Specialised Tools for Automating Data Mining for Hospital Management.

J. Alapont[1], A. Bella-Sanjuán[2], C. Ferri[2], J. Hernández-Orallo[2],
J. D. Llopis-Llopis[2], M. J. Ramírez-Quintana[2].

[1]Dimensión Informática
Av. Cataluña, 11, 46020 Valencia, Spain
(E-mail: j.alapont@dimension-informatica.es)

[2]Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, Camí de vera s/n, 46022 Valencia, Spain
(E-mail: {abella, cferri, jorallo, jdaniel, mramirez}@dsic.upv.es)

Abstract: This paper presents a research project which is directed to the partial automation of Data Mining (DM) in hospital information systems (HIS). We concentrate on hospital management applications and information systems such as emergencies and ward management, human resources (services, night duties, etc.), physical resources (beds, intervention theatres, etc.), etc. We have realised how the business objectives are usually the same across several hospitals and so is the information which is gathered in several HIS (even using different DBMS). This means that although the models extracted highly differ between hospitals, data mining processes are highly similar across different hospitals. We argue how a tool can be constructed in such a way that it automates many DM processes and that can be ported to other hospitals which could benefit more quickly of a first DM experience.

Our work plan covers all the stages in the process of Knowledge Discovery from Databases (KDD): data cleansing, extraction and integration from the HIS and external data, construction of tasks and minable views, model generation, and finally a module to carry out and interpret their predictions. We also consider a module to perform simulations and to integrate the models extracted by the previous modules with other decision support systems as well as model monitoring.

Keywords: Data Mining, KDD, Hospital Information Systems, Hospital Management, Model reuse**.**

1. Introduction

The growing quality demand in the hospital sector makes it necessary to exploit the whole potential of stored data efficiently, not only the clinical data, in order to improve diagnoses and treatments, but also on management, in order to minimise costs and improve the care given to the patients. In this sense, Data Mining (DM) can contribute with important benefits to the health sector [7] [10], as a fundamental tool to analyse the data gathered by hospital information systems (HIS) and obtain models and patterns which can improve patient assistance and a better use of resources and pharmaceutical expense [3][4].

Data Mining [7] is the fundamental stage inside the process of extraction of useful and comprehensible knowledge, previously unknown, from large quantities of data stored in different formats, with the objective of improving the decisions of companies, organisations or institutions where the data have been gathered. However, data mining and the overall process, known as Knowledge Discovery from Databases (KDD), is usually an expensive process, especially in the stages of business objectives elicitation, data mining objectives elicitation, and data preparation. This is especially the case each time data mining is applied to a hospital: many meetings have to been held with the direction of the hospital, area coordinators, computer scientists, etc., to establish the objectives, prepare the data, the mining views and for training the users to general DM tools.

From our experience, we have seen that, given a business area, in our case, healthcare, many data mining implementations repeat the same business objectives, data mining objectives, needs of external data, feature construction, etc., than previous implementations. When implementing a data mining programme in a hospital, especially when using the same people involved in previous projects, the time required to deliver the project is shorter than for the first project. However, most of the work is still manual and hence most of the work involved in a previous project is not reused for subsequent projects. We see that models can be very different between different hospitals, but the process from data to rules is almost the same for every hospital.

In this paper we analyse which parts of a data mining project for hospital management are equal or highly similar across different hospitals (at least in the same national healthcare system). This allows us to design several data mining modules which can be portable across several hospitals, thus dramatically reducing the time to implement a data mining programme in a new hospital. These specialised tools must be accompanied by some degree of adaptation in any new hospital (especially, to integrate internal and external data sources, depending on the category and the geographical area the hospital covers), but the data mining models are "re-trained" on each new hospital and deployed much more seamlessly.

The paper is organised as follows. In section 2 we use the stages of the CRISP-DM standard to illustrate which stages are almost identical (and hence reusable) across different projects. Section 3 discusses the business objectives in hospital management and their translation into data mining objectives, which are common to most hospitals we have analysed. Section 4 is devoted to data integration, where most of the differences appear (since many hospitals use different HIS and are located in quite different areas). Section 5 discusses data preparation, data transformation and feature construction, in particular, which is frequently the same in different hospitals. Section 6 centres on modelling, the definition of a DM task and minable view, which allow the DM module to generate the models. Finally, section 7 closes the paper with the conclusions of this work and some other future work.

2. Structure of an Automated Tool for Hospital Management

CRISP-DM [2] (*Standard CRoss-Industry Process for Data Mining*), is a consortium of companies (initially granted by the European Commission) which has defined and validated a data mining process that is applicable to several industry sectors. The following Figure 1 shows the different stages of this process:

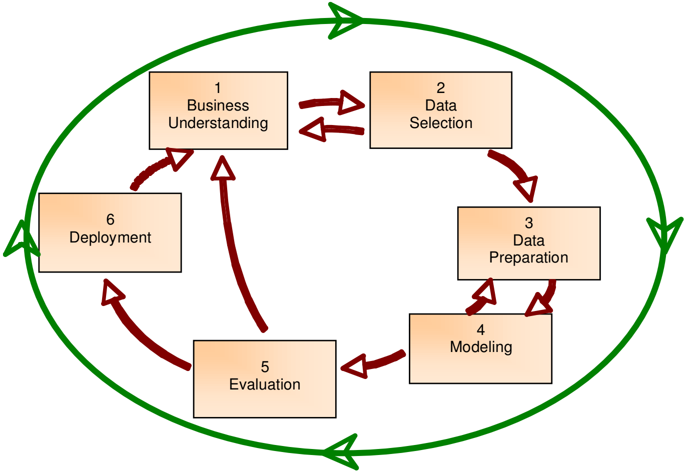

Figure 1: Life cycle of a data mining project.

The initial stage (*Business Understanding*) focuses on identifying the problems we are trying to solve through DM (i.e., the business objectives are defined). In our area of interest (healthcare), some hospital management objectives might be: to improve the use of hospital resources, to avoid bed occupation greater than 100% or to plan the schedule for using the operating theatre more intensively. These objectives are defined by the people in charge of the hospital management, and then they have to be converted into data mining objectives. For instance, some data mining objectives defined from the business objectives mentioned above are: to obtain a predictive model of hospital bed occupation, to predict the stay time of a patient depending on their disease, to establish models for estimating operations with higher cancellation or delay probability, etc. Objectives like these are of general interest for improving the management of any hospital independently of whether it is a general or a specialised health centre. So these objectives could be included as an initial set of generic objectives in an automated data mining tool specially developed for this area.

Something similar occurs with respect to the data that could be relevant for the hospital management: they are usually gathered for every centre. For instance, admission date, admission cause, discharge data, medical service assigned at the admission time, etc. The main difference between hospitals is the format in which this information is stored in the DBMS. This fact makes it possible to (semi-)automate the rest of the life cycle stages. Hence, for stage 2, we only need to characterise the data load process from the particular HIS to the data warehouse (D.W.) for collecting all data needed for the data mining process. Likewise, regarding the data preparation stage, the same transformation processes (construction of new attributes, grouping continuous data in ranges, etc…) will be applicable for any HIS since all of them work with the same kind of data. In general, stages 4 to 6 can also be done in an automated way since those generated models which are of interest for a hospital probably are also of interest for another one, and so on.

Taken all of these considerations into account, we propose the following general scheme for an automated data mining tool for hospital management (Figure 2).

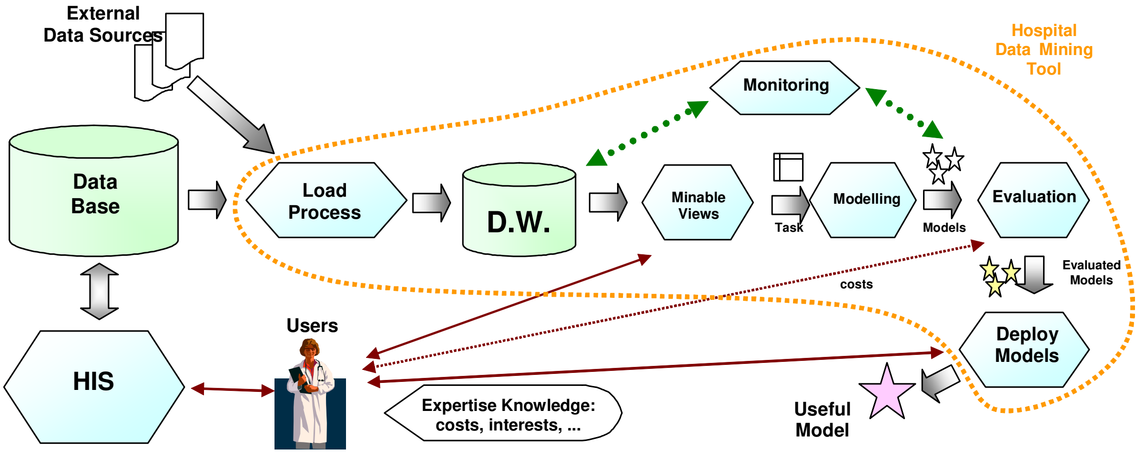

Figure 2: Data Mining Tool for Hospital Management.

The tool is composed by several processes (modules) that correspond to the stages described in Figure 2. Thus, the *load process* corresponds to stages 2 and 3 (as we have discussed before). The *Minable View process* integrates the business objectives in order to select from the D.W. the data to be used for constructing the models. Finally, the following processes (*Modelling*, *Evaluation* and *Deploy Models*) represent stages 4, 5 and 6, respectively.

3. Business and Data Mining Objectives

We have seen at Section 2 that the first step in a data mining project is to establish the business objectives. In hospital management situations we are dealing with objectives such as:

- To optimise bed occupation.
- To improve the use of operating theatres, avoiding the cancellation of operations.
- To know how emergencies affect to the administration of the hospital departments or services (cancellation of operations, etc).
- To optimise the allocation of human and material resources to wards and shifts.
- To detect the influence of certain diseases in the hospital's services.
- To find clusters of patients.

Most of these objectives are related to emergency hospitalisations since it is a special service whose medical treatments and procedures cannot be usually delayed. Also, these objectives are interrelated. For example, if the bed occupation is closer 100%, it is necessary to cancel operations previously planned. If the operations are frequently cancelled, then the waiting lists are increased.

Now, the previous objectives have to be transformed into Data Mining objectives, such as:

- To carry out global models about pressure emergencies by different time periods (daily, by shifts of work, by day of the week, etc).
- To generate a model for predicting the number of daily hospitalisations coming from emergencies.
- To obtain predictive models of global and partial use of beds by hospital service.
- To construct models for estimating how the resources of a hospital are affected by a certain disease (for instance, influenza).
- To carry out models to cluster patients (by age, by area, by pathology class, etc).

In this paper, we will show implementation examples on the emergency area, since this part of the project is one which is more portable across hospitals.

4. Data integration

For solving the data mining objectives such as those shown in Section 3, we need two kinds of information: internal (contained in the HIS) and external (not contained in the HIS). Internal information changes from one hospital to another, but for example, all of them collect general data from patients and their treatments. External data are not easy to obtain, because they are not gathered in any database.

In the area we're focussing on in this paper, emergencies, we implemented the following integration:

- For internal data, our system gathers the personal patient details which are usually present in any hospital.: sex, birthday date, country and living area. It is also fed by information about the patient workflow: admission date and time, reason of admission, discharge date and time, discharge code from emergencies, code of the medical service assigned at the admission time, initial diagnosis, final diagnosis, etc.
- For the external data, we gather the following data (different for each hospital, since this is geographically dependent): meteorological data (temperature, quantity of rain, wind speed, etc), lunar stage, character of the day (holiday, before holiday or after holiday, and also the festivals in the city, etc., important events, for example, football matches.

5. Data Preparation

One of the main problems to apply data mining for improving the management of a hospital is the bad quality of the source data. In many cases, the collected data contain missing or anomalous values. This can be due to a wide range of reasons: many patients do not have enough time (or they are not conscious) for filling the admission form patients do not have documents when they arrive at the hospital, illegible data, bad transcriptions, repetition of values, etc. Therefore, in these contexts, a thorough data preparation stage is very important for a successful data mining process. Some processes in the data extraction phase have been adapted to particular hospitals, but many other data cleansing/preparation processes (detection of missing or anomalous values, attribute transformation, feature creation, etc.) are the same across hospitals.

On the other hand, in many cases we will find attributes containing text, for instance, an initial description of the pathology of the patient. Since this kind of attributes cannot be directly dealt with classical learning methods, we could employ retrieval information techniques to transform the text attributes in one or more discrete attributes. For instance, we could transform the attribute with the initial description of the patient's pathology into a discrete attribute with a value for the most common pathologies (flu, traumatisms…), and a value "unclassified" for the rest of cases.

Part of this preparation stage is reused from hospital to hospital, through the automation of all these processes in a data preparation module.

We implemented scripts for extracting data from the different hospitals into the Data Warehouse (DW). These scripts must be slightly different from hospital to hospital. From the DW, since the data definition (multidimensional schema) is the same for every hospital, we used SQL scripts to generate the minable views, which are exactly the same. For instance, the minable view for the emergency pressure must integrate the number of admissions per day (or per shift) and calculate means for admission numbers of the previous week. Additionally, the number of non-working days before and after must be computed in order to get the attributes for the minable new. All these complex SQL queries are highly time-consuming. With our approach, these complex queries are 100% portable from one hospital DW to another, and all this effort is reused.

From the minable views, the data is converted into a standard format (the arff format of WEKA) by means of Python scripts. In this way, using the command-line option in WEKA, we can generate, evaluate and export the models. Then the models are applied to new data. All this process is automated. Additionally, in some cases, the predictions can be integrated into the HIS.

6. Learning the models

Once the data have been properly filtered, cleaned and transformed, we can proceed with the induction of the prediction models. For this purpose, we employ the suite WEKA [14], and we make our modules work with it. This suite integrates many of the most known learning techniques, as well as, several pre-processing and post-processing tools. Additionally, WEKA has been released as open source, so, if it is required, we can adapt this software for our particular requirements.

The key point for using WEKA is the proper construction of the minable view in such a way that could be directly used by the learning methods. A standard format (arff) has been defined as a data and model file repository in WEKA. So, the idea is to generate the data in this format, and in this way we can employ all the different leaning techniques integrated in this suite.

In our case, we generated different minable views for some areas of hospital, although in this paper we only show the results obtained for predicting the number of emergency admissions. The minable views were constructed by considering: the number of admissions in the seven previous days, holidays or celebrations, important sport events, meteorological data (rain and temperature of the seven previous days), etc. These data belong to a hospital from 2000 to 2004, both years inclusively.

Table 1 shows the minable view that we used from our experiments.

Attribute	SQL Type	Description/Values
day_week	nvarchar	Day of the week
type_day	nvarchar	Type of the day: F=Holiday, VFF=Before holiday, etc.
sport_events	nvarchar	TRUE/FALSE
average_temperature	float	Average temperature
rain	float	Amount of rain in mm^3.
numEmerg-1	integer	Number of admissions from the previous day.
numEmerg-2	integer	Number of admissions from the two previous days.
numEmerg-3	integer	Number of admissions from the three previous days.
numEmerg-4	integer	Number of admissions from the four previous days.
numEmerg-5	integer	Number of admissions from the five previous days.
numEmerg-6	integer	Number of admissions from the six previous days.
numEmerg-7	integer	Number of admissions from the seven previous days.
monthNumEmerg	integer	Average number of admissions in the same month of the year before.
dayWeekYearNumEmerg	integer	Average number of admissions in the same day of the week of all the year before.
daysBefHoli	integer	Number of holidays before the day.
daysAftHoli	integer	Number of holidays after the day.
numEmerg (class)	integer	Number of admissions in this day.

Table 1: Initial minable view

In summary, this minable view has a total of 1459 rows with 17 attributes where the attribute to predict is the number of admission in a particular day.

With this initial minable view, we used different learning methods included in WEKA: LinearRegression, LeastMedSq, SMOreg, MultilayerPercepton, Kstart, LWL, Tree DecissionStump, Tree M5P and IBK. We used 10-fold cross validation in the experiments. The method which obtained the best results was the linear regression and tree M5P. We used a statistical model to compare with linear regression model. The Statistical model is just an average of admissions per day. Table 2 shows the improvement from statistical to data mining models.

	Statistical model	Linear regression model	Tree M5P model
Mean absolute error	36.996	23.8988	24.0446
Relative absolute error	100 %	64.59 %	64.98 %

Table 2: Statistical model vs. Data mining models.

We show the model obtained with the WEKA Linear Regression method for the ‘pilot’ hospital (years 2000-2004). It weights significant attributes with a positive or negative weight to obtain the number of admissions in this day.

```
numEmerg =
    58.7937 * day_week=Monday +
    14.806  * type_day=DFF,LL,NLF,VFF,DFL +
   -11.7541 * type_day=LL,NLF,VFF,DFL +
    32.3177 * type_day=NLF,VFF,DFL +
     0.387  * average_temperature +
    -1.6026 * rain +
     0.2572 * numEmerg-1 +
     0.076  * numEmerg-2 +
     0.0678 * numEmerg-3 +
     0.0748 * numEmerg-4 +
     0.0473 * numEmerg-5 +
     0.0702 * numEmerg-6 +
     0.1193 * numEmerg-7 +
     0.1496 * dayWeekYearNumEmerg +
    -3.6549 * daysBefHoli +
    50.438
```

Hospitals are used to work with the averages (statistical model). With our data mining models, we obtained an improvement in mean absolute error of 13.1 admissions per day for the linear regression model, and 12.95 admissions per day for the tree M5P model. Furthermore, considering that the average value of number emergencies admission is 493 admissions per day, these mean absolute errors are small in comparison with the average value, and allow the hospital manager to better adjust resources. We did similar models per day, per service, per shift, which in some cases showed higher difference with the by default (statistical) models.

The final goal, however, is not to solve the management problems of a single hospital, but to port these results to other hospitals. The study performed in the 'pilot' hospitals are crucial for this. According to these, we extracted which minable views, which attribute selection, and which learning methods are best from these data. Consequently, we implemented in the automated system the best minable views and just two models: the linear regression and the tree M5P. Given a new hospital with its standard datawarehouse, the system makes the complex minable views (with the relevant attributes), exports the data to Weka, run the algorithms and evaluate the models, and show the results, allowing the managers of the new hospital to apply them.

7. Conclusions and future work.

Data mining is still below its full potential in many areas. Healthcare, especially public healthcare, is one of these areas. In this paper, we have analysed the adequacy of designing specialised modules for data mining for hospital management and we have also identified which are the stages in the KDD process which could be reused and automated across different hospitals. The success of this project could turn data mining into an available technology to many hospitals which cannot afford a complete data mining programme from scratch. Additionally, as long as hospital processes and patient flows and forms become more standard across countries, especially in the European Union, the data integration part would become more seamless, and, hence, the costs and application time of these specialised tools could even be smaller.

Additionally, as future work, we like to extend the modules to modify or define new data mining objectives, not only the predefined data mining objectives identified in general. The idea is to be able to implement the programme initially with the by default models but being able to add more models and objectives which fit specific needs of a particular hospital.

Acknowledgements

This work has been supported by the Valencian Government through an IMPIVA project. Additionally, from the several hospitals we used as "reference" hospitals, we want to especially thank the Hospital Clínic Universitari de València, their managers, clinical analysts and computer scientists, which also took part of the IMPIVA project and allowed us to work with their data for the first minable views and for the first stage of this project.

8. References

[1] Andrei Gagarine, MD, John D. Urschel, MD, John D. Miller, MD,W. Frederick Bennett, MD, and J. Edward M. Young, "Preoperative and Intraoperative Factors Predictive of Length of Hospital Stay after Pulmonary Lobectomy".

[2] CRISP-DM, http://www.crisp-dm.org

[3] David Riaño, Susana Prado (2001), "Improving HISYS1 with a Decision-Support System", The Eighth European Conference on Artificial Intelligence in Medicine. Pp: 413 - 416 Cascais (Portugal).

[4] David Riaño, Susana Prado (2000), "A data-mining alternative to model hospital operations: clinical costs and predictions", Lecture Notes in Computer Science 1933, pp. 293-299. Rüdiger W. Brause and Ernst Hanisch (Eds.), Springer-Verlag, ISBN: 3-540-41089-9.

[5] Fu-ren Lin, Shien-chao Chou, Shung-mei Pan, Yao-mei Chen (2001), "Mining time dependency patterns in clinical pathways", International Journal of Medical Informatics 62.

[6] HIPAA, http://www.etsu.edu/irb/HIPAA_KIT.htm

[7] José Hernández Orallo, M^a José Ramírez Quintana, Cèsar Ferri Ramirez (2004), "Introducción a la minería de datos", Prentice Hall.

[8] Margaret R. Kraft, Kevin C. Desouza, Ida Androwich (2002). "Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population", Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03).

[9] Nada Lavrac (1999), "Selected techniques for data mining in medicine", Artificial Intelligence in Medicine 16 3–23.

[10] L. Goodwin, M. VanDyne, S. Lin, S. Talbert (2003),"Data mining issues and opportunities for building nursing knowledge" Journal of Biomedical Informatics, 36, 379-388.

[11] Pofahl W, Walczal S, Rhone E, Izenberg S (1998). "Use of an artificial neural network to predict length of stay in acute pancreatitis". *Am Surg* 1998; 64: 9: 868-72.

[12] Shusaku Tsumoto (2000), "Application of Knowledge Discovery Process to Hospital Information System", AMIA 2000: Annual Symposium, 2000.

[13] Sun-Mi Lee, Patricia A. Abbott (2003), "Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers", Journal of Biomedical Informatics 36.

[14] WEKA, http://www.cs.waikato.ac.nz/ml/weka/