Estimating the Surveillance of Liver Disorder using Classification Algorithms

A.S.Aneeshkumar
Research Scholar,
PG and Research Department of Computer Science, Presidency College,
Chennai, India – 600 005

C.Jothi Venkateswaran, PhD.
Research supervisor & Dean,
Department of Computer Science & Applications,
Presidency College,
Chennai, India – 600 005.

ABSTRACT
Data mining is an activity of extracting some useful knowledge from a large data base, by using any of its techniques. In this paper we are using classification, one of the major data mining models, which is used to predict previously unknown class of objects. Unlike other diseases, liver disorder prediction from common symptoms is typically difficult job for medical practitioners. Most of the features or symptoms are seen in many other fever related diseases and so it is not free from false assumptions. In most cases the opportunity of liver disease will not identified because of the domination of other diseases.

General Terms
Data mining, Classification, Liver Disease.

Keywords
Preprocessing, Naive Bayesian, C4.5 decision tree.

1. INTRODUCTION
Data mining is a process of knowledge discovery from the large database and it comprises of several distinct algorithms and explained statistical techniques. All the data mining techniques are used for identify the yet valuable knowledge from vast database with different kinds of data. The mined information is represented as a model of semantic structure of the data set and it might be possible to employ the model in the prediction and classification of new data [1]. So Data mining has an interactive role between the user and database, because interesting data patterns are showed to the user [2]. There are different approaches have been proposed for Classification of various diseases from its symptoms. Generally Classification consists of three phases, which are rule generation, rule pruning and classification. The performance, however, might different depending on the algorithm employed in any of these three phases [3]. The Classification is based on supervised learning algorithm and the selection of appropriate algorithm is also a very difficult task, because the selected method should be highly effective for the particular data set.

Liver is the largest gland in the body with about 3 lb (1.36 kg) weight. It is reddish brown in colour and is divided into four lobes of unequal size and shape [4]. Several disease states can affect the liver functions and its disorder can varies from simple reaction of medicines to liver cancer. Now a day's liver diseases may seen in all category of peoples without any age differentiation in India, due to the lack of physical activity, modern improper food habits, smoking, alcohol consumption, multiple sexual partnerships and injected drug usage. Sadly we can say these all are the part of Indian's life style.

The ultimate goal of risk factor prevention, detection, and control is to prevent acute events. In India, most of the disease having multiple symptoms and most patients admitted in hospitals with multiple diseases, in such cases it is very complex to diagnosis each disease and treats separately. So they may get treated for major disease and the upcoming diseases may not be identified or neglected. Moreover than that, patients are not ready to spend more time in the hospital and they need to be cure immediately. So in this paper, we try to classify the cases which may suffer liver disease with the help of cofactors. Risk factors are divided into two categories-major and contributing. Major risk factors are here used to prove the increase of liver disease risk. Contributing risk factor are those that doctors think can lead to an increased risk of liver disease, but their exact role has not applicable in all cases of diagnosis, because it can fluctuate with patients lifestyle and other history or presence of other diseases. The more risk factors you have, the more likely to have liver disease development. Some risk factors can be changed or irrelevant to the particular disease in some cases. But as a whole we consider it as evidence. But in case of earlier identification most of the risk factors can be changed or treated and that is useful to control much of other related risk.

2. DATA DESCRIPTION
Total of 2453 real medical data with 15 attributes were collected from a Public Charitable Hospital in Chennai. The data record contain the common symptoms noted by the physical examiner for liver related and non-liver related patients with similar symptoms of those who were admitted in male and female medical ward. Some occasional symptoms were eliminated here as a preprocessing stage. The collected fields are Age, Sex, Frequent alcoholic consumption (FAC), Obesity, Fever, Vomiting, Abdomen pain(AP), Yellowish Urinary Discharge (YUD), Loss of appetite (LA), Disturbance in abdomen (DA), Pale stools, Chills, rigor, Head ache, Acting differently(AD) and a class for diagnosis the Liver Disease (LD) with the help of blood test results. Here the study of age and sex factors are not dependent for general analysis and feature prediction of LD, because the disease may affect any age group and seen in any sex group except in cases of Alcoholic Fatty Liver Disease and Non-alcoholic Fatty Liver Disease. The given symptoms are very common in any liver related disorders.

Table 1. Sex based classification of the data

Sex	Yes	No
Male	1387	127
Female	843	96
Total	2230	223
Grand Total	2453	

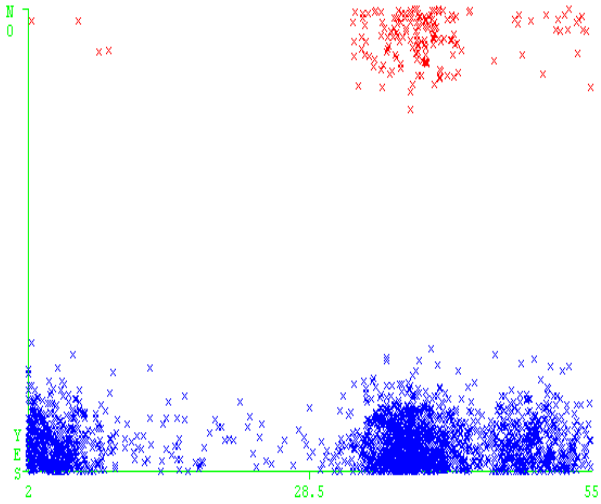

Fig 1: Age factor representation

3. METHODOLOGY

The methodology used here is, effective classification of Liver Disease and Non-liver Disease (NLD) patients with the help of symptoms. Before classifying the data, we have to preprocess it to avoid anomalies.

3.1 Data Preprocessing

Incomplete and noisy data are common in a real world data set, because the attribute was not important at the time of entry, misunderstanding of field values, duplications or usage of the data for other purposes. The action comprised in the pre-processing of a data set are the removal of duplicate records, normalizing the values used to represent information in the database, accounting for missing data points and removing unneeded data fields [5]. Here most of the Fever related diseases having the same symptoms. So initially, we preprocess the data set to avoid various inconsistent and missing values for our study.

3.2 Classification Models

In this study we used two classification methods, which are Naive Bayesian and C4.5 decision tree. The intention of using two algorithms is to identify the improvement of the algorithm for this particular data set. Both of these classification models follow different methods to evaluate its algorithm.

3.2.1 Naive Bayesian Classification

Bayesian Classifiers are the statistical classifiers and can predict class membership probability that a given tuples belong to a particular class. It assumes that the effect of an attribute value on a given class is independent of the value of the other attributes [6]. In other words a Naive Bayesian classifier calculates the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature [7].

Table 2. Mean & Standard Deviation

S. No:	Attributes	Mean		S.D.	
		YES	***NO***	***YES***	***NO***
1	Age	31.01	38.39	16.83	7.48
2	F A C	0.07	0.43	0.25	0.50
3	Obesity	0.20	0.30	0.40	0.46
4	Fever	0.94	0.99	0.23	0.17
5	Vomiting	0.60	0.84	0.49	0.37
6	A P	0.52	0.45	0.50	0.50
7	Y U D	0.86	0.41	0.35	0.49
8	L A	0.83	0.74	0.38	0.44
9	D A	0.57	0.48	0.50	0.50
10	pale stools	0.64	0.50	0.48	0.50
11	chills	0.27	0.70	0.44	0.46
12	rigor	0.28	0.43	0.45	0.50
13	head ache	0.30	0.58	0.46	0.49
14	A D	0.32	0.53	0.47	0.50

3.2.2 C4.5 Decision Tree

The decision trees are used in the analysis and application of classifying various cases into low-mind or high risk groups, creating rules in order to predict the future events, definition for relations of the certain sub groups, and getting the most effective decision by the help of medical observations [8].

C4.5 extracts rule from an unpruned tree, and then prunes the rules using a pessimistic approach. The training tuples and their associated class labels are used to estimate rule accuracy. In addition, any rule that does not contribute to the overall accuracy of the entire rule can also be pruned. C4.5 adopts a class-based ordering scheme. It groups all rules for a single class, and then determines a ranking of this class rule sets [6].

classification. In C4.5 the size of the tree is 23 and number of leaves are 12. The mean absolute error (MAE) and root mean square error (RMSE) is also 0.01 and 0.08 with best accuracy, which are defined as,

$$MAE = \frac{\sum_{k=1}^{n} |e_i|}{\sum i\,(e)}$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{n} |e_i{}^2|}{\sum i(e)}}$$

Where, e_i is the difference between the actual and forecast values.

Table 3. Performance of methods in datasets

Classification	Ratio	Accuracy (%)	Time Taken (Seconds)	Mean absolute Error	Root Mean Square Error	TP Rate	FP Rate
Naive Bayesian	50-50	88.38	0.03	0.14	0.29	0.92	0.49
	75-25	88.57	0.03	0.15	0.31	0.90	0.34
	90-10	89.60	0.03	0.14	0.30	0.93	0.42
C4.5	50-50	99.11	0.16	0.02	0.09	1	0.09
	75-25	99.03	0.16	0.02	0.09	1	0.10
	90-10	**99.20**	0.16	**0.01**	**0.08**	**1**	**0.08**

4. ANALYSIS AND RESULT

Generally LD can be diagnosis in a routine blood testing. But there are many scrutinizing factors for the diagnosis of the accurate type and its reasons, which makes the physician's job knotty [9] [10]. In this paper we used 2453 datasets for analysing the performance of the proposed classification methodology. Table 1 illustrate that, from the total dataset, 2230 cases having liver disease and 223 cases are not having. Figure 1, represents the age group differences of collected data, where the first cluster of data which is near to 2 years age category are probably with hepatitis A infection and most of the second and third cluster of peoples are belongs to alcoholic and non-alcoholic related liver disorder. Table 2 shows the average and standard deviations of each factor of both classes in this study. In table 3, the total data sets have been divided into the ratio of 50-50, 75-25, 90-10 and evaluated the accuracy. 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

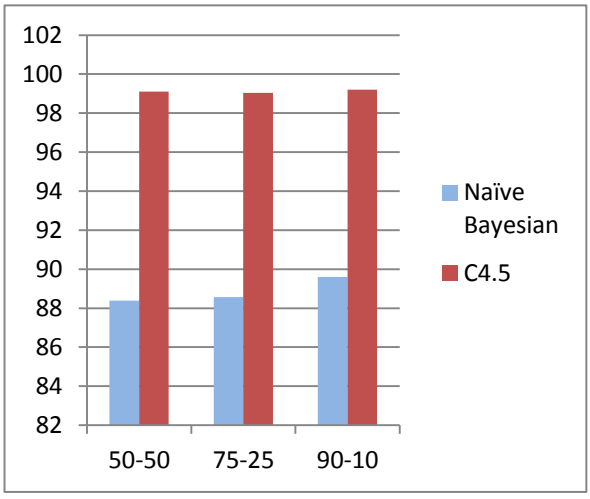

Fig 2: Accuracy evaluation chart

Figure 2 demonstrates the accuracy of both algorithms, where we can see the overall performance of C4.5 decision tree is better than Naive Bayesian.

5. BENEFITS AND LIMITATIONS

The developed model can serve as a decision maker tool for Liver disease related diagnosis. This is an effective classification method developed with 15 fields, which shown in table 2 and table 3 for the collected data. The field may expand or change according to the situation and influence of associated diseases.

6. CONCLUSION

Clinical screening of data is employing a crucial part in diagnosis process. Recent days every hospital is rich with large amount of medical records but without effective analysis in that. According to this methodology, it may useful for experts to identify the chances of disease and conscious prescription of further medical examinations and treatment. In developing countries the average time for a patient to reach into a hospital in any emergency situation is more than one hour. So medical practitioners are demonstrating awareness of evidence based treatment and therefore this application will give more support to such society for their future work and assessments.

7. ACKNOWLEDGEMENTS

We express our thanks to the Director Dr.(Capt.) K. J. Jayakumar M.S., M.N.A.M.S., F.A.I.S. and Chief Manager Dr. R. Rajamahendran, B.Sc., M.B.B.S., D.M.C.H., D.H.H.M., P.G.D.H.S.C.(Diab), F.C.D., Sir Ivan Stedeford Hospital, Chennai for providing permission to collect data. We are grateful to the chief Manager for his guidance and also would like to thank other hospital staffs for their valuable suggestions and help throughout this study.

8. REFERENCES

[1] Sally Jo Cunningham and Geoffrey Holmes, "Developing innovative applications in agriculture using data mining", In the Proceedings of the Southeast Asia Regional Computer Confederation Conference, 1999.

[2] Sengul DOGAN and Ibrahim TURKOGLU, "Iron-Deficiency Anemia Detection from Hematology Parameters by using Decision trees", International Lournal of Science & Technology, Volume 3, No.1, 85-92, 2008.

[3] Asha.T, Dr.S. Natarajan and Dr.K.N.B.Murthy, "A Study of Associative Classifiers with Different Rule Evaluation Measures for Tuberculosis Prediction", IJCA Special Issue on "Artificial Intelligence Techniques-Novel Approaches & Practical Applications", AIT, 2011

[4] P.Rajeswari and G.Sophia Reena, "Analysis of Liver Disorder Using Data mining Algorith", Global Journal of Computer Science and Technology, vol.10 issue 14(ver. 1.0) November 2010.

[5] Shantakumar B. Patil and 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.

[6] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Published by Elsevier, second edition – 2006.

[7] V.Ramesh and K.Ramar, "Classification of Agricultural Lands Soil: A Data Mining Approach", Agricultural Journal 6(3): 82-86, 2011, ISSN: 1816-9155.

[8] Bishop, C.M. (1996). "Neural networks for Pattern Recognition", Clarendon Press, Oxford.

[9] Huda Yasin, Tahseen A. Jilani and Madiha Danish, "Hepatitis-C Classification using Data Mining Techniques", International Journal of Computer Applications (0975-8887), Volume 24-No.3, June 2011.

[10] Polat K. and Gunes S., "Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation", Digital Signal Processing 16, 2006, pp. 889-901.

[11] Das S K, Mukherjee S, Vasudevan D M and Balakrishnan V, "Comparison of haematological parameters in patients with non-alcoholic fatty liver disease and alcoholic liver disease", Singapore Med J 2011: 52(3): 175.

[12] K. Rajeswari, Dr. V.Vaithiyanathan and Dr. P. Amirtharaj, "Prediction of risk score for heart disease in India using machine Intelligence", 2011 International Conference on Information and Network Technology, IPCSIT vol.4(2011) IACSIT Press, Singapore.

[13] "Heart disease burden in the next two years", http://www.medicalnewstoday.com / articles / 105302 .php.

[14] A.Sudha, P.Gayathri and N.Jaisankar, "Effective Analysis and Predictive Model of Stroke Disease using Classification Methods", International Journal of Computer Applications (0975-8887) Volume 43-No. 14, April 2012.

[15] Peiman Mamani Barnaghi, Vahid Alizadeh Sahzabi and Azuraliza Abu Bakar, "A Comparative Study for Various Methods of Classification", 2012 International Conference on Information and Computer Networks, IPCSIT vol.27 (2012) @ IACSIT Press, Singapore..

[16] "India's no.1 killer: Heart disease", http://indiatoday. intoday.in/story/ India's+no.1 +killer:+ Heart= disease/ 1/92422.html.

[17] Torky I. Sultan, Ayman Khedr and Samir Sabry, "Biochemical Markers of Fibrosis for Chronic Liver Disease: Data mining-based Approach", International Journal of Engineering Research and Development, Volume 2, issue 2 (July 2012), PP. 08-15.