IJCAT - International Journal of Computing and Technology, Volume 7, Issue 6, June 2020
ISSN (Online) : 2348-6090
www.IJCAT.org
Impact Factor: 0.835

Ensemble of Data Mining Classifiers for Classification of Cancer Dataset

Pushpalata Pujari

Department of CSIT, Guru Ghasidas Vishwavidyalaya, Bilaspur, C.G, India

Abstract - This paper proposes an ensemble model for classification of Cancer dataset. Ensemble models are used to improve the classification accuracy of a system by combining the outcomes of individual classifiers. In this paper a number of data mining classifiers like C5.0 (C5.0 Decision Tree), CART(Classification and Regression Tree), CHAID(Chi-squared Automatic Interaction Detection), QUEST(Quick Unbiased Efficient Statistical Tree), ANN (Artificial Neural Network) and SVM(Support Vector Machine) are used as individual classifier for classification purpose. The outcomes of the individual classifiers are evaluated using performance measures like accuracy, specificity, sensitivity, gain charts and response chart. A comparative analysis is carried out among the individual classifiers. Further to improve the classification accuracy of the system the outcomes of individual classifiers are combined using confidential voting scheme to develop the ensemble model. The performance of the ensemble model is evaluated and compared with the individual classifiers. From experiment it is found that the ensemble model developed exhibit well as compared to the individual classifiers.

Keywords - *Classification, C5.0, CART, CHAID, QUEST, ANN, SVM and Ensemble model.*

1. Introduction

The management and analysis of information and using existing data for correct prediction has been an important and challenging research area for many years. Information can be analyzed in various ways. Classification of information is an important part of decision making tasks. In the context of data mining, data mining [2], [4], [9] is classified into supervised and unsupervised concept learning methods. Supervised learning depends on predefined classes to build classification models by forming concept definitions from set of predefined data. Unsupervised learning does not depend on the predefined classes to build models. It uses a clustering system where instances are grouped together based on a similarity scheme. Identification of target attributes is the requirement of supervised learning. The supervised learning tries to find patterns between independent attributes (predictors) and the dependent attribute. It builds a model that best represents the functional relationships. Typically, for the data mining process, the data is separated into two parts; one for training and another for testing. The initial model is built using the first sample of the data and then the model is applied to the second sample to evaluate the accuracy of the model predictions. All the classification models [4] are built in 4 steps. The first step is identifying a set of subjects with a known behaviour. In this step all inputs and their target classes are well known in advance. The second step is preparation of data which includes cleaning of data, selection of most important features and transformation of data. The third step is training the model. This process uses about two third of all the subjects identified in the first step to identify the relationships between the inputs and the target data. Classification algorithms used plays an important role in finding these relationships. The fourth step is testing the model. This step uses the remaining one third of subjects to test the relationships identified in the previous step. The efficiency of the model how accurate it is tested in this step.

In this paper an ensemble model is proposed for classification of cancer dataset. The ensemble model is built using individual classifier like C5.0, CART, CHAID, QUEST, ANN and SVM. The paper is organized into seven sections. Section 2 deals with background detail. Section 3 deals with the dataset used. Section IV describes individual classifiers and ensemble model used. Performance measures are described in section V. Section VI deals with experimental results followed by conclusion.

2. Background Details

Some of the research works related to classification of cancer dataset are as follows.

In [6] Alaa M. Elsayad investigated three different data mining methods; multilayer perceptron neural network, C5.0 decision tree and linear discriminate analysis in order to build an ensemble model to the problem of differential diagnosis of these erythemato-squamous diseases. The classification of ensemble model was found to provide

greater accuracy. In [18] the authors have used different data mining techniques, neural networks and association rule mining, for anomaly detection and classification. Both the techniques achieved classification accuracy over 70% percent. In [19] G.Sujatha et.al have used ID3,C4.5 and CART classifiers to obtain better accuracy and execution time for the decision tree construction. It is observed that C4.5 performs well for tumor dataset. For the enhanced data set of primary tumor C4.5 decision tree classifier is found to be the best one. Both ID3 and C4.5 exhibit well for enhanced Colon tumor data set and obtained equal classification accuracy. Five different classification algorithms namely Naive Bayes, Support Vector Machines (SVM), Radial Basis Neural Networks (RB-NN), Decision trees J48 and simple CART have been proposed in [20] by Aruna et.al. The classification results were analyzed.

A comparative study has been carried out among the classifiers on the Wisconsin Breast Cancer dataset. In [21] Aik Choon Tan et.al focused on C4.5 decision tree, and bagged and boosted decision trees supervised machine learning techniques for cancer classification on seven publicly available microarray data. They observed that the performance of ensemble learning (bagged and boosted decision trees) is better as compared to the individual decision trees in classification task. In [22] D.Lavanya et.al proposed a hybrid method to enhance the classification accuracy of Breast cancer data sets. In this feature selection method is used to eliminate those attributes that have no significance in the application process. From the experimental study the hybrid approach with the combination of preprocessing, bagging with CART enhanced classification accuracy to a greater extent. Delenet.al have compared ANN, decision tree and logistic regression techniques for breast cancer survival analysis [23] .

They have used the SEER data's twenty variables in the prediction models. The decision tree with 93.6% accuracy and ANN with 91.2% were found more superior to logistic regression with 89.2% accuracy. In [24] Jacob et.al, explored the performance of classification algorithms on the Breast Cancer dataset through Data mining algorithms. The authors found that the Random Tree algorithm and the Quinlan's C4.5 algorithm produced 100 percent classification accuracy. An ensemble classifier
approach has been proposed in [26] using base classifiers like Naive Bayes classifier, Random Forest classifier, SVMs and Logistic Regression for twitter sentiment analysis. In [27] an ensemble approach is proposed based on the integration of rule-based classifier with machine learning techniques for detection of spam reviews in Arabic online sources.

3. Dataset Description

The data set used in this study is taken from UCI machine learning repository [8]. The dataset contains 32 numbers of attributes out of which 31 attributes are taken as input attributes and the rest one is taken as output attribute. Each sample of the dataset is classified into two categories: Benign and Malignant. The target attribute is a binary valued variable, whether the patient shows signs of cancer disease. All total the dataset contains 222 numbers of attributes. Out of which 106 numbers of instances belongs to class Benign and 116 attributes belongs to class Malignant.

Models [3] are developed in two phases: training and testing Training refers to building a new model by using historical data, and testing refers to trying out the model on new unseen data to find the efficiency and performance of the model. Often a large subset of the whole data sample is considered for training the model and the remaining subset is used for testing the model. Once a model obtained from training subset, it is applied on testing subset to find the accuracy of the model. Two mutually exclusive datasets [10], a training dataset comprising 70% of the total cancer dataset, and a testing dataset of 30% is created by using partitioning node and balanced node portioning techniques. Classification techniques are applied on this dataset. In all there are 222 numbers of instances in the cancer dataset out of which 151 instances are taken as training set and 71 instances are taken as testing set by using balanced node concept. Out of 151 training instances 72 instances belongs to class 'Benign' and 79 instances belongs to class 'Malignant'. Out of 71 testing instances 34 instances belongs to class 'Benign' and 37 instances belongs to class 'Malignant'. Table 1 shows the number of instances taken for training and testing data set.

Table 1: Number of Instances for Training and Testing Dataset.

CLASS	TRAINING	TESTING	TOTAL
BENIGN	72	34	106
MALIGNANT	79	37	116
TOTAL	151	71	222

4. Proposed Model

The proposed model includes six data mining classifiers C 5.0, CART, CHAID, QUEST, ANN and SVM for diagnosis. Around 70% of the dataset is taken as training subset and 30% of the dataset is taken as testing subset. Training set and testing sets are applied on each classifier. The performance of all classifiers are combined together to form the ensemble model. The performances of the classifiers and the ensemble model are evaluated and compared with various statistical measures. For measuring the performance statistical measures accuracy, specificity, sensitivity, gain chart and response chart are used.

4.1 C5.0 classifier

C 5.0 classifier [14] works by finding the field that provides maximum information gain. Basing on the field it split the whole sample. The subsample formed from the first split is again split based on other field. The processed is repeated again and again until no more split is possible further. At last the terminal nodes are reexamined. The terminal nodes which do not contribute significantly are pruned or removed. It first finds the independent attribute that best separates the tuples of different classes from each other at each stage of the construction of the classifier. Different classes are defined as having different values for the target attribute. The goal of each split is to obtain a new set of tuples containing as many as possible belonging to one class and as few as possible belonging to other classes. This process of splitting the data set is recursively repeated for each subset until a subset contains only instances from a single class or until a further split is not expected to result in any improved prediction accuracy. Only categorical targets are predicted by C5.0 classifier. There must be one categorical output field and one or more input fields of any type for training C5.0 classifier. C5.0 can handle missing data and large number of inputs efficiently. It requires very less training time period to estimate. In this paper C5.0 is viewed in terms of a set of rules derived from the model having very simple interpretation. It also offers the powerful boosting method to increase accuracy of classification. C5.0 algorithm learns by using simple decision rules inferred from the input attributes to predict the value of a target variable. The rule sets generated by C5.0 classifier for classification are as follows

Rules for M - contains 4 rule(s)
Rule 1 for M
if variable 26 > 803.700
and variable 30 > 0.110
then M
Rule 2 for M
if variable 30 > 0.110
and variable 31 > 0.354
then M
Rule 3 for M
if variable 20 <= 0.011
and variable 30 > 0.110
then M
Rule 4 for M
if variable 16 > 41.180
then M
Rules for B - contains 3 rule(s)
Rule 1 for B
if variable 16 <= 41.180
and variable 30 <= 0.110
then B
Rule 2 for B
if variable 10 <= 0.029
and variable 30 <= 0.110
then B
Rule 3 for B
if variable 26 <= 803.700
and variable 31 <= 0.354
then B
Default: B

Where 'M' represents malignant case and 'B' represents benign case.

4.2 CART (Classification and Regression Tree) classifier

The Classification and Regression Tree [7] model generates a decision tree to predict or classify future observations. For the spilt of training subsets with similar output fields values CART classifier uses recursive partitioning methods. The spilt by CART is based on the reduction in an impurity index that results from the spilt. It examines the input fields to find the best split. The generated subsets from the first split are spilt again and the process is repeated until the stopping criterion is reached. All splits are binary (only two subgroups).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. Based on more complex criteria this method allows the tree to grow large before pruning into smaller trees with better cross-validation properties. To train CART model there should be one or more input fields and exactly one output field. CART classifier accepts range or categorical predictor and target fields. Fields set to both or none are ignored by CART. 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).The rule sets generated by CART classifier for classification are as follows

Variable 30 <= 0.109 [Mode: B]
Variable 4 <= 21.720 [Mode: B] => B
Variable 4 > 21.720 [Mode: M]
Variable 3 <= 14.805 [Mode: B] => B
Variable 3 > 14.805 [Mode: M] => M
Variable 30 > 0.109 [Mode: M]
Variable 26 <= 805.250 [Mode: M]
Variable 31 <= 0.358 [Mode: B]
Variable 24 <= 32.920 [Mode: B] => B
Variable 24 > 32.920 [Mode: M] => M
Variable 31 > 0.358 [Mode: M] => M
Variable 26 > 805.250 [Mode: M] => M

Where 'M' represents malignant case and 'B' represents benign case

4.3 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. Each input or predictor field is significant for the split. A chi-square independent test is used by CAHID classifier for significance of the predictor. First it examines the cross tabulations between each of the predictor variables and the outcome. All the predictor fields are merged which do not produce significant differences in target fields. 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. According to the chi-square test the predictor which forms grouping that differentiate the most is considered 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. CHAID can generate non-binary trees. Unlike the binary growing methods it tends to create a wider tree. CHAID works for all types of predictors, and it accepts both case weights and frequency variables. The rule sets generated by CHAID classifier for classification are as follows

Variable 30 <= 0.099 or Variable 30 is missing [Mode: B]
Variable 16 <= 67.340 [Mode: B] => B
Variable 16 > 67.340 [Mode: M] => M
Variable 30 > 0.099 and Variable 30 <= 0.150 [Mode: M]
Variable 26 <= 768.900 [Mode: B] => B
Variable 26 > 768.900 [Mode: M]
Variable 7 <= 0.082 [Mode: M] => B
Variable 7 > 0.082 [Mode: M] => M
Variable 30 > 0.150 [Mode: M]
Variable 16 <= 20.200 [Mode: M] => M
Variable 16 > 20.200 [Mode: M] => M

Where 'M' represents malignant case and 'B' represents benign case.

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

QUEST [8] is a binary classification method for building decision trees uses a sequence of rules, based on significance tests, to evaluate the predictor variables at a node. The split in QUEST is based on quadratic discriminate analysis. It uses the selected predictor on groups formed by the target categories. It separates splitting predicate selection into variable selection and split point selection. Instead of impurity function it uses statistical significance tests.

The rule sets generated by QUEST classifier for classification are as follows

Logistic Variable 30 <= 0.123 [Mode: B]
Variable 16 <= 40.665 [Mode: B] => B
Variable 16 > 40.665 [Mode: M] => M
Variable 30 > 0.123 [Mode: M] => M

Where 'M' represents malignant case and 'B' represents benign case

4.5 ANN (Artificial Neural network)

A neural network, [4] sometimes called multilayer perceptron, is basically a simplified model of the way the human brain processes information. It works by simulating a large number of interconnected simple processing units that resemble with the neurons of human brain. Neural network is organized into three layers: an input layer, hidden layer and output layer. The input layer represents the input fields. The number of hidden layers can be one or more. The output layer represents the output field. Output layer can contain one or more neurons. The three units are connected with varying connection strengths called weights. At first inputs are presented to the input layer. The values are propagated from each neuron to every neuron in the next layer of the network. Finally the result is obtained at the output layer. Individual samples are

examined by the ANN for determining the class label of each sample. In case of in correct prediction it adjusts the weights to make correct prediction. The network improves its prediction capability by repeating the process a number of times until a stopping criterion is reached. Initially all weights of the ANN are initialized randomly. The training samples with known outputs are presented to the network. The predicted outputs are obtained at the output layer. The predicted outputs are compared with the target outputs to generate the error terms. The error terms are propagated back to adjust the weights of the network. This process is repeated again and again until the desired solution is reached. Once a model is developed from the training samples, the network is applied on unknown data for finding its performance.

In this paper the input layer consisting of 31 input, hidden layer consisting of 4 neuron and output layer consisting of 1 neuron is developed for classification purpose.

4.6 Support Vector Machine (SVM)

Support Vector Machine [4] (SVM) is a robust classification technique that maximizes the predictive accuracy of a model without over fitting the training data. For categorization of data points SVM transforms the input data into a high dimensional feature space. It generates a hyper plane as decision surface to separate the data points using support vectors. This hyper plane acts as the margin of separation between the data points. The structural risk minimization principle is used for this purpose.In this paper polynomial kernel of degree 7 is used for building the model.

4.7 Ensemble Model

An ensemble model [13], [15], [25] is a collection of models whose individual predictions are combined to improve the classification accuracy of a system. Ensemble methodology, builds a classification model by integrating multiple classifiers, can be used for improving prediction performance. The multiple classifiers are known as base classifiers. Ensemble models are always more accurate than the individual classifiers. It increases the accuracy of a classification to a significant level. Ensemble model removes a biased decision by integrating the predictions of base classifier. Hence it reduces the chances of over training. The ensemble model presented in this paper combines the prediction of C5.0, CART, CHAID, QUEST, ANN and SVM using confidential weighted voting scheme

5. Performance Measurement

Performances of each classifier are evaluated by using very well known statistical measures [5] classification accuracy, sensitivity and specificity. These measures are defined by true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Table 2 represents a matrix showing number of TP, TN, FP and FN.

Table 2: Matrix for Actual and Predicted cases

	P'(PREDICTED	N'(PREDICTED)
P(ACTUAL)	TRUE POSITIVE(TP)	FALSE NEGATIVE(FN)
N(ACTUAL)	FALSE POSITIVE(FP)	TRUE NEGATIVE(TN)

Based on the above table following statistical performance measures [15] [16] are evaluated.

5.1. 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

Classification accuracy = Total number of correct prediction hits / Total number of cases
= (TP + TN)/ (P+N) ... (1)

5.2 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 using the following formula.
Sensitivity = Positive hits / Total positives

= TP/ (TP+FN) ... (2)

5.3. 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 of the desired one. It is calculated as follows

Specificity = Negative hits / Total Negatives

Table 3: Confusion matrices of all models for training and test dataset.

Model	Desired Output	Training Data		Testing Data	
		Benign	Malignant	Benign	Malignant
C5.0	Benign	67	2	36	1
	Malignant	3	77	3	33
CART	Benign	67	2	35	2
	Malignant	3	77	1	35
CHAID	Benign	67	2	33	4
	Malignant	3	77	1	35
QUEST	Benign	63	6	30	7
	Malignant	5	75	1	35
ANN	Benign	67	2	36	1
	Malignant	1	69	1	35
SVM	Benign	68	1	35	1
	Malignant	11	69	4	32
Ensemble	Benign	67	2	36	1
	Malignant	2	78	0	36

$$= TN / (TN+FP) \quad \ldots (3)$$

6. Experimental Results

The experimental work is carried out by using Clementine Software [17]. The dataset contains 222 dataset with class distribution: Benign, Malignant. Whole dataset is divided for training the models and test them by the ratio of around 70%: 30 % respectively. The data set is initially partitioned into training and test sets. The classifiers are trained with the training dataset .The test dataset is used to evaluate the generalization capability of the classifiers. The predictions from the individual classifiers are combined to build the ensemble models and compared with the individual 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 with statistical measures. Table 3 shows confusion matrices of all models for training and test data partition. Table 4 shows the value of three statistical parameters [12] (sensitivity, specificity and total classification accuracy) of all models for training and testing dataset. Table 5 shows total number instance showing correct and wrong classification for testing and training dataset for all models.

Table 4: The value of statistical measures for all models with training and test dataset.

Measures %				
Model	Partition	Accuracy	Sensitivity	Specificity
C 5.0	Training	96.64	97.10	97.29
	Testing	94.52	96.25	91.66
CART	Training	96.64	97.10	94.59
	Testing	95.89	96.25	97.22
CHAID	Training	96.64	97.10	89.18
	Testing	93.15	96.25	97.22
QUEST	Training	92.62	91.30	81.08
	Testing	89.04	93.75	97.22
ANN	Training	97.99	97.10	97.29
	Testing	97.26	98.57	97.22
SVM	Training	91.95	98.55	86.25
	Testing	91.78	97.22	88.88
Ensemble	Training	97.32	97.10	97.29
	Testing	98.63	97.29	100

Table 5: Total number of correct and wrong classification of all models with accuracy for training and testing dataset.

Model	Cases	Training Data		Testing Data	
		Number of Instances	Accuracy (%)	Number of Instances	Accuracy (%)
C 5.0	Correct	144	96.64	69	94.52
	Wrong	5	3.36	4	5.48
CART	Correct	144	96.64	70	95.89
	Wrong	5	3.36	3	4.11
CHAID	Correct	144	96.64	68	93.15
	Wrong	5	3.36	5	6.85
QUEST	Correct	138	92.62	65	89.04

	Wrong	11	7.38	8	10.96
ANN	Correct	146	97.99	71	97.26
	Wrong	3	2.01	2	2.74
SVM	Correct	137	91.95	67	91.78
	Wrong	12	8.05	6	8.22
Ensemble	Correct	145	97.32	72	98.63
	Wrong	4	2.68	1	1.37

These above results show that the accuracy of ensemble mode is higher as compared to individual models.

6.1 Gain Chart

The gains chart [7] plots 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 [11] always start at 0% and end at 100% as we go from left to right. For more accurate model, the chart rises steeply towards 100% and level off. Models which are not more accurate follow the diagonal from lower left to upper right. The steeper the curve the higher is the gain. Fig.1 shows the gain chart of all models for the class Malignant and Fig.2 shows the gain chart of all models for the class Benign.

6.2 Response Chart

The response chart [14] plots the values in the Response (%) column of the table. The response is a percentage of records in the increment that are hits, using the following equation:
(Responses in increment / records in increment) x 100 … (5)

Response charts [10] 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 more accurate model, the line starts 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.3 shows the response chart of all models for the class Malignant and Fig.4 shows the gain chart of all models for the class Benign.

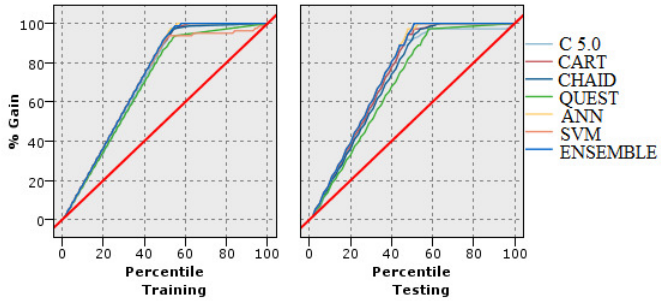

Fig.1: Gain chart of Ensemble model and individual Classifier for Malignant Class

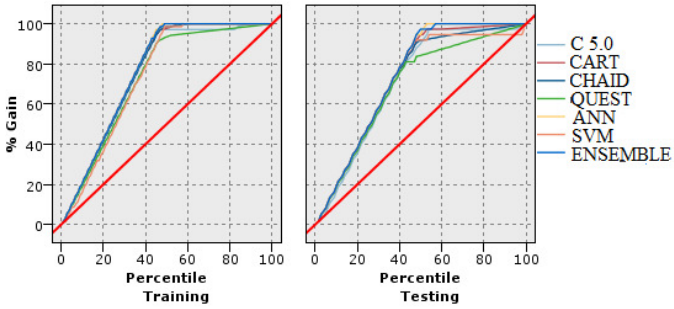

Fig.2: Gain chart of Ensemble model and individual Classifier for Benign Class.

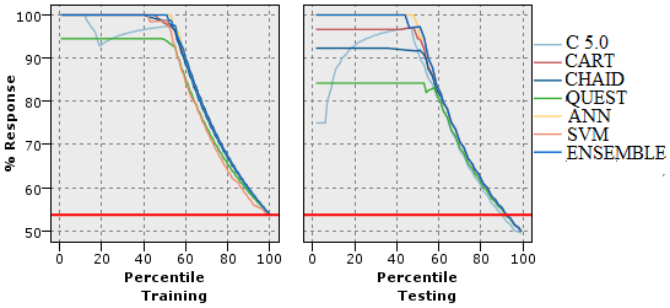

Fig.3: Response chart of Ensemble model and individual Classifier for Malignant Class

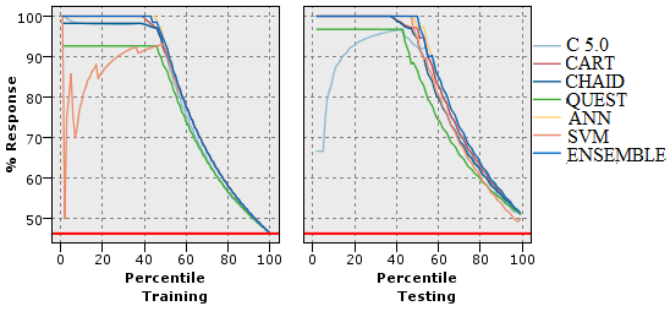

Fig.4. Response chart of Ensemble model and individual Classifier for Benign Class

7. Conclusion

The main goal of this study is to show the effectiveness of ensemble model. The performance of individual classifiers

C5.0, CART, CHAID, QUEST, ANN, SVM and ensemble models are analyzed on the cancer dataset. The performance of all models is investigated by using statistical performance measures like accuracy, specificity and sensitivity. The performance of each classifier is also investigated with the help of gain chart and response chart for both training and testing set. The accuracy of C5.0, CART, CHAID, QUEST, ANN and SVM is found be 94.52, 95.89, 93.15 ,89.04,97.26 and 91.78 respectively on test dataset. The accuracy of the ensemble model built using individual classifier is found to 98.63 on test data set. It is observed that performance of ensemble is higher than the individual models. Thus the proposed ensemble models can be a competitive technique for the classification of cancer dataset.

References

[1] Usama M. Fayyad. "Data mining and knowledge discovery: Making sense out of data". IEEE Expert: Intelligent Systems and Their Applications, 1996, Vol. 11(5), pp: 20–25.

[2] Jiwaei Han, Kamber Micheline, Jian"Pei, Data mining: Concepts and Techniques", Morgan Kaufmann Publishers (Mar 2006).

[3] Cabena, Hadjinian, Atadler, Verhees, Zansi "Discovering data mining from concept to implementation" International Technical Support Organization, Copyright IBM corporation 1998.

[4] S.Mitra, T. Acharya "Data Mining Multimedia, Soft computing and Bioinformatics", A john Willy & Sons, INC, Publication, 2004.

[5] Alaa M. Elsayad "Predicting the severity of breast masses with ensemble of Bayesian classifiers" journal of computer science, 2010, Vol. 6 (5), pp: 576-584.

[6] Alaa M. Elsayad, "Diagnosis of Erythemato-Squamous diseases using ensemble of data mining methods", ICGST-BIME Journal Volume 10, Issue 1, December 2010.

[7] SPSS Clementine help file. http//www.spss.com

[8] UCI Machine Learning Repository of machine learning databases. University of California, School of Information and Computer Science, Irvine. C.A. http://www.ics.uci.edu/~mlram,?ML.Repositary.html

[9] Michael J. A. Berry, Gordon Linoff, "Data Mining Techniques ", John Wiley and Sons, Inc.

[10] Hota H.S, Pujari P; "A comparative study of Decision tree based data mining algorithm and its ensemble method for classification of data" ,Proceeding of international conference on Emerging trends in soft computing and ICT (SCICT-2011) ,(pp:41-44),Organized by Dept of CSIT,GGV ,Bilaspur, India on 16-17 March

[11] Jozef Zurada and Subash Lonial "Comparison of The Performance of Several Data Mining methods for Bed Debt Recovery in The Health Care Industry".

[12] Matthew N Anyanwu & Sajjan G Shiva "Comparative Analysis of serial Decision Trees Classification Algorithms",(IJCSS), Volume (3) : Issue (3).

[13] Mahesh Pal, "Ensemble Learning With Decision Tree for Remote Sensing Classification", World Academy of Science, Engineering and Technology 36 , 2007.

[14] Kelly H. Zou, PhD; A. James O'Malley, PhD; Laura Mauri, MD, MSc "ROC Analysis for Evaluating Diagnostic Test and Predictive Models".

[15] "Ensemble Data Mining Model for Classification of Pima Indian Diabetes Data set", Proceeding of International Joint Conference on Advance Engineering & Technology, ISBN: 978-93-81693-88-22, Raipur, (PP: 16-22)9th-10th, April 2013.

[16] Pujari P, Gupta J.B; "Estimation and Comparison of Classification Models by Using Numeric Predictor on Iris Data Set (AICON-13)", All India Conference on "Global Innovations in Computer Science and Engineering and Information Technology", Organized by CSIT, Durg, (C.G), on April 12-13, 2013., (pp. 1.30-1.39) ISBN: 978-81-923288-1-2.

[17] Sharma D.K, Hota H.S, Pujari P., "Neural network, support vector machine and its ensemble model for prediction of different categories of dermatology data set", Proceedings of Academy of Information and Management Sciences, Volume 16, Number 1, Allied Academies International Conference New Orleans, Louisiana,4-6 April 2012.

[18] Maria-Luiza Antonie, Osmar R. Zaıane, Alexandru Coma, .Application of Data Mining Techniques for Medical Image Classification. Proceeding of second International workshop on Multimedia data mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference.SAN FRANCISCO, USA, AUG 26, 2001.

[19] Sujatha, Dr.K.Usha Rani, "Evaluation of Decision Tree Classifiers on Tumor Data sets",IJETTCS,Vol2,Issue4,July-aug2013,pp.418-423.

[20] Aruna, Dr S.P. Rajagopalan and L.V. Nandakishore, 2011 Knowledge Based Analysis Of Various Statistical Tools In Detecting Breast Cancer.

[21] Tan,Gilbert, " Ensembling machine learning on gene expression data for cancer classification", Proceedings of New Zealand Bioinformatics Conference, Te Papa, Wellington, New Zealand, 13-14 February 2003.

[22] D.Lavanya and Dr.K.Usha Rani, "Ensemble Decision Tree Classifiers for Brest Cancer Data", International Journal of Information Technology Convergence and Services, Feb 2012 Vol.2, No.1, pp.17-24

[23] Delen Dursun, Walker Glenn and Kadam Amit, "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial Intelligence in Medicine, June 2005, vol. 34, Pg. no: 113-127.

[24] Shomona Gracia Jacob, Dr.R.Geetha Ramani, P.Nancy (2011 b), "Feature Selection and Classification in Breast Cancer Datasets through Data Mining Algorithms", Proceedings of the IEEE International Conference on Computational Intelligence and

Computing Research (ICCIC'2011), Kanyakumari, India,, IEEE Catalog Number: CFP1120J-PRT, ISBN: 978-1-61284-766-5. pp. 661-667

[25] Eva Volna and Martin Kotyrba, "Enhanced ensemble-based classifier with boosting for pattern recognition," *Applied mathematics and computations*, October 2017, vol. 310, pp. 1-14.

[26] Ankit, Nabizath Saleena, "An Ensemble Classification System for Twitter Sentiment Analysis", Science Volume, 2018, pp.937-946.

First Author: Pushpalata Pujari, is working an Assistant Professor and Head in the CSIT department, Guru Ghasidas Vishwavidyalaya, Central University, Bilaspur, Chhattisgarh, India. She has received her M.C.A Degree from Berhampur University, Berhampur, Odisha, India in 1998. She has received her Ph.D degree from the department of Computer Science and Information Technology, Guru Ghasidas Vishwavidyalaya, Central University, Bilaspur, Chhattisgarh, India. Her areas of interest include Character Recognition, Pattern Recognition, Soft Computing, Evolutionary Computing and Data Mining