6^{th} International Conference on Applied Informatics
Eger, Hungary, January 27–31, 2004.

Application of Data Mining Methods in Health Care Databases*

Ágnes Vathy-Fogarassy

Department of Mathematics and Computer Science,
University of Veszprém, Hungary
vathya@almos.vein.hu

Abstract

By the spreading of the information systems a huge amount of data has been aggregated in these systems up to the present. Since strategically important information can be hidden in this mass of data, these pieces of information may be very valuable. With the help of data mining and knowledge discovery methods we can extract the hidden knowledge from these large amounts of data. These methods can be applied to numerous areas, for example commerce, telecommunication, finance and health care, too.

By now the hospital information systems are widespread over the world, also in Hungary. These systems store a great deal of data concerning the patients' physical parameters, lifestyle, laboratory values, case history and treatment modality. With the application of data mining methods to the medical and health data we can discover unknown relationships among these parameters concerning the examined population, which is the Hungarian population here. This process includes forming groups characterizing the patients from the point of view of clinical outcome, identifying the risk factors, analyzing the trends of the changes of clinical parameters, etc.

In this paper we discuss the preparation steps that must be taken before analyzing the medical data. We also deal with data mining methods that are practical to use for different purposes. Finally we discuss the applicability of these tools in a particular area of health care, namely in the research of osteoporosis.

Key Words and Phrases: data mining, methods, health care, osteoporosis

1. Introduction

From the late 1980s to the present the dominant research area in the information technology has been knowledge discovery, including data mining methods

*Research presented in this article is supported by the project IKTA 142/2002.

and data warehouses. Data mining itself can be viewed as a result of the natural evolution of information systems. After having solved the problem of creation and design of databases, a huge amount of the collected data has been aggregated in these database systems up to the present. That is how it works in the area of medical sciences, too. All over the world there are many research projects that are based on the application of data mining methods in various fields of science. When considering the Hungarian medical database systems, we can see that these information systems are big and diverse enough to extract some valuable hidden information from them.

Researches based on data mining that cover health care domain have complex goals. On the one hand these projects can examine the applicability of data mining methods in health care, how the general algorithms can be improved by building in the expertise knowledge, and what the typical problems and mistakes are on which focus must be placed during this work.

On the other hand by the application of these methods in Hungarian health care database systems some hidden pieces of information might be discovered, which can be used in medical practice, for example improving treatment protocols or analyzing risk factors.

2. Knowledge discovery

The concept of data mining is often used as a synonym of knowledge discovery (KDD), however data mining is only an essential step of the knowledge discovery process. This process includes the following major steps: learning the application domain; creating the target data set; choosing the data mining functionalities and the proper algorithms; pattern evaluation; knowledge presentation; assessing of the discovered information.

Data mining work in an unknown domain always starts with the understanding of the application domain and with the specification of the problem that we would solve. In the domain of health care it means that we need to get familiar with the major medical terms, and we need to talk over the aims of application with doctors or other experts.

Then we must preprocess the available data, this includes data selection and aggregation, data cleaning, and data reduction and transformation. Chapter 4 discusses these activities in detail. After collecting and preprocessing the data according to the goals of the application, we need to choose the functionality of our data mining activity, and find the best data mining algorithms. Data mining functionalities include creating concept or class descriptions, classification, clustering, association analysis, and evolution analysis. The choice between these possibilities is mainly influenced by the limits of the mining system in use. After running the data mining algorithms the discovered patterns need to be visualized for the experts. For this purpose we can use charts, tables, diagrams, decision trees, rules, etc. By the evaluation of these results the desired new knowledge is obtained, which the end-users can utilize during their work.

3. The source data

Medical data is arising from different sources. Two types of databases are available in medical domain. The first type of medical data comes from medical experts. For example it can be diagnosis, status, medication and so on. It is typical of this type of data that the number of records is small, but the number of attributes for each record is relatively large if compared with the number of records and in this kind of data we do not find missing values frequently.

The other type of medical data is coming mainly from Hospital Information Systems (HIS). This data is automatically stored in databases without any specific further purpose. For example laboratory test information is classified to this group.

The source systems of medical data are mostly the Hospital Information Systems and flat files, but in some special cases data warehouses can also provide this kind of data. Unfortunately in many cases important data is stored in paper form only. The data needed for data mining examinations must be integrated before analysis. Most often the target of the integrated data is a relational database or a data warehouse.

Examining the medical data it may be often seen that the base data is not in the suitable form, and/or is dirty, and some data transformation activities may needed to be performed on them. So before running the data mining algorithm it must be selected, cleaned, integrated, and transformed.

4. Data preprocessing

Analyzing dirty, wrong data never provides useful information. Before starting any data mining task, the data must be preprocessed. This activity includes solving the problem of dirty data, the problem of missing values, handling redundant data, dealing with unstructured information and other data preprocessing activities, such as creation new features, data normalization or data generalization.

In medical information systems it often occurs that some fields are empty. The reason for this may be that data isn't available (for example the examination wasn't performed), or data isn't stored (for example physical parameters or lifestyle). To improve the discovery process it is suggested to get and fill in the missing data. For this purpose we can use questionnaires filled by the patients, or other databases, too. Otherwise the tuple must be ignored. Generally there are some other possibilities, for example using a global constant, or using the attribute main or the most probable values to fill in the missing values. Replacing the missing value with a global constant (for example "unknown") is not a good choice, because the data mining algorithms may operate with this value as a new concept. In medical domain it is neither suggested to replace the missing values with the attribute main or the most probable values of the field, as it can happen that this parameter would predict an illness or an adverse event which we are just analyzing.

The problem of noisy data can occur for a number of reasons: random fault during recording, diverse unit of measures of laboratory values or leaving the default value in the field. The default values in many cases can cause problems, because for example seeing a 0 value in the field of a laboratory parameter, it cannot decide whether this means the absence of the examination or the absence of a substance. Data outside the domain can be corrected manually, or deleted. The outliers can be detected for example by clustering. Outliers in medical databases may draw the attention of the analyzer for example to unusual responses to various medical treatments. Some possibilities, such as smoothing by binning or by regression are not recommended if the data mining algorithm works on medical data.

Redundant data is mostly generated by the aggregation of several different database systems. For example, physical parameters of patients are usually stored in more than one database that we intend to integrate. In this case we must choose one to be the source system, and ignore the rest. Comparing the correspondent data of the different databases, we may find inconsistency. The difference between the values may also derive from a temporal change. In this case new information can be obtained from time series data. For this purpose each piece of information can be placed in a new database or a data warehouse accompanied with a timestamp, and then evolution analysis can be executed for finding hidden patterns, trends.

The major problem of data mining in the health care domain is that a huge amount of data is stored as simple, unstructured text. That is how physical status or case history is stored, for example. The analysis of textual data requires significantly different algorithms than the ones used for the analysis of continuous, interval-scaled, binary, ordinal or nominal data. So it is recommended to transform this data into some structured form. This can only be achieved with the help of doctors or other experts, because of the complexity of the terminology used.

Similar to other data pre-processing activities, data mining applications working on medical data also require some data transformation steps. In situations, where we are not concerned with the accurate value, only the characteristics of that data we have to generalize. Such a typical situation may result, for example, from blood-pressure values or from laboratory values. Likewise we can construct new features from the given set of attributes (for example calculate BMI value from height and weight), or carry out normalization (for example normalization of the laboratory parameters, which are stored in different units of measure).

5. Type of mined knowledge

What kind of patterns can be mined in the domain of health care? There are several different types of medical data, and the questions they imply are equally diverse. As a result of this we can say that all types of knowledge can be mined in order to answer these questions. For example illnesses can be characterized or discriminated by mining class or concept descriptions. Association analysis is used to identify relations between parameters, such as identifying possible risk factors or effects of medications. Using classification algorithms patients can be grouped based on their risk factors. With these results evolving of an illness can be predicted at new patients from their parameters. Cluster analysis is useful for grouping the

Hungarian population, and with these results international treatment proposals can be revised. Evolution analysis is one of the most useful data mining functionalities considering this area. In hospital information systems a lot of time-related data has been aggregated up to the present. Using this data we can model regularities or trends in course of illnesses, changing of laboratory parameters, and so on.

6. Research on osteoporosis

Based on these principles we are carrying out researches in the area of osteoporosis. So far we have collected data of about 20,000 persons, who were referred for examination with suspicion of osteoporosis. Naturally, a number of these patients later proved to be healthy, so we have data of healthy and ill persons, too. The examined persons were coming from the region of Veszprém county and Fejér county. From this huge amount of data 1000 representative patients were selected for further examination.

The available data of patients is diverse. We have data concerning the personal and familial history of the patients, for example birth weight, fractures of bones, medication, previous illnesses, and illnesses of relations. We can also use many possible risk factors, such as lifestyle (alcohol, sport, smoking), parameters of present and previous work and composition of drinking-water. Probably the most valuable data is the results of densitometry examinations for years back. Moreover we have data of laboratory results, data about proposed treatments and follow-up information. In the future we would like to examine the DNA of the patients in connection with osteoporosis.

All this data was stored earlier in paper form. So after getting familiar with the application domain our first task was to provide possibility for recording this data in a database system. Parallelly with this recording the preprocessing of pattern discovery process has also started.

Seeing the methods, we are performing association analysis for finding out the connection of osteoporosis and the possible risk factors, and the connection of densitometry values and fractures. With the classification algorithms we are grouping the patients into three categories, namely osteoporosis, osteopenia and healthy status. In this project we base the clustering methods on phenotype, genotype and examination results. We are planning a number of evolution analysis, including the examination of the change of density of bones in time, and we are searching for other regularities and trends.

7. Summary

In the last decades the quantity of data stored in various information systems increased considerably. Data mining is one of the most popular methods to analyze this huge amount data. Hospital information systems and other medical databases also store valuable data, raising a need for knowledge discovery. Medical data is diverse and offer numerous analysis possibilities. So we can mine class or concept description of medical terms, searching for association rules, classifying patients and predict medical events at new patients, searching for clusters from different points of view, and carry out evolution analysis based on time-series data. As we have seen, this process starts with the analysis of the available data. Which analysis differs significantly in the medical domain from the general preprocessing principles used in other areas.

References

[1] Jiawei Han and Micheline Kamber: Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers, August 2000. ISBN 1-55860-489-8

[2] H. Galhardas, D. Florescu, D. Shasha, E Simon, C-A. Saita: Declarative Data Cleaning, Language, Model, and Algorithms, Proc of the 27th VLDB, pages 307-316, Rome, Italy, 2001

[4] Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE.: Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse, Proc AMIA Annu Fall Symp. 1997

[5] Tsumoto, S.: Knowledge discovery in clinical databases, Proceedings of the 11th International Symposium on Foundations of Intelligent Systems, 1999.

[6] Tsumoto S.: Clinical Knowledge Discovery in Hospital Information Systems: Two Case Studies, PKDD2000, Springer Verlag, pp.652-656, 2000.

[7] M. Last, O. Maimon, A. Kandel: Knowledge Discovery in Mortality Records: An Info-Fuzzy Approach, Medical Data Mining and Knowledge Discovery, Vol. 60, 2001

[8] P. Fazi, D. Luzi, F. L. Ricci, m Vignetti: The Conceptual Basis of WITH, a Collaborative Writer System of Clinical Trials, ISMDA 2002 p. 86-97.

[9] Fogarassyné Vathy Á., Fogarassy Gy.: Egészségügyi adatok előkészítése elemzések céljából, Informatika és Menedzsment az Egészségügyben, 2003/8, p. 36-41

Postal address

Ágnes Vathy-Fogarassy
Department of Mathematics and Computer Science
University of Veszprém
8200 Veszprém, Egyetem u.10.
Hungary