The New Navigators: from Professionals to Patients
R. Baud et al. (Eds.)
IOS Press, 2003

Analysis of hospitalised patient flows using data-mining

Thierry Dart, Yongmei Cui, Gilles Chatellier, Patrice Degoulet

Santé Publique et Informatique Médicale (SPIM), university of Paris 6, France

Abstract

Face to the development of hospital information system in the "Hôpital Européen George Pompidou" (HEGP), computerized patients records made medical data easier to analyse than before. We use data-mining technology to analyse intra-hospital patients' paths with one year of PMSI data (a French medical information system similar to Diagnosis Related Group). Methods: 1."sequential patterns mining" was used to analyse the most frequent patients' paths, 2. an integrated framework of "association rules mining" and "classification rule mining" was used to build prediction rules of patients' paths. Result: we construct a rule based prediction model, which gives the tendency of the patient's paths between the different medical units.

Keywords:
Hospitals; Length of Stay; Patient path; Data-mining.

1. Introduction

Patient path presents a multidisciplinary care map, in which the procedure of the therapeutic and the potential relationships between the different medical units is sequenced on a timeline with a point of departing (the beginning of the hospitalisation) and a point of arriving (the end of the hospitalisation). In the context of exceptional resource and competition between the hospitals, the competition advantage will be not only on increasing the medical quality, but also on reducing the length of hospital stay. A deep understanding of the patients' paths becomes more and more important for the physicians and the administrators today. Benefit of Information System of HEGP makes it possible to answer these concerns. Data-mining technology makes the procedure of analysing this great deal of data easier and more effective than other technologies.

Among the method described to analyze patient's paths, we could find statistical and data-mining methods. Markov chain analysis is a classical statistical way to analyze patient flow [1]. Data-mining methods are more recent. A first example of data-mining analysis of patients' paths used many data-mining methods: Bayesian methods, decision tree and neural network [2]. In this work, a pathway was created for each patient with the different medical events and the decision tree shows the most discriminative variables to characterize a patient path. An other interesting approach is the use of "rough set theory" for the extraction of association rules from a hospital database [3]. Rough set allows the extraction of association rules from a database.

Objectives

This paper presents two data-mining methods to describe, predict and visualize patient paths into a hospital. The purpose of this project is to:

1. find and describe the most frequent patient paths inside of the hospital, their probabilities and the average length of stay (descriptive model);

2. determine the most predictive variables of patients paths and build a model to predict patient paths using rules (predictive model);
3. make these models accessible via the hospital intranet.

2. Material and methods

Data source

We used one year (2001) of data extracted from the PMSI database (Programme de Médicalisation du Système d'Information). PMSI is the French casemix–based funding for public hospital, derived from the DRG (Diagnosis Related Group). Each patient path is described by the following fields: patient-id, age, sex, principal diagnosis, medical unit-id, length of stay in each unit.

Data-mining methods

Data mining is an interdisciplinary area of tools (statistics, machine learning, neural network...) to extract models from large database. Although some specific problems [4] may emerge, the use of data-mining in healthcare is growing rapidly with the growth of database [5]. Many medical areas are concerned: genetic, bioinformatic (DNA micro array), patient flow, adverse effect of drugs, prediction of death ... Among all the different techniques of data-mining [5, 6], we choose two methods to analyze patient paths: sequential patterns mining [7] and classification of association rules (CARs) [8].

Sequential Patterns mining (Prefixspan)

To find the most frequent patient paths inside the hospital (objective 1), we choose the "sequential patterns mining" method [7], which was first introduced by Agrawal [9], using an algorithm called "Apriori". We used an other algorithm: "PrefixSpan" [10] (Prefix projected sequential pattern mining), which explores prefix projection in sequential pattern mining. The problem of mining sequential patterns is to find the maximal sequences among all sequences that have a certain user-specified minimum support.

An example of "PrefixSpan" algorithm is shown in Figure 1. We try to find the most frequent patient paths among four patients (Pat. 1 to Pat. 4) who stayed in four different medical units (A to D). Patient 1 goes from unit A to unit B, from B to unit D and goes back to unit B. Using data mining terminology, he goes thought a sequence of four units: <ABDB>. Using a minimal support of 3 stays, the "Prefixspan" algorithm starts to find the most frequent items: A (4 times), B (3 times) and D (4 times) (C is eliminated because of its frequency is lower than the support). <A>, <B> and <D> are stored into the "prefix projection". For the prefix <A>, the algorithm tries to find the most frequent items: AB(3 times), AD(4 times) (AC is eliminated). The <AB> and <AD> items become two new prefix; and, the algorithm goes on for the prefix <AB>, <AD> and <B>. Using this recursive algorithm, we find two most frequent patient paths: <ABD> and <BD> as shown in Figure 1.

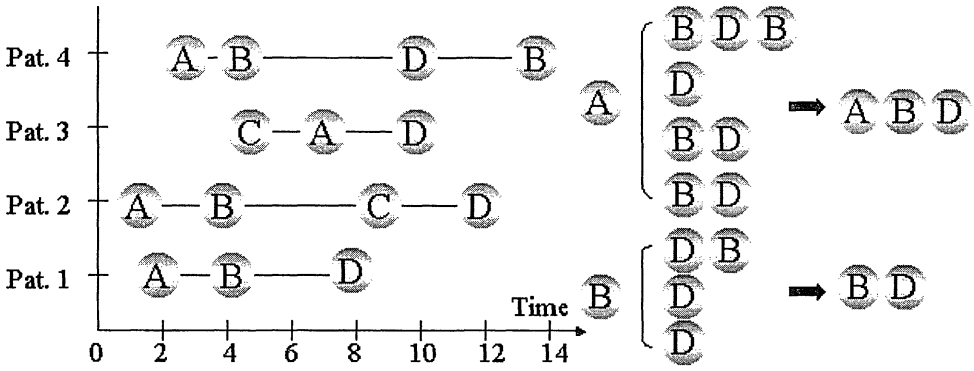

Figure 1 – Finding frequent paths using the "PrefixSpan" algorithm

Classification of Association Rules (CARs)

To find prediction rules of patient paths (objective 2), we use the CARs method (Classification of Association Rule) [8], which is a hybrid method combining association rule mining and classification rule mining. Using the frequent patient path found by "Prefixspan", CARs is a classifier for building association rules.

"Association rules mining" methods are useful to search interesting relationships among items in a large database [11]. The simplest form of this technique, called "market basket analysis", is widely used by data manager to find the buying habits of their customers. Many algorithms have been developed (Apriori, AprioriAll, AprioriSome, FPTree ...).

Steps of analysis

We carried out the following steps for building the models:

- Step 1: extraction and preprocessing of the data. The preprocessing of the data is a significant phase, which determines some performances of the models. After data extraction from the PMSI database (relational database), we clean them by adding some constraints.
- Step 2: construction of the descriptive and predictive models. We extract the sequential patterns using the algorithm "PrefixSpan" to identify the most frequent patients' paths (descriptive model). Then we build the rules for patients' paths prediction, using the CARs method (predictive model).
- Step 3: validation of the predictive model. This step consists in validating these rules, in order to estimate the performance of the models, in particular with regard to the exactitude of prediction on new data. This step uses the "K-fold cross-validation" method.
- Step 4: datamart and intranet integration. These rules of the predictive model have been integrated into a relational database called a "datamart".

3. Results

Characteristic of patients' stays (descriptive model)

A simple statistic analysis of our data shows that 17628 patients were hospitalized during the year 2001. Among these 17628 patients, 22867 patients' stays were registered. Multiple

units' stays represents 2368 (10%): 1834 stays in 2 units, 422 in 3 units, 76 in 4 units, 28 in 5 units, 8 in 6 units, 2 in 7 units and 1 in 8 units.

Table 1 - comparison of two types of patients' paths

	Simple units' stays	Multiple units' stays
Male (%)	45.0 %	45.4 %
Age class (years)		
- 0 – 14	2.4 %	7.9 %
- 15 - 34	13.0 %	19.2 %
- 35 - 64	16.7 %	11.5 %
- 65- 104	4.8 %	2.8 %
Admission mode		
- From home	93.8 %	91.0 %
- From other unit of HEGP	6.2 %	8.7 %
- From other hospital	0.0 %	0.3 %
Most frequent first stay unit	Orthopaedic ward: 16.5% Cardiology: 12.4%	Emergency ward: 45.2% Cardiology: 7.9%

Table 1 shows a comparison of two types of patients' stays (multiple versus simple patients' stays). This suggests that younger patients are more likely to follow a multiple unit stay than a single unit stay. The length of stay is more likely to be longer in a multiple unit stay than in a single unit stay. Among the multiple units' stays, we find 63 "frequent patients' paths": 53 in 2 units, 9 in 3 units and 1 in four units.

Predictive model of patients 'paths

This model is composed of prediction rules. The medical unit is the predicted variable (dependant variable), and the four others are the predictive variables (independent variables: age, sex, admission mode and principal diagnosis).

Building the rules based predictive model

We use the CARs method to classify these four variables associated with the "most frequent patients paths". The rules are in a format "If <condition (age, sex, principal diagnosis, admission mode)> Then <unit>". We found 52 rules for the multiple units' stays and 229 rules for the single unit stays. An example of these rules is showed in Figure 2.

```
ICD = S, ENTRY_MODE = 8, CL_AGE = 5, SEX = F
->    unit = 5 (confidence: 93.02%)
```

Figure 2 -one prediction rule of patients' paths
ICD: diagnosis (chapter of ICD-10), CL_AGE: age class

Validation of this predictive model

Using the "K cross-validation" method, we found that this model correctly classifies only 49.9 % of new patients.

Data mart of patients' paths

The rules have been integrated into a relational database to create a datamart. This system let us visualize the predictive and the descriptive models using the hospital's intranet.

4. Discussion

By using an administrative database and two different techniques of data mining, we were

able 1) to propose an analysis of patients' path 2) to describe a predictive model of a given path according to a set of predictive variables and 3) to propose a web application for end-users such as physicians or hospital managers.

In medical field, the use of data-mining techniques is still not usual, but it should be developed with the growth of size and complexity of medical databases. The databases resulting from the DRGs or insurance systems are two examples of very large databases exploitable by these techniques. Although statistical method exits (Markov chain analysis), data-mining methods are more flexible with massive amount of data continuously collected and stored in large database. We chose to test these data-mining methods on a relatively small amount of data (one year), which however comprise the data concerning more than 20000 patients' stays.

Comparing to the closest work [3], we used two different methods of data-mining: "Prefixspan" and CARs. "PrefixSpan" decreases the requirements in memory while avoiding reading again several times the whole of the database. Other methods of data-mining based on "Apriori-like" algorithms [12, 13] consume much memory to generate sequences candidates. The "Class Association Rules" (CARs), by combining classification rule mining and association rule mining, is a more accurate classifier [8, 14] for prediction than other classification system (like C4.5).

Our principle limit is the low prediction rate (49.9%). This could be explained by the low power of discrimination of the four variables (sex, age, admission mode, principal diagnosis) and the low number of data (limited to one year). Therefore, the models are less relevant to predict new patients' paths.

It remains indeed much more to analyze patients' paths. The finality of this work is to optimize the resources available (number of beds, personnel, reducing length of stay [15]...). A better model could be a start point of computer simulations to solve hospital capacity planning problems [16]. At least three directions should be explored:

1. Compare the different data-mining methods to improve prediction;
2. Combine statistics and data-mining methods. For example, factorial analysis method may help a better identification of the predictive variables;
3. A better use of the data richness by analyzing patients' paths into a unit and by including more predictive variables from other databases (significant medical procedures, costs, etc...).

5. Conclusion

Medical data analysis using data-mining technologies has certainly to be developed. Our work is limited to the analysis of the patients' paths within a hospital. Our result is encouraging, and suggests that work should be continued. Physicians frequently use models to predict a disease or the length of stay among their patients, for triage or to optimize bed occupancy. Our work suggests that a new data analysis method could add a new tool to help resource allocation within a medical unit or a whole hospital. Moreover, quality may also be analyzed using this technique.

6. Acknowledgement

We would like to thank Drs. Catherine Karanfilovic and Catherine Rovani for their help using the PMSI database.

References

[1] Weiss EN, Cohen MA, Hershey JC. An iterative estimation and validation procedure for specification of semi-Markov models with application to hospital patient flow. *Oper Res* 1982;30(6):1082-104.

[2] Le Duff F, Happe A, Burgun A, Levionnois S, Bremond M, Le Beux P. Sharing medical data for patient path analysis with data mining method. *Medinfo* 2001;10(Pt 2):1364-8.

[3] Pagnoni A, Parisi S, Lombardo S. Analysis of patient flows via data mining. *Medinfo* 2001;10(Pt 2):1379-83.

[4] Cios K, William Moore G. Uniqueness of medical data mining. *Artif Intell Med* 2002;26(1-2):1.

[5] Hobbs GR. Data mining and healthcare informatics. *Am J Health Behav* 2001;25(3):285-9.

[6] Smyth P. Data mining: data analysis on a grand scale? *Stat Methods Med Res* 2000;9(4):309-27.

[7] Srikant R, Agrawal R. Mining sequantials patterns: generalizations and performance improvements. In: Fith Int'l Conference on Extending Database Technology (EDBT); 1996 March 1996; Avignon, France; 1996.

[8] Liu B, Hsu W, Ma Y. Integrating Classification and Association Rule mining. In: KDD'98; 1998 Aug 27-31, 1998; New York; 1998.

[9] Agrawal R, Imielinski T, Swami A. Mining associations between sets of items in massive databases. In: ACM-SIGMOD 1993, Int'l Conference on Management of Data; 1993 May 1993; Washingon, DC; 1993. p. 207-16.

[10]Pei J, Han J, Mortazavi-Asl B, Pinto H. PrefixSpan : mining sequential patterns efficiently by prefix-projected pattern growth. In: Int. conf. On Data Engineering (ICDE'01); 2001 April 2001; Hedelberg, Germany; 2001.

[11]Han J, Kamber M. Mining association rules in large database. In: Data mining: concepts and techniques. San Francisco: Morgan Kaufmann publishers; 2001. p. 225-77.

[12]Agrawal R, Srikant R. Mining sequential patterns. In: Int'l Conference on data engineering (ICDE); 1995 Mars 1995; Tapei, Taiwan; 1995.

[13]Han J, Dong G, Yin Y. Efficient mining of partial periodic patterns in time series database. In: Int. Conf. Data Engineering (ICDE'99); 1999 1999; Australia; 1999.

[14]Han J, Kamber M. Classification based on concepts from association rule mining. In: Data mining: concepts and techniques. San Francisco: Morgan Kaufmann publishers; 2001. p. 311-4.

[15]Fernandes CM, Christenson JM. Use of continuous quality improvement to facilitate patient flow through the triage and fast-track areas of an emergency department. *J Emerg Med* 1995;13(6):847-55.

[16]Isken MW, Rajagopalan B. Data mining to support simulation modeling of patient flow in hospitals. *J Med Syst* 2002;26(2):179-97.

Address for correspondence

Thierry Dart, thierry.dart@spim.jussieu.fr, laboratoire SPIM, UFR Broussais Hôtel-Dieu

15, rue de l'école de médecine 75270 Paris cedex 06; Tel: (33) 1 42 34 69 83, Fax: (33) 1 53 10 92 01