TECH-CHRONICLE
AN INTERNATIONAL E-JOURNAL ON
EMERGING TRENDS IN SCIENCE, TECHNOLOGY AND MANAGEMENT

A Comparative Study of Data Mining Tools for Predicting the Liver Disorder

A Raju Samanthula
Department of CSE, Raghu Engineering college, Vishakapatnam,Andhra Pradesh, India
s.a.raju.cse@gmail.com

Adidela Daveedu Raju
Professor ,Department of CSE, Ramachandra College of Engineering,Vatluru, Eluru, Andhra Pradesh, India
123.davidjoy@gmail.com

G M Murali Krishna
Department of Computer Science, St Theressa Engg College, Garividi, Andhra Pradesh, India.
mmkriss@gmail.com

***Abstract*-Chronic liver disease was the leading cause of death for all and the second leading cause of death for men, ages 35-44. The liver in the body has its own importance due to its significant functionalities .It is responsible for functions vital to life. The main functions of the liver are to process nutrients from food, make bile, remove toxins from the body and build proteins. The loss of functionality will cause serious damage to the body. Some of the causes for this disorder are long-term, excessive alcohol consumption, severe reactions to certain medications, fatty liver caused by obesity etc. Various features are considered in the data set related to liver disorder that mainly differentiates the disordered and non disordered patients. This paper focuses on comparative study of using data mining techniques in predicting the disease. Two models are developed for comparative study. The first model uses directly the classification algorithms on the data and the second model uses the same classification techniques after adopting the clustering technique on the data. The accuracies are measured for the two models; The second model gives the accurate results for almost all the classification algorithms. This second proved model for accuracy is further used for predicting the disease as a virtual doctor.**

***Keywords*- classification algorithms, clustering, data mining, liver disorder**

1. INTRODUCTION:

The rise of health care cost is one of the world's most important problems. Due to increase in world population, the health care industries are facing many challenges and issues based on patients severity is to reduced and detect it earlier in a more effective way. Liver Disorder is one of the severe problems in the world [1]. The liver is the largest glandular organ of the body it weight's about 1.36 kg .it is reddish brown in color and is divided into four lobes of unequal size and shape. An early diagnosis of liver problems will increase patient's survival rate. Liver disease can be diagnosed by analyzing the levels of enzymes in the blood[2]. When the liver becomes diseased, it may have serious consequences. Liver diseases is also called hepatic disease. Jaundice caused by increasing the levels of bilirubin in the system. The bilirubin results from the breakup of the hemoglobin of dead red blood cells [2]. The symptoms related to liver dysfunction include both physical signs and a variety of symptoms related to digestive problems, blood sugar problems, immune disorders, abnormal absorption of fats, and metabolism problems. The mal absorption of fats may lead to symptoms that include indigestion, reflux, deficit of fat soluble vitamins, hemorrhoids, gall stones, intolerance to fatty foods, intolerance to alcohol, nausea and vomiting attacks, abdominal bloating, and constipation Nervous system disorders include depression, mood changes, especially anger and irritability, poor concentration and "foggy brain",

overheating of the body, especially the face and torso, and recurrent headaches (including migraine) associated with nausea[6].

This paper describes a comparative study of data mining tools for predicting the liver disorder by using classification algorithms with and without clustering. Accuracies are measured for both the phases and compared the results.

Due to the widespread of computerization and also due to affordable storage facilities, there is an enormous wealth of information embedded in huge data bases belonging to different enterprises. These data bases further facilitated to share one another by new emerging technologies like cloud computing and performing wide verity of data mining techniques to extract the knowledge and obtaining meaningful patterns and rules. Knowledge Discovery in Data bases (KDD) is the process of identifying a valid, potentially useful and ultimately understandable structure in data. This process involves selecting or sampling data from a data ware house, cleaning and preprocessing it, transforming or reducing it if necessary, applying a data mining component to produce a structure, and evaluating the derived structure. Data mining is a step in the KDD process. Data mining is a technique that discovers the reliable, intelligent information from raw data. It involves scientists from wide range of disciplines including mathematics, computer scientists and statisticians, as well as those working in the fields such as machine learning, artificial intelligence, bio informatics, biotechnology, information retrieval and pattern recognition.

Most of the raw data is noisy and discrepant or in complete. This should be preprocessed before further analyzing and extracting the information. If there is no proper pre processing that leads to production of inaccurate and vague results. The preprocessing step includes managing the three basic difficulties carried by raw data. The liver data which is used here are preprocessed for finding missing values and inconsistency of data. Some of the algorithms can manage the missing values by incorporating some methodologies like ignoring the tuple or replace the missing value with respective attribute mean. Data normalization and attribute selection also performed in the preprocessing step. In the data normalization attribute data are scaled so that they fall within the specified small range say 0 to 1.

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. The values should be normalized before clustering to avoid the domination of high value attribute to low value attribute.

The preprocessed data is transferred to further steps of this model i.e. clustering and classification. The paper uses k-means clustering technique for grouping the data in to different clustering. It starts with a random initial partition and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centers until a convergence criterion is met (e.g., there is no reassignment of any pattern from one cluster to another, or the squared error ceases to decrease significantly after some number of iterations). The k-means algorithm is popular because it is easy to implement, and its time complexity is O(n), where n is the number of patterns. A major problem with this algorithm is that it is sensitive to the selection of the initial partition and may converge to a local minimum of the criterion function value if the initial partition is not properly chosen. There are several methods available [3] to select the number of clusters which perfectly find the genuine number of clusters.

Data classification is a two-step process, in the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (training phase) and second phase is classification phase where the model is used for classification and its performance is analyzed with the testing set of tuples. Test data is used to estimate the accuracy of the classifier. Data mining uses several classification algorithms like .Naive Bayes, Decision Trees, neural network, genetic algorithm,C4.5 ,back propagation algorithm, fuzzy logic algorithms etc. Among these C4.5 is used in this paper. Some care to be taken while using the formed clusters after preprocessing for next step of classification process by C4.5 algorithm [4]. C4.5 is extension to ID3 algorithm [5]. The C4.5 algorithm is widely used classification algorithm for its simplicity and non biasing of attributes by using "gain ratio" rather than information gain that is used in ID3. Algorithm for generating decision tree is given below

Input:

1.Data partition, D, which is set of training tuples and their associated class labels;

2.*attribute_list*,the set of candidate attribute;

3.*Attribute_selection_method*,a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes.

Output: A decision tree.

C4.5, a successor of ID3,uses an extension to information gain known as gain ratio. It applies a kind of normalization to information gain using a "split information "value defined with Info(D) as

$SplitInfo_A(D) = -\sum_{j=1}^{v} |Dj|/|D| * \log_2(|D_J|/|D|)$.

This value represents the potential information generated by splitting the training data set, D ,into v partitions, corresponding to the v outcomes of a test on attribute A.The gain ratio is defined as

GainRatio(A)=Gain(A)/SplitInfo(A).

The attribute with the maximum gain ratio is selected as the splitting attribute. Note, however, that as the split information approaches 0, the ratio becomes unstable. A constrained is added to avoid this, whereby the information gain of the selected must be large-at least as great as the average gain over all tests examined.

2. LITERATURE REVIEW:

Liver is the largest internal organ in the human body, and it is known that the organ is responsible for more than one hundred functions of human body. The complexity of this organ makes it easily affected by disease of disorder. So diagnosing liver disorder disease is a high interest to researchers of data miners, and decision trees have been a good data mining method to diagnose the disease. Decision trees have been considered one of good data mining tools with respect to understandability and transformability. But, weakness of decision trees arises due to the fact that their branching criteria give higher priority for major classes. BUPA liver disorder data set that is our interest for data mining is relatively small and has high error rate so that it may be vulnerable due to the property of decision trees [6]. In order to overcome the problem of disdaining minority classes of the data set in decision tree generation algorithms, we used over-sampling technique for minor classes. Experiments with two representative decision tree algorithms, C4.5 and CART, showed very good results so that we may recommend oversampling for the data set to generate decision trees[7]. Michael J Sorich [8] reported that SVM classifier produces best predictive performance for the chemical datasets. Lung-Cheng Huang reported that Naïve Bayesian classifier produces high performance than SVM and C 4.5 for the CDC Chronic fatigue syndrome dataset [9]. Paul R. Harper, A review and comparison of classification algorithms for decision making [10] reported that there is not necessary a single best classification tool but instead the best performing algorithm will depend on the features of the dataset to be analyzed

3. DATA SET:

The data, BUPA liver disorders [7], is collected from the UCI machine learning data base. The data set is donated by Richard S. Forsyth. Previous usage of the data is given in PC/BEAGLE User's Guide written by Richard S. Forsyth. The data has total 7 attributes. The first 5 variables are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each line in the bupa.data file constitutes the record of a single male individual. The total number of instances in the data base is 345. The seventh attribute is the class label by which reference classification of the data taken place. This paper employed data mining techniques preprocessing, clustering and classification on this data set elected from UCI machine learning database. Attributes and their abbreviated terms are given in table 1.

TABLE 1
ATTRIBUTES AND THEIR ABBREVIATED TERMS

Attribute number	***Name***	***Abbreviated variable name***
1	Mcv	mean corpuscular volume
2	Alkphos	alkaline phosphotase
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	selector field used to split data into two sets	

4. METHODOLOGY:

All the data mining tools are well work and produce high accuracy depending up on the degree of

preprocessing the data. The first step of the model uses three preprocessing methods.. This includes missing value treatment, normalization of data and attribute selection. The missing values are globally replaced by mean or mode of the respective attribute. The normalization is needed for prevent the data attributes which has the high range of values from out weighing attributes with initially smaller range of values. This normalization is needed while performing the data mining techniques like neural network back propagation, distance measures and clustering techniques. Further in pre processing each attribute is evaluated by using information gain attribute evaluator which evaluates the worth of an attribute by measuring the information gain with respect to the class. By using the ranker algorithm attributes by their individual evaluations. Use in conjunction with attribute evaluators (ReliefF, GainRatio, Entropy etc). The number of features that you we want to select from the feature vector can always be defined. Omit the features one at a time that have lower ranks and see the predictive accuracy of the classifier. Weights put by the ranker algorithms are different than those by the classification algorithm. There is a constraint that up to what point one can omit the features from ranking list. There are ups and downs in the learning curves The learning curve always reaches the global minimum. After reaching the global minimums the error rate will goes high and high never reach the minimum.

The second step of the process carries two phases of evaluation. In the first phase of the study, the pre processed data is directly given as input to the classifier, C4.5. The algorithm uses 66 percent of the data for training the system and 34 percent of the data for testing the classifier. A 10 fold cross validation is performed to find the accuracy of the classifier.

In the second phase of the study, the data is clustered by using very well known clustering algorithm k-means algorithm. This employs a square error criterion [11]. The number of clusters built start from 2 to square root of the volume of the data set i.e. 18. The optimal number of clusters is used to build the decision tree classifier. This paper proposes that the average accuracy of the classifiers for the optimal number of clusters is more than the accuracy of the classifier which obtained from whole data from phase one.

5. RESULTS AND ANALYSIS:

The liver disorder data set contains 365 total data items. In the first step data is preprocessed by missing value treatment, normalization and attribute ranking. The sample data after preprocessing the raw liver disease data is given in table 2 after replacement of the missing values.

TABLE 2
NORMOLISED DATA

Mcv	Alkphos	Sgpt	Sgot	gammagt	drinks	class
0.526316	0.6	0.271523	0.285714	0.089041	0	1
0.526316	0.356522	0.364238	0.350649	0.061644	0	2
0.578947	0.408696	0.052 98	0.298701	0.017123	0	2
0.868421	0.278261	0.059603	0.155844	0.041096	0	2
0.605263	0.339 13	0.105 96	0.155844	0.013699	0.025	1
0.605263	0.382609	0.112583	0.077922	0.020548	0.025	1
0.710526	0.269565	0.119205	0.194805	0.006849	0.025	1
0.657895	0.356522	0.377483	0.350649	0.027397	0.025	1
0.552632	0.469565	0.139073	0.181818	0.044521	0.025	1
0.815789	0.382609	0.165563	0.194805	0.020548	0.025	1

0.6 84 21 1	0.4 78 26 1	0.1 05 96	0.3 37 66 2	0.0 44 52 1	0. 0 2 5	1
0.6 31 57 9	0.3 82 60 9	0.1 25 82 8	0.1 42 85 7	0.0 17 12 3	0. 0 2 5	1

The attributes of the data ranked and the attribute mcv has the lowest rank, the reduced attribute vector contains 6 attributes, alkphos, sqpt, sgot, gammagt, drinks and class.

In the first phase the data is supplied to the classifier, C4.5 and the data is classified and produces the decision tree, a sample decision tree is in figure 1. The accuracy of the classifier is given in comparative table 2.

In the second phase, the data is clustered by k-means, the average accuracy of the classifier for 5 clusters is reasonably good. The accuracy of the classifier is given in table 3.

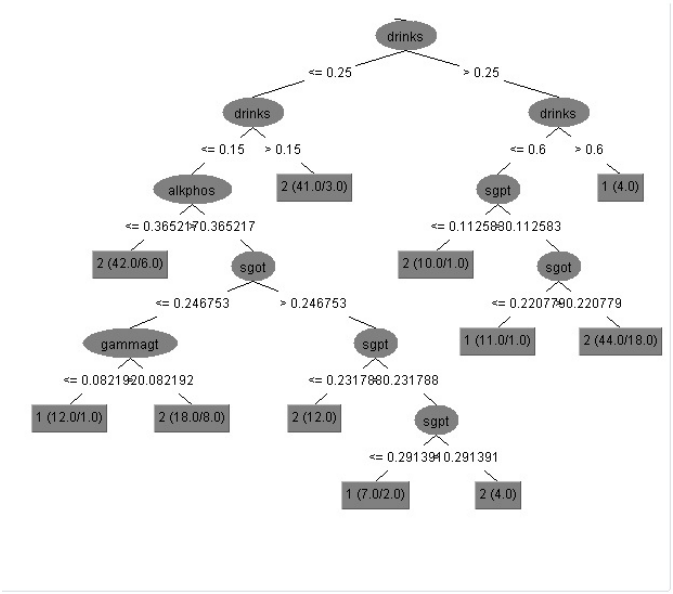

Fig.1: Decision tree produced by C4.5 classifier

TABLE 3
COMPARATIVE STUDY TABLE OF CLASSIFIER ACCURACY WITH AND WITH OUT CLUSTER

Cluster0	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.694	0.263	0.714	0.694	0.704	0.699	1
	0.737	0.306	0.718	0.737	0.727	0.699	2
Weigted average	0.716	0.285	0.716	0.716	0.716	0.699	
Cluster1	.0.889	.133	0.8	0.889	0.842	0.849	1
	867	0.111	0.929	0.867	0.897	0.893	2
Weigted average	0.875	0.119	0.88	0.875	0.876	0.893	
Cluster2	0.889	0.133	0.8	0.889	0.842	0.893	1
	0.867	0.111	0.929	0.867	0.897	0.893	2
Weigted average	0.875	0.119	0.88	0.875	0.876	0.893	
Cluster3	0.429	0.094	0.667	0.429	0.522	0.77	1
	0.906	0.571	0.784	0.906	0.841	0.77	2
Weigted average	0.761	0.426	0.748	0.761	0.744	0.77	

In the other phase of the model the data is grouped in to optimal 4 clusters and produced respective classification trees, by using k-means and C4.5 algorithms respectively. The accuracy of the classifiers is much more improved than the non cluster classification. The similar procedure further carried out by using Expectation Maximization clustering algorithm, naïve Bayesian classification technique and obtained the satisfactory results.

6. CONCLUSION AND FUTURE SCOPE:

The proposed model used the preprocessed data in two phases, the accuracy of predicting the liver disorder is more accurate in the second phase where clustering and classification used than the one where only classification performed. K-means and C4.5 algorithms are used for clustering and classification of data. The model concludes that classification by division produces the more accurate prediction. The extension of this work includes attribute subset selection, adopting weights for attributes by gradient descent technique, adopting fuzification algorithms in classification.

7. REFERENCES:

[1] A.Sudha P.Gayathri N.Jaisankar Utilization of Data mining Approaches for Prediction of Life Threatening Diseases Survivability

[2] P.Rajeswari , G.Sophia Reena, Analysis of Liver Disorder Using Data mining Algorithm

[3] Zahid Ansari, M.F. Azeem, Waseem Ahmed,A.Vinaya Babu,Quantitative Evaluation of Performance and Validity Indices for Clustering the Web Navigational Sessions,World of Computer Science and Information Technology Journal, vol. 1, no. 5, pp. 217-226, 2011.

[4]..J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kauffman, San Francisco, 1993.

[5] J.R. Quinlan, Induction of decision trees, Machine Learning 1,pp. 81– 106, 1986.

[6] Hyonti Sug,Improving the prediction Accuracy of Liver Disorder with Oversampling

[7] BUPA Liver Disorders Dataset. UCI repository of machine learning databases. Available from ftp://ftp.ics.uci.edu/pub/machine- learningdatabases/ liverdisorders/bupa.data,

[8] Michael J. Sorich,† John O. Miners,*,‡ Ross A. McKinnon,† David A. Winkler,§ Frank R. Burden,| and Paul A. Smith‡ Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP-Glucuronosyltransferase Isoforms

[9] Lung-Cheng Huang, Sen- Yen Hsu and Eugene Lin, A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data (2009)].

[10] Paul R. Harper, A review and comparison of classification algorithms for decision making

[11] McQueen, J. 1967. Some methods for classification and analysis of multivariate observations, In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281–297.