DATA MINING IN MEDICINE

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

ABSTRACT

Every day vast quantities of data are generated in the health care industry. It is these mountains of information that have been collected by hospitals concerning their patient's physical diagnoses, which has formerly been overlooked and untapped. By examining medical data using relationship and pattern analysis, various data mining techniques are now being implemented to assist physicians and clinical researchers in the war against diseases. This paper will examine what is data mining, explore the new emerging variations of data mining techniques, examine the uses of artificial intelligence (AI), and any negative issues associated with data mining processes.

INTRODUCTION

Today the computer has become an integral tool for the collection and retrieval of data, and while technology has brought substantial changes into the area of data analysis, there is an "illusion that improved availability of data, automatically leads to improved knowledge about the world."[1] According to a quote from Mark Last and Abraham Kandel "we still live in a society where "data is *rich* and knowledge is *poor*" [1]. Thus while we have major sources of information, it is through new emerging variations of Data mining, which has enabled the search for trends and patterns in disease progression.

What is Data Mining?

A data query answers specific questions about historical data, data that describes events that have already happened. Data mining on the other hand, predicts answers to questions based on partial data. Data mining can also answer questions you might not think to ask by finding significant "hidden" categories of events in your data, which you might not have seen, as significant. Data mining uses software to filter through large amounts of raw data in the hope of acquiring new or previously unknown information. Before the overall data mining process begins all collected data must first go into a data warehouse. As the data pours in from so many different sources, before it can be used the data must be unified into a single data store and then cleansed of any errors or redundant data. Once the data in the warehouse is ready, the next step is to extract from it a data mart. Here a subset of the data warehouse is focused and optimized for the most efficient discovery of a particular type of information. After the formation of the data mart, data is now ready to be analyzed through mining applications.

Data mining is simply the act of building a model (through a set of examples or a mathematical relationship) based on data from situations where the answer is known and then applying the model to other situations where the answers aren't known. Data mining applications build and apply these models using algorithms, or mathematically based problem-solving methods. The types of modeling necessary to extract results are:

Decision-tree; this is a direct mining approach, which is good at classifying records into a small number of predefined categories. This model uses a series of questions that branch off depending on the answers.

Rule induction; is the extraction of useful if-then rules from data, based on statistical significance.

Genetic algorithms; are optimized techniques based on the concepts of genetic combination, mutation, and natural selection.
Clustering (or nearest neighbor); a direct mining approach, in which records are classified into smaller predefined categories, by finding groups of records that are similar to one another.
Artificial neural networks; Non-linear predictive models that learn through training and resemble biological neural networks in structure.
Data mining is an ongoing process that involves a lot of analysis and refining along the way, and thus no one single model offers the optimal solution. As some models are better than others depending on the type of data, by using more than one type of method, this can yield the best results.

New Data Mining techniques

At this part we examine the data mining techniques that are currently applied to the health industry.
Text Data Mining: The first data mining process for medical applications is called Text Data Mining (TDM). According to Marti Hearst "In TDM the researcher seeks relationships between the content of multiple texts and then sets about linking this information together to form a testable hypothesis about new information" [3] Marti Hearst also says that "TDM ought to be able to help researchers find all possible linkages in published research findings, even across disciplines." [3] Researching medical journals for new hypothesis of cause and effect is an ideal goal of what text and data mining should be able to accomplish.
Intelligent Data Mining: The second technique is called Intelligent Data Mining or IDM; this is a combination of data mining and knowledge acquisitions from experts. IDM can be broken down into two parts, Interest-driven, and Data-driven data mining.
Interest-driven data mining is composed of seven core phases. The first phase is the acquisition of domain knowledge through structured expert interviews. The next phase is the formulation of business questions, necessary to structure the data mining results. The third phase is the refinement of business questions through relevant concepts, which correspond to attributes in the database. The fourth phase is the transformation of business questions into data mining queries, through the mapping of an object, to one or many data mining methods or statistical tests. The last three phases are the execution of the data mining queries, the processing of data minding findings, and the transformation of data mining results into answers.
Interest-driven analysis tends to overlook unexpected patterns in the data, and to avoid this problem the Data-driven data mining is also required. Through the combination of Interest-driven and Data-driven data mining, an Intelligent Data Mining process is formed for the evaluation and modification of measures and discovery of new measures. [6]
Image-content database query: The third data mining process presently being applied in the health care industry is called Image-content database query. This is a new technique that enables a computer to quickly sift through both medical images and patient records. Here Data-mining techniques capable of exploring this raw data can discover pixel patterns and correlate health factors useful to medical researchers seeking cures and clues to breast cancer. [7]
As this data mining process seeks to discover common and indicative pixel patterns among benign and the malignant tumors recorded in medical archives, further data mining should discover other useful clues for the diagnosis and screening of breast cancer.
Knowledge Discovery in Database & Knowledge Quality Assurance: The fourth technique is a combination of two processes, Knowledge Discovery in Databases (KDD), and Knowledge Quality Assurance (KQA). KDD begins with a mass of raw data striving to generate knowledge and according to Matheus, Chan, and Piatetsky-Shapiro KDD is defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". [8]

As physicians strive to integrate scientific based results (i.e. laboratory results) with subjective clinical information and patient history, a new emphasis is placed on medical decision support and medical reasoning through Knowledge quality assurance (KQA).

Artificial Intelligence in Medicine (AIM): As computers are very good at following rules, if we can explain our rules and patterns to computers, then we can design systems which can follow those rules. While it is much more difficult to teach computers how to find a pattern, however with limited boundaries, one can teach computers to identify patterns, extract rules, and implement them. Presently different AIM systems are being developed in an effort for medical knowledge and diagnosis. Today there is a mixture of technologies used to design intelligent agents, and all of them use some combination of statistical operations, artificial intelligence, machine learning, inference, neural networks, and information technologies. These medical agents can monitor patient data and sensors over time and decide if a situation is becoming critical before the threshold is reached. In other words, they can avert a crisis by identifying it as it starts to develop.

Pros and Cons of New Medical Processes

While the advancement of new data mining processes is providing benefits, there are some negative aspects such as quality data, privacy, and cost.

The first item of concern in any data mining operation is the quality of raw data. While a data warehouse does perform data unification and cleansing activities, it is a management responsibility to insure that medical forms, such as protocols, do not contain misleading information or missing information. Another important component of medical data is images. A medium size hospital handles about one million images per year [9], and it stands to reason that with that large number, errors will happen. It is the reduction of errors that is required to eliminate false data, or maybe even false diagnosis.

Another area of concern is the threat of "loss of privacy". Here with new technologies requiring more and more data, there are many negative concerns such as:

- Who has access to my medical records?
- How will my privacy be maintained?
- Will I be diagnosed with someone else's records?

The last but not least concern is the cost involved. In today's economy with more people unable to afford health insurance, who is going to pay for all these technological advances? Potential requirements for implementing quality medical service through data mining are:

- New larger data warehouses, to store all the raw data.
- Computers and all related machinery will require huge investments.
- The expense of software, from leasing or upgrading is another huge cost.
- The cost of hiring and training IT workers is enormous.
- The cost of insuring adequate quality control will be requiring another huge investment of dollars.
- The rising cost of healthcare industry, may make these new technological diagnosis and procedures available only for the rich.

Conclusion

By utilizing data mining techniques we are now able to sift through the mountains of medical research records that have accumulated over the last hundred years. As the potential benefits of new data mining processes are explored, the negative cost incurred will also have to be balanced. we have only mined the tip of the mountains of raw data, and the new discoveries awaiting us are worth the cost.

References are available upon request.