Medical Data Mining Applications and its Uses

Gayathri. P, Dr. N. Jaisankar
School of Computing Science and Engineering
VIT University, Vellore – 632014, Tamil Nadu, India.

Abstract

Data mining is considered as a new method for the analysis of data. It is used to discover new useful knowledge. Data mining help the researcher to have deep insight of understanding of large information repositories. Data mining can reveal new health care knowledge and in turn can support strong clinical and administrative decision making from large clinical data bases. This paper first summarizes about data mining in general. Then speaks about data mining in KDD process and then about the uniqueness of data mining in health care applications compared with other applications. Then briefly summarizes various data mining algorithms and their usage guidelines. Then introduces the usage of data mining in health care. Successful data mining application in health care can support both clinical decision making and administrative decision making. This paper concludes with the benefits associated with the clinical use of data mining by medical experts.

Key words: clinical data bases, Health care, decision making

1. Introduction

The practice of using concrete data and evidence based medicine has existed for many centuries. Data mining applications in medicine and public health is relatively young field of study [1]. Data mining is considered as a new data analysis method. It can discover some useful knowledge from large data bases. Now-a-days lots of data is being collected and stored at enormous speed. Traditional data analysis techniques are infeasible. Human analyst may take time to discover useful information and much of the data is never analyzed at all. Data mining is considered as automated analysis of massive data sets. Compared with other data mining areas, medical data mining has some unique characteristics [1]. Since medical files are related to human subjects, privacy concern is taken more seriously than other data mining tasks.

Data mining originated as interdisciplinary field of statistics and machine learning and then advanced from these beginnings to include artificial intelligence, pattern recognition, high performance computing, data base technology, visualization etc [1]. The term data mining can be defined in many ways [3]. One way of defining the term data mining is “easy extraction of hidden, true, previously unknown and worthy information from data in large data repositories”. Data mining can be referred by many other names like knowledge discovery in data bases (KDD), knowledge extraction, data analysis, data archeology, data dredging, information harvesting etc [3].

2. Data mining in KDD process

Researchers believe data mining as one of the step in KDD process (Fig – 1). Knowledge discovery in databases (KDD) is organized into data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation [1, 3].
3. Data mining in health care and its applications

Health care data are primarily generated from the delivery of patient care. Therefore, medical data mining involves both privacy issues and legal issues. Data mining in health care is essential because nature of medical data is incomplete as different person suffering from same disease may not undergo same kind of test and diagnostic procedure because of the factors like age, family history, work environment etc. Too many diseases are now available for decision making and there is increased demand for health care services like creating disease awareness. The onset of disease has to be detected in a cost-effective, non-invasive and painless way. The quality of medical data is inferior because of missing values [11] and Hospital Information System (HIS) or database designed for financial or billing purpose.

Today, health care industry generates large amounts of data about patients, hospital resources, diagnosis of diseases, electronic patient records, medical devices etc. This large amount of data is to be processed for knowledge extraction that enables support for cost-savings and decision making. Data mining introduces a set of methods that can be applied to this preprocessed data to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making clinical and administrative decisions. But however at last the decision rests with health care professionals.

3.1 Medical Applications:

Data mining application in health care is mainly used for clinical decision making and administrative decision making. This section describes the usage of data mining for various medical applications.
3.1.1 Diagnosis: Data mining can assist in decision making with a large number of inputs and in stressful situations. It can perform automated analysis of Pathological signals (ECG, EEG and EMG) and Medical images (mammograms, ultrasound, X-ray, CT, and MRI). Examples: Heart attacks, Chest pains, Rheumatic disorders, Myocardial ischemia using the ST-T ECG complex and Coronary artery disease using SPECT images.

3.1.2 Therapy: Based on modeled historical performance, data mining can select best treatment plans. Examples: Using patient model, predict optimum medication dosage: e.g. for diabetics and Data fusion from various sensing modalities in ICUs to assist overburdened medical staff.

3.1.3 Prognosis: Accurate prognosis and risk assessment are essential for improved disease management and outcome. Examples: Survival analysis for AIDS patients, Predict pre-term birth risk, Determine cardiac surgical risk, Predict ambulation following spinal cord injury, Breast cancer prognosis.

3.1.4 Biochemical/Biological Analysis: Data mining can automate analytical tasks for urine and blood analysis, tracking the level of glucose, determining level of ion in body fluids, pathological condition detection. Examples: Survival analysis for AIDS patients, Predict pre-term birth risk, Determine cardiac surgical risk, Predict ambulation following spinal cord injury, Breast cancer prognosis.

3.1.5 Epidemiological Studies: Study of health, disease, morbidity, injuries and mortality in human communities. It can discover patterns relating outcomes to exposures, study independence or correlation between diseases and analyze public health survey data. Examples: Assess asthma strategies in inner-city children and Predict outbreaks in simulated populations.

3.1.6 Hospital Management: Optimize allocation of resources and assist in future planning for improved services. Examples: Forecasting patient volume, ambulance run volume, etc. and predicting length-of-stay for incoming patients.

4. Data mining algorithms and guidelines

Researchers must have a deep insight of understanding of various data mining algorithms and their functionality. Data mining algorithms can be classified into two categories. First category is called as descriptive or unsupervised learning and second category is called as predictive or supervised learning [3, 5, 7, 8, 9]. In descriptive data mining algorithms class labels of training data are unknown. Clustering and Association rule mining comes under this category. In predictive data mining algorithms training dataset are accompanied by class labels and new data is classified based on training set. Classification and Regression methods come under this category.

Three of the most widely used data mining algorithms in health care are classification method, clustering and association rule mining. Each of these methods are described below along with the guidelines for their respective use.

4.1 Classification

Classification method is used to predict the class label of a categorical attribute. Class is the dependent attribute in which users are most interested. The value of this dependent attribute is predicted by using the set of independent attributes. In the medical field, classification method can be used to define the diagnosis procedure and prognosis prediction based on the symptoms and health conditions of the patient.
Classification method consists of two steps. They are labeled as learning step and classification step (Fig – 2). Learning step takes training data as input and builds a classifier which generates the classification rules. In classification step, test data is used to check the accuracy of the classifier. Predictive accuracy of the classifier is estimated. Accuracy of the classifier is the percentage of test data whose class labels are correctly classified by the classifier. If the accuracy of the classifier is considered acceptable then it is used to classify the class labels of future data tuples.

4.1.1 Classification Guidelines

Before applying classification algorithm to a data set, it is must to identify the relevant and irrelevant attributes. For example age and date of birth are relevant attributes. That is, one can be derived from other. Such attributes need to be identified and one among them has to be removed from the dataset. The totally irrelevant attributes should be identified because these attributes may act as noise and slow down the efficiency and accuracy of classification method. For example sex attribute on prostate cancer data mining is irrelevant.

Classification methods have high prediction power. That is why it is preferred in medical data mining. Many classification algorithms exist. But there is no one best algorithm for any data set. If the number of attributes is changed by the attribute selection method, then the classification algorithm for a data set is also changed [12, 13]. Thus it is preferable to apply many classification algorithms to the training data set. The output of the classification method should be tested with the testing data to measure the accuracy of the model. At last the best algorithm should be selected for future prediction activities.

4.2 Clustering

Cluster is the collection of similar data objects. Objects in a cluster are similar to each other and dissimilar to the objects in another cluster (Fig – 3). Cluster analysis finds similarity between data objects based on some characteristics and grouping them into clusters. It is an unsupervised learning process that occurs by observing only independent attributes in the data set. Clustering doesn’t use class concept. That is why it is preferred for the studies that encompass large amount of data but very little information is known about data.
Good clusters have the property of high intra-class similarity and low inter-class similarity. Quality of the cluster is determined by the method of the similarity measure used. Quality of the clustering method is measured by its ability to discover some or all hidden patterns.

4.2.1 Clustering Guidelines

Clustering is generally preferred when very little information about data is known or nothing is known about the data. That is why clustering is widely applied to study micro array data sequence. Almost every clustering algorithm can handle only numerical attributes. However, most health care data bases have number of attributes of categorical type. Although conversion of such attributes into numeric is possible, such conversion may distort the distance between categories. Some algorithms can handle the categorical attributes without converting them into numerical attributes. For example FarthestFirst and Two-step Cluster analysis methods can handle the categorical data attributes [1].

4.3 Association Rule Mining

Association rule mining otherwise called as frequent item set or pattern mining is often employed to find frequent patterns, associations, correlations among set of items in the data base. In the health care industry, one can use this method to discover underlying relationships among disease symptoms, relationship among health conditions of different patients suffering from same disease and relationship among various diseases. Researchers can derive evidence-based hypotheses about health conditions and symptoms contributing to a disease or complications [1].

Association rule mining algorithms are not evaluated based on its accuracy because every association algorithm mines all association rules. Association mining algorithms are evaluated based on its efficiency to mine hidden rules from large dataset. Efficiency can be improved in many ways. Association mining algorithm efficiency can be improved by using hash functions and tables, by transaction reduction methods [14] and by using sampling methods [15].

4.3.1 Association Guidelines

Concept hierarchy can be used to compress lot of raw association rules like {smoking, drinking, helicobacter pylori infection} \(\rightarrow\) {cancer} into manageable size. The main focus of classification algorithm is towards class label whereas association mining algorithms are mainly used to discover relationship among all attributes. Good association mining method should ignore association rules
which are of meaningless. For example \{cervical cancer\} → \{male\} is a meaningless rule to be ignored by association algorithm. Raw association rules have to be organized by using concept hierarchy.

5. Uses of health care mining

Many have initiated the research work in health care mining. Some of the uses of health care mining are listed below

5.1 Hosting of safety issues and prevention of hospital errors: When health care industry apply data mining methods on their existing data bases, they can discover true, new, useful and life saving information. Data preprocessing can reveal unknown medical errors. By mining patient’s records many safety issues can be flagged and addressed by the hospital management.

5.2 Policy recommendations in public health: Data mining can lead to policy making in public health. Data mining can explore the cause of failing to implement strict sanitary and sterilization measures. It can also explore the advantages of implementing such policies.

5.3 Cost savings at minimal extra cost: Data mining allows health care management and other organizations to get more out of existing at minimal extra cost. For example data mining can be applied to discover fraud in credit cards and insurance claims.

5.4 Early detection and prevention of diseases: Medical practitioners can identify patterns and outliers better by using data mining method rather than just by looking at the tabulated data. For example classification algorithms can help medical experts in early detection of heart diseases.

5.5 Cost effective and painless diagnosis: Some disease diagnosis methods and lab test procedures are painful, costly and invasive to the patients. For example conducting biopsy in women to detect cervical cancer is costly and painful procedure. Medical experts can first use data mining algorithms to decide whether or not to recommend a biopsy for a patient who is suspected to have cervical cancer.

5.6 Adverse Drug Events: Data mining can be used to discover knowledge about drug side effects. Some drugs that have been approved as non-harmful are later detected to have harmful side effects. Data mining algorithms can be used to detect the side effects caused by a drug well in advance.

6. Conclusion

Data mining can help the medical experts to discover new life saving information. When medical institutions apply this on their existing data base, they can uncover new, valid and useful knowledge that otherwise would have remained inert in their data bases. More over data mining can lead to better medical policies like strict sterilization and sanitation, vaccination planning etc. But however before using data mining algorithms an medical institution must formulate clear policies on the privacy, confidentiality and security of patient records. Efficient screening tools reduce demand on costly health care resources. Data mining provides better insight into the medical survey data and help the physicians cope with the information overload. Health care insurers, health care organizations, physicians and patients get more out of this data mining technique. Healthcare insurers can detect fraud in the medical insurance claims, Healthcare organizations can make clear customer relationship management decisions, Physicians can identify cost effective and efficient treatment methods and Patients can receive better and more affordable health care services. With the help of data mining technique, health care professionals can provide better service to the society. However, at last clinical decisions rests with the medical expert. Data mining in health care can change the world.
References


