Disease Prediction in Data Mining Technique – A Survey

S.Vijiyarani
Assistant Professor
Department of Computer Science
School of Computer Science and Engineering
Bharathiar University
Tamil Nadu, India

S.Sudha
M.Phil Research Scholar
Department of Computer Science
School of Computer Science and Engineering
Bharathiar University
Tamil Nadu, India

ABSTRACT
Data mining is defined as sifting through very large amounts of data for useful information. Some of the most important and popular data mining techniques are association rules, classification, clustering, prediction and sequential patterns. Data mining techniques are used for variety of applications. In health care industry, data mining plays an important role for predicting diseases. For detecting a disease number of tests should be required from the patient. But using data mining technique the number of test should be reduced. This reduced test plays an important role in time and performance. This technique has an advantages and disadvantages. This research paper analyzes how data mining techniques are used for predicting different types of diseases. This paper reviewed the research papers which mainly concentrated on predicting heart disease, Diabetes and Breast cancer.

Keywords
Association rules, Breast Cancer, Classification Clustering, Diabetes, and Heart disease.

1. INTRODUCTION
Data Mining is the process of extracting hidden knowledge from large volumes of raw data. The knowledge must be new, not obvious, and one must be able to use it. Data mining has been defined as "the nontrivial extraction of previously unknown, implicit and potentially useful information from data. It is "the science of extracting useful information from large databases". It is one of the tasks in the process of knowledge discovery from the database. [1]

Data Mining is used to discover knowledge out of data and presenting it in a form that is easily understand to humans. It is a process to examine large amounts of data routinely collected. Data mining is most useful in an exploratory analysis because of nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. There are two primary goals of data mining tend to be *prediction* and *description. Prediction* involves some variables or fields in the data set to predict unknown or future values of other variables of interest. On the other hand *Description* focuses on finding patterns describing the data that can be interpreted by humans.

The Disease Prediction plays an important role in data mining. There are different types of diseases predicted in data mining namely Hepatitis, Lung Cancer, Liver disorder, Breast cancer, Thyroid disease, Diabetes etc… This paper analyzes the Heart disease, Diabetes and Breast cancer disease predictions.

The rest of this paper is organized as follows. Section 2 describes the heart disease prediction by using various data mining techniques. Section 3 describes the breast cancer prediction. Section 4 describes about prediction of diabetes. Conclusions and References are given in Section 5 and 6.

2. HEART DISEASE PREDICTION
Medical data mining has high potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis for widely distributed in raw medical data which is heterogeneous in nature and voluminous. These data should be collected in an organized form. This collected data can be integrated to form a hospital information system. Data mining technology provides a user oriented approach to novel and hidden patterns in the data. From the analysis of World Health Organization, they estimated 12 million deaths occur worldwide, every year due to the Heart diseases. Half the deaths occur in United States and other developed countries due to cardio vascular diseases. On the above discussion, it is regarded as the primary reason behind deaths in adults. Heart disease kills one person every 34 seconds in the United States. The following paper reviewed about predicting of heart disease using data mining technique.

Jyoti Soni et. al [3] proposed three different supervised machine learning algorithms. They are Naïve Bayes, K-NN, and Decision List algorithm. These algorithms have been used for analyzing the heart disease dataset [14]. Tanagra data mining tool is used for classifying these data. These classified data is evaluated using 10 fold cross validation and the results are compared. **Decision tree** is one of the popular and important classifier which is easy and simple to implement. It doesn't have domain knowledge or parameter setting. It handle huge amount of dimensional data. It is more suitable for exploratory knowledge discovery. The results attained from Decision Tree are easier to interpret and read [1]. **Naïve Bayes** is a statistical classifier which assigns no dependency between attributes. To determine the class the posterior probability should be maximized. The advantages are one can work with the naïve bayes model without using any Bayesian methods. Here Naïve Bayes Classifiers performs well. [1] ***K-nearest* neighbor's algorithm (*k*-NN)** is the one of the important method for classifying objects based on closest training data in the feature space. It is simplest among all machine learning algorithm but, the accuracy of k-NN algorithm can be degraded by presence of noisy features. This observation is performed using training to consist 3000 instances with 14 different attributes. The dataset is divided into two testing and training i.e. 70% of data are used for training and 30 % is used for testing. The authors concluded that Naïve Bayes algorithm performs well when compared to other algorithms.

Jyoti Soni et.al [3] proposed for predicting the heart diseases using the association rule data mining technique. In their work, unfortunately they have produced a large number of rules when association rules are applied to medical dataset. Most of the rules are medically irrelevant to the data. In [15], the authors proposed four constraints to reduce the number of rules i.e., item filtering, attribute grouping, maximum item set size and antecedent/consequent rule filtering. The important issue is without validation, the association rules are mined on the entire dataset. To solve these limitations, the author introduced an algorithm that uses search constraints to decrease the number of rules. The training set searches association rules and test set to check the validation. Here, a new parameter 'lift' is used instead of support and confidence. Lift has been used as the metrics to evaluate the reliability and medical significance of association rules. To validate the results the two basic statistics sensitivity and specificity are used by medical doctors. The chance of correctly identifying sick patients are defined by sensitivity and chance of correctly identifying healthy individuals is defined by specificity. To find predictive association rules in medical dataset the algorithm has three steps: [15]

(i) In medical dataset both the categorical and numeric attribute are transformed into transaction dataset.
(ii) To find the predictive association rules with medically relevant attributes the search process should be incorporate with the above mentioned four constraints.
(iii) To validate the association rules the train and test approach should be used.

Genetic algorithm have been used in [5], to reduce the actual data size to get the optimal subset of attributed sufficient for heart disease prediction. Classification is one of the supervised learning method to extract models describing important classes of data. Three classifiers e.g. Decision Tree, Naïve Bayes and Classification via clustering have been used to diagnose the Presence of heart disease in patients. **Classification via clustering:** Clustering is the process of grouping same elements. This technique may be used as a preprocessing step before feeding the data to the classifying model. The attribute values need to be normalized before clustering to avoid high value attributes dominating the low value attributes. Further, classification is performed based on clustering. Experiments were conducted with Weka 3.6.0 tool [13]. Data set of 909 records with 13 attributes. All attributes are made categorical and inconsistencies are resolved for simplicity. To enhance the prediction of classifiers, genetic search is incorporated. Observations exhibit that the Decision Tree data mining technique outperforms other two data mining techniques after incorporating feature subset selection but with high model construction time. Naïve Bayes performs consistently before and after reduction of attributes with the same model construction time. Classification via clustering is not performing well when compared to other two methods.

In the survey of [6] Naïve bayes have been used to predict attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease. The clinical dataset is having been collected from one of the leading diabetic research institute in Chennai. The records of 500 patients are taken. The data is analyzed and implemented in WEKA ("Waikato Environment for Knowledge Analysis") tool. Data mining finds out the valuable information hidden in huge volumes of data. Weka tool is a collection of machine learning algorithms for data mining techniques, written in Java. It consists of data pre-processing, classification, regression, association rules, clustering and visualization tools. We have used Naïve bayes method to perform the mining and classification process. We have used 10 folds cross validation to minimize any bias in the process and improve the efficiency of the process. From the experiment the result of bayes model was able to classify 74% of the input instances correctly. It exhibited a precision of 71% in average, recall of 74% in average, and F-measure of 71.2% in average. The results show clearly that the proposed method performs well compared to other similar methods in the literature, taking into the fact that the attributes taken for analysis are not direct indicators of heart disease.

ISSUES AND CHALLENGES

Applying data mining in the medical field is a very challenging task in medical profession. In medical research the data mining begins with a hypothesis and results are adjusted to fit the hypothesis. This differs from standard data mining practice, which simply starts with datasets without an apparent hypothesis. [11]Patterns and trends in dataset are mainly concerned with traditional data mining, but in medical data mining they are not conformed. According to the doctor intuition the clinical decision are often made. The quality of service provided to patients is affected due to unwanted bias, errors and excessive medical cost. Data mining have the capacity to generate a knowledge-rich environment. It can help to improve the significant quality of clinical decision. [5]

In the survey of [3] the three supervised machine learning algorithms are used. These algorithms have been used for analyzing the heart disease dataset. The Classification Accuracy should be compared for this algorithm. This work should be extended to predict the heart disease with reduced number of attributes. In the survey of [3] the heart disease is predicted by using association rule data mining technique. The author introduced an algorithm that uses search constraint to decrease the number of rules. In future this work should be extended by using fuzzy learning models to find the accuracy of time to decrease the number of rules. In the survey of [4] the author proposed a new concept that uses weighted association rule for classification. In future this work can be extended by using association rule hiding technique in data mining. In the survey of [5] the author proposed the minimal subset of attributes for predicting heart disease. In future this work can be expanded and enhanced for the automation of heart disease prediction. Real data should be collected from health care organizations and agencies are taken to compare the optimum accuracy with all data mining technique. In the survey of [6] the author predicts attributes of a diabetic patient getting a heart disease. Weka tool is performed as a result bayes model was able to classify 74% of the input instances correctly. In future this work is extended by using other data mining techniques.

3. BREAST CANCER PREDICTION

Breast cancer has become a common cancer in women. For instance, it affects one in every seven women in the United State [16]. The mammography is the traditional method for breast cancer diagnosis. However, the radiologists show considerable variability in how they interpret a mammogram. Moreover, Elmore indicated that 90% of radiologists

recognized fewer than 3% of cancers and 10% recognized about 25% of the cases. The fine needle aspiration cytology is another approach adopted for the diagnosis of breast cancer with more precise prediction accuracy. However, the average correct identification rate is around 90% [17]. Generally, the purpose of all the related research is identical to distinguish between patients with breast cancer in the malignant group and patients without breast cancer in the benign group. There are three predictive focus of cancer prognosis: 1) prediction of cancer susceptibility (risk assessment), 2) prediction of cancer recurrence and 3) prediction of cancer survivability. The accepted prognostic factor for breast cancer is the American Joint Commission on Cancer (AJCC). It is staging system based on the TNM system (T, tumor; N, node; M, metastasis) [7] and survival is considered as any incidence of breast cancer where the person is still living from the date of diagnosis. The objective is to handle cases for which cancer has not recurred (censored data) as well as case for which cancer has recurred at a specific time. This section describes various technical and review articles on data mining techniques applied in breast cancer prognosis.

C4.5 is a well known classification technique in decision tree induction which has been used by Abdelghani Bellaachia and Erhan Gauven [8] along with two other techniques i.e. Naïve Bayes and Back-Propagated Neural Network. They conduct an analysis of the prediction of survivability rate of breast cancer patients using above data mining techniques and it is used in the new version of the SEER Breast Cancer Data. The preprocessing data set consists of 151,886 records, which are available in 16 fields from the SEER database. They have adopted a different category in the pre-classification process by including three fields: STR(Survival Time Recode), VSR(Vital Status Recode), and COD(Cause Of Death) and it is used by Weka toolkit to experiment these three data mining algorithms. Some of the experiments were conducted using these algorithms. The attained prediction performances are compared to existing techniques. However, the author found the model generated by C4.5 algorithm for the given data has a much better performance than the other two techniques.

In [9] M. Lundin et al has applied ANN on 951 instances dataset of Turku University Central Hospital and City Hospital of Turku. To evaluate the accuracy of neural networks in predicting 5, 10 and 15 years breast cancer specific survival. From the experiment the values of ROC curve for 5 years was evaluated as 0.909, for 10 years 0.086 and for 15 years 0.883, these values were used as a measure of accuracy of the prediction model. They author compared 82/300 false prediction of logistic regression with 49/300 of ANN for survival estimation and found ANN predicted survival with higher accuracy.

In [10] Delen et al compared ANN, decision tree and logistic regression techniques for breast cancer prediction analysis. They used the SEER data of twenty variables in the prediction models. From the experiment the author found that the decision tree with 93.6% accuracy and ANN with 91.2% are more superior to logistic regression with 89.2% accuracy.

ISSUES AND CHALLENGES

In [8] the author discussed and resolved the algorithms, issues and techniques for the problem of breast cancer prediction. This analysis does not include records with missing data so in future this work is enhanced by including the missing data. In [9] by analyzing the artificial neural network, trained on a number of clinic pathological variables of patients with breast cancer, predicted survival with high accuracy. The author concluded that the consistent accuracy over time and the good predictive performance of a network trained without information on nodal status. It shows that neural networks are valuable tools in cancer survival prediction. In future the study should concentrate on collecting data from a more recent time period and find new potential prognostic factors to be included in a neural network model. In the survey [10], the study is based on multiple prediction models for breast cancer survivability using large datasets along with 10 fold cross validation method. It provides a relative prediction ability of different data mining methods. In future this work is extended by collecting real dataset in the clinical laboratory.

4. DIABETES

Insulin is one of the most important hormones in the body. It aids the body in converting sugar, starches and other food items into the energy needed for daily life. However, if the body does not produce or properly use insulin, the redundant amount of sugar will be driven out by urination. This disease is referred to diabetes. The cause of diabetes is a mystery, although obesity and lack of exercise appear to possibly play significant roles. Based on the American Diabetes Association [4] in November 2007, 20.8million children and adults in the United States (i.e., approximately 7% of the population) were diagnosed with diabetes. In early the ability to diagnose diabetes plays an important role for the patient's treatment process.

In [18] the author predicts whether a new patient would test positive for diabetes. This paper studied a new approach, called the Homogeneity- Based Algorithm (or HBA) to determine optimally control the over fitting and overgeneralization behaviors of classification on this dataset (Pima Indian diabetes data set). The HBA is used in conjunction with classification approaches (such as Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), or Decision Trees (DTs)) to enhance their classification accuracy. Some expermental results seem to indicate that the proposed approach significantly outperforms current approaches. From the experiment the author concluded that it is very important both for accurately predicting diabetes and also for the data mining community, in general.

In [19] Data mining algorithm is used for testing the accuracy in predicting diabetic status. Fuzzy Systems are been used for solving a wide range of problems in different application domain Genetic Algorithm for designing. Fuzzy systems allows in introducing the learning and adaptation capabilities. Neural Networks are efficiently used for learning membership functions. Diabetes occurs throughout the world, but Type 2 is more common in the most developed countries. The author implemented in Genetic Algorithm. The steps involved in this algorithm namely selection, crossover, mutation, fitness and population statistics. As a result the author concluded that the optimization of chromosome using GA is obtained and it is based on the rate of old population diabetes can be restricted in new population to get chromosomal accuracy.

ISSUES AND CHALLENGES

In [18] the author proposed a new algorithm Homogeneity-Based Algorithm to determine over fitting and

overgeneralization behavior of classification. The algorithms used in this paper are Support Vector Machine, Decision Tree and Artificial Neural Networks. In future this work is enhanced by using any optimization techniques. In [19] for predicting diabetic status the author uses data mining algorithm for testing the accuracy. The author implemented using genetic algorithm. In future this work is extended by using other optimization technique.

5. CONCLUSION

In this survey paper the problem of summarizing the different algorithm of data mining are used in the field of medical prediction are discussed. The main focus is on using different algorithm and combination of several targets attributes for different types of disease prediction using data mining. First we discuss about the heart disease prediction, in that machine learning algorithms namely naïve bayes, K-NN, Decision List. Of these the classification accuracy of the naïve bayes algorithm is better when compared to other algorithm. In Weighted Associative Rule Classifier, the GUI has been designed to enter the patient record and the presence of Heart Disease for a patient is predicted by using the rules stored in the rule base. Next we discuss the feature subset selection using genetic algorithm. In this attributes are reduced using genetic search. Here the accuracy is compared to the three classifiers namely Decision Tree, Naïve bayes and classification via clustering. Association rule discovery is mainly based on four constraints namely item filtering, attribute grouping, maximum item set size and antecedent/consequent rule filtering. To find predictive rules in medical data set the three important steps are generated in this algorithm [3, 5, 15]. The heart disease is diagnosed for diabetic patients using naïve bayes technique [6]. Of these the author concluded that naïve bayes classify 74% of input instances correctly. Next we discuss about the breast cancer prediction. It is performed by using various data mining techniques namely C4.5, ANN and fuzzy decision trees. By using C4.5 the author discussed and resolved the issues and algorithms of the problem. Using ANN the author concluded that the consistent accuracy over time and good performance of the network is trained. The fuzzy decision tree survives by using 10 fold cross validation method. Finally we discuss about diabetes prediction, by using homogeneity based algorithm the author find over fitting and overgeneralization behavior of classification. By using genetic algorithm the author predicts accuracy of the class.

In future the work can be expanded and enhanced for the automation of various types of disease prediction. It also extended to find other types of diseases with the uses of these attributes.

6. REFERENCES

1]. Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006

[2] "Data mining: Introductory and Advanced Topics" Margaret H. Dunham

[3]. Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" IJCSE Vol. 3 No. 6 June 2011

[4]. Carloz Ordonez, "Association Rule Discovery with Train and Test approach for heart disease prediction", IEEE Transactions on Information Technology in Biomedicine, Volume 10, No. 2, April 2006.pp 334-343.

[5]. M. ANBARASI, E. ANUPRIYA, N.CH.S.N.IYENGAR, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370-5376

[6]. G. Parthiban, A. Rajesh, S.K.Srivatsa "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method"

[7]. Choi J.P., Han T.H. and Park R.W., "A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis", J Korean Soc Med Inform, 2009, pp. 49-57

[8] Bellaachia Abdelghani and Erhan Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques,"Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining," 2006.

[9] Lundin M., Lundin J., BurkeB.H.,Toikkanen S., Pylkkänen L. and Joensuu H. , "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer", Oncology International Journal for Cancer Resaerch and Treatment, vol. 57, 1999.

[10] Delen Dursun , Walker Glenn and Kadam Amit , "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine ,vol. 34, pp. 113-127 , June 2005.

[11]. Ruben D. Canlas Jr.,"DATA MINING IN HEALTHCARE: CURRENT APPLICATIONS AND ISSUES", August 2009

[12] Michael Feld, Dr. Michael Kipp, Dr. Alassane Ndiaye and Dr. Dominik Heckmann "Weka: Practical machine learning tools and techniques with Java implementations"

[13] K.P Soman, Shyam Diwakar, V.Vijay "Insight into Data mining theory and practice"

[14] Asha Rajkumar, G.Sophia Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm", Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver. 1.0 September 2010.

[15] Shantakumar B.Patil, Y.S.Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009.

[16] Wingo PA, Tong T, Bolden S, "Cancer statistics", 1995, CA Cancer J Clin 45 (1995), no. 1, 8-30.

[17] Fentiman IS, "Detection and treatment of breast cancer", London: Martin Duntiz (1998).

[18] Huy Nguyen Anh Pham and Evangelos Triantaphyllou "Prediction of Diabetes by Employing a New Data Mining Approach Which Balances Fitting and Generalization" Department of Computer Science, 298 Coates Hall, Louisiana State University, Baton Rouge, LA 70803

[19] Ms.S.Sapna, Dr.A.Tamilarasi "Data mining – Fuzzy Neural Genetic Algorithm in predicting diabetes" Department Of Computer Applications (MCA), K.S.R College of Engineering "BOOM 2K8" Research Journal on Computer Engineering, March2008.