ISSN:0975-9646

# Development of an Efficient Classifier for Classification of Liver Patient with Feature Selection

Harsha Pakhale[1], Deepak Kumar Xaxa[2]

*Dept of CSE,*
*Mats University,*
*Raipur, Chhattisgarh, India*

**Abstract- Diagnosis of health conditions is a very challenging task in field of medical science. In medical science, day by day data is increasing continuously and creates problem to identify the accurate diseases. Data mining based classification plays very important role in classification of data. In this research work we have used various data mining based classification technique to develop the classifier for classification of liver and non liver patient. We have used techniques like C4.5, Random Forest (RF), Multilayer Perceptron(MLP), Classification and Regression Technique (CART) and applied all these techniques on liver patient data collected from UCI repository.In this paper we have used ensemble model to develop the robust classification model which gives higher classification accuracy compare to its individual model. We have also used Information gain feature selection technique is applied on best model ensemble of C4.5, Random Forest and CART which gives 76.02% of accuracy with 3 numbers of features.**

**Keywords : Classification, Feature Selection, Ensemble.**

## I. INTRODUCTION

Now a days the data is increasing day by day in every organization. One of the most important organizations is medical science where every day lots of patient data are stored. Due to large amount of data, the data mining based classification plays very important role for classification of data. In this research work we have worked on liver patient data classification. There are various authors who have worked in the field of classification of liver patient data. **Hoon Jin et.al. [1]** have suggested various classification algorithms such as Naïve Bayes, Decision Tree, Multilayer Perceptron, k-NN, Random Forest and Logistic for classification of liver patient data set. These algorithms were compared in several kinds of evaluation criteria like precision, recall, sensitivity, specificity. Logistic and Random Forest gives highest and second highest precision and recall value respectively as compared to other previous four algorithms. **Pankaj Saxena et.al. [2]** have proposed different clustering algorithms such as COBWEB clustering algorithm, DBSCAN clustering algorithm, Hierarchical clustering algorithm and K-means clustering algorithm on ILPD dataset. Results show that k-means clustering algorithm is simplest and fastest algorithm as compared to other clustering algorithms. **AnjuGulia et.al.[3]** have used J-48 classifier, Multilayer Perceptron classifier, Random Forest classifier, Support Vector Machine classifier and Bayesian Network classifier for classification of liver patient data. The results obtained show that Support Vector Machine algorithm gives better performance with an accuracy of 71.3551% as compared to other algorithms when evaluated without feature selection and Random Forest algorithm gives better performance with an accuracy of 71.8696% as compared to other algorithms when evaluated after feature selection. **Sina Bahramirad et.al. [4]** have suggested different classification algorithms such as Logistic, Bayesian Logistic Regression, Logistic Model Trees(LMT), Multilayer Perceptron, K-star, RIPPER, Neural Net, Rule Induction, Support Vector Machine(SVM) and CART. **Ersaa M. Hashem et. al. [5]** have used Support Vector Machine (SVM) classification technique for classification of liver patient data to achieve improved performance. Two dataset are used for performance evaluation. The first dataset is BUPA dataset and the second dataset is ILPD dataset. Both dataset are obtained from UCI Repository. Error Rate, Sensitivity, Prevalence, Specificity and Accuracy are used to evaluate the performance of Support Vector Machine (SVM). The result obtained shows that the Specificity at first 6 ordered features are best for BUPA dataset compared to ILPD dataset while the Sensitivity, Error Rate, Accuracy and Prevalence at first 6 ordered features are best for ILPD dataset as compared to BUPA dataset.

## II. METHODOLOGY

Classification plays very important role for classification of data. In this research work we have used various classification techniques for classification of liver patient data. They are described below:

### *A. Decision Tree*

Decision tree induction [7] is the learning of decision trees from class labelled training tuples. A decision tree is a flow chart like tree structure, where each internal node denote a test on an attribute, each branch represent an outcome of the test, and each leaf node hold a class label. The topmost node in a tree is the root node. Decision tree can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate to human. The learning and classification steps of decision tree induction are simple and fast. Decision tree algorithm is simple and fast. These tree classifiers have good accuracy. Decision tree are built, many of the branches may reflect noise or outliers in the training data. In this research work we have used CART, C4.5 and Random Forest(RF) for classification of liver patient data. CART [9] builds a binary decision tree by splitting the record at each node according to a function of a single attribute. CART uses the gini index for determining the best split. C4.5 [9] is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees and rule derivation. Random Forest (or RF) (Parimala, R. et al., 2011a) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. Random Forests are often used when we have very large training datasets and a very large number of input variables.

### *B. Bayesian Net*

Bayesian Net [7] is statistical classifier which can predict class membership probabilities, such as the probability that a given tuple belong to a particular class. Let X be a data sample whose class label is unknown. Let H be some hypothesis, such as that the data sample X belongs to a specified class C. For classification problems, we want to determine P(H|X), the probability that the hypothesis H holds given the observed data sample X.P (H|X) is the posterior probability, or a posteriori probability, of H conditioned on X.

### *C. Multilayer Perceptron (MLP)*

MLP [9] is a development from the simple perceptron in which extra hidden layers (layers additional to the input and output layers, not connected externally) are added. More than one hidden layer can be used. The network topology is constrained to be feed forward, i.e., loop-free. Generally, connections are allowed from the input layer to the first (and possible only) hidden layer, from the first hidden layer to the second and so on, until the last hidden layer to the output layer.

### *D. Ensemble Model*

Two or more models combined to form a new model is called an ensemble model. An ensemble model is a combination of two or more models to avoid the drawbacks of individual models and to achieve high accuracy. Bagging and boosting [7] are two techniques that use a combination of models. Each combines a series of *k* learned models (classifiers), *M*1, *M*2,…..*Mk*, with the aim of creating an improved composite model, *M*. Both bagging and boosting can be used for classification.

### *E. Feature Selection*

Feature subset selection [11] is an important problem in knowledge discovery, not only for the insight gained from determining relevant modeling variables, but also for the improved understandability, scalability and possibly accuracy of the resulting models. In Feature selection, the main goal is to find a feature subset that produces higher classification accuracy. Feature selection consists of detecting the relevant features and discarding the irrelevant ones, with the goal of obtaining a subset of features that describes the given problem properly with a minimum degradation of performance.

## III. Performance Measurement

Performance of each classifier can be evaluated by using some very well-known statistical measures like accuracy, sensitivity and specificity. These measures are defined by true positive (TP), true negative (TN), false positive (FP) and false negative (FN). With the help of confusion matrix we can calculate the performance measures [7].

**Classification Accuracy** measures the proportion of correct predictions considering the positive and negative inputs. It is calculated as follows:
Classification accuracy = (TP+TN)/N (1)

**Sensitivity** measures the proportion of the true positives, that is, the ability of the system on predicting the correct values in the cases presented. It is calculated as follows:
Sensitivity=TP/(TP+FN) (2)

**Specificity** measures the proportion of the true negatives, that is, the ability of the system on predicting the correct values for the cases that are the opposite of the desired one. It is calculated as follows:
Specificity =TN/ (TN +FP) (3)

## IV. Data Description

The Indian Liver Patient Dataset (ILPD) data set is collected from UCI repository [8] which is classified under two class liver and non-liver. This data set is binary classification problem. This data set consists of 10 attributes and 1 class. This data set also consists of 583 instances. In which 416 are liver patient records and 167 non liver patient records.

## V. Experiment Results

This experiment done in WEKA environment [6] with NET Beans as editor, 4GB RAM and i5 machine. In this experiment, we have used WEKA library and used various data mining based

classification techniques shown in table 1 to develop the classifier. The experiment is done into two steps: First develop the individual and ensemble model and second feature selection applied on best model.

Various Classification techniques have applied on liver patient data set and check the accuracy of model. Table 1 shows that accuracy of various individual and ensemble model with 75-25% training-testing model. Table 1 shows that accuracy of individual models are not sufficient for classification of model that means individual techniques are not capable to classify the liver patient data with high accuracy. We have used ensemble model like C4.5+RF, C4.5+CART, RF+CART etc. to classify the liver patient data which gives the higher classification accuracy compare to its individuals models. We have recommended ensemble of C4.5, CART and RF model which is robust for classification of liver patient data. Table 1 shows the accuracy of different individuals and ensemble models. Figure1 shows the accuracy of different model with different individuals and ensemble techniques.

Table1: Accuracy of model in case of 75-25% of liver patient data set

| **Techniques** | **Accuracy in Percentage** |
|---|---|
| C4.5 | 69.86 |
| RF | 71.23 |
| MLP | 69.86 |
| BayesNet | 68.49 |
| CART | 71.91 |
| C4.5+RF | 72.60 |
| C4.5+CART | 71.91 |
| MLP+BayesNet | 70.54 |
| RF+CART | 72.60 |
| C4.5+CART+RF | 75.34 |

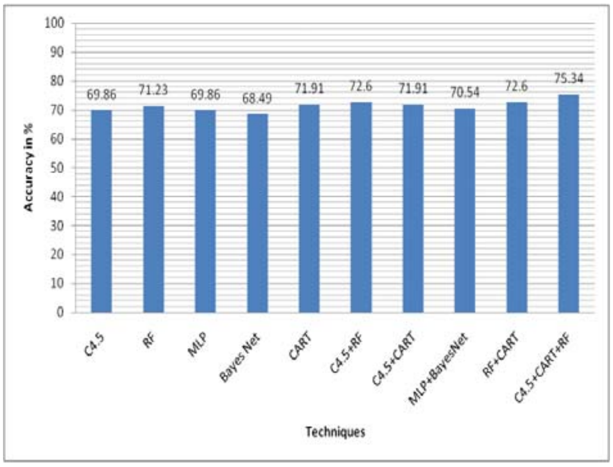

Figure 1. Accuracy of different models

In this research work, we have also used raking based feature selection technique i.e. Information gain that can be used to rank the feature. Basically feature selection technique can be used to computationally increase the performance of model. Table 1 shows that ranking of features in ascending order after applying information gain feature selection technique. In this work, we have applied the information gain feature selection on best ensemble model i.e. ensemble of C4.5, CART and Random Forest (RF).Table 2 shows that rank of feature in ascending order. We have eliminated features one by one from less important feature, applied data set into best model and calculate the accuracy with different feature subsets. Table 3 shows the accuracy of model with different feature subsets. Our proposed ensemble model gives 76.02% of accuracy with 3 feature subset which is computationally efficient and robust model for classification of liver patient data. Finally we have recommended C4.5+CART+RF+InfoGain model for classification of liver patient. Figure 2 shows the accuracy of best model with different feature subsets. Table 4 and Table 5 shows the confusion matrix of best model with 10 and 3 features respectively. With the help of this confusion matrix, we can calculate other performance measures like sensitivity and specificity. Other performance measures like sensitivity, specificity and accuracy are shown with feature set in table 6. The accuracy, sensitivity and specificity are 75.02%, 100% and 41.46% respectively in case of 10 features while accuracy, sensitivity and specificity are 76.34%,88.57% and 2.27% respectively in case of 3 features.

Table 2: Ranking of features using Information gain feature selection techniques

| **Feature Selection Techniques** | **Ranking of Feature in Ascending order** |
|---|---|
| Information Gain | 8,2,9,1,10,5,7,4,6,3 |

Table 3: Accuracy of best model with reduce number of feature subsets

| **Number of Features** | **Accuracy** |
|---|---|
| All features | **75.34** |
| 9 | 63.69 |
| 8 | 71.23 |
| 7 | 73.28 |
| 6 | 67.80 |
| 5 | 70.54 |
| 4 | 69.86 |
| 3 | **76.02** |
| 2 | 73.28 |

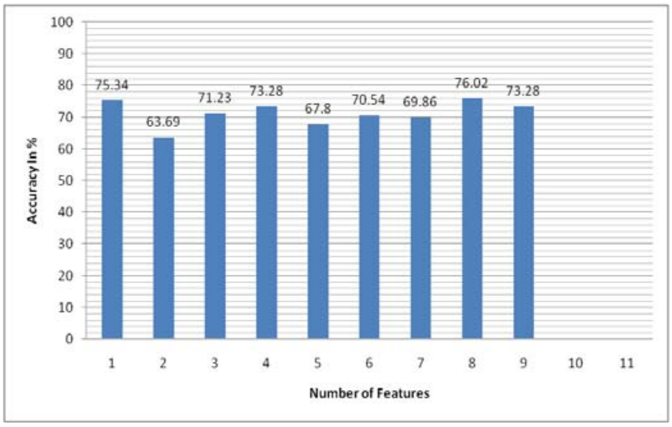

Figure 2: Accuracy of best model with different feature subsets

Table 4: Confusion matrix of best model with 10 features

| **Actual Vs. Predicted** | **Liver** | **Non Liver** |
|---|---|---|
| Liver | 93 | 12 |
| Non Liver | 24 | 17 |

Table 5: Confusion matrix of best model with 3 features

| **Actual Vs. Predicted** | **Liver** | **Non Liver** |
|---|---|---|
| Liver | 110 | 0 |
| Non Liver | 35 | 1 |

Table 6: Performance measures with feature sets

| **Number of Features** | **Accuracy** | **Sensitivity** | **Specificity** |
|---|---|---|---|
| 10 | 75.34 | 88.57 | 41.46 |
| 3 | 76.02 | 100 | 2.77 |

## VI. CONCLUSION

In medical science diagnosis of health condition is very challenging task . This research work focuses on classification of liver and non-liver patient with high accuracy. In this research work, we have used various classification techniques like decision tree, MLP and bayes net for classification of liver patient. The individual model does not satisfy the classification accuracy of model so we have ensemble the individual models to develop the robust classifier. An ensemble of C4.5, CART and RF gives better accuracy 75.34% compare to individuals and other ensemble models. We have also applied the information gain feature selection technique to comptaionally increase the performance of model. In case of 3 feature subset our ensemble model gives 76.03% of accuracy as robust model.

## REFERENCES

[1]. Decision Factors on Effective Liver Patient Data Prediction. International Journal of Bio-Science and Bio-Technology Vol.6, No.4, pp.167-178.

[2]. Pankaj Saxena and Sushma Lahre,(2013). Analysis of various clustering algoriths of data mining on health informatics, International Journal of Computer & Communication Technology , Volume-4, Issue-2, pp.108-112

[3]. AnjuGulia, Dr. RajanVohra, Praveen Rani (2014). Liver Patient Classification Using Intelligent Techniques,

[4]. International Journal of Computer Science and Information Technologies, Vol. 5 (4) ,pp. pp. 5110-5115.

[5]. SinaBahramirad, et al. (2013), “Classification of Liver Disease Diagnosis: A Comparative Study”, IEEE, pp. 42-46.

[6]. Ersaa M. Hashem and Mai S. Mabrouk (2014), “A Study OfSopport Vector Machine Algorithm For Liver Disease Diagnosis”, American Journal of Intelligent Systems, pp. 9-14

[7]. WEKA Data Mining Tools: http://www.cs.waikato,ac.nz/~ml/weka/ (Browsing date: Oct 2015).

[8]. Han, J.,&Micheline, K. (2006). Data mining: Concepts and Techniques, Morgan Kaufmann .Publisher.

[9]. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets.html].

[10]. Pujari, A. K. (2001), Data Mining Techniques. Universities Press (India) Private Limited. 4th ed., ISBN: 81-7371-380-4.

[11]. Parimala, R. and Nallaswamy, R. (2011 a), A Study of Spam e-mail Classification using Feature Selection Package. Global Journal of Computer Science and Technology. 11, ISSN: 0975-4172.

[12]. Wang, J. (2003). Data Mining: opportunities and challenge,IdeaGroup, USA.