Improving Classification Accuracy by using Feature Selection and Ensemble Model

Pushpalata Pujari, Jyoti Bala Gupta

Abstract*— Classification is an important technique of data mining. In this paper feature selection technique and an ensemble model is proposed to improve classification accuracy. Feature selection technique is used for selecting subset of relevant features from the data set to build robust learning models. Classification accuracy is improved by removing most irrelevant and redundant features from the dataset. Ensemble model is proposed for improving classification accuracy by combining the prediction of multiple classifiers. Three decision tree data mining classifiers CART, CHAID and QUEST are considered in this paper for classification. The ionosphere dataset investigated in this study is taken from UCI machine learning repository which is classified under two category “Bad” and “Good”. The proposed ensemble model combines the classifiers CART, CHAID and QUEST by using confidential-weighted voting scheme. A comparative study is carried on the performances of the classifiers before and after carrying out feature selection. The performance of each classifier and ensemble model is evaluated by using statistical measures like accuracy, specificity and sensitivity. Gain chart and R.O.C (Receiver operating characteristics chart) are also used for measuring performances. It is found that with feature selection the ensemble model provides a greater accuracy of 93.84% than any of the individual model. Experimental results show that the proposed ensemble model with feature selection is quite effective for the task of classification of ionosphere dataset.

***Index Terms*—Classification, Ensemble Model, Ionosphere Dataset, Feature Selection.**

I. INTRODUCTION

Classification model [1] [2] [3] is an analysis technique used to describe data classes. Based on supervised learning the process automatically creates a classification model from a set of records called a training set. The induced model consists of patterns essentially generalization over the records, that are useful for distinguishing classes. Once a model is induced, it can be used automatically to classify records belonging to a small set of class that is predefined called a testing set. Training refers to building a new model by using historical data and testing refers to trying out the model on new, previously unseen data to determine its accuracy. Training is typically done on a large proportion of the total data available, where as testing is done on some small percentage of the data.The training dataset is used to train or build a model. Once a model is built on training data, the accuracy of the model on unseen data can be found. In this paper feature selection technique is applied to the data set to retrieve more important attributes. The proposed system uses an ensemble model of three decision tree data mining classifiers CART, CHAID and QUEST. The idea of the ensemble model is to employ multiple models to do better than a single individual model. The three models are combined by using confidential weighted voting scheme .Classification techniques of data mining such as CART, CHAID, QUEST and ensemble model is analyzed on ionosphere dataset. Each sample of the dataset is classified into a bad or a good group. A comparative study is carried out among the three models and its ensemble model for the prediction of radar condition on ionosphere dataset. The performance of individual models and ensemble model is evaluated by using different statistical measures including classification accuracy, specificity and sensitivity.

II. RELATED WORKS

Some of the recent research works related to improving classification accuracy by using feature selection and ensemble model are as follows.

Shu-Ting Luo et al. [14] have used two feature selection methods, forward selection (FS) and backward selection (BS) , to remove irrelevant features for improving the results of breast cancer prediction. They showed that feature reduction is useful for improving the predictive accuracy. In addition they applied decision tree (DT), support vector machine —sequential minimal optimization (SVM-SMO) and their ensembles were applied to solve the breast cancer diagnostic problem .Their results demonstrated that ensemble classifiers are more accurate than a single classifier.

R. Nithya et al. [15] have developed a CAD (Computer Aided Diagnosis) system based on neural network and a proposed feature selection method. They proposed Maximum Difference Feature Selection for detection of breast cancer (MDFS). They showed that neural network based model with proposed feature selection method improved the classification accuracy to a large extent.

Alexey Tsymbal et al. [16] proposed two new sequential-search-based strategies for ensemble feature selection, and evaluated them by constructing ensembles of simple Bayesian classifiers for the problem of acute abdominal pain classification and compared the search strategies with regard to achieved accuracy, sensitivity, specificity, and the average number of features they select.

Thomas Abeel et al. [17] have analyzed the robustness of a biomarker selection algorithm and conducted a large-scale

Manuscript received on April 26, 2012.

Pushpalata Pujari, Computer Science & Information Technology Department , Guru Ghasid Das Central University , Bilaspur, India, Mobile No- 94252-62192, (e-mail: pujari.lata@rediffmail.com).

Jyoti Bala Gupta, Information Technology Department, C.V.Raman University, Bilaspur, India, Mobile No-993867210,
(e-mail: jyoti_jbg@yahoo.co.in).

analysis of the recently introduced concept of ensemble feature selection, where multiple feature selections are combined in order to increase the robustness of the final set of selected features. They focused on selection methods that are embedded in the estimation of support vector machines (SVMs). Their feature selection extensions also offered good results for gene selection tasks. They showed that the robustness of SVMs for biomarker discovery can be substantially increased by using ensemble feature selection techniques, while at the same time improving upon classification performances.

Gidudu. A [18] showed that classification accuracy increased more as the number of features per base classifier increases than as the number of base classifiers increases. He also showed that classification accuracy increases with additional features up to a given limit beyond which increasing the number of features per base classifier did not significantly increase classification accuracy.

Zili Zhang et al. [19] proposed a multi-objective Genetic algorithm (GA) and ensemble classifiers to improve the overall sample classification accuracy identifying the most important features in the data set of interest. They showed that the GA ensemble model outperformed other algorithms in comparison and found to be the best method for classification.

III. DATA SET DESCRIPTION

The ionosphere dataset [7] investigated in this study is taken from UCI machine learning dataset. The radar data was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere. Received signals were processed using an auto correlation function whose arguments are the time of a pulse and the pulse number. There are 17 pulse numbers for the system. Instances in this database are described by 2 attributes per pulse number, corresponding to the complex values returned by the function resulting from the complex electromagnetic signal. All 34 predictor's attributes are continuous. The 35th attribute is either "Good" or "Bad" according to the definition summarized above. Classification techniques are applied on this data set.The ionosphere data set contains all total 351 instances. Out of 351 instances 126 instances are categorized under "Bad" class and 225 instances are categorized under "Good" class. Table 1 shows the attributes of ionosphere dataset. Two mutually exclusive datasets, a training dataset comprising 60% of the total ionosphere dataset, and test dataset of 40% is created by using partitioning node and balanced node partitioning techniques. Out of 351 instances 205 instances are taken as training data set and 146 instances are taken as testing data set. Table.2. shows the instances of training and testing data set. Fig.1. shows the proportion of ionosphere dataset for training and testing.

Table.1. attributes of ionosphere dataset

Attributes	Types	Values
P01	Range	[0,1]
P02	Range	[-1.0,1.0]
P03	Range	[-1.0,1.0]
.	.	..
P34	Range	[-1.0,1.0]
P35(Target)	Flag	g/b

Table.2 Instances of training and testing data set

Class	Training	Testing	Total
Bad	74	52	126
Good	131	94	225
Total	205	146	351

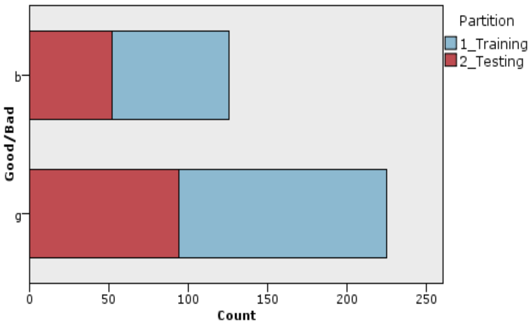

Fig.1. Proportion of training and testing dataset

IV. METHODOLOGY

A. Feature selection

Feature selection [9] helps to identify the fields that are most important in predicting a certain outcome. Feature selection consists of three steps. Screening: It removes unimportant and problematic predictors and records or cases, such as predictors with too many missing values or predictors with too much or too little variation to be useful. Ranking: Sorts remaining predictors and assigns ranks based on importance. Selecting: It identifies the subset of features by preserving only the most important predictors and filtering or excluding all others.From a set of hundreds or even thousands of predictors, the Feature Selection screens, ranks, and selects the predictors that are most important. The predictors which contribute less in prediction can be skipped from the data set. Ultimately, it ends up with a quicker, more efficient model that uses fewer predictors, executes more quickly, and easier to understand. In this piece of research work importance of attributes are ranked based on Pearson Chi-square measure. The unimportant features are skipped and the performances are compared against the performances of the classifiers before carrying out feature selection. Table.3. shows the list of important and unimportant attributes after carrying out feature selection technique with their rank and values.

Table.3. List of important attributes and unimportant attributes with their ranks and values.

Important Attributes			Unimportant Attributes		
Rank	Attributes	Values	Rank	Attributes	Value
1	P03	1.0	1	P27	0.92
2	P01	1.0	2	P16	0.87
3	P05	1.0	3	P17	0.81
4	P07	1.0	4	P10	0.78
5	P09	1.0	5	P22	0.69
6	P33	1.0	6	P04	0.69
7	P29	1.0	7	P26	0.47
8	P21	1.0	8	P28	0.44
9	P23	1.0	9	P23	0.41
10	P31	1.0	10	P20	0.22
11	P08	1.0	11	P24	0.13
12	P15	0.99	12	P30	0.4
13	P25	0.99	13	P32	0
14	P06	0.99	14	P02	0
15	P13	0.98			
16	P19	0.97			
17	P12	0.96			
18	P14	0.96			
19	P11	0.96			
20	P18	0.96			

B. CART (Classification and Regression Tree) classifier

The Classification and Regression (C&R) Tree model [8], [11] generates a decision tree to predict or classify future observations. CART builds a binary tree by splitting the records at each node according to a function of a single input field. The measure used to evaluate a potential splitter is diversity. The best splitter is the one that decreases the diversity of the record sets by the greatest. This method uses recursive partitioning to split the training records into segments with similar output field values. The CART tree node starts by examining the input fields to find the best split, measured by the reduction in an impurity index that results from the split. The initial split produces two nodes, each of which is attempted to split in the same manner as the root node. In this way all the input fields are examined to find candidate splitters. If the field only takes on one value, it is eliminated from consideration. The best field for each of the remaining fields is determined. When no split can be found that significantly decreases the diversity of a given node, it is labeled as a leaf node. CART trees gives the option to first grow the tree and then prune based on a cost-complexity algorithm that adjusts the risk estimate based on the number of terminal nodes. This method, which allows the tree to grow large before pruning based on more complex criteria, may result in smaller trees with better cross-validation properties. Increasing the number of terminal nodes generally reduces the risk for the current (training) data, but the actual risk may be higher when the model is generalized to unseen data. To train CART model [6] there should be one or more In fields and exactly one Out field. Target and predictor fields can be range or categorical. Fields set to both or none are ignored. Fields used in the model must have their types fully instantiated, and any ordinal fields used in the model must have numeric storage (not string). If necessary, the Reclassify node can be used to convert them.

C. CHAID (Chi-squared Automatic Interaction Detection) classifier

CHAID, or Chi-squared Automatic Interaction Detection [8], is a classification method for building decision trees by using chi-square statistics to identify optimal splits. In CHAID each of the input or predictor fields is considered as a potential splitter. CHAID first examines the cross tabulations between each of the predictor variables and the outcome and tests for significance using a chi-square independence test. In the first step, all the predictor fields that do not produce statically significant differences in the target field values are merged. In the second step, each group of three or more predictors is re-spilt by all possible binary division. If any of these splits yields a statically significant difference in outcomes, it is retained. Once each of the predictor field has been grouped to produce the maximum possible diversity of classes in the target field, the chi-squared test is applied to the resulting groupings. The predictor that generates the groupings that differ the most according to this test is chosen as the splitter for the current node. Target and predictor fields [6] can be range or categorical; nodes can be split into two or more subgroups at each level. Any ordinal fields used in the model must have numeric storage (not string). CHAID can generate non-binary trees. It therefore tends to create a wider tree than the binary growing methods. CHAID works for all types of predictors, and it accepts both case weights and frequency variables.

D. QUEST (Quick, Unbiased, Efficient Statistical Tree) decision tree classifier

QUEST is a binary classification method [6] for building decision trees. It uses a sequence of rules, based on significance tests, to evaluate the predictor variables at a node. For selection purposes, as little as a single test may need to be performed on each predictor at a node.Spliting predicate in QUEST are determined by running quadratic discriminate analysis using the selected predictor on groups formed by the target categories. It separates splitting predicate selection into variable selection and split point selection. It uses statistical significance tests instead of impurity function. Predictor fields can be numeric ranges, but the target field must be categorical. All splits are binary. Weight fields cannot be used. Any ordinal fields used in the model must have numeric storage (not string). If necessary, the reclassify node can be used to convert them.

E. Ensemble Model

An ensemble model [12] is defined as a set of individually trained classifier whose predictions are combined when classifying a new data. Ensemble combines the output of several classifier produced by weak learner into a single composite classification. It can be used to reduce the error of any weak learning algorithm. The purpose of combining all these classifier together is to build an ensemble model which will improve classification accuracy as compared to each individual classifier. The three models are combined by using confidential weighted voting scheme [5] where weights are weighted based on the confidence value of each prediction. Then the weights are summed and the value with highest total is again selected. The confidence for the final selection is the sum of the weights for the winning values divided by the number of models included in the ensemble model. If one model predicts no with a higher confidence than the two yes predictions combined, then no wins. It not only increases classification accuracy but also reduce chances of over training since it avoids a biased decision by integrating the different predictions from individual classifiers. The ensemble model presented in this paper combines the prediction of CART, CHAID and QUEST decision tree classifier. Fig. 2 shows the architecture of the proposed model. Fig. 3 shows the block diagram of the proposed model. In the proposed model ionosphere data set is partitioned into 60 % of training set and 40 % of testing set. Then feature selection technique is applied to the data set to skip unimportant attributes. The data set with reduced number of attributes is then applied to the three classifiers and their ensemble model. The training data set is applied to train the model and the testing data set is applied to test the model. The performance of the classifiers and ensemble model is measured by using statistical measures like classification accuracy, sensitivity and specificity. Gain chart and R.O.C chart are also used for performance measurement.

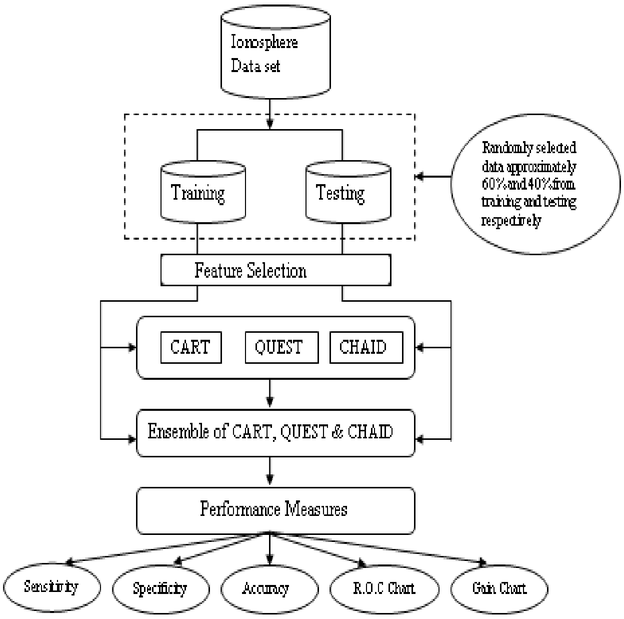

Fig.2. Architecture of the proposed model

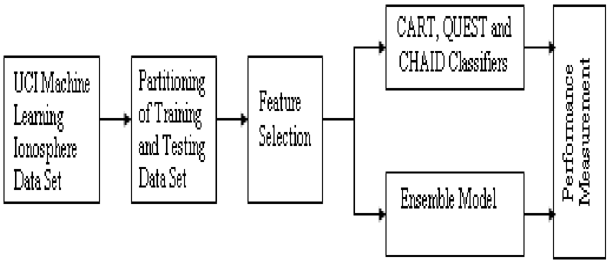
UCI Machine Learning Ionosphere Data Set
Partitioning of Training and Testing Data Set
Feature Selection
CART, QUEST and CHAID Classifiers
Ensemble Model
Performance Measurement

Fig.3. Block diagram of the proposed model

V. PERFORMANCE MEASUREMENT

The performance of each classification [4] model is evaluated by using three statistical measures: classification accuracy, sensitivity and specificity. These measures are defined by true positive (TP), true negative (TN), false positive (FP) and false negative (FN) cases. Say we test some people for the presence of a disease. Some of these people have the disease, and our test says they are positive. They are called true positives. Some have the disease, but the test claims they don't. They are called false negatives**.** Some don't have the disease, and the test says they don't **-** true negatives**.** Finally, we might have healthy people who have a positive test result false positives**.** Table.4. represents a matrix showing number of TP, TN, FP, and FN cases.

Table.4. Matrix for Actual and Predicted cases

	P'(predicted)	N'(predicted)
P(Actual)	True Positive	False Negative
N(Actual)	False Positive	True Negative

A. *Classification Accuracy*

It measures the proportion of correct predictions considering the positive and negative inputs. It is highly dependent of the data set distribution which can easily lead to wrong conclusions about the system performance. It is calculated as follows

$$ACC = \text{Total Hits} / \text{Number of entries in the set} = (TP+TN)/(P+N) \qquad \text{... (1)}$$

B. *Sensitivity*

It 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

$$SENS = \text{Positive hits} / \text{Total Positives} = TP/(TP+FN) \qquad \text{... (2)}$$

C. *Specificity*

It 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 to the desired one. It is calculated as follows

$$SPEC = \text{Negative hits} / \text{Total negatives} = TN/(TN+FP) \qquad \text{... (3)}$$

VI. RESULTS AND DISCUSSION

The experimental work is carried out by using Clementine Software.The ionosphere dataset contains 351 dataset with class distribution: Bad, Good. The whole dataset is divided for training the models and test them by the ratio of 60: 40 % respectively. The data set is initially partitioned into training and test sets. Feature selection technique is carried out on the data set. The classifiers are trained on the former. The test set is used to evaluate the generalization capability of the classifiers. The predictions from the classifiers are combined to build the ensemble model and compared with the original classes to identify true positive, true negative, false positive and false negative values. These values have been computed to construct the confusion matrix [1]. A comparative study on the performance of each classifier and ensemble model is carried out before and after feature selection. Table.5. shows confusion matrices of different model of training and test data partition before feature selection. Table.6. shows confusion matrices of different model of training and test data partition after carrying out feature selection.

Table.5. Confusion matrices of different model of training and test data partition before feature selection

Model	Desired Output	Training Data		Test Data	
		Bad	Good	Bad	Good
CART	Bad	67	7	47	5
	Good	4	127	9	85
CHAID	Bad	70	4	41	11
	Good	9	112	18	76
QUEST	Bad	60	14	44	8
	Good	4	127	6	88
Ensemble	Bad	68	6	46	6
	Good	4	127	6	88

Table.6. Confusion matrices of different model of training and test data partition before feature selection

Model	Desired Output	Training Data		Test Data	
		Bad	Good	Bad	Good
CART	Bad	65	9	46	6
	Good	3	128	7	87
CHAID	Bad	69	5	42	10
	Good	5	126	7	87
QUEST	Bad	60	14	44	8
	Good	4	127	6	88
Ensemble	Bad	65	9	46	6
	Good	4	127	3	91

Table.5 and Table.6 show the computed confusion matrices. Each cell [10] contains the row number of samples classified for the corresponding combination of desired and actual model output. The prediction are compared with original classes to identify true positive, true negative, false positive and false negative. Table.7. and Table.8.show the values of three statistical parameters (sensitivity, specificity and total classification accuracy) for different models of training and test data partition before and after feature selection respectively.

Table.7. Values of statistical measures of different models for training and test data partition before feature selection.

		Measures %		
Model	Partition	Accuracy	Sensitivity	Specificity
CART	Training	94.63	90.54	96.94
	Test	90.41	90.38	90.42
CHAID	Training	93.66	94.59	93.12
	Test	80.14	78.84	80.85
QUEST	Training	91.22	81.08	96.94
	Test	90.41	84.61	93.61
Ensemble	Training	95.12	91.89	96.94
	Test	91.78	88.64	93.61

Table.8. Values of statistical measures of different models for training and test data partition after feature selection.

		Measures %		
Model	Partition	Accuracy	Sensitivity	Specificity
CART	Training	94.15	87.83	97.7
	Test	91.1	88.46	92.55
CHAID	Training	95.12	93.24	96.18
	Test	88.36	80.76	92.55
QUEST	Training	91.22	81.08	96.94
	Test	90.41	84.61	93.61
Ensemble	Training	93.66	87.83	96.94
	Test	93.84	88.46	96.8

These results show that the accuracy and sensitivity of CART is better than the other two individual models. Specificity of QUEST is better than the other two models. The ensemble method has achieved a better result for both training and testing sample. Another way to compare the performance of different classifier is gain chart and ROC chart.

A. Gain Chart

The gains chart plots [6] the values in the Gains % column from the table. Gains are defined as the proportion of hits in each increment relative to the total number of hits in the tree, using the following equation:

(Hits in increment / total number of hits) x 100% … (4)

Cumulative gains charts always start at 0% and end at 100% as we go from left to right. For a good model, the gains chart will rise steeply toward 100% and then level off. A model that provides no information will follow the diagonal from lower left to upper right. The steeper the curve the higher is the gain. Fig.4. shows the gain chart for three models and ensemble model before feature selection for training and testing data set respectively. Fig.5. shows the gain chart for three models and ensemble model after feature selection for training and testing data set respectively.

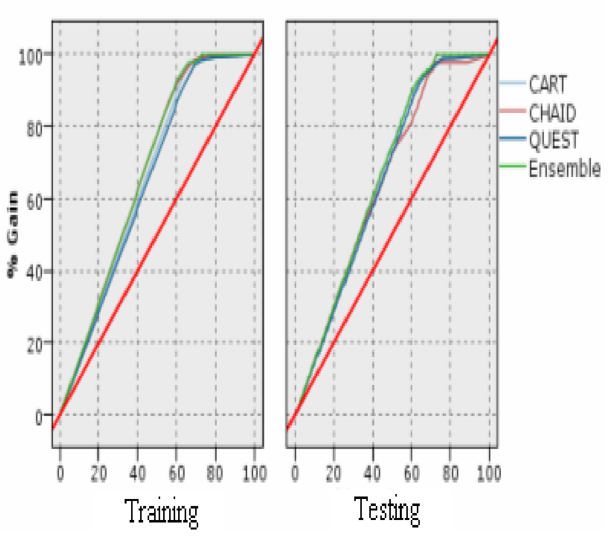

Fig.4. Gain chart for three models and ensemble model before feature selection for training and testing data set respectively

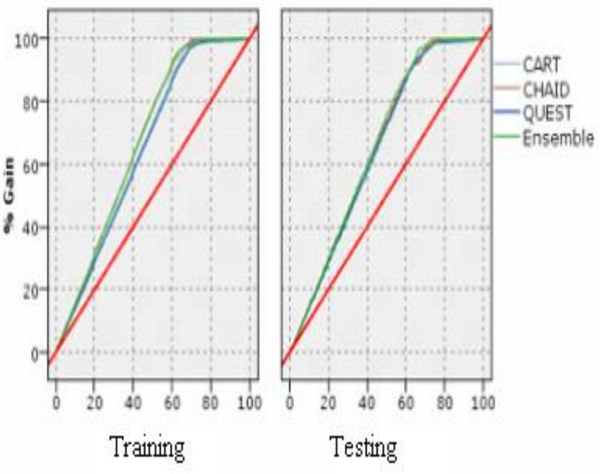

Fig.5. Gain chart for three models and ensemble model before feature selection for training and testing data set respectively

B. R.O.C chart

The response chart [13] plots the values in the Response (%) column of the table. ROC curve, is a graphical plot of the sensitivity, or true positives, vs. (1 - specificity), or false positives, for a binary classifier system

The response is a percentage of records in the increment that are hits, using the following equation:

(Responses in increment / records in increment) x100%... (5)

Response charts usually start near 100% and gradually descend until they reach the overall response rate (total hits / total records) on the right edge of the chart. For a good model, the line will start near or at 100% on the left, remain on a high plateau as you move to the right, and then trail off sharply toward the overall response rate on the right side of the chart. Fig. 6 shows the ROC chart for three models and ensemble model before feature selection for training and testing data set respectively. Fig. 7 shows the ROC chart for three models and ensemble model before feature selection for training and testing data set respectively.

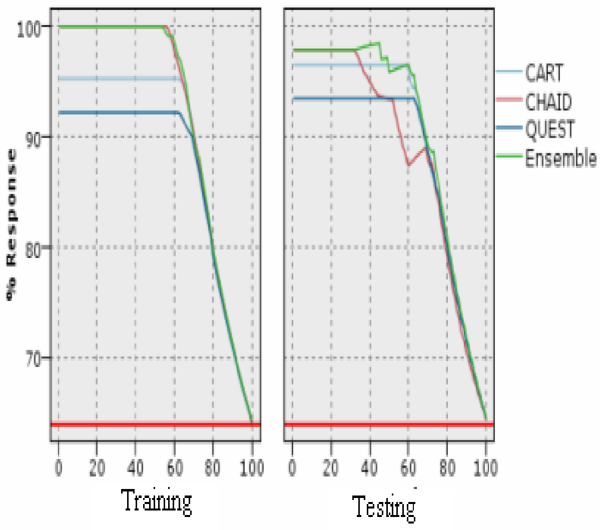

Fig.6. ROC chart for three models and its ensemble model before feature Selection for training and testing data set respectively

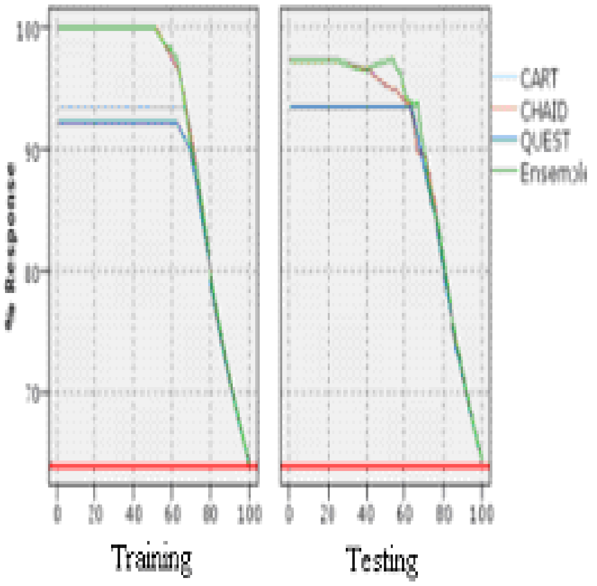

Fig.7. ROC chart for three models and its ensemble model after feature selection for training and testing data set respectively

VII. CONCLUSION

The main goal of this study is to show the effectiveness of feature selection technique and ensemble model in improving classification accuracy. The performance of three different classifiers CART, CHAID and QUEST and its ensemble model is analyzed on ionosphere dataset. The performance of all classifier is investigated by using statistical measures like accuracy, specificity and sensitivity. Also the performance of each classifier is investigated with the help of gain chart and ROC chart for both training and testing set. Table.6 and 7 show the classification accuracy for training and testing data set for the three models and its ensemble model before and after carrying feature selection respectively.

The accuracy of CART, CHAID and QUEST are found to be 94.63, 93.66, 91.22 on training dataset and 91.1, 88.36, 90.41 on test dataset before feature selection. The accuracy of CART, CHAID and QUEST are found to be 94.15, 95.12, 91.22 on training dataset and 90.41, 80.14 and 90.41 on test dataset after feature selection, where as the accuracy of the ensemble model is found to be 95.12 and 91.78 on training and testing dataset respectively before feature selection. After feature selection the accuracy of ensemble model is found to be 93.66 and 93.84 respectively on training and testing data set. In overall the ensemble model with feature selection has achieved a remarkable performance with highest accuracy of 93.84 on test data set. In all respect ensemble model is performing well, hence this model can be recommended for the classification of ionosphere data set.

REFERENCES

1. Jiwaei Han, Kamber Micheline, Jian Pei "Data mining: Concepts and Techniques", Morgam Kaufmann Publishers (Mar 2006).
2. Cabena, Hadjinian, Atadler, Verhees, Zansi "Discovering data mining from concept to implementation" International Technical Support Organization, Copyright IBM corporation 1998.
3. S.Mitra, T. Acharya "Data Mining Multimedia, Soft computing and Bioinformatics, A john Willy & Sons, INC , Publication, 2004.
4. Alaa M. Elsayad "Predicting the severity of breast masses with ensemble of Bayesian classifiers" journal of computer science 6 (5): 576-584, 2010, ISSN 1549-3636
5. Alaa M. Elsayad " Diagnosis of Erythemato-Squamous diseases using ensemble of data mining methods" ICGST-BIME Journal Volume 10, Issue 1, December 2010
6. SPSS Clementine 12.0, 2007. Data mining workbench software. Product
7. Of SPSS, Inc. http://www.cad100.net/247_dataminingworkbench-SPSS-Clementine-12.html
8. UCI Machine Learning Repository of machine learning databases.University of California, school of Information and Computer
9. Science, Irvine. C.A. http://www.ics.uci.edu/~mlram,?ML.Repositary.html
10. Michael J.A .Berry Gordon Linoff "Data Mining Techniques for Marketing, Sales and Customer Support ", John Wiley & Sons publishers, 1997
11. P.Nancy and R.Geetha Ramani, "A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data", International Journal of Computer Applications (0975 – 8887), Volume 32– No.8, October 2011.
12. Milan Kumari and Sunila Godara "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCSt Vol. 2, ISSue 2, June 2011, IJCSt Vol. 2, ISSUE 2, June 2011 ISSN: 2229-4333 (Print) | ISSN: 0976-8491 (Online)
13. Matthew N Anyanwu &Sajjan G Shiva "Comparative Analysis of serial Decision Trees Classification Algorithms", (IJCSS), Volume (3): Issue (3)
14. Mahesh Pal "Ensemble Learning With Decision Tree for Remote Sensing Classification", World Academy of Science, Engineering and Technology 36 2007.
15. Kelly H. Zou, PhD; A. James O'Malley, PhD; Laura Mauri, MD, MSc "ROC Analysis for Evaluating Diagnostic Test and Predictive Models"
16. Shu-Ting Luo & Bor-Wen Cheng, "Diagnosing Breast Masses in Digital Mammography Using Feature Select ion and Ensemble Methods" J Med Syst, DOI 10.1007/s10916-010-9518-8, Springer Science+Business Media, LLC 2010.
17. R.Nithya, B.Santh "Mammogram Classification using Maximum Difference Feature Selection Method", Journal of Theoretical and Applied Information Technology, 30 Th November 2011. Vol. 33 No.2, ISSN: 1992-8645, E-ISSN: 1817-3195.
18. Alexey Tsymbal, Pádraig Cunningham, Mykola Pechenizkiy, Seppo Puuronen "Search Strategies for Ensemble Feature Selection in Medical Diagnostics" Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems (CBMS'03) 1063-7125/03 © 2003 IEEE
19. Thomas Abeel, Thibault Helleputte, Yves Van D Peer etc, "Robust biomark identification for cancer diagnosis with ensemble feature selection methods" Oxford Journals, Bioinformatics , Volume 26 , Issue 3, PP :392-398.
20. Gidudu. A, "Random ensemble feature selection for land cover mapping", Geo Science and remote Sensing Symposium, 2009 IEEE , International IGRSS 2009, Volume: 2, On Page(s): II-840 - II-842
21. Zhang, Zili and Yang, Pengyi 2008, "An ensemble of classifiers with genetic algorithm Based Feature Selection", The IEEE intelligent informatics bulletin, vol. 9, no. 1, pp. 18-24.

AUTHORS PROFILE

Pushpalata Pujari has received her master degree in Computer Application from Berhampur University, India. Currently she is working as an assistant professor in the department of computer science and IT, Guru Ghasi Das Central University, India. Pushpalata Pujari has published several papers in national and internal conferences.All focusing in classification, data mining, and soft computing. Her current research interest involves improving classification accuracy of data mining algorithms.

Jyoti Bala Gupta works as an assistant professor in the department of IT, Dr. C.V Raman University, India. She has received her master degree from Guru Ghasi Das University, Bilaspur, India. She has published several papers in national, international conferences and journals. Her current research interest involves image processing and data mining.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication