Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 5, May 2016, pg.483 – 488

A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder

D.Sindhuja
Department of Computer Science, Bishop Heber College (Autonomous), Trichy-17, Tamilnadu
sindhujai666@gmail.com

R. Jemina Priyadarsini
Assistant Professor in Computer Science, Bishop Heber College (Autonomous), Trichy-17, Tamilnadu
jemititus@gmail.com

Abstract— Data mining is the process of extracting meaningful information from large database. In Medical field the problem may arise in the era data mining has vital role to predict and diagnosis the disease in early stage with the use of machine learning tool. Liver is the largest internal organ in the human metabolism plays an important role in human body and doing several vital functions. Liver disease may cause symptoms like Jaundice, Tendency to bruise, Bleed easily, Ascites, Impaired brain function, General failing health. The Liver disease is caused by the person who takes more alcohol in short time. Types of liver disease are Acute liver failure, Hepatitis, Liver Cancer, and Cirrhosis. In India many men's were affected by liver disease disorder. This paper describes survey on classification techniques in data mining for analyzing liver disease disorder.

Keywords— Data Mining, Liver Disease Disorder Data Set, Data Mining in Medical, Classification Techniques.

I. INTRODUCTION

Liver is the largest glandular organ of the body, it weights 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. The blood is carried out to the liver through two large vessels called the hepatic artery and the portal vein. Liver tissue is composed of thousands of lobules, and each lobule is made up of hepatic cells the basic metabolic cells of the body. [1]

Various types of stress and irregular eating habits as well as inhalation of alcohol and ongoing toxic gas, indigestion of contaminated food, excessive consumption of pickled food and drug intake, enables liver disease patients to grow up year by year.[2] The person who cause by liver disease has some more symptoms they are dark urine, pale stool, bone loss, easy bleeding, itching, spider like blood vessel visible in the skin, enlarged spleen, fluid in abnormal cavity, chills pain from the biliary track or pancreas, and enlarged gallbladder.[3]. Alcohol abuse generally leads to three pathologically distinct liver disease they are fatty liver, hepatitis and alcoholic cirrhosis. One or all of the three can occur at the same time and in the same patient. [4]

Data mining is process of analyzing of bulk amount of data to automatically discover the interesting regularities or associations which in turn lead to improved understanding of the original processes [5].To find the useful classes or patterns using decision making. There are two categories of data mining are: 1.Data mining in Descriptive.2.Data Mining in Predictive. Descriptive data mining, it generalizes or summarizes the general properties of the data in the database. Predictive data mining is searched to inference on the present data to make predictions .Classification maps data into predefined groups, it is often referred to as supervised learning as the classes are determined prior to examining the data. Classification algorithms usually require that the classes be defined based on the data attribute values.

Classification is the technologies used for classify the data and predict the accuracy for the future work with the use of behind and present data. The main aim of the classification techniques is to analyze the input data. Data mining plays a vital role in medical field to find the relationship between patient data and medical data set from the large database. Here, specifically takes the liver disease disorder data from the medical database. This paper shows a survey about the liver disease disorder from various papers and gives the idea for the future work, that which data mining and diagnosis the liver disease disorder. The rest of this paper is organized as follows: Section II as Literature Survey, Section III as classification techniques in data mining, Section IV as Conclusion and references.

II. LITERATURE SURVEY

Bendi Venkata Ramana [6] the author evaluate the different types of liver dataset that is AP liver dataset and UCLA dataset and then he evaluate the performance of the classification techniques from precision, accuracy, specificity and sensitivity. The author said, AP liver dataset is better than the UCLA liver dataset. Using classification algorithm, they are support vector machine, C4.5, Back propagation neural network algorithm, and Naive bayes classifier.

Aneeshkumar.A.S[7],in his study there is a methodology used to effective classification of liver and non-liver disease dataset. 15 attributes of real medical data are collected from dataset. Classification techniques are used to classify the data. Pre-processing method is used to cleansing the data for effective classification, after cleansing the data. C4.5 and Naive Bayes are the two algorithms used in his study. Datasets are divided into three different types of ratio based on average and standard deviation of each factor of both class and evaluated the accuracy. After evaluate the accuracy he said C4.5 is gives better accuracy than Naive Bayes, because it gives more accuracy with the minimum time taken.

Bendi Venkata Ramana and M.Surendra Prasad Babu [8] in his paper, Modified Rotation Forest algorithm were proposed to calculate the accuracy of the liver classification techniques in UCI liver dataset using the combo of feature selection technique and selected classification technique algorithm.

Ratnamala Kiruba.H [9] from her paper, she used intelligent agent based system to hike a precise and accurate of diagnosis system. C4.5 decision tree algorithm and Random tree algorithm are used to predict. Two different types of liver disease disorder dataset are combined and predict the accuracy of the disease. And then conclude these both algorithms gives very good accuracy for diagnosing liver disease disorder.

Dhamodharn.S [10] in his paper he reviewed the classification technique algorithm in data mining techniques for liver disease disorder. Particularly, compared two decision tree algorithms that is FT growth and Naïve Bayes and found out which algorithm gives better accuracy. From his study he said Naïve Bayes is better than FT growth algorithm with the use of machine learning because, Naïve Bayes (75.54%) gives more accuracy than FT growth algorithm (72.66) using WEKA Tool. This comparison happened among 29 datasets with 12 different attributes.

Rajeswari.P, Sophia Reena.G [11], in his paper, said BUPA liver disorder dataset for early diagnosis the disease. Classification technique algorithms such as Ft tree Naïve Baiyes and Kstar are used to predict the liver disease disorder with evaluate using 10-fold cross validation. Then the results which got from using these algorithms are compared. Generally, accuracy and time taken to build the algorithms are compared and finally she said that from comparing all those algorithms, FT tree gives better role for increasing the accuracy of the dataset in classification technique algorithm.

Gunasundari S and Janakiraman S [12] from his study, said many article which is using various textual analysis method for liver disease disorder classification from abdominal Computed Tomography scans and finally conclude conversional image processing operations, neural networks and Genetic algorithm gives successful result for liver disease disorder diagnosis. In future liver disease disorder diagnosis extended in many directions. Such as using effective algorithms and more texture feature technique algorithms.

Hyontai Sug [13] in his study, he compensate the insufficiency of liver disease disorder data usefully, he said a method based oversampling in minor classes. Decision tree algorithm does not give high priority for minor classes for that reason using duplication in BUPA liver disease disorder dataset, increase the number of instances of minor class and proceed with two decision tree algorithms. Such that CART and C4.5 both algorithms are gives good result with oversampling for liver disease disorder dataset, but in future work can reduce the minor class increment to smaller percentage.

RONG-HO LIN [14] in his study, author used Case Based Reasoning (CBR) and Classification and Regression Tree (CART) techniques which are useful to detect the liver disease. The data sets were taken from the medical center in Taiwan from 2005 to 2006.

CK Ghosh [15] in his study, the liver abscess was the commonest cause of hepatomegaly and it was due to amoebiasis, followed by fatty liver, congestive cardic failure, hepatocellular carcinoma, and viral hepatitis seen only in few patients. The most common disease was hepatic steatosis, followed by cirrhosis, portal traditis and chronic hepatitis.

Newton Cheung[16] he found, using data mining classification techniques he found various results using C4.5 algorithm gives 65.59%, using Naive Bayes gives63.39%, using BNND(Bayesian Network with Naïve Dependence)gives 61.83%, using BNNF (Bayesian Network with Naïve Dependence & Feature Selection) gives 61.42%.

S.NO	ATTRIBUTES	DESCRIPTION
1	Alkphos	Alkaline phosphotase
2	ALB	Albumin
3	DA	Disturbance in abdomen
4	Drinks	Alcoholic beverages drunk per day
5	FAC	Frequent alcoholic consumption
6	Gammagt	Gamma-glutamyl transpeptidase
7	Mcv	Mean corpuscular volume
8	Sgpt	Alamine aminotransferase
9	Selector	Used to split data into two sets
10	Sgot	Asparatate aminotransferase
11	TB	Total Bilirubin
12	YUD	Yellowish Urinary Discharge

III. CLASSIFICATION TECHNIQUES IN DATA MINING

C4.5

In Classification techniques, C4.5 algorithm is used to generate decision tree. Improvements from ID.3 algorithm is C4.5 algorithm. Using the concept of information entropy, C4.5 builds decision trees from a set of training data. C4.5 follows a post pruning approach. Information gain is normalized from the splitting criteria. Using divide and conquer algorithm, C4.5 first grows an initial tree. This algorithm performs well in noise free data.

Advantage:

C4.5 algorithms construct trees and grow its branches.

The attribute with the highest normalized information gain is chosen to make the decision.

This algorithm is used to handle continuous and discrete values.

Disadvantage:

The C4.5 algorithm contains empty branches.

The insignificant branches not only reduce the usability of decision.

Over fitting happens in C4.5 algorithm.

Naive Bayes Classifier:

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.

Rule:

P(Y|X1__Xn) = P(X1___Xn|Y)P(Y)/ P(X1___Xn)

P(X1____Xn|Y) is referred as likelihood,

P(Y) is referred as prior,

P(X1_____Xn) is referred as normalization constant.

*Advantage**:***

Training is very easy and fast.

Naive bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification.

Naive bayes classifiers have worked quite well in many complex real world problems.

Naive bayes algorithm affords fast highly scalable model building and scoring.

Naive bayes can be used for both binary and multiclass classification problems.

Disadvantage:

Naive bayes considering each attribute in each class separately.

Violation of independence assumption.

Zero conditional probability problem.

Decision Tree:

The Decision Tree [18] consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. Decision tree are like those used in decision analysis where each non-terminal node represents a test or decision on the data item considered. Depending upon the outcome of the test, one chooses a certain branch. Decision tree can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules.

Advantage:

The estimation criterion [19] in the decision tree algorithm is the solution of an attribute to test at each decision node in the tree.

Disadvantage:

Decision tree is that trees use up data very rapidy in the training process. They should never be used with small data sets. They are also highly sensitive to noise in the data and they try to fit the data exactly.

Back Propagation Neural Network:

The Back Propagation algorithm uses supervised learning, which means that we provide the algorithm with examples of the inputs and outputs we want the network to compute and then the error is calculated. The main idea of the back propagation algorithm is to reduce the error, until the neural networks learns the training data. One of the most important neural network algorithms is back propagation algorithm; this algorithm is stopped when the value of the error function has become sufficiently small. Neural network model could be created to help with classifying the new data.

Advantage:

Using high accuracy neural networks are able to approximate complex non-linear mappings.

The noise tolerance in neural network is very flexible with respect to incomplete, missing and noisy data. Neural networks can be updated with present data, making them useful for dynamic environments, because it is ease of maintenance.

Disadvantage:

There are no general methods to determine the optimal number to solving any problem.

It is difficult to select a training data set which fully describes the problem to be used.

Support Vector Machine:

The main aim of support vector machine is to find the accurate classification technique to differentiate between members of the two classes in the training data. In support vector machine technique the optimal boundary is known as hyper plane. The vectors that are placed near the hyper plane are called supporting vectors. If the space is not linearly separable there may be no separating hyper plane.

Advantage:

Support vector machine can be easily extended to perform numerical calculations.

Support vector machine is very useful for general pattern recgonization, regression and classification.

Support vector machine can be used for pattern classification [20] which has multi layer perceptions and radial-basis function network.

Disadvantage:

Support vector machine is computational inefficiency.

To finding an approximate minimum enclosing balls a set of instances.

Classification and Regression:

Classification and Regression Tree supports high speed deployment. Classification and Regression Tree is used to display important data relationships that could remain hidden using other analytical tools very quickly. Classification and Regression Tree also denoted as CART. Classification and Regression Tree builds multivariate decision binary trees.

Advantage:

Classification and Regression Tree does not require specification of any functional form, it is non parametric.

This algorithm will itself identify the most significant variables and eliminate the non-significance ones. It does not require variables to be selected in earlier.

Classification and Regression Tree can easily handle outliers and noisy data. Classification and Regression Tree is flexible and has an ability to adjust in time.

Disadvantage:

Classification and Regression Tree may have unstable decision tree. The decision tree is increase or decrease of tree complexity, changes splitting variables and values.

Classification and Regression Tree splits only by one variable.

IV. CONCLUSION

The study surveyed some data mining techniques to predict the liver disease at earlier stage. The study analyzed algorithms such as C4.5, 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. It is seen that C4.5 gives better results compare to other algorithms. In future an improved C4.5 could be derived with various parameters. This paper gives generalization of various data mining techniques to diagnosis the liver disease at earlier stage.

References

[1] P.Rajeswari and G.Sophia Reena, " Analysis of Liver Disorder Using data mining AAlgorithms", Global Journal of Computer Science and Technology, Volume.10 issue 14(version 1.00 November 2010) PP 48.

[2] Hoon Jin, Seoungchon Kim, Jinhong Kim,"Decision Factors on effective liver patient Data Prediction", International journal of Bio-Science and Bio-Technology, Volume6 No 4(2014) PP 167-168.

[3] Cassangnou M, Boruchowicz A, Guillemot F "Hepatic Steatosis revealing celiac disease: A case complicated by transistory liver failure.AMJ Gastroenterol 1996;91: PP 1291-1292.

[4] Bal MS,Singh SP, Bodal VK, Oberoi SS, Surinder K, "Pathological findings in liver autopsy", Jounal of Indian Academy of Forensic Medicine 2004; 26(2) PP 971-973.

[5] Yao,H.,Hamilton, H.J., Buzz, C.J.,"A foundational Approach to mining itemset utilities from databases", In 4th SIAM International Conference on Data Mining, Florida USA(2004).

[6] Bendi Venkata Ramana, M.rendra Prasad Babu and N.B. Venkaeswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis" , International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011.
[7] A.S.Aneeshkumar and C.Jothi Venkateswaran, "Estimating the Surveillance of Liver Disorder using Classification Algorithms", International Journal of Computer Applications (095-8887), Volume 57-No.6, November 2012.
[8] Bendi Venkata Ramanaland Prof.M.Surendra Prasad Babu, " Liver Classification Using Modified Rotation Forest", Internationa Journal of Engineering Research and Development ISSN: 2278-067X, Volume 1, Issue 6 (June 2012), PP.17-24.
[9] H.Ratnamala Kiruba and Dr G. Tholkappia arasu, " An Intelligent Agent based Framework for Liver Disorder Diagnosis Using Artificial Intelligence Techniques", Journal of Theoretical and Applied Information Technology, Vol.69 No.1, 10th November 2014.
[10] S.Dhamodharan, "Liver Disease Prediction Using Bayesian Classification", 4th National Conference on Advanced Computing, Applications & Technologies, Special Issue, May 2014.
[11] P.Rajeswari and G.Sophia Reena, " Analysis of Liver Disorder Using data mining AAlgorithms", Global Journal of Computer Science and Technology, Volume.10 issue 14(version 1.00 November 2010 PP 48-52.
[12] Gunasundari S and Janakiraman S, "A Study of Textural Analysis Methods for the Diagnosis of Liver Disease from Abdominal Computed Tomography", International Journal of Computer Applications (0975-8887), Volume 74-No.11, July 2013.
[13] HYONTAI SUG**,** " Improving the Prediction Accuracy of Liver Disorder Disease with Oversampling", Applied Mathematics in Electrical and Computer Engineering, American-MATH 12/CEA12 proceedings of the 6th Applications and proceedings on the 2012 American Conference on Appied Mathematics 2012 PP 331-335.
[14] R.H.Lin, " An Intelligent model for liver disease diagnosis", Artificial Intelligence in Medical Volume 47, Issue 1, 2009 PP 53-62.
[15] CK Ghosk, F.Islam, E.Ahmed, DK Ghosh, A Haque, QK Islam, "Etiological and clinical patterns of Isolated Hepatomegaly" Journal of Hepato-Gastroenterology 2012 Jan-June 2(1); PP 1-4.
[16] N. Cheung, "Machine learning techniques for medical analysis", School of Information Technology and Electrical Engineering, BsC thesis, University of Queenland, 19 Oct. 2001.
[17] Http://WWW.liver.ca/liver-disease/types/fatty-liver.aspx.
[18] Lior Rokach, "Decision tree", Department of Industrial Engineering Tel-Avir University, Oden Maimon, Department of Industrial Engineering Tel-Avir University.
[19] H.Hamilton E.Gurak, L.Findlater W.Olive,"Overview of Decision Trees".
[20] A.H.Rosalina and A. Noraziah(2010). "Prediction of Hepaitis Prognosis Using Support Vector Machine and Wrapper Method", IEEE, PP 2201-2211.

