MEDICAL DATA MINING AND KNOWLEDGE DISCOVERY

*Nafisseh Heiat, College of Business, Montana State University-Billings, 1500 University Drive,
Billings, MT 59101. 406 657-2224, nheiat@msubillings.edu*

ABSTRACT

Improved data and information handling capabilities have contributed to the rapid development of new opportunities for knowledge discovery. Interdisciplinary research on knowledge discovery in databases has emerged in this decade. Data mining, as automated pattern recognition, is a set of methods applied to KDD that attempts to uncover patterns that are difficult to detect with traditional statistical methods. Medical data has a lot of information buried within it that will reveal patterns relating to successes and failures in clinical operations. Data mining by discovering these patterns could provide new medical information. This paper will provide an overview of data mining methods and their application to healthcare. This will include discussion of data preparation, methods, constraints, and challenges as well as consideration of the opportunities that data mining may provide for improved prediction, explanation, and discovery of causal structures.

INTRODUCTION

Knowledge Discovery in Databases (KDD) may be defined as the process of finding potentially useful patterns of information and relationships in data. More and more healthcare organizations are storing large amounts of data about patients and their medical conditions. As the quantity of clinical data has accumulated, domain experts using manual analysis have not kept pace and have lost the ability to become familiar with the data in each case as the number of cases increases. Data visualization techniques can assist in the manual analysis of data, but ultimately the human factor becomes a bottleneck as an organization using a large database can receive hundreds or even thousands of matches to a simple query.

Patterns are evaluated for how well they hold on unseen cases. Databases, data warehouses, and data repositories are becoming ubiquitous, but the knowledge and skills required to capitalize on these collections of data are not yet widespread. Innovative discovery-based approaches to healthcare data analysis warrant further attention.

There are situations where healthcare organizations would like to search for patterns but human abilities are not well suited to search for those patterns. This usually involves the detection of "outliers", pattern recognition over large data sets, classification, or clustering using statistical modeling. Most databases are not set up to allow the data manipulation that these types of tasks require. Serious computational and theoretical problems also exist with performing data modeling in high-dimensional spaces and with massive amounts of data.

The need to discover the knowledge buried in clinical and non-clinical data has led to the concept of data warehousing. This is where clinical data is archived into a database dedicated to providing healthcare providers with online data for medical analysis. Data mining takes clinical analysis one-step further by automating the process of discovering patterns or knowledge in a data warehouse.

Knowledge Discovery Life Cycle

Data Preparation-The most important requirement for data mining applications in healthcare is domain knowledge or understanding the complexities of data. The quality of the results will depend on the quality of the data and what data are being included or excluded in the models.

Pre-Mining- To avoid many of the issues associated with data preparation, some authors suggest "pre-mining" the data first in an attempt to understand the data before warehousing it. Pre-mining can provide an understanding of what we are going to warehouse and how we might use it before we actually warehouse it. Then, once the data is warehoused, we are able to mine the data and analyze it in depth. This can be accomplished through "concentric design" and by pre-mining with prototypes.

Data Selection- The data warehouse may contain data not needed for data mining. Therefore the next step in data mining is to select the target data.

Data Transformation- After data selection, the user needs to perform transformations on the data. Transformation methods are the organization of data in a particular manner. They may include conversion of one type of data to another and the definition of new attributes.

Data Mining- In this step the miner uses one or more techniques to extract the desired information. This step includes searching for patterns of interest. The techniques used may be classification, rule induction, regression analysis, clustering, sequence modeling, dependency, or link analysis or a combination of two or more of these techniques

Result Interpretation- The final step in the data mining process is the analysis of the mined information with respect to the end-user's goals. During this step the miner must also decide how to best present the results to the decision-maker if the miner is not the decision-maker.

Types of Data Mining

Predictive versus Descriptive- The key to data mining is the building of data models. Data mining utilizes two types of data models, predictive and descriptive models. Predictive models can use historical data to predict results or outcomes using other concurrent data. Descriptive data mining models describe patterns in existing data that are useful to guide decisions.

Verification-driven data mining- Verification-driven data mining methods are limited to verifying a user's hypothesis. The three data mining operations associated with this method are, query and reporting, multidimensional analysis, and statistical analysis.

Discovery-driven data mining- Discovery-driven methods find new patterns in the data. The discovery-driven data mining operations include predictive modeling, database segmentation, link analysis, and deviation detection.

Data Mining Tools

A data-mining tool should provide a range of knowledge discovery techniques in order to address the widest range of problems. In addition to assessing a tool's effectiveness based on algorithms, the following criteria should also be applied:

- Effective support for all the stages of a data mining life cycle; data preparation, pattern discovery and pattern deployment.
- Support for multiple pattern discovery techniques.
- Connectivity, scalability and performance.

Essentially, the data mining techniques will fall into three basic groups; discovery, prediction, and forensic:

- Discovery association rules involve analyzing exiting data to find natural groupings. This is a process where you find the rules that enable you to correlate the presence of one set of items with another set of items.
- Prediction techniques provide you the opportunity to give an example of the relationship you want to predict or model. Common data-driven modeling techniques include regression, neural networks, rule induction, and decision trees. These types of algorithms essentially "learn" the relationship among the inputs and outputs from example data.
- Forensic techniques allow the user to apply extracted patterns to find unusual data elements. To discover the unusual, the first step is to find the norm and then detect the deviations from the norm within a given threshold

CHALLENGES AND ISSUES IN HEALTHCARE DATA MINING

Information Technology Infrastructure- In general, the healthcare industry lags far behind other industries in terms of information technology expenditures. Lack of information technology sophistication and some historical clinician skepticism have hindered the ability to analyze data adequately. As healthcare continues to become more complex, the industry needs to find an effective means of evaluating its large volume of clinical, financial, demographic, and socioeconomic data. Creating clinical data warehouses or data marts, to make data accessible for analysis, would be the first step in the right direction.

Data- As with any large data warehousing and mining endeavor, the degree to which an organization reaps the benefits of outcomes measurement depends on how it resolves a host of issues. For instance, one of the biggest strengths of outcomes measurement, the ability to view data in the aggregate, can be a pitfall if it discourages consideration of cases on an individual level. Any time you make broad, generalized statements about data, you're liable to miss specific cases. In the medical field, overemphasizing aggregate data can have dire consequences for a patient

Quality Assurance-Since the quality of the data in the data warehouse affects the quality of the decisions being made, it is essential to use data quality management methods in the prototyping phase before building the warehouse. In addition, it is important to have an ongoing data quality program.

Segmenting versus Sampling- The issue is whether to use all the data, a sample of the data or a segment of the data in your analysis. Both of these techniques reduce the data sizes you work with and you may end up with entirely different results because of information loss and distortion

Privacy and Access- Any time an organization deals with customer data, privacy and security become paramount. That is especially true in the healthcare industry where medical records are highly sensitive.

End Users Getting buy-in from the end user can be another thorny issue for organizations implementing a data mining project. Convincing users to surrender peacefully their standard modes of operation for a new technology is never easy. Doctors might be even less tolerant to forced change than most users since they're accustomed to a high degree of professional autonomy.

Inadequate tool support- Most data mining tools support only one of the core discovery techniques. The tools must support the full knowledge discovery process and provide a user interface suitable for business users.

Scalability- Current tools cannot handle vast quantities of data. Progress is being made toward using massively parallel and high-performance computing systems to help deal with large databases.

CONCLUSIONS

This paper has provided an introduction to the concepts of knowledge discovery and data mining in healthcare industry. The review of literature presented is intended to highlight the factors perceived as contributing to the success of the knowledge discover approach. It is evident from the literature that the success of knowledge discovery is largely determined prior to the actual data mining activity, i.e. during the activities performed in the production of the cleaned data. Healthcare organizations should therefore pay special attention during the pre-mining and selection activities in order to ensure that the target dataset actually contains relevant and usable data, and that these data are in a form suitable for use.

It is also apparent from the literature, that if the selected data does not contain appropriate data within the context of the domain of the investigation, the use of data mining tools cannot generate useful information. However, the use of data mining tools may allow this conclusion to be reached more quickly than might ordinarily be the case.

Finally, I must emphasize that; data mining approach should be used in conjunction with traditional approaches, not in direct competition with them. The results obtained from data mining must always be thoroughly investigated and tested by appropriate statistical methods.

References available upon request.