International Journal of Advanced Technology in Engineering and Science
Vol. No.4, Issue No. 02, February 2016
www.ijates.com
ijates
ISSN 2348 - 7550

APPLICATION OF DATA MINING IN HEALTH CARE

Kavita[1], Ms Priyanka Mahani[2], Prof. (Dr.) Neelam Ruhil[3]

[1]MTech(CS), Banasthali University, Jaipur

[2]Assistant Professor, [3]HOD, Electronics and Computer Engineering Department,

Dronacharya College of Engineering, Gurgaon (India)

ABSTRACT

In today's world Healthcare area is growing faster with huge amount of data present. The data in healthcare includes patient information, treatment given to them resources provided and much more relevant information. The information present is very large and considerable. Hidden knowledge and pattern from healthcare can be discovered by using various data mining techniques. In this paper we intend to apply some of the classification technique over the healthcare data. WEKA machine learning tool is used to test data set of Hepatitis disease. For performance evaluation accuracy percentage of classifier is considered. The technique with high accuracy is chosen for that particular dataset.

Keywords: Classification technique, Data Mining, Healthcare, Hepatitis disease, KDD, Machine Learning Tool

I. INTRODUCTION

Data mining is to extract meaningful and useful information from the large databases. Data Mining has attracted great attention from various fields due to wide and large data present in these fields. To convert these large data into useful information and knowledge data mining is required. The information and knowledge gained by data mining and their applications can be used in various areas including market analysis, Business and E-Commerce fraud detection, customer retention, production control, Scientific, Engineering, and Health Care etc. Various data mining techniques can be applied in various fields.

This study mainly discusses Data mining applications and techniques in medical field. A detailed survey on data mining applications in the healthcare sector, various data mining tasks, algorithms and techniques in healthcare is carried out in this work. Data mining algorithms can be very helpful in the field of healthcare for prediction and diagnosis of various diseases more efficiently. There are various applications of Data Mining in the healthcare sector and medical related areas.Another question that arises is why data mining tools are required in this area. The data mining tools are required because the method used traditionally by hospitals were too complex and time consuming. To find predictive information for any disease they prepare large databases which are very difficult to handle and manage manually. Thus data mining tools attracted a great attention in this area for discovery of useful information from huge collection of databases. In this paper we mainly focus on data mining techniques that can be applied in healthcare sector for prediction and diagnosis of diseases. For experimental work we collect data available online. Database for Hepatitis disease is selected. The collected data is initially converted into acceptable form so that data mining techniques can be efficiently applied on the dataset. In this work we mainly focus on classification techniques although other data mining techniques like

clustering can also be applied on the dataset. But for now experimental work is carried out only for classification tasks. There are various data mining techniques and tools that are available and currently used in healthcare sector.

II. APPLICATIONS OF DATA MINING IN HEALTHCARE SECTOR

Today in healthcare sector large amount of data is generated that includes patient personal information, hospital resources, diagnosis of disease, patient history, treatment provided etc. These data collected that present in large amount are key resources that can be processed and analyzed for the extraction of knowledge for the purpose of decision-making and cost saving. Data mining applications can be divided into various categories

2.1 Treatment Effectiveness

Data mining applications can be developed for the evaluation of effectiveness of medical treatments. Data mining analysis can be delivered by comparing and contrasting symptoms, causes; course of treatment for a group of patients which were treated for same condition or disease but with different drug regimens to check which treatment or drug is more effective.

2.2 Healthcare management

To better identify and track persistent disease states and high-risk patients, design proper interventions, and decrease the number of admissions in hospital and claims to support healthcare management data mining applications can be developed.

2.3 Customer relationship management

To manage interactions between commercial organizations- banks and retailers and their customers, customer relationship management is a core approach. It is also very important in context of healthcare. With the help of call centers, physician's offices, inpatient settings, billing departments, and ambulatory care settings customer interaction may occur.

2.4 Fraud and abuse

To identify fraud and abuse data mining applications often set up norms and then recognize unusual patterns of claims by physicians, clinics, laboratory or some others. These data mining applications can also throw a light on unsuitable prescriptions or referrals and false insurance and health claims.

2.5 Medical Device Industry

One important point of healthcare system is medical device. This is mostly used for best communication work. Mobile healthcare applications supply a convenient, constant and safe way for monitoring of vital signs of patient. Thus mobile communications and low cost of wireless biosensors have lined the way for development of these applications.

2.6 Pharmaceutical Industry

To manage pharmaceutical firms their inventories and for development of new product and services the technology is being used. For a competitive position of firm and organizational decision-making a bottomless understanding of knowledge hidden in the pharmacy data is essential.

2.7 Hospital Management

A vast amount of data is collected and generated by organizations which include modern hospitals. Thus data mining is applicable for development of Hospital management system. Hospital management involves: services for hospital management, medical staff and patients.

2.8 System Biology

A wide variety of data types with rich relational structure frequently is contained by Biological databases. Thus multi-relational data mining techniques are applied to biological data commonly.

III. TECHNIQUES USED FOR DATA MINING

Data mining techniques are very useful in healthcare; these techniques are helpful in diagnosis and treatment of diseases, resource management in healthcare, fraud detection, customer relationship management etc. Two strategies are used in data mining i.e. supervised learning and unsupervised learning. In supervised learning a training set is there with the help of which model parameters are learned. On other hand there is absence of training set in unsupervised learning, no training set is present therefore learning is modeled with unknown target parameter. The models are in descriptive form which describes the interesting and valuable information present in data. Descriptive and predictive are the two categories in which data mining tasks are classified. The goal of descriptive tasks is to review the data and construction of entire model and find out human interpreted forms and associations. While in predictive task the goal is focused to find out the interesting outcomes. Also it find out there is any relationship present between dependent and independent variables. The descriptive and predictive tasks can be mainly classified as:

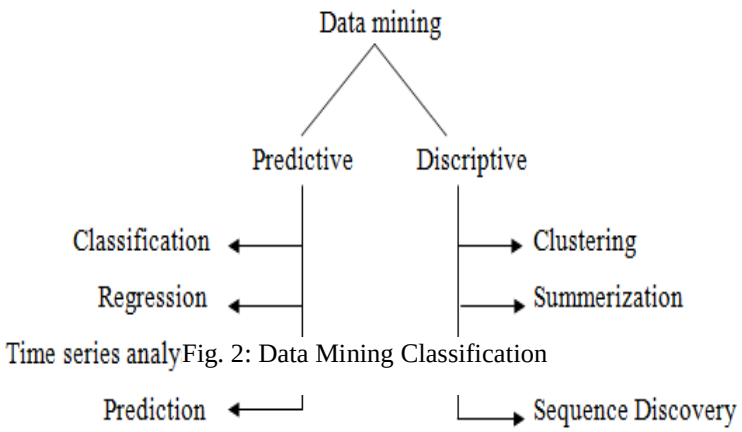

Fig. 2: Data Mining Classification

Data mining classes are as follows:

- **Classification** –It is the task of generalizing a well-known structure to apply to new data. It is the process of identification of a set of categories on the basis of a training set of data containing observations whose

category of membership is known. For example, some of e-mails are classified as “legitimate" and other are classified as "spam" on the basis of content present or on the basis of some other characteristic.

- **Clustering** –Cluster analysis or clustering is the task of alliance more similar objects in same group (known as cluster). It is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- **Association rule learning** – It is an admired and well researched technique for discovering attention-grabbing relations between variables in large databases. It searches for relationships between variables.
- **Regression** – It is the process to find a function which models the data with the least error. It includes various techniques for modeling and analyzing several variables. The task mainly spotlight on the correlation between a dependent variable and one or more independent variables.
- **Anomaly detection** – Also known as Outliner detection or Deviation detection. The task involves the discovery of unusual data records, measures or annotations that might be exciting or data errors that require further exploration.
- **Summarization** –Automatic summarization is the process of reducing a text document using a computer program in order to generate a summary that retains the most significant points of the original document. It provides a more dense demonstration of the data set which includes visualization and report generation. The interest in automatic summarization has increased nowadays due to the increase in information overload and the quality of data has increased.
- **Time Series Analysis**- Time series analyses consist of methods for analyzing time series data consecutively to dig out meaningful statistics and other characteristics of the data. Time series forecasting is to utilize a model to forecast upcoming values based on formerly observed values. Data of Time series includes natural temporal ordering. Additionally, time series models will frequently take advantage of the usual one-way ordering of time so that values for a known period will be expressed as deriving in several ways from past values, rather than from future values. Time series analysis can be applied to continuous data, real-valued, discrete numeric data, or discrete symbolic data.
- **The prediction task**- It is a supervised learning task which works on direct data and there is no explicit model for new instance of class value prediction. Some of the approaches which are used for prediction task are:
    - ❖ Instance-based (nearest neighbor)
    - ❖ Statistical (naive bayes)
    - ❖ Bayesian networks
    - ❖ Regression (a kind of concept learning for continuous class)
- **Sequence Discovery**- Also known as Sequential Pattern mining is a topic of data mining concerned with discovery of statistically significant patterns among data examples where the values are conveyed in a sequence. It is usually assumed that the values are discrete. Sequential pattern mining is a special case of structured data mining.

Each and every technique has its own importance. All of these tasks can be efficiently used in healthcare field. Many of the researchers are currently working on these techniques for various purposes.

IV. DATABASE AND TOOLS USED IN EXPERIMENT

We have practiced Hepatitis disease datasets taken from UCI Machine Learning Repository. WEKA Machine learning tool is used for classification tasks. This study will help the researchers to determine the better results from the available data within the datasets.

Weka consist a variety of machine learning algorithms for various tasks of data mining. These algorithms are useful in two ways whether user can apply the algorithm directly on the dataset or user is free to call own Java code. Weka includes tools for data pre-processing, classification, clustering, regression, association rules, and visualization. It is also compatible for developing new schemes for machine learning. Weka is free open source software having the GNU General Public License.

V. EXPERIMENT RESULT & ANALYSIS

The experimental result and analysis for this study is given in this section. Different classification techniques for this experiment have been applied on Hepatitis disease healthcare datasets taken from UCI repository.

Performance is measured using WEKA which gives the output in terms of TP, TN, FP, and FN. Accuracy is interpreted from the given formula.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Where TP, TN stands for True Positive and True Negative and FP, FN stands for False Positive and False Negative. TP+TN signify percentage of correctly classified instances and TP+FP+TN+FN signifies total of correctly and incorrectly classified instances. In this study three different algorithms are selected for classification. The percentage of accuracy for classification techniques is used as the measurement parameters for analysis. In our work we compare performance of these algorithms on the basis of their accuracy. Although our comparison is completely depends on current dataset of Hepatitis disease on which we perform experiment. The results may vary on different datasets. Table 1 shows classification accuracy based on different techniques applied.

Table 1- Comparison of Accuracy of Classifier

Classifier	Accuracy
Naive Bayes	84.5 %
Decision Table	76.12%
J48	83.9%

VI. CONCLUSION

The experimental results have shown depending on the nature of their attributes and size the classification techniques behave differently on different datasets. At last, the classification technique which has shown the highest accuracy rate and lowest error rate over a dataset has been selected as the best classification technique for the dataset. By knowing the best classification technique over a dataset a set of rules can be generated for

that particular dataset and these rules will complement the healthcare researchers‟ study for intelligent decision making. For future work it is suggested that we can do more experiments on different healthcare datasets using different parameters and techniques.

REFERENCES

[1] Boris Milovic, Milan Milovic, “Prediction and Decision Making in Health Care using Data Mining”, International Journal of Public Health Science, Vol. 1, No. 2, December 2012, pp. 69-78

[2] Dr. D. P. Shukla, Shamsher Bahadur Patel, Ashish Kumar Sen,” A Literature Review in Health Informatics Using Data Mining Techniques”, International Journal of Software and Hardware Research in Engineering, Volume 2 Issue2, February 2014

[3] Hlaudi Daniel Masethe, Mosima Anna Masethe,” Prediction of Heart Disease using Classification Algorithms",Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II,WCECS 2014, 22-24 October, 2014, San Francisco, USA

[4] M. Durairaj, V. Ranjani, “*Data Mining Applications in Healthcare Sector: A Study*”, International Journal of Scientific & Technology Research, Volume 2, Issue 10, October 2013

[5] Monali Dey, Siddharth Swarup Rautaray,” Study and Analysis of Data mining Algorithms for Healthcare Decision Support System”, International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 470-477

[6] N. Abirami, T. Kamalakannan, Dr. A. Muthukumaravel, “A Study on Analysis of Various Datamining Classification Techniques on Healthcare Data”, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 7, July 2013

[7] Parvez Ahmad, Saqib Qamar and Syed Qasim Afser Rizvi, “Techniques of Data Mining In Healthcare: A Review”, International Journal of Computer Applications (0975 – 8887) Volume 120 – No.15, June 2015

[8] Prakash Mahindrakar, Dr. M. Hanumanthappa,” Data Mining In Healthcare: A Survey of Techniques and Algorithms with Its Limitations and Challenges”, Int. Journal of Engineering Research and Applications, Vol. 3, Issue 6, Nov-Dec 2013, pp.937-941

[9] Shubpreet Kaur and Dr. R.K.Bawa,” Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System”, International Journal of Energy, Information and Communications Vol.6, Issue 4 (2015), pp.17-34

[10] UCI Machine Learning Repository. [Online] Available: http://archive.ics.uci.edu/ml/

[11] Weka, “Data Mining Machine Learning Software, [Online] Available: http://www.cs.waikato.ac.nz/ml/weka/