An Overview of Data Mining In Medical Field

Reema Arora , Sandeep Jaglan

Abstract— **Data mining is a upward area of research that intersects with many disciplines such as Artificial Intelligence (AI), databases, parallel computing , statistics, visualization, and high-performance. The target of data mining is to turn data that are specifics, records, or text which can be processed by a computer into knowledge. Nowadays, the reliance of health care on data is mounting. Therefore, this paper aims to allow the readers to comprehend about data mining and its importance in medical systems. The ambition of studying data mining techniques for the diagnosis and prognosis of various diseases is to identify the well-performing data mining algorithms used on medical databases. The subsequent algorithms have been acknowledged: Decision Trees, Artificial neural networks and their Multilayer Perception model, Naïve Bayes. Analyses show that it is very thorny to name a single data mining algorithm as the most suitable for the diagnosis or prognosis of diseases. At times some algorithms execute superior than others, but there are cases when a grouping of the best properties of some of the algorithms mentioned above collectively results more valuable.**

Index Terms— **Naïve Bayes, Multilayer Perception, C4.5, medical data mining, medical decision support.**

I. Introduction

*Data mining***: -** The term data mining refers to the finding of relevant and helpful information from databases. Data mining and knowledge discovery in the database is a new interdisciplinary field, integrating facts from statistics, machine learning, database and parallel computing .Researchers have defined the term data mining in many ways:
1. Data mining and knowledge discovery in databases, as it is also well-known is the non-trivial pulling out implicit, previously unfamiliar and potentially valuable information from the data. This encompasses a numeral of methodological approaches, such as clustering, data summarization, classification, discovering dependency networks, analysing changes, and detecting anomalies.

Data retrieval in its common sense in database attempts to retrieve data that is stored visibly in database and present it to the user in a way that user can understand it. It does not attempt to extract implicit information
Data mining is the seek for the relationships and global patterns that exist in the large databases but are hidden among enormous amount of data, such as the bond between patient information and their medical diagnosis. This relationship represents important knowledge about the database and the objects in database, if the database is a reliable mirror of the real world registered by the database.

II. Data Mining Process

The data mining process may be intricate and can be divided into the following steps:

1. Domain analysis and data understanding
2. Data selection
3. Data analysis and pre-processing
4. Data reduction and transformation
5. Important attributes selection
6. Reduction of the number of dimensions
7. Normalization
8. Aggregation
9. Selection of data mining method
10. Data mining process
11. Visualization
12. Evaluation
13. Knowledge utilization and evaluation of the results to an appropriate target.

The master's thesis focuses on the ninth- the selection of data mining methods. This selection is based on an in-depth analysis of the methods' performance measured with the use of several metrics, like ROC curves, false or true, optimistic or pessimistic rates.

III. Data Mining Techniques

There are several major data mining techniques being developing and using in data mining projects including *association*, *classification*, *sequential patterns* ,*clustering* and *decision tree*. We will look at those data mining technique in brief as follows:

Classification:

It is the demonstration of data in given classes which is called supervised classification, it uses given class labels to order the objects in the data collection. Classification consider as an imperative task of data mining. Using this approach data must be defined as class label (target attribute). In binary classification, the target attribute has only two possible values: for example, high or low. Multiclass targets have more than two values: for example, low, medium, high, or unknown. Classification can applied into Business modeling, marketing, credit analysis, biomedical and drug response modeling.

Clustering

Clustering is the depiction of data in classes. However, not like classification, in clustering class labels are mysterious and it is up to the clustering algorithm to determine acceptable classes. This is called unsupervised classification. Clustering

Reema Arora, Computer Science and Engineering, N.C. College of Engineering, Israna Panipat, India
Sandeep Jaglan, Computer Science and Engineering, N.C. College of Engineering, Israna Panipat, India

is a collection of data objects, similar data are taking in the same cluster, dissimilar data are taking in different clusters.

Association:

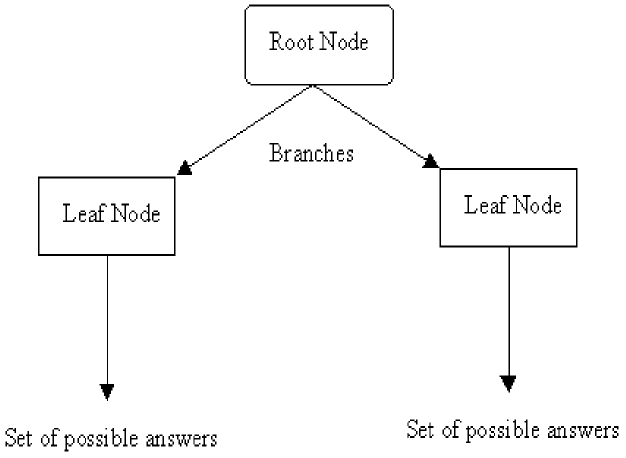
Association analysis is the unearthing of association rules. It depends on the occurrence of transactional data occur together in database, also depends on a threshold called support, and identifies the frequent item sets. Association data mining designed to find association between attributes, generate rules from data sets. The association rule mining role is to arrive at all rules having support≤ minsup (minimum support) threshold and confidence ≤minconf (minimum confidence) threshold.

Sequential Patterns

Sequential patterns analysis is another data mining technique that seeks to discover or classify similar patterns, regular events or trends in transaction data over a business period.

Decision trees:

Decision tree is one of the most used data mining techniques because its model is easy to comprehend by users. In decision tree technique, the root of the decision tree is a simple question or condition that has numerous answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it.



IV. Data Mining Algorithms Applied In Medical System

Data mining also identified as Knowledge Discovery in databases is very often utilized in the field of medicine. The practice of supporting medical diagnoses by robotically searching for valuable patterns undergoes noticeable improvements in terms of precision and response time. This paper shortly describes the most common data mining algorithms and explains the use cases of each of them. The usefulness of the following methods was verified by medical personnel and confirmed by independent experts. The algorithms are as follows:

The C4.5 algorithm

C4.5 is an extension of the well-known ID3 algorithm. The extension includes avoiding data over fitting by shaping how deeply a tree can grow. The C4.5 algorithm is capable of managing continuous attributes, which are vital in case of medical data (e.g. blood pressure, temperature, etc.). Other very familiar aspect – missing values – was also taken into consideration in C4.5. Moreover the algorithm handles attributes with differing costs.

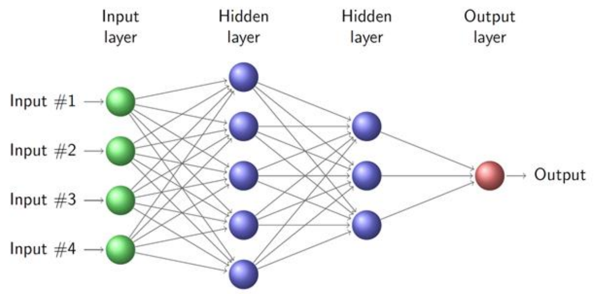
The utility of C4.5 algorithm was widely proven in medicine . This algorithm suits medical data because it copes with missing values. What is more the algorithm handles continuous data which are common among medical symptoms. The efficiency of C4.5 was shown e.g. in breast cancer and prostate cancer classification to generate a decision tree and rules which may be helpful in medical diagnosing process.

A Naïve Bayes

A naïve bayes classifier assumes that the occurrence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For illustration, a fruit may be assumed to be a tomato if it is red in colour, round in shape, and about 2.5" in diameter. This classifier takes all these description to contribute separately to the probability that this fruit is a tomato, whether or not they're in fact related to each other or to the existence of the other features. The Bayes theorem is as follows: Let X={x1, x2,.....,xn} be a set of n attributes. In Bayesian, X is assumed as proof and H be some hypothesis means the data of X belongs to precise class C. To determine P (H|X), the probability that the hypothesis H holds specified facts i.e. data sample X. According to Bayes theorem the P (H|X) is expressed as P (H|X) = P (X| H) P (H) / P (X) As Naïve Bayes classifiers depends on the precise nature of the probability model , so it can be trained very efficiently in a supervised learning setting. Here independent variables are considered for the principle of prediction or occurrence of the event. It has been shown that Naïve Bayes classifiers often work much better in many complex real world situations. An assistance of the Naïve Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

V. Neural Networks

Artificial neural networks (ANN's) are based on study of a human brain. Human brain is a very complicated web of neurons. Analogically, the ANN is an interconnected set of neurons – simple units which are of three types: input, concealed and output ones. The path of the network symbolizes the data flow during the course of prophecy. The attributes that are passed as input to the net form a first layer. In medical diagnosis the input to the artificial neural network are the patient's symptoms. The set ***S*** and an output is the analysis: the set ***D***. The concealed neuron practices the outcomes of earlier layers. The connections between neurons have weights assigned to them. Their values are accustomed with the use of appropriate algorithms, like back propagation. It uses a gradient descent to tune network's parameters to best fit the values of the output. The veiled layers add nonlinear characteristics to the network. The output layer may have more than one output node (prediction of different diseases). One of the main advantages of Artificial Neural Networks is their high performance. The drawback of this method is its complexity and difficulty in understanding the predictions.



Multilayer perceptions (MLPs)

Multilayer perceptions are feedforward neural networks skilled with the paradigm backpropagation algorithm. They are called as supervised networks because they require a desire reply to be skilled. They gain knowledge of how to transform input data into a desired response,, therefore they are widely used for pattern classification. With single or two hidden layers, they can fairly accurate virtually any input-output map. They have been revealed to approximate the performance of optimal statistical classifiers in difficult problems. Most neural network applications involve MLPs.

VI. Medical Decision Support System in Data Mining

Data mining applications are presently being applied to two main branches in health care and medicine: Medical decision support system, and policy planning/decision making. A. Medical decision support system MDSS is an interactive Decision support system (DSS) Computer Software, which is intended to lend a hand to physicians and other health professionals with judgment making tasks, such as determining diagnosis of patient data. . The main purpose of modern MDSS is to help clinicians at the point of care. It means, a clinician would interact with a MDSS to help determine diagnosis, analysis, etc. of patient data. It is a decision-support system program that offers employees in detail, purpose, custom-made, and current information on all healthcare conditions. Employees receive the information, implements and support they need from incorporated web, phone, and print based materials. This helps employees formulate more informed healthcare decisions while working with their own physician. An example of how a MDSS might be helpful in medicinal gather from the subset of Medical Decision Support System and Diagnosis Decision Support Systems. A DDSS would obtain the patients data and recommend a set of correct diagnoses. The doctor then takes the output of the DDSS and point out which are relevant and which are not. Another important classification of a MDSS is based on the timing of its use. Doctors apply these systems at point of care to aid them as they are handling a patient, with the time of use as either pre-diagnoses, during diagnoses, or post diagnoses. Pre-diagnoses MDSS systems are used to help the physician prepare the diagnoses. MDSS helpful during diagnoses in reviewing and filtering the physician's preliminary diagnostic choices to improve their final results. And post-diagnoses MDSS systems are used to mine data to derive connections between patients and their past medical history and to predict future events. Features of a Knowledge-Based MDSS Most MDSS are divided into three parts, the knowledge base, inference engine and mechanism to communicate. The knowledge base possessed the IF-THEN rules. The inference engine gather the rules from the knowledge base with the patient's data. The communication mechanism will permit the system to show the results to the user as well as have input into the system. Features of a non-Knowledge-Based MDSS Two types of non-knowledge-based systems are neural networks and genetic algorithm. Neural networks use nodes and weighted connections between them to analyze the patterns found in the patient data to develop the associations between the symptoms and a diagnosis. Genetic Algorithms are based on basic evolutionary processes using directed selection to achieve optimal MDSS results. The MDSS features associated with success include the following:

- It is incorporated into the health care workflow rather than as a separate log-in or screen.

- It is electronic unlike paper-based templates.

- It gives decision support at the time and location of care rather than prior to or after the patient encounter.

- It gives(active voice) recommendations for care, not just assessments.

VII. Characteristics of Medical Decision Support Systems

The Medical DSS's are the type of computer programs that help out physicians and medical staff in Medical decision making tasks. Most of the Medical decision support systems (MDSS's) are equipped with diagnostic assistance module, therapy critiquing and planning module, medications prescribing module, information retrieval subsystem (for instance formulating accurate clinical questions) and image recognition and interpretation section (X-rays, CT, MRI scans) Interesting examples of MDSS's are machine learning systems which are able to create new healthcare knowledge. By analyzing healthcare cases a Medical Decision Support System can produce a detailed description of input features with a unique characteristic of healthcare conditions. It supports may be priceless in looking for changes in patient's health condition. These systems may improve patients' safety by reducing errors in diagnosing. They may also get enhanced medications and test ordering. Furthermore, the quality of care gets better due to the lengthening of the time clinicians spend with a patient. It may be an effect of application of proper guidelines, up-to date healthcare evidence and improved documentation. Moreover, the efficiency of the health care delivery is improved by reducing costs through faster order processing or eliminated duplication of tests.

Examples of Medical *Decision Support Systems*

There exist several Medical Decision Support Systems (MDSS's). They help in early detection of diseases. In this survey a few of the most significant systems are accessible. They are utilized in hospitals. To provide you the idea of Medical Decision Support Systems three sample ones are described: HELP, DX plain and ERA.

HELP

One of the most accepted and advanced Medical Decision Support System is called HELP. It helps the clinicians in interpreting healthcare information, diagnosing the disease of patients, maintaining healthcare protocols and other tasks .In 2003 a new version was released called HELP II. It is equipped with a knowledge database which stores about 32000 emergency cases and a Medical decision support engine. This system contains two assistants called antibiotic assistant and pneumonia diagnostic assistant. The idea of the former is to find the pathogens causing the infection and to suggest the cheapest therapy for patients with e.g. allergies or renal functions. DX plain It is a Medical Decision Support System (MDSS) available through the World Wide Web that assists clinicians by generating stratified diagnoses based on user input of patient signs and symptoms, laboratory results, and other healthcare findings Each healthcare finding entered into DX plain is assessed by determining the importance of the finding and how strongly the finding supports a given diagnosis for each disease in the knowledge base.By Using this criterion, DX plain produces ranked differential diagnoses with the most likely diseases compliant the lowest rank. ERA (Early Referrals Application) The Early Referrals Application (ERA) is one of the newest and most promising Medical Decision Support Systems. This clarification is devoted to detection of different types of cancers in their early stage.

VIII. Data Mining Challenges in Healthcare

One of the most significant challenges of the data mining in healthcare is to obtain the quality and relevant medical data. It is not easy to acquire the exact and complete healthcare data. Health data is complex and heterogeneous in nature because it is collected from various sources such as from the medical reports of laboratory, from the conversation with patient or from the review of medical doctor. For healthcare provider, it is crucial to maintain the quality of data because this data is useful to offer cost effective healthcare treatments to the patients. Health Care Financing Administration helps in maintaining the minimum data set (MDS) which is recorded by every hospital. In MDS there are 200 questions which are answered by the patients at register time. But this practice is difficult and patients face problems to respond the entire questions. Due to this MDS face some difficulties such as missing information and incorrect entries. Without quality data there is no useful results. For successful data mining, complication in medical data is one the significant hurdle for analyzing medical data. So, it is essential to maintain the quality and accuracy data for data mining to making effective decision. Another difficulty with healthcare data is data sharing. Healthcare organizations are unwilling to share their data due to privacy concern. Most of the patients do not want to disclose their health data. Therefore, the Health Maintenance Organization and Health insurance Organization are not distributing their data for preserving the privacy of patient. This poses hurdle in the fraud detection studies in health insurance. The startup cost of data warehouse is very high. Before applying data mining techniques in healthcare data it is important to accumulate and record the data from different sources into a central data warehouse which is a costly and time overriding process. flawed data warehouse design does not supply effective data mining.

IX. Conclusion

Nowadays, massive amount of data is gathered in medical databases Such databases may contain important information contained in nontrivial dependencies among symptoms and diagnoses. By using medical systems the process of revealing such relationships in historical data is become much easier to conduct. This knowledge can be valuable in diagnosis of future cases.

The main purpose of the research was to recognize the most common data mining algorithms, implemented in modern Medical Decision Support Systems, and evaluate their performance on several medical datasets. Three algorithms were chosen: C4.5, Multilayer Perception and Naïve Bayes. For the evaluation five UCI databases were used: heart disease, dermatology diseases, hepatitis, breast cancer and diabetes datasets. Several performance metrics were utilized: percent of accurate classifications, *True or False optimistic* rates, AUC, *Precision*, *Recall*, *F-measure* and a set of errors. The underlying cause for such a research was the fact that no work was found which would analyze these three algorithms under identical conditions.

References

[1] Aftarczuk K., Kozierkiewicz A., The method of supporting medical diagnosis based on consensus theory. Report of Institute of Information Science & Engineering, University of Technology. Wroclaw, 2006 Series

[2] Herron P., Machine Learning for Medical Decision Support: Evaluating Diagnostic Performance of Machine Learning Classification Algorithms, INLS 110, Data Mining, 2004.

[3] W. Lord and D. Wiggins, "Medical Decision Support Systems Advances in Health care Technology Care Shaping the Future of Medical." vol. 6, G. Spekowius and T. Wendler, Eds., ed: Springer Netherlands, 2006, pp. 403-419

[4] J. Han and M. Kamber, Data Mining, concepts and techniques, 1st ed.: Academic Press, 2001.

[5] M. Kantardzic, Data mining: concepts, models, methods, and algorithms: Wiley-IEEE Press, 2003

[6] Chen, M.S., Han, J. and Yu, P. (1996), ``Data mining: an overview from a database perspective.

[7] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996), ``From data mining to knowledge discovery: an overview'', in Fayyad, U., Piatestsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds), Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA.

[8] Amandeep Kaur Mann, Navneet Kaur, Survey Paper on Clustering Techniques, IJSETR, 2278 – 779. [2] Pavel Berkhin, A Survey of Clustering Data Mining Techniques, pp.25-71, 2002

[9] Oded Maimon, Lior Rokach, Data Mining AND Knowledge Discovery Handbook, Springer Science + Business Media.Inc, pp.321-352, 2005.

[10] Jenn-Lung Su, Guo-Zhen Wu, I-Pin Chao (2001). The Approach of Data Mining Methods For Medical Database. IEEE. p1-3.

[11] Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare."Journal of Healthcare Information Management— Vol 19.2 (2011): 65.