ANALYSIS AND PREDICTION OF VARIOUS HEART DISEASES USING DNFS TECHNIQUES

[1]S.Prabhavathi, [2]D.M.Chitra

[1] Research Scholar, Department of Computer Science, Padmavani Arts &Science College for Women, Salem
[2]Assistant professor, Department of Computer Science, Padmavani Arts &Science College for Women, Salem

Abstract: Data mining techniques have been applied magnificently in many fields including business, science, the Web, bioinformatics, and on different types of data such as textual, visual, spatial and real-time and sensor data. Medical data is still information rich but knowledge poor. There is a lack of effective analysis tools to discover the hidden relationships and trends in medical data obtained from clinical records. This paper reviews the state-of-the-art research on heart disease diagnosis and prediction. DNFS stands for Decision tree based Neural Fuzzy System. Specifically here present an overview of the current research being carried out using the data mining techniques to enhance heart disease diagnosis and prediction including decision trees, Naive Bayes classifiers, K-nearest neighbour classification (KNN), support vector machine (SVM), and artificial neural networks techniques. Results show that SVM and neural networks perform positively high to predict the presence of different types heart diseases. Still the performance of data mining techniques to detect heart diseases is not encouraging.

1. INTRODUCTION

Health care related data are huge in nature and they arrive from various birthplaces all of them not wholly suitable in structure or quality. These days the utilization of knowledge and experience of copious specialists and medical screening data of patients collected in a database during the diagnosis process has been widely accepted. In this paper we have presented an efficient approach for fragmenting and extracting substantial forms from the heart attack data warehouses for the efficient prediction of heart attack.

The diagnosis of diseases is a difficult but critical task in medicine. The detection of heart disease from

"various factors or symptoms are a multi-layered issue which is not free from false presumptions often accompanied by unpredictable effects". Thus can use patients data that have been collected and recorded to ease the diagnosis process and utilize knowledge and experience of numerous specialists dealt with the same symptoms of diseases. Providing invaluable services with less cost is a major constraint by the healthcare organizations (hospitals, polyclinics, and medical centers). According to "valuable" quality service denotes the accurate diagnosis of patients and providing efficient treatment. Poor clinical decisions may lead to disasters and hence are seldom entertained". Besides, it is essential that the hospitals decrease the cost of clinical tests. Using professional and expert computerized systems based on machine-learning and data mining methods should help in one direction or another with achieving clinical tests or diagnosis at reduced risks. K-nearest neighbour classification algorithm is a well-known method for classifying an unseen instance using the classification of the instances closest to it. Basic KNN classification algorithm works by finding K training instances that are close to the unseen instance using distance measures such as Euclidean, Manhattan, maximum dimension distance, and others. Then, the algorithm decides the class for the unseen instance by taking the most commonly occurring class in the nearest K instances. Here using a decision tree based, support vector machine and naive bayes classification techniques and fuzzy logic used for analyzing the data sets.

2. RELATED WORK

The difficult of recognizing constrained association rules for heart illness prediction was studied by Carlos Ordonez. The data mining techniques have been engaged by various works in the works to analyze various diseases, for instance: Hepatitis, Cancer, Diabetes, Heart diseases. Frequent Item set Mining (FIM) is measured to be one of the basic data mining difficulties that expects to discern collections of items or values or forms that co-occur regularly in a dataset. Integrating Decision Tree and K-Means Clustering with

Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients says developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. IHDPS 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. IHDPS is Web-based, user-friendly, scalable, reliable and expandable. It is implemented on the .NET platform.

Intelligent Heart Disease Prediction System Using Data Mining Technique says Represent the use of artificial neural networks in predicting neonatal disease diagnosis. The proposed technique involves training a Multi Layer Perceptron with a BP learning algorithm to recognize a pattern for the diagnosing and prediction of neonatal diseases. The Backpropogation algorithm was used to train the ANN architecture and the same has been tested for the various categories of neonatal disease. About 94 cases of different sign and symptoms parameter have been tested in this model. This study exhibits ANN based prediction of neonatal disease and improves the diagnosis accuracy of 75% with higher stability.

An Artificial Neural Network Model for Neonatal Disease Diagnosis Proposed research contains data mining classification techniques RIPPER classifier, Decision Tree, Artificial neural networks (ANNs), and Support Vector Machine (SVM) are analyzed on cardiovascular disease dataset. Performance of these techniques is compared through sensitivity, specificity, accuracy, error rate, True Positive Rate and False Positive Rate. 10-fold cross validation method was used to measure the unbiased estimate of these prediction models. As per our results error rates for RIPPER, Decision Tree, ANN and SVM are 2.756, 0.2755, 0.2248 and 0.1588 respectively. Accuracy of RIPPER, Decision Tree, ANN and SVM are 81.08%, 79.05%, 80.06% and 84.12% respectively. The analysis shows that out of these four classification models SVM predicts cardiovascular disease with least error rate and highest accuracy.

Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks has been proposed a Decision Support System for diagnosis of Congenital Heart Disease. The proposed system is designed and developed by using MATLAB's GUI feature with the implementation of Back propagation Neural Network,
The Back propagation Neural Network used in this study is a multi layered Feed Forward Neural Network, which is trained by a supervised Delta Learning Rule. The dataset used in this study are the signs, symptoms and the results of physical evaluation of a patient. The proposed system achieved an accuracy of 90%.

Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction says K-means clustering with the decision tree method to predict the heart disease. In their work they suggested several centroid selection methods for k-means clustering to increase efficiency. The 13 input attributes were collected from Cleveland Clinic Foundation Heart disease data set. The sensitivity, specificity, and accuracy are calculated with different initial centroids selection methods and different numbers of clusters.

For the random attribute and random row methods, ten runs were executed and the average and best for each method were calculated. When comparing integrating k-means clustering and decision tree with traditional decision tree applied previously on the same data set, integrating k-means clustering with decision tree could enhance the accuracy of decision tree in diagnosing heart disease patients. In Addition, integrating k-means clustering and decision tree could achieve higher accuracy than the paging algorithm in the diagnosis of heart disease patients. The accuracy achieved was 83.9% by the enabler method with two clusters.

Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network says a proficient methodology for the extraction of significant patterns from the heart disease warehouses for heart attack prediction has been presented. Initially the data warehouse is pre-processed in order to make it suitable for the mining process. Once the preprocessing gets over, the heart disease warehouse is clustered with the aid of the K-means clustering algorithm.

Consequently the frequent patterns applicable to heart disease are mined with the aid of the MAFIA algorithm from the data extracted. In addition, the patterns vital to heart attack prediction are selected on basis of the computed significant weightage. The neural network is trained with the selected significant patterns for the effective prediction of heart attack.

Development of a Data Clustering Algorithm for Predicting Heart says that existing research mainly focused in performance analysis and comparison. The quality of the service is very essential in analyzing field that means the accuracy result. This research is mainly focuses on Prediction of Heart Disease using K-Means Clustering in the context of data mining. The cost function like predictive stable accuracy evaluated in prediction of heart disease using K-Means Clustering technique in data mining. Similarly complementary measures like compactness and connectedness of clusters are treated as two objectives for cluster analysis. The grouping is done by minimizing the sum of squares of distances between data using Euclidean distance formula and the corresponding cluster centroid. Thus the purpose of K Means Clustering is to classify the data with predictive stable accuracy.

Heart Disease Prediction System using Naive Bayes says that classification of the given data into differerent categories and also predicts the risk of the heart disease if unknown samsple is given as an input.The system can be served as training tool for

medical students.Also,it will be helping hand for doctors.As we have developed generalised system,in future we can use this system for analysis of different datasets by only changing the name of dataset file which is given for training module.

Survey on prediction of heart morbidity using data mining techniques says that various existing techniques, the issues and challenges associated with them. The discovered knowledge can be used by the healthcare administrators to improve the quality of service and also used by the medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. In this paper we discuss the popular data mining techniques namely, Decision Trees, Naïve Bayes and Neural Network that are used for prediction of disease.

Detection of health care using datamining concepts through web says that prediction of the risk levels of heart attack from the heart disease database. The heart disease database consists of mixed attributes containing both the numerical and categorical data. These records are cleaned and filtered with the intention that the irrelevant data from the database would be removed before mining process occurs. Then clustering is performed on the preprocessed data warehouse using K means clustering algorithm with K value so as to extract data relevant to heart attack. Subsequently the frequent patterns significant to heart disease diagnosis are mined from the extracted data using the MAFIA algorithm. Finally, we used the ID3 as training algorithm to show the effective risk level with decision tree.\

3. PROPOSED WORK

3.1 Extraction of the structure information from dataset:

Construct a neighborhood graph to connect each object to its K-Nearest Neighbors (KNN); Estimate a density for each object based on its proximities to its KNN; Objects are classified into 3 types: Cluster Supporting

Object (CSO): object with density higher than all its neighbors; Cluster Outliers: object with density lower than all its neighbors, and lower than a predefined threshold; the rest.

3.2 Local/Neighborhood approximation of fuzzy memberships:

Initialization of fuzzy membership: Each CSO is assigned with fixed and full membership to itself to represent one cluster; All outliers are assigned with fixed and full membership to the outlier group; The rest are assigned with equal memberships to all clusters and the outlier group; Then the fuzzy memberships of all type 3 objects are updated by a converging iterative procedure called Local/Neighborhood Approximation of Fuzzy Memberships, in which the fuzzy membership of each object is updated by a linear combination of the fuzzy memberships of its nearest neighbors.

3.3 Neural Networks

Neural networks are biologically inspired highly interconnected cells that simulate the human brain. The perception is the simplest architecture which has one neuron and a learning method. More sophisticated architecture is multi-layer neural networks (MLP) which one or more neurons connected at different layers. Neural networks can be trained to learn a classification task and to predict disease

Architecture Diagram

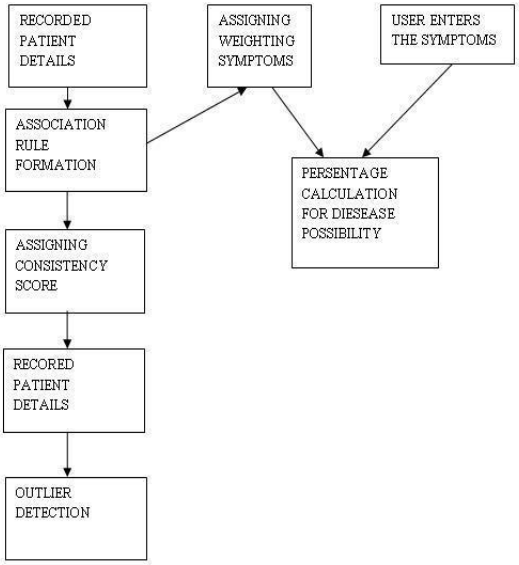

Figure 1: Architecture Diagram

4. PROJECT DESCRIPTION

The objective of this research is to create an intelligent & cost effective system which will overcome the limitations of existing system and improve its performance.

5. CLASSIFICATION TECHNIQUES

A way to enhance the performance of a model that combines neural network and fuzzy logic for classification is the application of genetic algorithm to improve the learning of neuro-fuzzy system. Various methods have been used for the detection of a potential medical problem.

Thus, a reliable method is required. The neuro –fuzzy system (NFS) model which combines the adaptability of fuzzy inputs with neural network is used to accurate prediction. And this kind of network is more efficient than the simple neural network.

5.1 Decision Tree Classification Algorithm

Decision trees are powerful classification algorithms used alternatively as decision/classification rules.

Popular decision tree algorithms include Quinlan's ID3, C4.5, C5, and Breiman *et al.*'s CART. As its name implies, this classification technique works by recursively constructing branches of the tree based on certain observations (or variables). A well-known algorithm to construct the branches is called top-down induction of decision trees (TDIDT). Decision trees are easily constructed with binary or categorical variables but the mission becomes harder with numerical variables.

A corresponding threshold value must be specified for the later based on some mathematical or observational considerations in order to be able to construct the branches of the tree. This step is repeated at each leaf node until the complete tree is constructed ending with leaves which gives one of the classes or predictions in the dataset.

The objective of the splitting algorithm is "to find a variable-threshold pair that maximizes the homogeneity of the resulting two or more subgroups of samples".

2. DATASET MODULES

Neuro fuzzy systems are fuzzy systems which use ANN's theory in order to determine their properties like fuzzy sets and fuzzy rules by processing data samples. The main objective of this research is to develop an Intelligent Heart Disease Prediction system. Here the dataset related to the cardiovascular disease is provided to the neuro-fuzzy system. The dataset consist of the patients symptoms of cardiovascular disease.

It consists of cardiovascular disease patients information. This Cleveland databases are publicly available. From the input data, association rules are formed using improved method Narita and

Kitagawa's algorithm

The process of Neuro-Fuzzy with GA presented below:

1. Initialized the process of predicting Cardiovascular Disease.
2. Extract the patient's details from dataset.
3. Assign the input and output to NFS.
4. Selection process start with assign weight randomly to each attributes.

5. Training is done using Back-propagation.

6. Compute output values.

7. Compute fitness using

Σ (Outputs- Targets)2 Mean Square Error = Number of samples

8. If MSE is less than error then go to step 10, otherwise go to step 9.

9. Select the parents and apply crossover and mutation.

10. Train NF with selected connection weights.

11. Study the performance of test data.

3. ASSIGNING CONSISTENCY SCORE

Each patient record is assigned with a score using consistency rule. By this rule, Consistency is evaluated using the created association rules. Patient detail which fulfills the rules is assigned with highest consistency value and the Patient detail which violates is assigned with lowest value.

4. ASSIGNING WEIGHTAGE SYMPTOMS

Then the scored Patient details are sorted in ascending order. Patient details with least consistent value are considered outlier.

Weightage is assigned to each symptom using sorted patient details and association rule. User enters the symptoms to know the percentage of cancer disease possibility. Percentage is calculated from the weightage of the symptom.

5. PERSENTAGE CALCULATION FOR DISEASE POSSIBILITIES

Percentage of possibility for cancer disease is calculated from the weight assigned to each symptoms of all considered cancer disease. With the more number of symptoms the accuracy of calculating the disease possibility will be higher.

5.1 C4.5 classification algorithm

One of the best decision tree algorithms is C4.5. This algorithm can manage continuous data in numerical forms using pruning algorithms which aim to simplify the classification rules without any loss of prediction accuracy. Only the most important features are kept whereby they lowered the error rates. One of its factors is denoted by *M* which indicates the minimum instances that a leaf should have. *C* means the confidence threshold which is considered for pruning. By changing these two factors, the accuracy of algorithm can be increased and the error can be decreased.

5.2 RIPPER classification algorithm

RIPPER stands for Repeated Incremental Pruning to Produce Error Reduction. This classification algorithm is based on association rules with reduced error pruning (REP), a very common and effective technique found in decision tree algorithms. To generate association rules using REP algorithm, the training data is divided into a growing set and a pruning set. The growing set is the initial association rules which can be generated purely from the dataset using some heuristic methods. The growing set contains a huge set of rules that should berepeatedly simplified to form the pruning set.

Thus the simplification is done using typical pruning operators which may allow deleting a term from any single rule or from different association rules. The pruning operator chosen for simplification should give the most accurate rule with the greatest reduction of error. The simplification process terminates when applying the pruning operator would increase the error value on the pruning set.

Support Vector Machine (SVM)

SVM is a state-of-the-art maximum margin classification algorithm rooted in statistical learning theory. It is a method for classification of both linear and non-linear data. The training data is converted into n-dimensional data using non-linear transformation method. Then, the algorithm searches for the best hyper-plane to separate the transformed data into two different classes. SVM performs classification tasks by maximizing the margin of the hyper-plane separating both classes while minimizing the classification errors.

5.3 Naïve Bayes Algorithm

One of the Bayesian methods is Naïve Bayes classifiers which uses the probabilistic formula: where *A* and *B* are two events (e.g. the probability that the train will arrive on time given that the weather is rainy). Such Naïve Bayes classifiers use the probability theory to find the most likely classification of an unseen (unclassified)

instance. The algorithm performs positively with categorical data but poorly if we have numerical data in the training set.

Several medical data mining techniques for heart diseases diagnosis focusing on three famous types namely CAD, CVD and CHD are discussed.

The effectiveness of different techniques

(i) C4.5 classifier performs better than other data mining techniques to diagnose CAD via stenosis of the LAD vessel, followed by KNN via the stenosis of LCX vessel. (ii) SVM and neural networks perform comparably and positively high; therefore, they can be utilized to predict the presence of CHD. (iii) Decision trees method after the reduction and optimization of features using GA is the best recommended classifier to diagnose CVD heart disease.

The performance evaluation metrics of data mining techniques for heart disease prediction. It is observed that the accuracy of various classification techniques for CVD diagnosis is highly encouraging (between 85% and 99%). Consequently, diagnosis systems that employs classifiers or clusters can assist the medical professionals in making decision about CVD early diagnosis. Moreover, data mining techniques perform positively well with diagnosing CHD (achieving accuracy between 82% and 92%). Still the performance of classification methods to detect CAD diseases is not encouraging (between 60%-75%) whether or not the LAD, LCX, and RCA vessels are considered separately. Therefore, further research should be carried out using more sophisticated features and hybrid algorithms to improve the prediction of CAD diseases. Due to the successful implementation of data mining techniques for heart diagnosis, various heart disease prediction systems that employed the aforementioned data mining techniques such as Intelligent Heart Disease Prediction System (IHDPS) have been proposed.

6. CONCLUSION

Early diagnosis of heart diseases may save humans from heart attacks. This paper reviews the state-of-the-art data mining techniques applied for diagnosing the different types of heart diseases namely CAD, CVD, and CHD. Among the famous and sever diseases is CAD which can be diagnosed via the stenosis of blood vessels. Such disease is data rich but unfortunately the obtained accuracy from CAD classifiers is poor. Data mining techniques applied for CVD and CHD are promising. Results showed that the optimization and feature reduction utilising GA or principle component analysis (PCA) for a certain disease may strongly increase the accuracy of a classifier. It is found that decision trees and Naïve Bayes classifiers are recommended for CVD diagnosis with an accuracy reaching more than 95%.

7. FUTURE WORK

Future works should focus on improving the predication of CAD diseases utilising more features and separatecombinations of vessel stenosis. Furthermore, feature reduction should be utilised in various ways to achieve better accuracy results with all diseases. New classifiers should be developed for other heart diseases and problems such as coronary microvascular diseases, pulmonary, and cyanotic heart diseases. This work can be further extended by working with different heart related datasets from health care organizations and agencies using all the available techniques and also using a combination of them.

REFERENCES

[1] Mai Shouman, Tim Turner and Rob Stocker,

[2] “Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease

[3] Patients”, Proceedings of the International Conference on Data Mining, 2012. _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.

[4] P .K. Anooj, _Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules_; Journal of King Saud University – Computer and Information Sciences (2012) 24, 27–40.

[5] Bala Sundar V, T Devi, N Saravanan, Development of a Data Clustering Algorithm for Predicting Heart, International Journal of Computer Applications (0975 – 888)Volume 48– No.7, June 2012.

[6] Shadab Adam Pattekari and Asma Parveen,

Prediction System For Heart Disease Using Naive Bayes, International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294

[7] Sellappan Palaniappan, Rafiah Awang, Intelligent 11. K.Srinivas, Dr. G.Raghavendra Rao and Dr.

[8] Heart Disease Prediction System Using Data Mining Technique, 978-1-4244-1968-5/08/$25.00 ©2008 IEEE.

[9] Dilip Roy Chowdhury, Mridula Chatterjee & R. K. Samanta, An Artificial Neural Network Model for Neonatal Disease Diagnosis, International Journal of Artificial Intelligence and Expert Systems (IJAE), Volume (2): Issue (3), 2011.

[10] Vanisree K, Jyothi Singaraju, Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks, International Journal of Computer Applications (0975 – 8887) Volume 19– No.6, April 2011.

[11] Milan Kumari, Sunila Godara, Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction, IJCST Vol. 2, Iss ue 2, June 2011.

[12] 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.

[13] Shantakumar B.Patil, Y.S.Kumaraswamy, A.Govardhan, Survey On Prediction Of Heart Morbidity Using Data Mining Techniques, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.3, May 2011

[14] Mounika Naidu P, Rajendra C, Detection of Health Care Using Datamining Concepts through Web, ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & TechnologyVolume 1, Issue 4, June 2012.

[15] Hnin Wint Khaing, Data Mining based Fragmentation and Prediction of Medical Data, ISBN: 978-1-61284-840-2, 2011.

[16] S Satapathy and S Chattopadhya, Mining Important Predictors Of Heart Attack, International Conference On Advances In Recent Technologies In Communication And Computing 2011.

[17] S. el Rafaie, Abdel-Badeeh M. Salem and K. Revett, On the Use of SPECT Imaging Datasets for Automated Classification of Ventricular Heart Disease, The 8th International Conference on INFOrmatics and Systems (INFOS2012).