RESEARCH ARTICLE OPEN ACCESS

Predictive Data Mining For Medical Diagnosis: An Overview Of Heart Disease Prediction

Ms. Priti V. Wadal*, Dr. S. R. Gupta**
*(Department of Computer Science, PRMIT & R, Badnera, Amravati
** (Department of Computer Science and Engineering, PRMIT & R, Badnera, Amravati)

ABSTRACT-
As large amount of data is generated in medical organizations (hospitals, medical centers) but this data is not properly used. There is a wealth of hidden information present in the datasets. The healthcare environment is still "information rich" but "knowledge poor". There is a lack of effective analysis tools to discover hidden relationships and trends in data. Advanced data mining techniques can help remedy this situation. For this purpose we can use different data mining techniques. This research paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction. This research has developed a prototype Heart Disease Prediction System (HDPS) using data mining techniques namely, Decision Trees, Naïve Bayes and Neural Network. This Heart disease prediction system can answer complex "what if" queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established.
Keywords **-** Data mining, Heart disease, Decision tree, Neural network, Naïve bayes, Classification, Clustering.

I. INTRODUCTION

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. A wide variety of areas including marketing, customer relationship management, engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, besides others utilize Data mining [1]. Numerous fields associated with medical services like prediction of effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data as well employ Data Mining methodologies. Therefore, data mining has developed into a vital domain in healthcare. It is possible to predict the efficiency of medical treatments by building the data mining applications. In the past decades, data mining have played an important role in heart disease research.

Data mining can deliver an assessment of which courses of action prove effective by comparing and evaluating causes, symptoms, and courses of treatments. The real-life data mining applications are attractive since they provide data miners with varied set of problems, time and again. Working on heart disease patients databases is one kind of a real-life application. The detection of a disease from several factors or symptoms is a multi-layered problem and might lead to false assumptions frequently associated with erratic effects [8]. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process [6]. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions. Decision Trees and Neural Networks use classification algorithms.

II. LITERATURE REVIEW

Although data mining has been around for more than two decades, its potential is only being realized now. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. Fayyad defines data mining as "a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database" [3]. Giudici defines it as "a process of selection, exploration and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of database" [4].

Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used (e.g., k-means clustering is unsupervised) [6]. Each data mining technique serves a different purpose depending on the modeling objective. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms [2].

For the prediction of heart disease we use three data mining algorithms.

1. Decision tree algorithm
2. Naïve Bayes algorithm
3. Neural Network

Decision Trees can handle high dimensional categorical data. It also handles continuous data (as in regression) but they must be converted to categorical data.

Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables.

Neural Networks consists of three layers: input, hidden and output units (variables). Connection between input units and hidden and output units are based on relevance of the assigned value (weight) of that particular input unit. The higher the weight the more important it is. Neural Network algorithms use Linear and Sigmoid transfer functions. Neural Networks are suitable for training large amounts of data with few inputs. It is used when other techniques are unsatisfactory.

III. DATA SOURCES

A total 1000 records we used in this prediction system with 15 medical attributes (factors) were obtained from the machine learning repository of UCI [7]. Table 1 lists the attributes. The records were split equally into two datasets: training dataset and testing dataset. To avoid bias, the records for each set were selected randomly. For the sake of consistency, only categorical attributes were used for all the three models. All the non-categorical medical attributes were transformed to categorical data. The attribute "Diagnosis" was identified as the predictable attribute with value "1" for patients with heart disease and value "0" for patients with no heart disease. The attribute "PatientID" was used as the key; the rest are input attributes. It is assumed that problems Such as missing data, inconsistent data, and duplicate data have all been resolved.

IV. DATA MINING ALGORITHMS FOR HEART DISEASE PREDICATION

In heart disease prediction System using data mining techniques we use three algorithms.

1. Decision Tree algorithm
2. Naïve Bayes algorithm
3. Neural Networks algorithm
4.

4.1 Decision tree algorithm

Decision trees are one of the most regularly used techniques of data analysis [8]. Decision trees are easy to visualize and understand and resistant to noise in data. Generally, decision trees are used to classify records to a proper class. Besides, they are applicable in both regression and associations tasks. In the medical field decision trees specify the sequence of attributes values and a decision that is based on these attributes.

One of the most popularly used decision tree algorithm is Iterative Dichotomized 3 (ID3). Quinlan introduced ID3 algorithm. The algorithm is based on Occam's razor, which means that the smaller trees are preferred. The Occam's razor is formalized using information entropy concept. The construction of a tree is top-down and start with the appropriate attribute for the root node.

TABLE 1: DESCRIPTION of ATTRIBUTE

Predictable attribute

1. Diagnosis (value 0: < 50% diameter narrowing (no heart disease); value 1: > 50% diameter narrowing (has heart disease))

Key attribute

1. PatientID – Patient's identification number

Input attributes

1. Sex (value 1: Male; value 0 : Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7. CA – number of major vessels colored by floursopy (value 0 – 3)
8. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach – maximum heart rate achieved
12. Oldpeak – ST depression induced by exercise relative to rest
13. Age in Year

The choice is tested and the procedure is repeated until all the attributes are used. The choice is based on calculating the entropy for each attribute. The entropy is a measure of information impurity. The ID3 uses an information gain as a measure of information carried by each of the attributes. The information gain measure is the reduction in entropy caused by the partition of the dataset [5]. It is a greedy algorithm. ID3 algorithm has a very significant advantage: it is less sensitive to errors as the decisions are based on all the instances not just the current one. It prefers short and small trees which have the attributes with the greatest information value closer to the root.

4.2 Naïve bayes algorithm

Naive bayes algorithm outperforms most of the sophisticated algorithms. It is a good tool in medical diagnosis. For example given a list of symptoms, it predicts occurrence of a disease. Naïve Bayes assumes its attributes to be conditionally independent. The classifier computes the probability of each attribute in a class. The result of the classification is the class with the highest posterior probability. Posterior probability is proportional to product of prior probability & like hood.

Naïve Bayes main strength is its simplicity, efficiency and good classification performance. It combines efficiency with good accuracy. Due to its good accuracy it is used in medical diagnosis. A small amount of training data is required for the estimation of variable values necessary for classification. Naïve Bayes is a very powerful technique in diagnosing diseases. It is used to provide efficient output with attributes independent to each other. The Naïve Bayes classifier needs a very large training set to obtain good results. Bayes' theorem can be used to compute the probability that a proposed diagnosis is correct, given that observation..

4.3 Neural networks

Artificial neural networks are analytical techniques that are formed on the basis of superior learning processes in the human brain. As the human brain is capable to, after the learning process, draw assumptions based on earlier observations, neural networks are also capable to predict changes and events in the system after the process of learning. Neural networks are groups of connected input/output units where each connection has its own weight. The learning process is performed by balancing the net on the basis of relations that exist between elements in the examples. Based on the significance of cause and effect between certain data, stronger or weaker connections between "neurons" are being formed. Network formed in this manner is ready for the unknown data and it will react based on previously acquired knowledge. One of the key advantages of Artificial Neural Networks is their high performance. The core function of Artificial Neural Networks is prediction. One of most popular algorithm of neural network is back propagation algorithm [6]. Rojas [2005] claimed that Back-Propagation algorithm could be broken down to four main steps. After choosing the weights of the network randomly, the back propagation algorithm is used to compute the necessary corrections.

V. RESULT OF HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

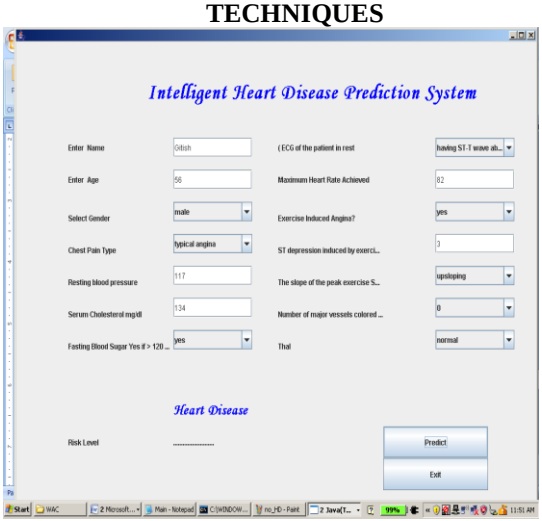

FIG 1: GUI FOR HEART DISEASE PREDICTION SYSTEM

By using three data mining algorithms namely decision tree, Naïve bayes, neural network we developed heart disease Prediction system as shown in Fig 1. All these three algorithms plays vital role in this prediction system. Each algorithm has unique as well as highest percentage of correct predictions for diagnosing patients with a heart disease. Different algorithms have different accuracy for predicting patient with heart disease. Finally at end when all information are added about patient in the form of attributes then we can diagnosis whether patient having heart disease or not. If entered patient data predicts heart disease then it will give you message as given patient is having heart disease, otherwise it will display patient having no heart disease.

VI. CONCLUSION

Health care related data are huge in nature and they arrive from various birthplaces all of them not wholly suitable in structure or quality. Data mining brings a set of tools and techniques that can be applied to the large amount of data in healthcare industry to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions. The need is for algorithms with very high accuracy as medical diagnosis is a significant task that needs to be carried out precisely and efficiently. A prototype heart disease prediction system is developed using three data mining classification modeling techniques such as Decision tree, Naïve bayes, Neural network. The system extracts hidden knowledge from a historical heart disease database. All three models are able to extract patterns in response to the predictable state. The most effective model to predict patients with heart disease appears to be Naïve Bayes followed by Neural Network and Decision Trees. Decision Trees results are easier to read and interpret. The drill through feature to access detailed patients' profiles is only available in Decision Trees. Naïve Bayes fared better than Decision Trees as it could identify all the significant medical predictors.

REFERENCES

[1]. Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh, "Knowledge Management, Data Mining, and Text Mining In Medical Informatics", Chapter 1, eds. Medical Informatics: Knowledge Management And Data Mining In Biomedicine, New York, Springer, pp. 3-34, 2005.

[2]. Charly, K.: "Data Mining for the Enterprise", 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295 304, 1998.

[3]. Fayyad, U, "Data Mining and Knowledge Discovery in Databases: Implications fro scientific databases", Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.

[4]. Giudici, P., "Applied Data Mining: Statistical Methods for Business and Industry", New York: John Wiley, 2003.

[5]. N. Abirami, T. Kamalakannan, Dr. A. Muthukumaravel, "A Study on Analysis of Various Data mining Classification Techniques on Healthcare Data", International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 7, July 2013.

[6]. R. Rojas, "Neural Networks: a systematic introduction", Springer-Verlag, 1996.

[7]. UCI Machine Learning Repository [homepage on the Internet]. Arlington: The Association; 2006 [updated 1996 Dec 3; cited 2011 Feb2]. Available from: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[8]. Sellappan Palaniappan , Rafiah Awang "Web-Based Heart Disease Decision Support System using Data Mining Classification Modeling Techniques", Proceedings of iiWAS2007. anagement of Data, pages 257{266, 1993