Diagnosis of Hepatitis using Decision tree algorithm

V.Shankar sowmien [#], V.Sugumaran [#], C.P.Karthikeyan[#], T.R.Vijayaram [#]

[#]School of Mechanical and building Sciences, VIT University, Chennai, India
[1]vshankar.sowmien2014@vit.ac.in
[2]cpkarthikeyan@vit.ac.in
[3]sugumaran.v@vit.ac.in
[4]vijayaram.tr@vit.ac.in

Abstract: This research paper proposes a prediction system for liver disease using machine learning. Researchers provided various data to identify the causes for Hepatitis. Here, Decision tree method is used to determine the structural information of tissues. The algorithm used to construct the decision tree is C4.5 that concentrates on 19 attributes such as age, sex, steroids, antivirals, spleen, fatigue, malaise, anorexia, liver big, liver firm, spiders, vilirubin, varices, ascites, ALK phosphate, SGOT, albumin, protime, and histology for the diagnosis of the disease. These features helped in determining the abnormalities of the patient which resulted in 85.81% accuracy.

Keywords: liver disease, decision tree, diagnosis of liver disease, attributes selection.

I. Introduction

Hepatitis is caused by the inflammation of the liver. It may occur with no symptoms which lead to yellow discoloration of the skin and enlargement of the spleen. The presence of jaundice indicates the advanced liver disease. It may also lead to weight loss. Common causes for hepatitis are bacterial, fungal, parasitic and decreased blood flow.

Viral hepatitis is caused due to hepatotoxic viruses such as hepatitis A, hepatitis B, hepatitis C, hepatitis D and hepatitis E. Alcoholic hepatitis is caused due to excessive alcoholic consumption, leads to liver failure. Toxic and drug induced hepatitis is caused due to intake of chemical agents. Paracetamol is the leading cause for liver failure, which results in damaging the cell and structural changes. Auto immune hepatitis is caused by immune response. This disease occurs mostly for young women. Insufficient blood or oxygen results in ischemic hepatitis which causes heart failure. Giant cell hepatitis occurs only in new-born babies.

Data mining is used to identify uncovering patterns from the stored data and that can be marked as insufficiency of data. Predictive models can be built and developed by using the above information. Besides, this can be used to evaluate the effectiveness of medical treatment by comparing and contrasting the symptoms and courses of treatment. Data obtained from the tested results can be converted into useful information as a decision tree.

Decision tree is used to find accurate and reliable results. The structure of a decision tree is evaluated in terms of accuracy, sensitivity, and confusion matrix. It consists of attribute nodes with two or more sub trees and decision nodes. Decision tree starts with two major divisions, one as a training set in which the data is stored and the other one is a testing set, where accuracy is obtained.

II. LITERATURE SURVEY

From the study, neural networks play an important role in medical diagnostic field. Standard feed forward networks and a hybrid network were discovered to identify the hepatitis disease by using different neural network architectures [1]. Multilayer Perceptron (MLP), Radial Basis Function (RBF) and Conic Section Function Neural Network (CSFNN) are the three neural network algorithms used to detect hepatitis disease in this paper. Results show that MLP does not have less classification accuracy however RBF gives good results. CSFNN which combines both MLP and RBF are more useful for detection.

The application of artificial intelligence for Hepatitis B diagnosis has been introduced. The expert system uses a logical interface along with a neural network architecture is used to identify hepatitis [2]. The detection of hepatitis B is done by using different data samples from different patients and the results have shown that artificial neural networks are equivalently good as the logical methods in the diagnosis of hepatitis B.

A mathematical model is introduced by including HBV, hepatocytes and the immune system. In this study, interaction of virus and immune system is described [3]. This also discusses on the other side of hepatitis B virus, which can supress the immune system by increasing the death rate of CTLs. So it is impossible for the immune system to remove the virus, and persistent infection takes place to cure chronic Hepatitis B, HBV viral load reduced firstly and develop immunity to the virus.

In this study, it has been concluded that the PETs like NBTree, CITree and CLLTree have an inherent barrier to achieve both high time complexity and equal weight attribute input, this issue can be overcome by combining with an attribute selection method [4]. The model provides good performance in both classification and ranking. It is concluded that WPETs model fits for liver cirrhosis and hepatitis TCM diagnosis.

The use of decision tree C4.5 algorithm, ID3 algorithm and CART algorithm are proposed to classify the diseases and compare the effectiveness, correction rate among them [5]. A CART decision tree algorithm is proposed where accuracy and time complexity is required CART algorithm performs better than the other two algorithms.

A hybrid intelligent syndrome diagnosis (HISD) model is proposed which helps to compare the traditional Chinese medicine (TCM). Instead of a single technique, the proposed method uses multiple methods, the model is more objective. The method increases the accuracy and helps better for clinical purposes [6].Hepatitis B has become a serious health problem which causes liver infection due to hepatitis B virus [7]. The detection of disease for a new patient can be performed on the basis of primary phase.

The details from the four diagnostic methods and information of patients are gathered, which are confirmed with Hepatitis B within a certain period of time span from three hospitals. A clinical data base is constructed, which can be used as the dataset for machine learning. Then an association rule mining algorithm is applied to analyse the data automatically. Testing the obtained data with the test data, the results shows that the accuracy rate is increased to 83.5% [8].

ADD……..

III. DESCRIPTION OF ATTRIBUTES

Lesser number of parameters are considered for the diagnosis of hepatitis. This can be done only by the selection of attributes. Proper selection of the attributes will lead to the prediction of the disease with greater accuracy. These attributes can then be recorded in each step during the diagnosis process. The attributes in the order of significance for classification by C4.5 decision tree algorithm is furnished below:

- Age: 10, 20, 30, 40, 50, 60, 70, 80
- Sex: Male, Female
- Steroid: no, yes
- Antivirals: no, yes
- Fatigue: no, yes
- Malaise: no, yes
- Anorexia: no, yes
- Liver big: no, yes
- Liver Firm: no, yes
- Spleen palbable: no, yes
- Spiders: no, yes
- Ascites: no, yes
- Varices: no, yes
- Vilirubin: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00
- ALK Phosphate: 33, 80, 120, 160,200, 250
- SGOT: 13, 100, 200, 300, 400, 500
- Albumin: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0
- PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, and 90
- HISTOLOGY: no, yes
- Class: DIE, LIVE

The final attribute considered here is “CLASS”. This defines the patient’s condition after the diagnosis process done by the decision tree algorithm. The step by step procedure involved in the algorithm is explained in the following section.

IV. DECISION TREE

Decision tree is used for extracting typical patterns like the sole reason causing hepatitis B in our scenario. Graph based induction method is used for extracting patterns of various sizes. Graph structured data is represented using particular sub graph (value of attributes).This works as, if the result is YES the graph contains sub graph and if NO, the resulting decision tree becomes a binary tree.

Chiba University Hospital, China has provided the hepatitis data using the time-series data of blood inspection and urinalysis. Generally, four types of experiments are carried out, in which during the first and second experiments, the stages of fibrosis are used as classes and the tree is constructed by discriminating the patients. Third experiment determines the type of hepatitis, which are used as classes and in the last experiment, the effectiveness of interferon therapy is used as class. The process of working with decision tree starts from the bottom of the tree thereby the sub-tree at each node is examined. While removing the sub-tree if it results error in testing data the corresponding sub-tree will be removed. This process continues till there is no further improvement in the tree. Each sample should have only one rule in a tree.

The majority of the work depends on the labelling of individual leaves. The formation of node in decision tree stops, when all instances for a particular node belongs to same class or if there is no remaining attributes on which the instances cannot be partitioned further. The algorithm used here for determining the decision tree is C4.5 which has more advantages than the nonlinear functions. The C4.5 algorithm attributes to two factors which are responsible for accuracy. They are;

- single coverage constraint
- fragmentation

To discover the decision tree from a gene expression, C4.5 algorithm is used. The data is represented by nineteen features and consists of 155 training samples. Each feature is considered as a gene by having continuous expressions.

This decision tree has a structure which is shown in the Figure 1. This tree structure is built by considering 8 attributes. Decision tree show your results only if the number of attributes selected is less. Hence, it shows accurate results for this diagnosis. This process is performed by using WEKA Explorer. This structure has 11 leave nodes and the size of the tree results with 21. So, the CCI (Correctly Classified Instance) value results with 84.5161% for 8 level attribute selection.

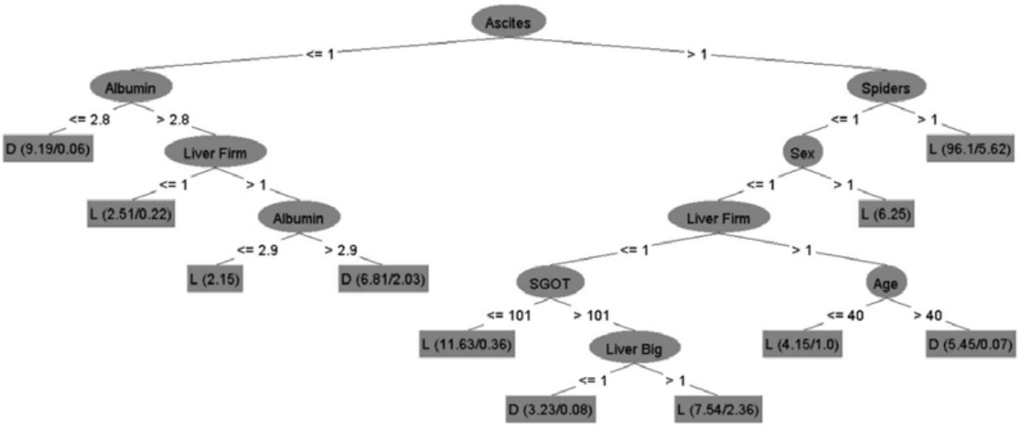

Figure 1: Decision tree with all attributes

This can be further established by considering only 2 attributes and come with better accuracy level. Totally, 155 samples have been considered. Out of these 155 samples, this tree has provided 100% accuracy for 71 samples and the remaining 85 samples show errors. Now, decision tree can be built by considering only 2 attributes. The modified tree structure is given in Figure 2. This structure uses only the CLASS attribute to build the leave node further. The sub division extends as, for node1(Ascites) (leaves& spruces),this results with the class LIVE only if its value is less than leave (spruce)1. Otherwise, the tree structure extends by considering other attribute with condition evaluation at each stage. This structure will results with greater accuracy of 85.8065% for 100 samples. The number of leaves is 5 and the size of the tree is 9.

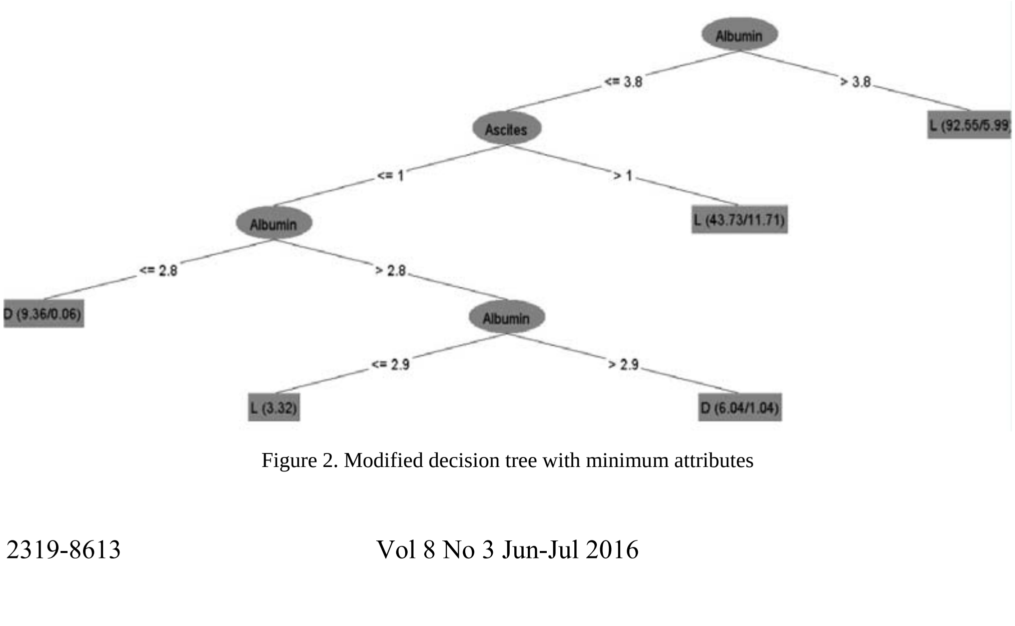

Figure 2. Modified decision tree with minimum attributes

V. RESULTS AND DISCUSSION

In this paper, 155 instances were taken, which includes both healthier and the persons affected by the liver disease. In order to evaluate the results in terms of CCI, some of the terms are considered. They are; (i) Number of attributes (ii) Minimum number of objects (m) (iii) confidence factor(c). Figure 3 represents the graph between the number of attributes and the correctly classified in terms of percentage.

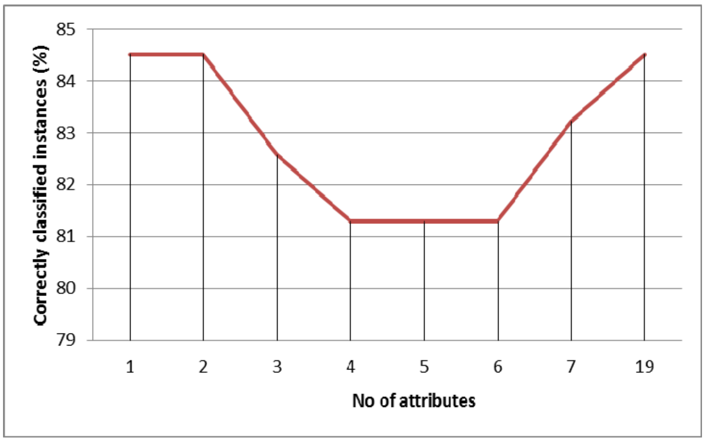

Figure 3. Graph showing CCI Vs No.of attributes

This shows that, if the number of attributes is lesser than the accuracy level for the diagnosis of the disease is higher. The graph shows maximum accuracy for attribute level 2. The value starts decreasing. Even if 19 attributes were considered, the accuracy level is also high. Less number of attribute selection will lead to reduction in the complexity of decision tree sturcture and the size can also be reduced.

Next consideration is minimum number of objects (m). This is interpreted by the graph in Figure 4. This is also based on the concept that accuracy level can be increased by considering lesser (m)value. Here, when m=2, gives the result of 85.8068%.

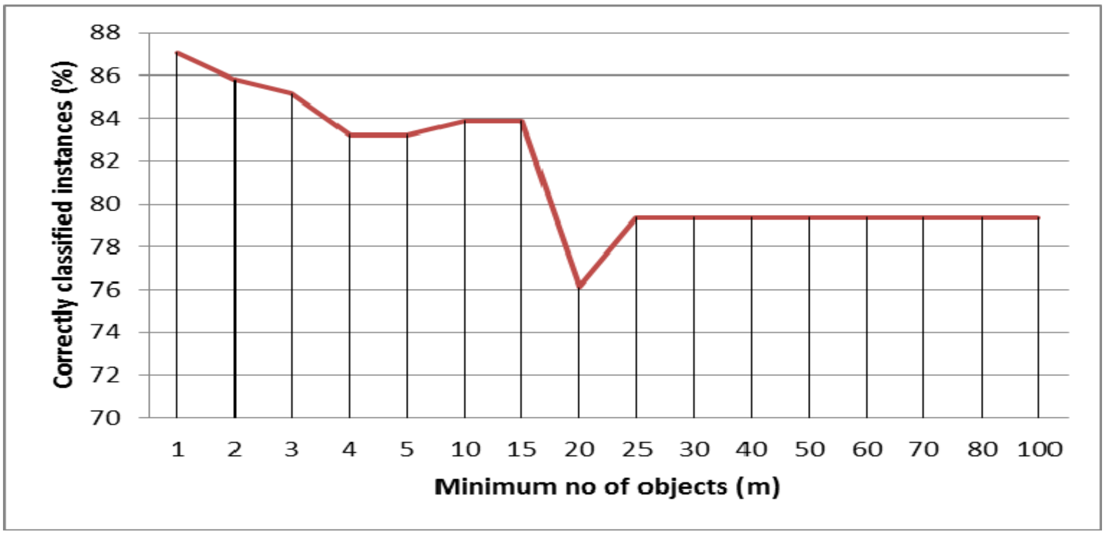

Figure 4. Graph showing CCI Vs Minimum no of objects(m)

2 minimum objects are required to correctly classify the instances C4.5 allows the decision tree to take the minimum objects to represent the features in a better way.

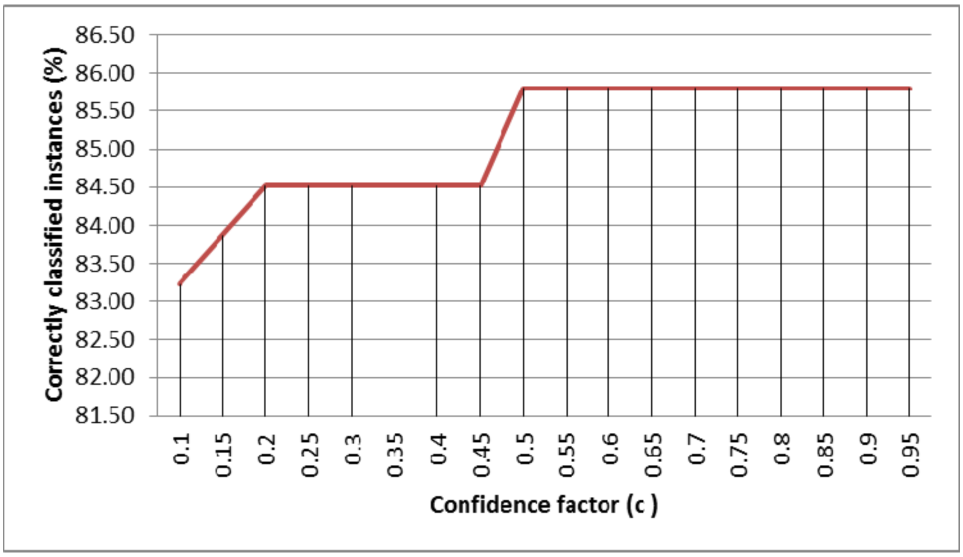

Figure 5. Graph showing CCI Vs Confidence factor

Figure-5 shown below infers that the coincidence factor, 0.25 is required to obtain the correctly classified instances. Sensitivity should be equal to 1. But 0.5 is obtained in this case which is acceptable with accuracy of 85.8051%.

Table 1: Detailed accuracy by class

	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
	0.976	0.594	0.863	0.976	0.916	0.769	L
	0.406	0.024	0.813	0.406	0.542	0.77	D
Weighted Avg.	0.858	0.476	0.853	0.858	0.839	0.769	

The detailed class wise accuracy gives better understanding of the classification. From the observation (TABLE 1), true positive rate (TP rate) and false positive rate(FP rate) are the factors which are of the most importance, in which the TP rate should be projected near the value of unity and FP rate should be closer to zero.

The classification accuracy of the C4.5 decision tree algorithm is represented in the form of confusion matrix (TABLE 2).

Table 2: Confusion matrix

A	B	Classified as:
120	3	**a** = L
19	13	**b** = D

From the inference, the following conclusions were derived:

- The correctly classified instances by the classifier are represented as the diagonal elements of the confusion matrix.
- The first element of the first row in the confusion matrix gives the number of data points belonging to the class or event "LIFE" i.e. ‘L’.
- The second element of the first row gives the number of data points belonging to class of "LIFE (A)", however misclassified under class of "DIE" i.e. ’D’.
- Similarly, number of misclassified instances in each class can be found individually. Computing the total number of misclassified instances the total error percentile of the classification is found to be 14.19%.

VI. CONCLUSION

Liver disease is one of the causes for death from last decade, which leads to hepatitis. Many researchers investigated on liver disease, in which they have used algorithms such as C5.0, PCL, J48, and fuzzy rule. In this paper, C4.5 algorithm is used which is an efficient one than the existing algorithms. Since, the number of atributes are lesser, the complexiety of the decision tree is reduced.Vilirubin and varices plays an important rule in determining the patients abnormalities 85.81% accuracy is obtained from the overall study. It may be used for real time applications. The future work of this paper is to expand research on the above algorithm and to bring the efficient discovery of emerging patterns.

VII. REFERENCES

[1] Lale ozyilmaz tulay yildirim,"Artificial Neural Networks for Diagnosis of Hepatitis Disease", Nov 2009.

[2] Ghumbre Shashikant Uttreshwar, Dr. A.A. Ghatol,"Hepatitis B Diagnosis Using Logical Inference And Generalized Regression Neural Networks" March 2009.

[3] Changjiang Long, Huan, "Modeling the virus and immune system of chronic hepatitis B", April 2014.

[4] Na Chu, Lizhuang Ma, "Attribute Weighting with Probability Estimation Trees for Improving Probability based Ranking in Liver Diagnosis", March2010.

[5] G.Sathyadevi, "application of cart algorithm inhepatitis disease diagnosis" June 2011.

[6] Na Chu, "An Intelligent Diagnosis Method for Chronis Hepatitis B in TCM" June 2013.

[7] Che Lijuan, Zhou Qiang,"Clinical Study on Data Mining of TCM Zheng in Chronic Hepatitis B", April 2014.

[8] C.Mahesh, K.Kiruthika, M.Dhilsathfathima, Assistant Professors, Department of Information Technology Veltech Dr.RR & Dr.SR Technical University. "diagnosing hepatitis b using artificial neural network based expert system", March 2014.

[9] Wen Shen, Zhihua Wei, Yunyi Li," Multiple granular analysis of TCM data with applications on hepatitis B", April 2015.

[10] http://www.hepb.org/patients/journal_articles.html

[11] Machine Learning Repository, www.archive.ics.uci.edu/ml/datasets.html.

[12] UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html.

[13] Na Chu, Ma Lizhuang, Zhiying Che, Min Zhou, and Xiaoyu Chen. A new hybrid PET algorithm for improving probability-based ranking for Diagnosis of Chronic Hepatitis in TCM. Advanced Science Letters, 2012, Vol.10, pp.544-548.

[14] Na Chu, Zhiying Che, Xiaoyu Chen, Min Zhou, Lizhuang Ma, and Yu Zhao. Hybrid Feature Selection based on Multi-view for Improved Diagnosis of Chronic Hepatitis. Chinese Journal of Integrative Medicine, Article ID 20110274.