-enabled health care management systems: capabilities, infrastructure, and decision-support,” *Expert Systems with Applications*, vol. 24, no. 1, pp. 59–71, 2003.

[11] S. Tsumoto and S. Hirano, “Data mining in hospital information system for hospital management,” presented at the ICME International Conference on Complex Medical Engineering, 2009. CME, 2009, pp. 1 –5.

[12] S. Tsumoto and S. Hirano, “Towards Data-Oriented Hospital Services: Data Mining-based Hospital Mangement,” presented at the The 10th IEEE International Conference on Data Mining Workshops, Sydney, Australia, 2011.

[13] K. Teow, E. Darzi, E. Foo, X. Jin, and J. Sim, “Intelligent Analysis of Acute Bed Overflow in a Tertiary Hospital in Singapore,” *Springer US*, 2012.

[14] E. Turban, R. Sharda, and D. Delen, *Decision Support and Business Intelligence Systems*, 9ª Edição. Prentice Hall, 2011.

[15] O. Maimon and L. Rokach, “Introduction to Knowledge Discovery and Data Mining,” in *Data Mining and Knowledge Discovery Handbook*, 2ª Edição., Springer, 2010.

[16] L. Torgo, *Data Mining with R: Learning with Case Studies*. CRC Press - Taylor & Francis Group, 2011.

[17] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, “Misc Functions of the Department of Statistics (e1071).” 2012.

[18] P. Cortez, “Simpler use of data mining methods (e.g. NN and SVM) in classification and regression.” 2013.

[19] E. Reis, *Estatística Descritiva*, 7ª ed. Edição Sílabo, 2008.

[20] I. Witten, E. Frank, and M. Hall, *Data Mining Pratical Machine Learning Tools and Techniques*, 3ª Edição. Morgan Kaufmann, 2011.

[21] P. Refaeilzadeh, L. Tang, and H. Liu, “Cross-Validation,” *Encyclopedia of Database Systems*, 5 vols. Springer, 2009.

[22] M. Kantardzic, *Data Mining Concepts, Models, Methods, and Algorithms*, 2ª Edição. Wiley - IEEE Press, 2011.

[23] A. Ben-Hur and J. Weston, “A User’s Guide to Support Vector Machines,” in *Data Mining Techniques for the Life Sciences*, O. Carugo and F. Eisenhaber, Eds. Humana Press, 2010.

[24] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3ª Edição. Morgan Kaufmann, 2012.