EASYMINER: DATA MINING IN MEDICAL DATABASES

Mohammad Saraee, George Koundourakis, Babis Theodoulidis

1. Introduction

Medical databases have accumulated large quantities of information about patients and their medical conditions. While technological advancements in the form of computer-based patient record software and personal computer hardware are making the collection of and access to health care data more manageable, few tools exist to evaluate and analyse this clinical data after it has been captured and stored. Evaluation of stored clinical data may lead to the discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management. Techniques are needed to search large quantities of clinical data for these patterns and relationships. In this study the techniques of data mining were used to search for relationships in large clinical databases.

Data mining, also referred to as Knowledge Discovery in Databases is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [1]. In other words, it is the search for relationships and global patterns that exist in large databases, but are "hidden" among the vast amount of data, such as a relationship between patient data and medical diagnosis. These relationships represent valuable knowledge about the database and objects in it and (if the database is a faithful mirror) of the real world registered by the database.

Data mining techniques have rarely been applied to medical domain. GTE Laboratories built a large data mining system that evaluated health-care utilisation to identify intervention strategies that were likely to cut costs. This system, however, is focused on cost analysis and not on identifying new associations or relationships within clinical data[2]. Data mining techniques are also used for improving birth outcomes[3] and automated detection of hereditary Syndromes[4].

We are currently in the process of experimenting a data mining project at UMIST using an extensive clinical database of stroke patients from East Lancashire[5], [6] to identify factors that contribute to this disease. EasyMiner is our data mining system designed and developed in Timelab research laboratory at UMIST for interactive mining of interesting patterns in time-oriented medical databases. This system implements a wide spectrum of data mining functions, including generalisation, relevance analysis, classification and discovery of association rules The eventual goal of this data mining effort is to identify factors that will improve the quality and cost effectiveness of patients care. In the following sections we describe briefly the EasyMiner data mining approaches.

2. Generalisation

EasyMiner uses generalisation rules as they are defined in [7]. By using these rules, values of the attributes are generalised at multiple concept levels. As a result of this process, a set of attributes is generated for every generalised attribute of the original data set. The generated attributes are instances of the same attribute but in different concept levels. In order to determine which of these attributes (concept levels) is most appropriate to the required data-mining task, relevance analysis is applied on them and only the one that is most relevant to the specific data-mining task is kept. In EasyMiner there are several procedures for the automatic generation of generalisation rules for numeric attributes.

3. Relevance Analysis

We perform relevance analysis as defined by Kemper et al.[8]. Examples of obvious relevancies amongst database attributes is 'Age' to 'Date of birth' and 'Title' to 'Sex' and 'Marital Status'. This kind of knowledge is qualitative and it is quite useful to mine from large databases that hold information about many objects (fields). For example a bank could look in its data and identify the factors that it should consider in order to give a credit limit to a customer. Applying Easy Miner to our credit history database revealed that Credit Limit is relevant to *Account Status*, *Monthly Expenses*, *Marital Status*, *Monthly Income* and *Gender*

4. Classification

Classification is an important area of research in data mining. Classification partitions massive quantities of data into sets of common characteristics and properties[9]. In classification a set of records, acting as *training set,* is analysed in order to produce a model of the given data. Each record is assumed to belong to a predefined class, as determined by one of the attributes, called *classifying attribute*. Table 1 shows a part from a sample training set of the medical database, where each record represents a patient and *Lived* is the *classifying attribute* of the training set.

Once derived, the classification model can be used to categorise future data samples, as well as providing a better understanding of the database contents. Classification is particularly useful when a database contains examples that can be used as the basis for future decision making, e.g. for assessing credit risks, for medical diagnosis, or for scientific data analysis.

.....	Sex	Date of Birth	Date of Stroke	Lived	.....
...	Male	9/21/1924	9/23/94	Died	...
...	Female	5/8/1921	12/10/94	Died	...
...	Male	1/18/34	2/7/94	Survived	...
...	Male	9/26/1925	11/13/94	Survived	...
...	Female	3/28/51	12/9/94	Died	...
...	Male	1/19/34	2/7/94	Survived	...

Table 1: Sample Training Set

The classification technique that we have developed in Easy Miner is based on the decision tree structure. By using a decision tree, untagged data sample can be classified by testing the attribute values of the sample data against the decision tree. A path is produced from the root to a leaf node, which has the class identification of the sample.

5. Association Rules

Agrawal describes discovery of association rules in large databases in [10]. The initial motivation for association rules was to aid in the analysis of large transaction databases, such as those collected by supermarkets. The discovery of associations between various line items can potentially aid decision making within organisations. Using the formalism provided by Agrawal, association rules can be defined as follows.

Let $I=\{i_1, i_2,..,i_m\}$ be a set of *items*. Let DB be a database of transactions, where each transaction T consists of a set of items such that $T \subseteq I$. Given an *itemset* $X \subseteq I$, a transaction T *contains* X only and only if $X \subseteq T$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \varnothing$. The association rule $X \Rightarrow Y$ holds in DB with *confidence c* if the probability

of a transaction in DB which contains X also contains Y is *c*. The association rule $X \Rightarrow Y$ has *support s* in DB if the probability of a transaction in DB contains both X and Y is *s*. The task of mining association rules is to find all the association rules whose support is larger than a *minimum support threshold* $ó_l'$ and whose confidence is larger than a *minimum confidence threshold* $ö_l'$. We use the semantic characteristic of decision trees to develop a method for finding association rules from a database. The rules that our method discovers have the form: $X \Rightarrow Y$, where X is a set of conditions upon the values of several attributes and Y a specific value of one attribute or a combination of several attributes' values. This attribute (or combination of attributes) Y is called *target attribute.* The method for mining association rules consist of 3 simple steps including Data Preparation, Generation of rules and Selection of strong rules.

References

1. Fayyad U, Piatetsky-Shapiro, Smyth Uthurusamy R., Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1995.

2. Hedberg The Data Gold Rush, Byte, October 1995, pp. 83-88.

3. Goodwin L, Prather J, Schlitz K, Iannacchione M, Hammond W, Grzymala J, DataMining Issues for Impproved Birth Outcomes, Biomed. Science Instrum, 34, 1997, pp. 291-296.

4. Evans S, Lemon S, Deters C, Fusaro R and Lynch H, Automated Detection of hereditary Syndromes Using Data Mining, Computers and Biomedical Research 30, 1997, pp. 337-348.

5. Du x, Cruickshank JK, McNamee R., Saraee M, Sourbutts J, Summers A, Roberts N, Walton E, Holmes S., Case-control study of stroke and the quality of hypertension control in north west England, The British Medical Journal (BMJ) 1997; 314:239-314, January 1997

6. Du x, Cruickshank JK, Sourbutts J, Summers A, Roberts N, Walton E, Holmes S, Saraee M, Theodoulidis B., Prevalence, Treatment, Control and Awareness of High Blood Pressure among First-Ever Stroke Patients in Northwest England, 16th Scientific Meeting of the International Society of Hypertension", 23-27 June 1996, Glasgow, UK. Also in Journal of Hypertension, vol. 14, 1996 (supp 1) pp. s224.

7. D. W. Cheung, Ada W. Fu, J. Han. Knowledge Discovery in Databases: A Rule-Based Attribute Oriented Approach. In Proc. of 1994 Int. Symp. On methodologies for intelligent systems, Charlotte, N.C. Oct 1994.

8. Kamber M, Winstone L, Gong W, Cheng S, Han J. Generalisation and Decision Tree Induction: Efficient Classification in Data Mining. In Proceedings of 1997 International Workshop on Research Issues on Data Engineering (RIDE'97), Birmingham, England, April 1997, pp 111-120.

9. Mehta M., Agrawal R., and Rissanen J. SLIQ: A Fast Scalable Classifier for Data Mining. In Proc. of the Fifth Int'l Conference on Extending Database Technology (EDBT), Avignon, France, March 1996.

10. Rakesh Agrawal and R. Srikant. Mining Quantitative Association Rules in Large Relational Tables. In Proc. of the ACM SIGMOD Int'l Conf. on Very Large Data Bases (VLDB), Zurich, Switzerland, September 1995.