International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600

http://www.ijcsjournal.com Volume 5, Issue 1, No 5, 2017 ISSN: 2348-6600
Reference ID: IJCS-177 PAGE NO: 1079-1083

Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

A SURVEY ON CLASSIFICATION TECHNIQUES IN DATA MINING FOR EARLY DETECTION OF LIVER DISEASE FROMINDIAN LIVER PATIENTS DATA

P.LauraJuliet1Dr.P.R.Tamilselvi2
1Assistant Professor of Computer Applications,Vellalar College for Women(Autonomous),Erode,India
laurapaulraj@gmail.com
2 Assistant Professor of Computer Science, Government Arts & Science College, Komarapalayam ,India
selvipr2003@gmail.com

ABSTRACT- Data mining techniques plays a very important role in health care industry to predict and diagnosis the disease in early stage with the use of machine learning tool. Liver disease is one of the growing diseases these days due to the changed life style of people. Various authors have worked in the field of classification of data and they have used various classification techniques like Decision Tree, Support Vector Machine, Naïve Bayes, Artificial Neural Network etc. These techniques are useful in timely and accurate classification and prediction of diseases and better care of patients. The main focus of this work is to analyse the use of data mining techniques by different authors for the prediction and classification ofliver disease. The dataset used in this work is Liver functional test data from Indian liver patients.

Keywords :Data Mining, Classification Techniques,liver disease,Liver Functional Test Data ,Indian Liver Patients.

INTRODUCTION

Liver is the largest glandular organ of the body, it weighs about 3lb (1.36kg). It is reddish brown in color and is divided into four lobes of unequal size and shape. The liver lies on the right side of the abdominal cavity beneath the diaphragm.[9]The functions of liver Helping to process fats and proteins from digested food.Making proteins that are essential for blood to clot.Processing many medicines which we consume.Helping to remove poisons and toxins from our body.Storing fuel (called glycogen) which is made from sugars.Liver function tests (LFTs) During the function of the liver, it makes chemicals that pass into the bloodstream. Various liver disorders alter the blood level of these chemicals. Liver function tests are commonly done on a blood sample. It helps to diagnose liver disorders and monitor the activity and severity of liver disorders.To screen for any potential liver disease.To check medicines, do not cause liver damage as a side-effect.[9]

DATA MINING AND ITS TECHNIQUES

The term data mining refers to the findings of relevant and useful information from databases. This encompasses a number of technical approaches, such as clustering, data summarization, classification, finding dependency network, analysing changes and detecting anomalies. Data mining and knowledge discovery is the new interdisciplinary field, merging ideas from statistics, machine learning, database and parallel computing. KDD

KDD is a process of extracting the hidden knowledge from data warehouse. Steps in, knowledge discovery data process are Selection, Pre-processing,data Transformation, data mining ,data integration, Evaluation and Visualization. There are various data mining techniques like classification, clustering, association rule mining and prediction.

2.2Classification Techniques in Data Mining

Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data Classification predicts categorical labels. Classification models are tested by comparing the predicted values to known target values. Classification is composed of two steps – training and testing.

A1.1 C4.5 [1] is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees and rules derivation. C4.5 is classification algorithm that can classify records that have unknown attribute values by estimating the probability of various possible results unlike CART, which generates a binary tree.

A.1.2 CHAID[9] attempts to stop growing the tree before over fitting occurs, whereas CART, ID3 and C4.5 generate a fully grown tree and then carry out pruning as post processing step. In this sense CHAID avoids the pruning phase.

A.1.3 Random Forest[2] is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of classes output by individual trees. Random forests are often used when we have very large training datasets and a very large number of input variables.

B. Support Vector Machine[5][6][7] divide the n dimensional space representation of the data into two regions using a hyper plane. This hyper plane always maximizes the margin between the two regions or classes. The margin is defined by the longest distance between the examples of the two regions or classes and is computed based on the distance between closest instances of both classes to the margin, which are called supporting vectors.

C.Naive Bayesian[1][2] Naive Bayes is a simple technique for constructing classifiers models that assign class labels to problem instances represented as vectors of feature values it is not a single algorithm for training such classifier. In medical diagnosis there is a list of symptoms, X are treated as features in naive bayes, then predict Y patient has disease.

D.Multilayer Perceptron(MLP)[8][10][4] is a development from the simple perceptron in which extra hidden layers are added. The presence of these layers allows an ANN to approximate a variety of non-linear functions. The transfer function generally a sigmoid function. Multilayer perceptron is a neural network that trains using back propagation.

LITERATURE REVIEW

BendiVenkataRamana et.al.,[4] proposed "Liver Classification Using Modified Rotation Forest".Modified Rotation Forest algorithm was proposed for accurate liver classification by analysing ten popular Classification Algorithms were considered with all the combinations of four feature selection methods for evaluating their classification performance in terms of accuracy in classifying two liver patient data sets. It concluded that Multi-layer perception classification algorithm and random subset feature selection method gives highest accuracy that is 74.7826% for BUPA liver data set. Nearest neighbour classification algorithm and correlation based feature selection method gives highest accuracy that is 73.0703% for ILPD liver data set.

A.S.Aneeshkumar et.al.[1]have proposed "Estimating the Survallience of Liver Disorder Using Classification Algorithms" . They have worked on Naïve Bayesian and C4.5 decision tree algorithms for classification of liver patient. For this research work, real medical data with 15 attributes were collected from a public charitable hospital in Chennai. This work gives us a binary classification i.e. either a person is a liver patient or not. As a result the maximum accuracy 99.20% lies in C4.5 decision tree with 90-10 splitting ratio. But when compare to Naive Bayesian, it is somewhat lazy, because the average time taken by C4.5 is 0.16 seconds. Naive Bayesian used only 0.03 seconds to achieve the same.

BendiVenkataRamana et.al.[3]proposed "A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis". In this study, the common attributes of the two data sets ALKPHOS, SGPT and SGOT are taken for one way analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA) and also observe any significant difference among the groups. Experiment 1 includes the analysis of all Patients that means both liver and non-liver patients, Experiment 2 includes the analysis of liver Patients and Experiment 3 includes the analysis of non-liver Patients. All these experiments shows that there exists more significant difference in the groups with all the possible attribute combinations.

AnjuGuliaet.al.,[2]proposed "Liver Patient Classification Using Intelligent Techniques" . This paper implements hybrid model with three phases using J-48, Multi Layer Perceptron, Support Vector Machine, Random Forest and Bayesian Network .In first phase, Applying Classification Algorithm without Feature Selection concluded that SVM algorithm is considered as the better performance with an accuracy of 71.3551%. In second phase, Applying Classification Algorithm after Feature Selection concluded that Random Forest algorithm is considered as the better performance with an accuracy of 71.8696%. In third phase, the results of classification algorithms with and without feature selection are compared with each other. The results obtained indicate that Random Forest algorithm outperformed all other techniques with the help of feature selection with an accuracy of 71.8696%.

Hoon Jin et.al.[8]proposed “Decision Factors on Effective Liver Patient Data Prediction”and evaluated on Indian Liver Patient data (ILPD) dataset.This research work analysis the classification algorithms such as Naïve Bayes , Decision Tree, Multilayer Perceptron and K-NN used in the previous study and additionally Random Forest, Logistic which was proposed by them. For comparison of selected classification algorithms, cross validation method was used and concluded that, in view of precision, Naive Bayes is preferable than others. In view of Recall and Sensitivity, Logistic and Random Forest took precedence over other algorithms in the performance of prediction test.

Esraa M. Hashem, et.al,[7] proposed A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis. In this work, Support vector machine is used for classifying liver disease using two liver patients datasets with different features combinations such as SGOT, SGPT and Alkaline Phosphates. Results show that the accuracy, error rate, sensitivity and prevalence at first 6 ordered features are the best for ILPD dataset compared to BUPA dataset which can help in diagnosis of liver disease.

Ch. Sanjeev Kumar Dash and PulakSahoo[5] proposed An Empirical Analysis of Evolved Radial Basis Function Networks and Support Vector Machines with Mixture of Kernels. They evolved radial basis function network with novel kernel and support vector machine with mixture of kernels are suitably designed for the purpose of classification of imbalanced dataset. The experimental outcome shows that support vector machine with mixture kernels is better than evolved radial basis function neural networks. DE-RBFNs work well for Indian Liver Patients with the accuracy73.10% and Bupa Liver Disorders datasets Patients with the accuracy 74.10%.

Dr. S. Vijayarani, Mr.S.Dhayanand[6], proposed ”Liver Disease Prediction using SVM and Naïve Bayes Algorithms”. The algorithms used in this work are Naïve Bayes and support vector machine (SVM) applied on Indian Liver Patient Dataset) taken from the UCI Repository. Outliers are predicted in this work based on the moderate range of the liver function test results. From the experimental results the SVM is a better classifier for predict the liver diseases with the accuracy 79.66%. Naïve Bayes performs with minimum period of execution time that is 1670.00ms than SVM.

MoloudAbdar et.al.,[9]proposed “Performance analysis of classification algorithms on early detection of liver disease”. This research work analysis the classification algorithms such as boosted C5.0 and CHAID algorithms that are relevant to the decision trees in order to discover hidden knowledge in the Indian liver patients data(ILPD) dataset in UCI repository. Comparing the performance using confusion matrix and various indicators such as specificity, sensitivity, precision and accuracy of assessment categories are calculated.The results obtained show thatC5.0 algorithm via Boosting technique has an accuracy of 93.75% and this result reveals that it has a better performance than the CHAID algorithm which is 65.00%.

PoojaShrivastava ,YuktiKesharwani[10] proposed An Efficient Classifier using Multilayer Perceptron for Classification of Liver patient . In this research work, They will use MLP to develop the robust classifier which can classify data as liver disease or non-liver disease. Proposed model MLP achieved 77.77% of accuracy in case of hidden layer 2 and learning rate 0.7 as robust model.

CONCLUSION

In medical science, diagnosis of health condition is very challenging task. This study surveyed some data mining techniques to predict the liver disease at earlier stage. The study analysed algorithms such as C4.5,C5,CHAID, Naive Bayes, Decision Tree, Support Vector Machine, Back Propagation Neural Network and Classification and Regression Tree Algorithms. These algorithm gives various result based on speed, accuracy, performance and cost. This paper provides the summary of the use of data mining techniquesto diagnosis the liver disease at earlier stage.

REFERENCES

[1] A.S. Aneeshkumar and C. JothiVenkateswaran, Estimating the Survallience of Liver Disorder using Classification Algorithms, International Journal of Computer Applcations, Vol 57- No.6, pp. 39-42, November 2012.

[2] AnjuGulia ,Dr.RajanVohra , Praveen Rani, Liver Patient Classification Using Intelligent Techniques, International Journal of Computer Science and Information Technologies, Vol. 5 (4) , pp. 5110-5115, 2014.

[3] BendiVenkataRamana, Prof. M. Surendra Prasad Babu, Prof. N. B. Venkateswarlu, A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis, International Journal of Engineering Research and Development, Volume 1, Issue 6, PP.17-24, June 2012.

[4] BendiVenkataRamana, Prof. M.Surendra Prasad Babu , "Liver Classification Using Modified Rotation Forest", International Journal of Engineering Research and Development, Volume 1, PP.17-24, Issue 6, June 2012.

[5] Ch. Sanjeev Kumar Dash and PulakSahoo, An Empirical Analysis of Evolved Radial Basis Function Networks and Support Vector Machines with Mixture of Kernels. International Journal onArtificial Intelligence Tools, Vol. 24, No. 4 ,2015.