Knowledge Discovery and Data Mining in Healthcare: Challenges and Issues

Abbas Heiat, College of Business, Montana State University-Billings
1500 N. 30th Street, Billings, MT 59101, 406-657-1627, aheiat@msubillings.edu

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.

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. In health care, pattern recognition has long been linked with expertise. 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. 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 health care 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. 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 knowledge.

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. Data mining tools enable goal driven knowledge discovery. For example, instead of the user asking for a report of patients with congestive heart failure, the provider can ask for patterns leading to a lower hospital admission rates for these patients.

Challenges and Issues

Infrastructure- In general, the healthcare industry lags far behind other industries in terms of information technology expenditures. Therefore, healthcare industry's information technology infrastructure is underdeveloped. Lack of information technology sophistication and some historical clinician skepticism have hindered the ability to analyze data adequately.

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. A doctor relying on clinical practice guidelines based on statistical analysis of outcomes might never think to check for glaucoma in a 29-year-old patient, because it's not standard procedure. Exacerbating that danger is the need of many large organizations to reduce all symptoms and descriptions to the lowest common denominator to make data mining work. Every hospital has its own system for coding and record keeping, which can cause difficulties when they attempt to merge their data. The quality of the outcomes measurement analysis also depends on the type of data being examined.

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. An intelligent database quality management system According to Parsaye, errors often manifest themselves as anomalies or exceptions to the expected patterns. He believes that a fair amount of the anomalies in many large databases will be due to errors in the data. How this is handled in practice is that anomalies are the flip side of rules. Once a set of rules is extracted from a large database, the rules can be applied to the same data or new data to see which data is anomalous. The literature points out the importance of using a moderate dose of each of the error detection components listed above in order to determine the key areas of concern. By neglecting a component, the most important problem in the data may be overlooked.

Segmenting versus Sampling-

Assume a health insurance company has a data warehouse and wants to find patterns for patient claims. It doesn't make sense to analyze the entire warehouse because there are numerous different types of illnesses and different types of treatments. Analyzing the entire warehouse may tell the user less than analyzing a segment of the warehouse in a case like this. The best way to analyze this type of situation is to use segmentation to analyze the claims for a given illness.

The need for segmentation becomes even clearer when predictive modeling is considered. This is because you probably want to base your prediction on what has been most similar in the past to what is being considered. The reason this works is that if a pattern holds strongly enough in the entire database, it will also hold strong in the segmentation of a database.

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. To protect themselves and their patients, some healthcare organizations have developed innovative security strategies. Aside from building firewalls and encrypting data, some health-care providers safeguard patient data by limiting who has access to it. Some healthcare organizations, for instance, includes patient identifiers in their data warehouse, but let only a team of doctors and administrators look at the data. If other doctors have a query, they must submit it to the quality improvement team. There are providers who sidestep the security issue altogether by stripping their databases of any patient identifiers. Only an account number is used to link elements of a patient's record. Leaving the data anonymous ensures that employees can access the database through the intranet. Doctors mine database themselves using query tools to answer what-if questions and determine the most appropriate treatments. Removing patient identifiers from the database makes sense for organizations interested in looking at their data in the aggregate. But it also limits what can be accomplished with the data. If no names are included, providers cannot find and alert patients whose lab results or vital signs deviate from the norm, indicating that they are at risk for certain medical conditions.

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. One way to foster user acceptance is to keep the data model as simple and easy to understand as possible. End users will never agree to change their procedures if they don't understand how the system works. But the surest way to circumvent end-user resistance in this is to include users from the beginning and listen to their feedback. After seeing the value of data mining for pneumonia patients, the caregivers in one healthcare organization embraced the concept. They were reassured that this is not about good and bad doctors, it's about good and bad processes. The premise is that physicians want to help patients. You need to convince the doctors that the project is good for patients.

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.

References available upon request