ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

Analysis of classification techniques for Medical data

P.Thangaraju
Associate Professor, Dept., of Computer Application
Bishop Heber College (Autonomous),
Tiruchirappalli, India

P.Aishwarya
Research Scholar, Dept., of Computer Application
Bishop Heber College (Autonomous),
Tiruchirappalli, India.

Abstract: **Data mining is widely used in many fields for analysing large data from different perspective and it helps us to extract and summarize useful information. Modern medicine generates large amount of information stored in the medical database. Analysing these data manually is a complex and tedious process. It is necessary to develop a model which helps to extract useful knowledge and provide scientific decision-making for the diagnosis. Early diagnosis of diseases is important need in healthcare industry for giving treatment. Data mining plays an important role in analysing medical data. Tools and various algorithms available in data mining help us to develop models that can assist us to make accurate and timely decisions. Classification is the task of generalizing known structure which can be applied to new data. In this paper, an analysis of various classification algorithms is done which are used for analysing medical data. The accuracy of the classification is mainly focused in this survey paper.**

Keywords— Datamining, Classification,ANN,Decision tree.

I. INTRODUCTION

With the increase in growth of population, there is a significant expansion in the health issues. Many new types of diseases and its symptoms have been identified. Numerous diseases are strongly associated with a common symptom which makes it complicated for the doctors to diagnose the exact diseases precisely on one go. This is where data mining comes to help. it helps in diagnosing the disease by analysing the patients data. Even-though the prediction is not extremely accurate, it gives the doctor a concise idea what the disease might be. Thus, Data mining is not a substitution to doctors as an alternative, it is a tool which support them to identify the diseases in advance stages [1].

Data mining plays an important role in analysing medical data. It helps us to create the models which assist us to analyse the raw data containing the patients symptoms and predict the disease.It can also improve the management quality of hospital. The medical information may be redundant, multi-attributed, inconsistent, incomplete and closely related with time. The key techniques of medical data mining involves in pre-processing of medical data, analysing different pattern and resource, applying mining algorithms and predicting the reliability of mining results[2].

II. LITERATURE SURVEY

R. Subhashini [3] et al., proposed a novel classification strategy to predict the chronic kidney disease using Optimal Fuzzy-K nearest neighbour technique. The performance of fuzzy is made optimum by tuning the membership functions utilizing the Bat optimization algorithm. Then the OF is utilised to measure the similarity in the KNN for the classification of disease. she compared the OF-KNN algorithm with ANN, SVM and KNN. she observed OF-KNN outperformed and given best result.

S. Vijayarani [4] et al., made the comparative analysis of classification algorithms such as Naïve Bayes and Support Vector Machine. She hasused the synthetic kidney function test dataset with six attributes and 584 instance for analysing kidney disease. The implementation is done using MATLAB. Based on the performance measures of classification accuracy, error rate and execution time it was observed that SVM is better when compared with Naïve Bayes.

Manish Kumar[5] et al., made a comparative analysis on various classification algorithm like Random Forest classifiers, Sequential Minimal Optimization, Naïve Bayes, Radial Basis Function and Multilayer Perceptron Classifier and Simple Logistic for the predicting Chronic Kidney Disease . He used UCI chronic kidney disease dataset with 25 attribute and 400 instance. The obtained result showed that the Random Forest classifier outperformed all other classifiers in terms of Area under the ROC curve (AUC), accuracy and MCC with values 1.0, 1.0 and 1.0 respectively.

Prerana [6] et al., proposed a systematic approach for earlier diagnosis of Thyroid disease using back propagation algorithm in neural network. He has used UCI dataset with 29 attributes. he has implemented the predictive neural model which works in two phase , one is back propagation and second phase is updation of weights to classify the thyroid disease in MATLAB Neural Network Toolbox software. He has taken FTI values as input to classify in three different classes with values 1,2,3. The Training performance plots for gradient descent training algorithm and Levenberg algorithm, presenting variation of MSE verses numbers of epochs and plots for the variation of error gradient values during training process . It has been observed that Levenberg Marquardt

All Rights Reserved © 2018 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

method has shown a better training performance for achieving the set target in 59 epochs and gradient decent is showing a poor performance as it is unable to achieve the set target value of 0.0001 in 1000epochs.

K. Saravana Kumar[7] et al., made a comparative analysis of K- Nearest Neighbor and Support Vector Machine in accuracy of predictions of Hypothyroid. He has applied SVM and KNN methods to the collected data to predict hypothyroid and observed the prediction accuracy is 94.4336 in SVM and 96.3430 accuracy in KNN. As the difference / variance is 1.9094. Therefore, he has concluded that KNN performs better than the SVM while predicting thyroid disease.

Anurag Upadhayay[8] et al., made an empirical comparison between two algorithms C4.5 and C5.0 of Decision Tree technique in predicting thyroid disease. He has used UCI dataset with 29 attributes and worked with 400 patients records. He observed that Running Time of C5.0 was Small as compared to C4.5,Tree size of C4.5 was very large when compared to C5.0, After Pruning C5.0 Tree generated more accurate rule set, Train error in case of c5.0 was small when compared to the C4.5,Rule set Generated by the C5.0 algorithm is 6 and the confidence level of the rules was more than 95%.so he concluded that C5.0 is better when compared to C4.5.

V Prasad[9] et al., proposed a Health Diagnosis Expert Advisory System on Trained Data Sets for predicting the level of Hyperthyroid in human body. The EAS system is developed by using Data Matching System which is applied on Training Data Set to identify the relevant disease according to the data of the symptoms specified in the knowledge base. Once the user enters the details of the symptoms and submits then it predicts the Human disease. One limitation in his work is that if the user enters wrong detail it may misguide them by predicting wrong disease.

Wei-Wen Chang[10] et al., combined the main concepts of estimation of distribution algorithms and immune algorithms to form a hybrid algorithm called immune-estimation of distribution algorithms (IEDA) and applied it To Classify UCI thyroid gland data set. They have compare the results between IEDA and traditional genetic algorithms. Based on the results, they concluded their research is better than traditional genetic algorithm including accuracy, type I error and type II error.

Hui-Ling Chen[11] et al., proposed a three-stage expert System (FS-PSO-SVM) based on a hybrid support vector machines approach for diagnosing thyroid disease. The first stage (FS) aimed at constructing diverse feature subsets with different discriminative capability. In second stage, the feature subsets obtained are used for training designed SVM classifier for training an optimal predictor model whose parameters are optimized using particle swarm optimization (PSO). Finally, the obtained optimal SVM model used for diagnosing the thyroid disease using the most discriminative feature subset and the optimal parameters. The proposed system has achieved the highest classification accuracy reported so far by 10-fold crossvalidation method, with the mean accuracy of 97.49% and with the maximum accuracy of 98.59%.

Ali Keles[12] et al.,proposed An Expert system for diagnosing thyroid disease(ESTDD).they found fuzzy rules by using neuro fuzzy method, which will be in ESTDD system. The accuracy of ESTDD is 95.33% while diagnosing thyroid diseases.

P. Thangaraju[13] et al., proposed a model to analyse the data of liver diseases using particle swarm optimization algorithm (PSO) with KStar Classification for classifying the existence of disease. He has used liver disorder UCI dataset with 3245 instance and seven attributes. The model used to find the chances of occurrence of liver diseases on the basis of input variables by building an intelligent system based on feature selection.

Amit Kumar Dewangan [14] et al., proposed CART-Info Gain and CART- Gain Ratio feature selection technique for classification of thyroid disease. He has chosen CART algorithm as best model as it provided highest accuracy of 99.47%. he has applied Info Gain and Gain Ratio feature selection technique to CART to increase its performance. Info Gain and Gain Ratio feature selection technique is used to reduce the irrelevant features from original data set. After observation, CART-Info Gain and CARTGain Ratio gave 99.47% and 99.20% accuracy with 25 and 3 feature respectively.

T.Karthikeyan[15] et al., proposed PCA-NB algorithm to improve the prediction accuracy of the classification. He has applied Principal Component analysis (PCA) as a feature evaluator and ranker for searching method. Naive Bayes algorithm is used as a classification algorithm. He has used hepatitis patients UCI dataset with 155 instances and 19 attributes. PCA-NB improved the accuracy of classification to 89%.

All Rights Reserved © 2018 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

Table 1: Summary of medical data mining techniques

Author	Techniques Applied	Data set	Result
R. Subhashini	Optimal Fuzzy-K nearest neighbour technique	Kidney disease	89%
S. Vijayarani	Naïve Bayes and Support Vector Machine	kidney function	SVM
Manish Kumar	Random Forest classifiers, Sequential Minimal Optimization, Naïve Bayes, Radial Basis Function and Multilayer Perceptron Classifier and Simple Logistic	chronic kidney disease	Random forest
Prerana	backward propagation algorithm in neural network	thyroid disease	achieved target in 59 epoch
K.Saravana Kumar	K-Nearest Neighbour and support vector machine	thyroid disease	96.34% accuracy with knn
Anurag upadhayay	compared C4.5 and C5.0	thyroid disease	95% accuracy with C5.0
Hui-Ling Chen	feature subset - PSO -SVM with 10 fold cross validation	thyroid disease	98% accuracy
Ali Keles	Fuzzy rules using neuro fuzzy method	thyroid disease	95.33%
Amit Kumar Dewangan	CART-InfoGain and CART-Gain ratio feature selection techniques	thyroid disease	99% accuracy
P.Thangaraju	PSO-KSTAR	Liver disease	100%
T.Karthikeyan	PCA-NB	Liver disease	89%

III.CLASSIFICATION TECHNIQUES

A. Decision tree: A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The decision tree can be linearized into decision rules, the contents of the leaf node is the outcomes, and the conditions along the path form a conjunction in the if clause. The rules have the general form:

if <condition1> and<condition2> and <condition3> then outcome.

Decision rules can be formed by generating association rules keeping target variable on right. They can also represent temporal or causal relations.[17]

Advantages:

- Decision trees are self–explanatory and easy to follow.
- Decision trees implicitly perform variable screening or feature selection.
- Set of rules can be constructed with the help of decision trees.
- Decision tree can be used to represent any discrete-value classifier as it can handle both type of attributes, nominal as well as numeric input attributes.[16].

Disadvantages:

- They are unstable as a small change in the data can lead to a large change in the structure of the optimal decision tree.[17]
- Calculations can get very complex, particularly if many values are uncertain and/or if many outcomes are linked.
- As decision trees use the divide and conquer method. Some algorithms like ID3 and C4.5 require the target attributes to have only discrete values.
- Performance of decision trees becomes low if there are more complex interactions among attributes.

All Rights Reserved © 2018 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

B. Naive Bayes Classifiers: Naive Bayes classifiers are based on Bayes' theorem with strong independence assumptions between the features. All naive Bayes classifiers believe that each feature is independent of other feature, given the class variable. Naive Bayes classifier calculates the probabilities for every feature. The features which gives the outcome with highest probability are selected.

Advantages

- Naïve Bayes Classifiers can be built with real-valued inputs.
- Naïve Bayesian classifier is the simplest algorithm among classification algorithms.
- It can readily handle a large data set with many attributes.
- The naive Bayesian classifier needs only small set of training data to develop accurate parameter estimations because it requires only the calculation of the frequencies of attributes and attribute outcome pairs in the training data set.[15]
- It can be applied for Real time Prediction, Text classification/ Spam Filtering, Recommendation System.

Disadvantages

- It cannot learn interaction between the features.
- Naive Bayes classifier makes assumption based on features independency which will be not be practically applicable to all data.

C. Artificial neural networks (ANNs): An ANN is based on a collection of connected units or nodes called artificial neurons. Each neurons can transmit a signal to one another through the connection between them. Each neuron receives the signal from other neuron and process it and send it as signal to other artificial neurons connected to it. Neural networks helps in recognizing patterns and making simple decisions.

Initially each neuron in artificial neural networks has random weights assigned to it. The ANN must be trained to solve the particular problem for. A back-propagation ANN is trained by humans to perform specific tasks. During the training period, By observing the pattern of ANN's output we can evaluate its correctness . If the output is correct the neural weightings that produced that output are reinforced; if the output is incorrect, those weightings responsible can be diminished.

Advantages

- They can work well even with incomplete data.
- When a neuron in network fails it continues to work without any problem.

Disadvantage

- Neural networks need training to work.
- Processing time is high for large network.

IV.CONCLUSION:

This paper provides a study and uses of various classification algorithm which can be applied in medical sector. From Literature survey this paper also provides various implementation ideas of classification algorithms suggested by other researcher. The accuracy of classification algorithm can be improved by pre-processing the raw data and combining feature selection techniques to it. As number of attributes reduces the accuracy level can be increased.

V.REFERENCE

[1] R. Naveen Kumar, M. Anand Kumar, Medical Data Mining Techniques for Health Care Systems, International Journal of Engineering Science and Computing, April 2016, ISSN 2321 3361

[2] Ibrahim M. El-Hasnony, Hazem M. El Bakry, Ahmed A. Saleh, Data Mining Techniques for Medical Applications: A Survey, Mathematical Methods in Science and Mechanics, 2017, ISBN: 978-960-474-396-4

[3] R. Subhashini, M.K. Jeyakumar, OF-KNN Technique: An Approach for Chronic Kidney Disease Prediction, International Journal of Pure and Applied Mathematics, ISSN: 1314-3395

[4] S. Vijayarani, S.Dhayanand, Data Mining Classification Algorithms For Kidney Disease Prediction International Journal on Cybernetics & Informatics, 2015.

[5] Manish Kumar, Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm, International Journal of Computer Science and Mobile Computing, 2016, ISSN 2320–088X.

[6] Prerana, Parveen Sehgal, Khushboo Taneja,Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network, International Journal of Research in Management, Science & Technology,2015 E-ISSN: 2321- 3264

[7] K. Saravana Kumar, Dr. R. Manicka Chezian ,Support Vector Machine And K- Nearest Neighbor Based Analysis For The Prediction Of Hypothyroid,International Journal of Pharma and Bio Sciences,2014, ISSN 0975-6299.

[8] Anurag Upadhayay, Suneet Shukla, Sudsanshu Kumar, Empirical Comparison by data mining Classification algorithms (C 4.5 & C 5.0) for thyroid cancer data set, International Journal of Computer Science & Communication Networks, 2013, ISSN:2249-5789.

[9] V Prasad,T Srinivasa Rao,A Veera Reddy,B Chaitanya ,Health Diagnosis Expert Advisory System on Trained Data Sets for

All Rights Reserved © 2018 IJARTET

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

Hyperthyroid, International Journal of Computer Applications,2014 *ISSN:0975 – 8887*

[10] Wei-Wen Chang , Wei-Chang Yeh, Pei-Chiao Huang ,A hybrid immune-estimation distribution of algorithm for mining thyroidgland data,2010,Expert Systems with Applications 37 (2010) 2066–2071.

[11] Hui-Ling Chen , Bo Yang , Gang Wang , Jie Liu ,Yi-Dong Chen , Da-You Liu,A Three-Stage Expert System Based on Support Vector Machines for Thyroid Disease Diagnosis, Springer Science+Business Media J Med Syst 2012,ISSN 1953–1963.

[12] Ali Keles, Aytu¨rk Keles , ESTDD: Expert system for thyroid diseases diagnosis,2008, Expert Systems with Applications 34 242–246

[13] P. Thangaraju1, R. Mehala, Performance Analysis of PSO-KStar Classifier over Liver Diseases, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) , 2015 ISSN: 2278 – 1323

[14] Amit Kumar Dewangan, Akhilesh Kumar Shrivas, Prem Kumar,Classification of Thyroid Disease with Feature Selection Technique, International Journal of Engineering and Techniques,June 2016 , ISSN: 2395-1303

[15] T.Karthikeyan, P.Thangaraju, PCA-NB Algorithm to Enhance the Predictive Accuracy, International Journal of Engineering and Technology (IJET), MARCH 2014,ISSN : 0975-4024.

[16] Parvez Ahmad, Saqib Qamar, Syed Qasim Afser Rizv, Techniques of Data Mining In Healthcare: A Review, International Journal of Computer Applications, June 2015,ISSN :0975 – 8887.

[17] https://en.wikipedia.org/wiki/Decision_tree

All Rights Reserved © 2018 IJARTET