I J C T A, 9(26) 2016, pp. 379-384

Liver Disease Analysis And Accuracy Prediction Using Machine Learning Techniques

D. Sindhuja[1] and R. Jemina Priyadarsini[2]

ABSTRACT

The liver disease is caused by the person who takes too much of alcohol, it is easily affected disease. Data miners have high interest to diagnosis the liver disease disorder, using the decision tree tool, the accuracy of the decision tree has been limited due to insufficient and small data set. Maximum number liver disease patients are in the age group of 41-55 years. The ratio between the men and women are6:1 ratio. To generate more accurate decision tree result for liver disease disorder , the paper suggested a method called sampling. In minor class to compensate the insufficiency of data is very effective. An experiment are done with two decision tree algorithms J48 and CART and the data set UCI machine learning repository for liver disease disorder and shows the validity of the method.

Keywords: Sampling, J48, CART, Decision tree, Classification Techniques, Liver disease.

1. INTRODUCTION

The liver is the largest glandular organ of the body. It weighs about 3 lb (1.36kg). It is reddish brown in color and is divided into four lobes of unequal size and shape [1]. Liver is vulnerable to a variety of metabolic, toxic, microbial and circulatory insults. The liver disease is curable in the earliest stage. Blood is carried to the liver two large vessels called the hepatic artery and the portal vein. The portal vein carries blood containing digested food from the small intestine. Liver tissue is composed of thousands of lobules, and each lobule is made up of hepatic cells, the basic metabolic cells of the liver. Alcohol is implicated in more than 50% of liver related deaths in the United States and complications of alcoholism contribute to a quarter of million deaths annually [2].

Fatty change is a very common finding both in biopsies and at post mortem examination. Liver cell involvement may be focal, diffuse, or zonal [3]. Non alcoholic fatty liver diseases include a spectrum of liver diseases ranging from simple steatosis to steatohepatitis, advanced fibrosis and cirrhosis [4]. Hepatocellular carcinoma and tumors arising from the bile duck epithelium are common tumors of the liver [5]. The liver CAD system consists of preprocessing, segmentation of liver, ROI analyse and classification, preprocessing is to decrease. Liver cancer is the third most common cancer in the world. Liver cancers can include Hepatocellular carcinoma and Cholangiocarnomia. 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 abnormalcavity, chills pain from the biliary track or pancreas, and enlarged gallbladder [6].

[1] Research Scholar, Computer Science, Bishop Heber College, Tiruchirapalli, Tamilnadu, India

[2] Assistant Professor, Computer Science, Bishop Heber College, Tiruchirapalli, Tamilnadu, India

Decision tree learning is the construction of a decision tree from class labeled training tuples. The top most node in the tree is the root node. Decision tree learning is a method commonly used in data mining. The process of top down induction of decision tree is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data [7]. The property of the decision is easily understandable by human. Another good point of decision tree is that it is straightforwatd to transform decision trees into rules so that the rules can be used for example to build expert systems [8,9].

Random sampling ia an issue because it may not have a perfect data mining dataset, similarly don't have detail about the property and knowledge. Due to data fragmentation it is called as the decision tree algorithms and more dependent upon the training data sets, while other machine learning algorithms like neural networks that do not divide the dataset during training are less dependent on [10].

In section 2, provide the related work to research, and in section 3 our method of experimentation. Section 4 method of experiment run and finally section 5 having conclusion.

2. RELATED WORK

Bendi Venkata Ramana [11] 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.

Aneeshkumar A.S. [12], in his study there is a methodology used to effective classification of liver and non-liver disease dataset. 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.

Dhamodharn.S [13] 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.

Rajeswari. P, Sophia Reena. G [14], in his paper, said UCI 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.

Gunasundari S and Janakiraman S [15] 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.In future liver disease disorder diagnosis extended in many directions. Such as using effective algorithms and more texture feature technique algorithms.

CK Ghosh [16] 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.

Newton Cheung[17] 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%

3. THE METHOD OF EXPERIMENTATION

In finding better decision trees for UCI liver disease data set. But the data set is small and have high error rate, to compensate the property of minor classes in decision tree algorithms. Decision tree algorithms not give high priority to splitting branches to minor classes, it is greatly possible that instances of minor classes

are treated in the lower part of the tree, and this treatment may increase misclassification rate for minor classes. Algorithm of decision tree to treat the instances of minor classes most importantly. In order to do this algorithm in decision tree, to increase the number of instances of minor classes by negative. The following content is the detailed description of the experimentation of the method.

INPUT: UCI Liver disease data set,

OUTPUT: Decision trees.

Begin

Do random sampling of size of 192, ten times.

For each sample data set Do

Generate a decision tree of the sample data;

Do while the accuracy of decision tree increases;

Generate a decision tree;

E**nd while;**

End Do;

End

In the algorithm of decision tree the instances of minor class by 100% the accuracy of decision tree data generated and the decision tree decreases. The sample data set size is half of the data set so that the large data set is suitable and useful for testing.

4. EXPERIMENTATION

Experiments are run using a UCI machine learning repository data set called liver disease disorder to watch the effect of the method. The number of instances is 380.There are 182 instances in 1st class and 178 instances in 2st class. 1st class is the minor class because its error rate is 87/182 = 47.8% while the error rate of class 2 is 60/178 = 33.7% based on 10 fold cross validation in J48. The overall error rate is 40.75% six continuous attribute is class attribute that have value of 1 or 2.table 1 for attributes description.

S.No	*Attribute Name*	*Meaning*
1.	Mcv	Mean corpuscular volume
2.	Alkphos	Alkaline phosphatase
3.	Sgpt	Alamine aminotransferase
4.	Sgot	Aspartate aminotransferase
5.	Gammagt	Gamma-glutamyl transpeptidase
6.	Drinks	Number of half-pint equivalents of alcoholic beverages drunk per day
7.	TB	Total bilirubin
8.	Selector	Used to split data into two sets
9.	YUD	Yellowish urinary discharge
10.	ALB	Albumin
11.	FAC	Frequent alcoholic consumption
12.	DA	Disturbance in abdomen

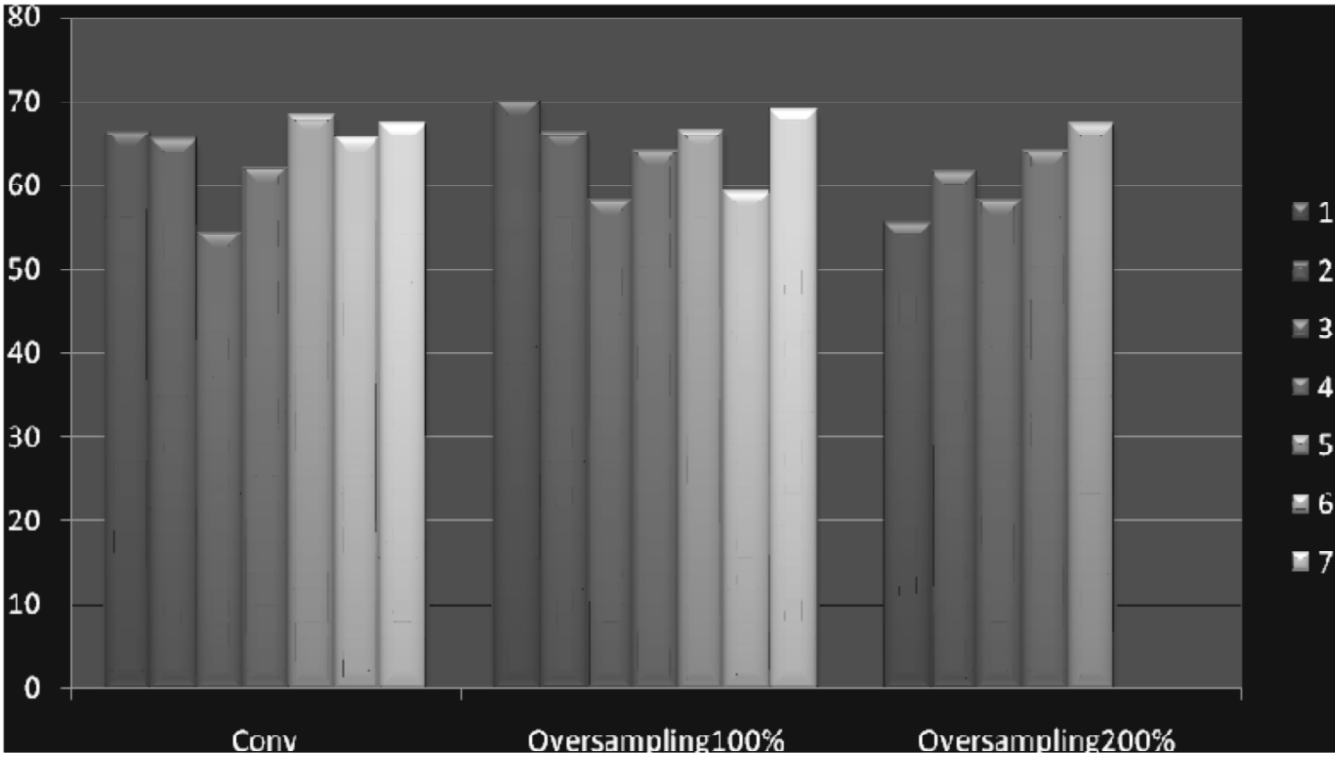

Figure 1

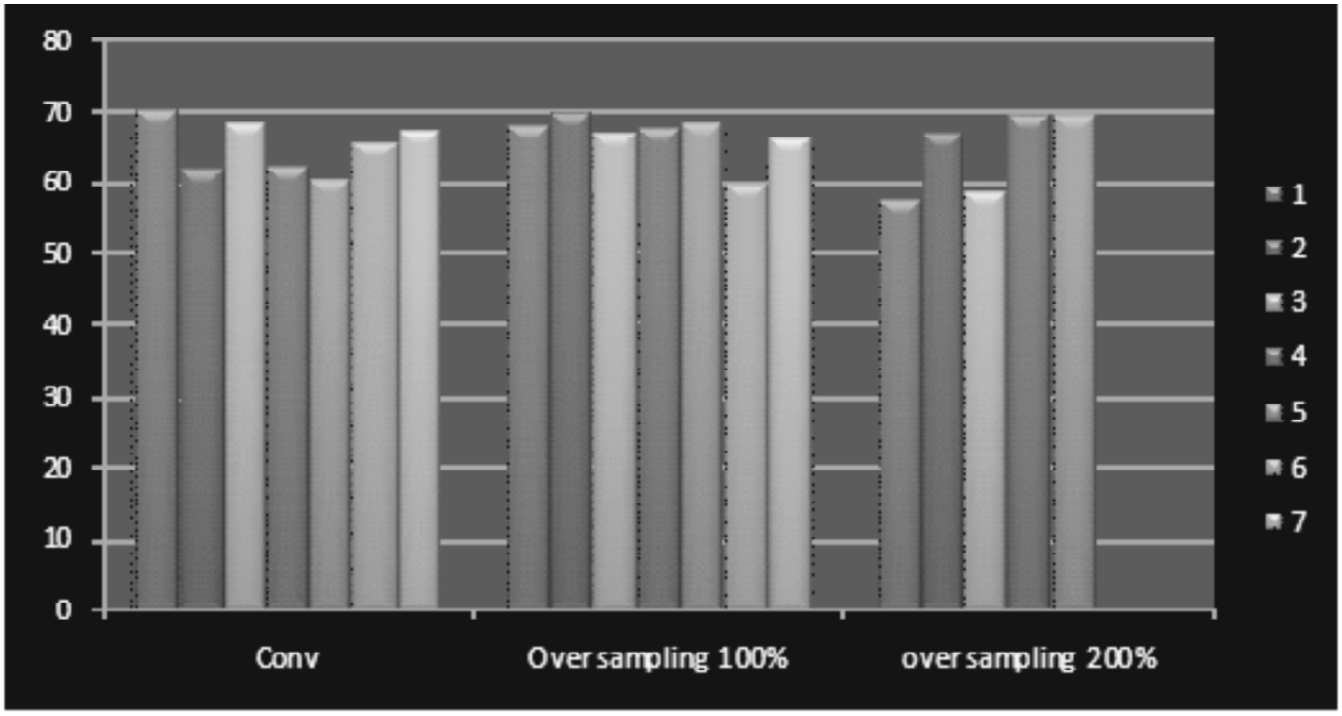

Figure 2

J48 and CART were used to generate decision trees for twelve random sample sets. Sample sets of size 192 were used. Remaining data were used for test. The following table 2 and 3 shows the accuracy for each decision tree algorithm with minor class over sampling. In the table numbers in bold characters represent the best results in the sample set.

Table 2
The accuracy of decision tree by J48 for sample sets

Sample Set #	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Con v	66.32%	65.47%	54.45%	62.12%	68.47%	65.58%	67.27%
Over. samp 100%	70.12%	66.43%	58.32%	64.27%	66.58%	59.23%	68.90%
Over. samp 200%	55.45%	61.69%	58.32%	64.16%	67.27%	NA	58.17%

If the table 2 have the better results with over sampling in 3 out of 7. In the table Conv means Conventional Sampling.

Table 3
The accuracy of decision tree by CART for sample sets

Sample set#	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Con v	70.32%	62.27%	68.82%	62.38%	60.80%	65.69%	67.43%
Over. samp 100%	68.58%	70.10%	66.92%	67.85%	68.94%	60.13%	66.32%
Over samp 200%	57.80%	66.96%	NA	58.76%	69.63%	NA	68.85%

If the table 3 have the better results with over sampling in 4 out of 7.

The over sampling method is effective in decision tree algorithm and the liver disease disorder data set, and it is more effective in CART decision tree algorithm.

5. CONCLUSIONS

In this paper, we have survived some data mining decision tree algorithms, J48 and CART algorithms using UCI Machine learning repository. Liver is one of the important organ in the human body. Decision tree has been considered as the nice and simplest tool, using this tool is easily understandable. The decision tree has a weakness because the dataset are small so the accuracy level decreases. UCI Dataset for liver disease disorder in data mining has high error rate and relatively small data this is due to the decision tree property. To overcome this problem taking decision tree algorithm, using oversampling techniques to mining the classes. Experiments with two algorithms, J48 and CART, showing good results for oversampling dataset to generate the decision trees.

REFERENCES

[1] P. Rajeswari and G. Sophia Reena, "Analysis of Liver Disorder Using Data mining algorithms," *Global Journal of Computer Science and Technology*, **10**, 48-49, 2010.

[2] Shah V.S., "Alcoholic liver disease in Hauser S Editors Mayo clinic gastroenterology and hespatology broad review," *4th edition New York Oxford University press*, 295-303, 2011.

[3] Sotoudehamanesh R Sotoudeh M, Asgari A, Abedi Ardakani B,Tavangar SM, Khakinejad A, "Silent Liver Disease in Autopsies from Forensic Medicine of Tehran," *Archives of Iranan Medicine*, 324-328, 2006.

[4] Sotoudehamanesh R Sotoudeh M, Asgari A, Abedi Ardakani B, Tavangar SM, Khakinejad A, "Silent Liver Disease in Autopsies from Forensic Medicine of Tehran," *Archives of Iranan Medicine*, **328**, 2006.

[5] Wight GD editor, "Systemic pathology liver billiary track and exocrine pancrease," *Great Britain churuchill living tone*, 3rd edition, 1-48, 1994.

[6] Cassangnou M, Boruchowicz A, Guilletmot F, "Hepatic steatosis revealing celiac disease A case complicated by transistory liver failure AMJ Gastroenterol," 1291-1292, 1996.

[7] Quinlan J.R., "Induction of Decision Trees Machine Learning," *Klumer Academic Publishers*, 81-106, 1986.

[8] D. Chiang, W. Chen, Y. Wang, L. Hwang, H.K. Chen, "Rules Generation from the Decision Trees," *Journal of Information Science and Engineering*, 325-339, 2001.

[9] T. Tamai, M. Fujita, "Development of an expert system for credit card application assesment," *Internation Journal of Computer Application in Technology*, 1-7, 1989.

[10] P. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining," *Addision-Wesley*, 2006.

[11] 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), **3(2)**, May 2011.

[12] A.S. Aneeshkumar and C. Jothi Venkateswaran, "Estimating the Surveillance of Liver Disorder using Classification Algorithms", *International Journal of Computer Applications* (095-8887), **57(6)**, November 2012.

[13] S. Dhamodharan, "Liver Disease Prediction Using Bayesian Classification", *4th National Conference on Advanced Computing, Applications & Technologies*, May 2014.

[14] P. Rajeswari and G. Sophia Reena, "Analysis of Liver Disorder Using data mining AAlgorithms", *Global Journal of Computer Science and Technology*, **10(14)**, 48-52, November 2010.

[15] 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)*, **74(11)**, July 2013.

[16] CK Ghosk, F. Islam, E. Ahmed, DK Ghosh, A Haque, QK Islam, "Etiological and clinical patterns of Isolated Hepatomegaly" *Journal of Hepato-Gastroenterology* **2(1)**; 1-4 2012 Jan-June.

[17] N. Cheung, "Machine learning techniques for medical analysis "*School of Information Technology and Electrical Engineering, BsCthesis, University of Queenland*, 19 Oct. 2001.