Clinical Knowledge Discovery in Hospital Information Systems: Two Case Studies

Shusaku Tsumoto

Department of Medicine Informatics, Shimane Medical University, School of Medicine,
89-1 Enya-cho Izumo City, Shimane 693-8501 Japan
E-mail: tsumoto@computer.org

Abstract. Since early 1980's, the rapid growth of hospital information systems stores the large amount of laboratory examinations as databases. Thus, it is highly expected that knowledge discovery and data mining(KDD) methods will find interesting patterns from databases as reuse of stored data and be important for medical research and practice because human beings cannot deal with such a huge amount of data. However, there are still few empirical approaches which discuss the whole data mining process from the viewpoint of medical data. In this paper, KDD process from a hospital information system is presented by using two medical datasets. This empirical study shows that preprocessing and data projection are the most time-consuming processes, in which very few data mining researches have not discussed yet and that application of rule induction methods is much easier than preprocessing.

1 Introduction

Since early 1980's, the rapid growth of hospital information systems (HIS) stores the large amount of laboratory examinations as databases (Van Bommel, and Musen, 1997). For example, in a university hospital, where more than 1000 patients visit from Monday to Friday, a database system stores more than 1 GB numerical data of laboratory examinations for each year. Furthermore, storage of medical image and other types of data are discussed in medical informatics as research topics on electronic patient records and all the medical data will be stored in hospital information systems within the 21st century. Thus, it is highly expected that data mining methods will find interesting patterns from databases as reuse of stored data and be important for medical research and practice because human beings cannot deal with such a huge amount of data.

In this paper, knowledge discovery and data mining (KDD) process (Fayyad, et. al, 1996) for two medical datasets extracted from a hospital information system is presented. This empirical study shows that preprocessing and data projection are the most time-consuming processes, in which very few data mining researches have not discussed yet and that application of rule induction methods is much easier than preprocessing.

2 KDD Process

2.1 Data Selection

In this paper, we use the following two datasets for data mining, which are extracted from two different hospital information systems. One is bacterial test data, which consists of 101,343 records, 254 attributes. This data includes past and present history, physical and laboratory examinations, diagnosis, therapy, a type of infectious disease, detected bacteria, and sensitivities for antibiotics. The other one is a dataset on the side effect of steroid, which consists of 31,119 records, 287 basic attributes. This data includes past and present history, physical and laboratory examinations, diagnosis, therapy, and the type of side effects. The characteristic of the second dataset is that it is a temporal database: although it includes 287 basic attributes, 213 attributes of which have more than 100 temporal records. These datasets are obtained through the first to third steps of KDD process: data selection, data cleaning and data reduction.

In the first step of KDD process, these databases are extracted from two different hospital information systems by simple query process. Table 1 gives results for data selection.

Table 1. Data Selection Results

Datasets	HIS size	Target Data	Time Required
Bacterlal Test	1,275,242(52GB)	361,932(14GB)	2.3 Days
Side-Effect	2,631,549(100GB)	135,749(6GB)	7.3 Days

2.2 Data Cleaning as Preprocessing

After data selection, data cleaning is required since the data obtained from the first step are not clean, including data records not suitable for data analysis. Even though these records are selected by matching the query condition, they do not include enough amount of information. For data cleaning, we define the two-fold cleaning steps: first we select the records which have no missing values in the pre-defined indispensable attributes. Then, we calculate how many attributes in the remaining attributes are used to describe the records in the first step. If the number of attributes used for a case is not sufficient, then this case will be removed. For those steps, the indispensable attributes and the threshold for the second selection are given a priori by domain experts. Table 2 summarizes results for data cleaning.

2.3 Data Reduction and Projection as Preprocessing

Data Projection for Bacterial Test Database. From the viewpoint of table processing, data cleaning can be viewed as cleaning steps in the direction of records,

Table 2. Data Cleaning Results

Datasets	Data	Cleaned Data	Time Required
Bacterlal Test	361,932(14GB)	101,343(3GB)	5.2 Days
Side-Effect	135,749(6GB)	31,119(1,5GB)	2.5 Days

that is, in the direction of row. On the other hand, data reduction can be viewed as cleaning steps in the direction of attributes, that is, in the direction of column. Although the data cleaning process is a time-consuming process, it takes much more time to reduct and project data in clinical databases due to the characteristics of biological science, including medicine. The tradition of classification in biology tends to have a large scale of classification systems. Thus, generalization of such over-classified attributes is required to discover rules which are easy for medical experts to interpret. For bacterial test databases, a simple concept hierarchy was used for generalization of values.

Data Reduction for Steroid Side-Effect Database. Since incorporating temporal aspects into databases is still an ongoing research issue in database area (Abiteboul, et. al., 1995), temporal data are stored as a table in hospital information systems (H.I.S.) with time stamps. The characteristics of medical temporal data are as follows(Tsumoto 1999): (1)The Number of Attributes are too many. (2) Irregularity of Temporal Intervals. (3)Missing Values are too many.

In order to deal with medical temporal databases is discussed in (Tsumoto, 1999). Tsumoto introduces extended moving average method, which automatically set up the scale of the temporal interval. Here, we applied this method to this dataset.

2.4 Total Time Required for Data Reduction

In summary, Table 3 gives the total time required for data reduction and projection for each data set, including knowledge acquisition process. The second column gives the type of preprocessing. The third column shows total time required for each process. Finally, the fourth column shows the time required for acquisition of knowledge from domain experts.

Table 3. Total Time Required for Data Reduction

Dataset	Preprocessing	Total Time	Time for Acquisition
Bacterial Test	Projection	15.25 Days	7.0 Days
Side-Effect	Summarization	2.3 Days	0

This table suggests the generalization of values in attributes should be a time-consuming process, especially when domain knowledge is given.

2.5 Rule Induction as Data Mining

After the third step, rule induction based on rough set model (Pawlak, 1991) was applied to two medical datasets. Tsumoto (1998) extends rough-set-based rule induction methods into probabilistic domain.

In this section, we skip this part due to the limitation of space. For further discussion, the readers may refer to (Tsumoto, 1998; Polkowski and Skowron, 1998). The algorithm introduced in (Tsumoto 1999) was implemented on the Sun Spaarc station and was applied to the above two medical databases, the information of which is summarized in Table 4. For rule induction, the thresholds for accuracy and coverage are set to 0.5 and 0.5, respectively.

Table 4. Summary of Data Mining

Data	Size	Attributes	Rules	Computational Time
Bacterial Test	101,343 (3GB)	254	24,335	60 hours (2.5 Days)
Side-Effects	31,119 (1.5GB)	287	14,715	18 hours (0.75 Days)

2.6 Interpretation of Induced Rules

After the data mining step, we obtain many rules to be interpreted by medical experts. Even if the amount of information is very small compared with the original databases, it still takes about one week to evaluate all the induced rules and only small part of rules were found to be interesting or unexpected to medical experts(Table 5).

Table 5. Summary of Induced Rules and Interpretation

	Induced Rules	Interesting Rules
Bacterial-Test	24,335	114 (0.47%)
Side-Effect	14,715	106 (0.72%)

3 Discussion

After the data interpretation phase, about one percent of induced rules are found to be interesting or unexpected to medical experts. In this section, the total KDD process is reviewed with respect to computational time.

Table 6 shows the total time required for KDD process. Each column shows two data sets for each process. Each row includes computational time required

for each process. Totally, it takes about one month and three weeks to complete the whole KDD process for bacterial test database and side-effect database, respectively. It is notable that more than 60% of the process is devoted to the three preprocessing processes: data selection, cleaning and reduction. Especially, as for the bacterial test dataset, 22.75 days (79.5%) are used for three processes. It is because domain knowledge should be acquired for generalization of data, as discussed in Section 3. On the other hand, only 4 to 8 percent of total time is spent for data mining process. In the case of bacterial test database, only 2.5 days (8%) is used for rule induction. Therefore, these empirical results suggest that the main KDD processes should be preprocessing rather than discovery of patterns from data.

Table 6. Total Time Required for KDD process

KDD Process	Bacterial Test Side-Effect	
Data Selection	2.3	7.3
Data Cleaning	5.2	2.5
Data Reduction	15.25	2.3
Data Mining	2.5	0.75
Data Interpretation	7.0	7.0
Total Time	32.25	19.85

References

1. Abiteboul, S., Hull, R., and Vianu, V. 1995. Foundations of Databases, Addison-Wesley, NY.
2. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. 1996. The KDD Process for Extracting Useful Knowledge from Volumes of data. CACM, 39: 27-34.
3. Pawlak, Z. 1991. Rough Sets, Kluwer Academic Publishers, Dordrecht.
4. Polkowski, L. and Skowron, A. 1998. Rough Sets in Knowledge Discovery Vol.1 and 2, Physica-Verlag, Heidelberg.
5. Tsumoto, S. 1998. Automated extraction of medical expert system rules from clinical databases based on rough set theory, Information Sciences, 112, 67-84.
6. Tsumoto, S. 1999. Rule Discovery in Large Time-Series Medical Databases. In: Proc. 3rd European Conference on Principles of Knowledge Discovery and Data Mining (PKDD), LNAI 1704, Springer Verlag, 23-31.
7. Van Bommel,J. and Musen, M. A.1997. Handbook of Medical Informatics, Springer-Verlag, NY.