A Comparative Result Analysis of Human Cancer Diagnosis using Ensemble Classification Methods

Jogendra Singh Kushwah
Department of Computer Science and Engineering
BUIT,Barkattuallah University
Bhopal, India

Divakar Singh
Department of Computer Science and Engineering
BUIT,Barkatullah University
Bhopal,India

ABSTRACT
Cancer research has been an interesting and challenging research area in the field of medical science. Classification techniques have been found useful in early diagnosis of cancer and better treatment. For diagnosis of cancer various classification methods are used but they suffer with one or more disadvantages. In this paper ensemble based classification methods which combine the prediction of individual classifiers to generate the final prediction are discussed. The methods discussed are Bagging, Boosting and Random Forest Algorithm. These ensemble methods have shown improvement in quality of result as compared to commonly used single classifier e.g. decision tree or neural network .The improvement in classification is however at the cost extra processing time and higher storage as decision tree or neural network are faster as compared to ensemble based techniques. The ideas for further improvement in this field are also discussed in this paper. Methods discussed in the paper are applied on human cancer data for appropriate cancer gene selection which leads to classification of cancer.

Keywords
Bagging, Boosting, Cancer Classification, Ensemble Classification Methods, Random Forest

1. INTRODUCTION AND MOTIVATION
Supervised learning algorithms are commonly described as performing the task of searching through a hypothesis space to fine a suitable hypothesis that will make good predictions with a particular problem. Even if the hypothesis space contains hypothesis that very well suited for a particular problem, it may be difficult to find a good one. Ensembles combine multiple hypothesis to form a better hypothesis. In other word, an ensemble is a technique for combining many weak learners in an attempt to produce a strong learner. The term ensemble is usually reserved for methods that generate multiple hypotheses using same base learner. Evaluating the prediction of an ensemble typically requires more computation than evaluating the prediction of a single model, so ensemble may be thought of as a way to compensate for poor learning algorithms by performing a lot of extra computation. For example fast algorithms such as decision tree sometime have relatively poor accuracy compared to other knowledge models like neural networks. In order to overcome this problem, a large number of decision trees are generated for the same data set, and used simultaneously for prediction. Random forest [6] is one such ensemble based method which is commonly used with decision trees.

There are often two main criticism of ensemble based classification research; the dearth of publicly available real data to perform the experiments on; and the lack of published well researched methods and techniques. To counter both of them, this paper garners all related literature for categorization ad comparison, selects some innovative methods and techniques for discussion; and point towards other data sources as possible alternatives.

There are basically two motivations behind building a ensemble of classifier.

1. Reduced variance: Results are less dependent on the peculiarities of a single training set.

2. Reduced bias: A combination of multiple classifiers may learn a more expressive concept class than a single classifier.

2. ENSEMBLE CLASSIFIERS
A classifier e is function that maps a vector of attribute value x (also called instance/example) to target classes c e {c1, c2, c3...cl}. An ensemble classifier consist of set of classifiers H= {C1, C2, C3...Cn} whose output is dependent on the output of the constituents classifiers (component classifiers). Further the reliability of a classifier e is denoted by re and in this study a estimate of the classification accuracy (or recognition rate) is the percentage of example that are classified. A typical structure of ensemble classifier is shown in fig.1. A diverse ensemble can be created by assigning each component classifier a different training set which is usually derived from original training set by re-sampling or other technique.

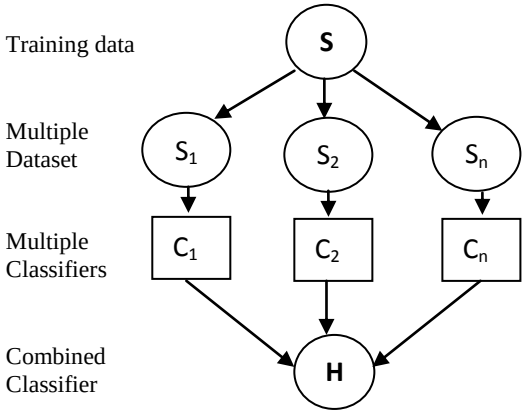

Fig.1: The structure of ensemble classifier.

3. WHY DO ENSEMBLE CLASSIFIERS WORK?

The uncorrelated errors of individual classifiers can be eliminated by averaging their outcome. Let us assume there are 25 base classifiers, each with the same error p=0.35. The probability that an ensemble classifier makes a wrong decision:

$$\sum_{i=0}^{25} \binom{25}{i} p^{i}=(1-p)^{25-i} = 0.06$$

It is very clear from the above result that classification error of ensemble classifier is comparatively low as compared to individual classifier, thus ensemble classifier always has a better accuracy in classification as compared to individual classifier.

4. METHODS OF CONSTRUCTING ENSEMBLE CLASSIFIERS

Several methods of constructing and combining an ensemble of classifiers have been proposed to improve the accuracy of learning algorithm. The most commonly studied methods which have drawn increased interest of researchers are:

1. Bagging
2. Boosting
3. Random Forest

Effectiveness of these methods especially, Bagging [5] and Boosting [7][5] has been demonstrated empirically, Breiman[1] has shown that Bagging can increase accuracy of CART decision trees on several real and artificial domain. Bauar and Kohavi [9] demonstrated the capability of Bagging and Boosting to improve C4.5 decision trees, based on a large number of data sets.

4.1 Bagging

Bagging is technique which combines predictions of independent base classifiers for arriving at final prediction. Given some database of training data, user can take t samples from this database with replacement. Using samples taken from the training example database, the underlying machine learning algorithm can be trained independently on each of these datasets. After the training has completed, user is left with Ct classifiers. When presented with some unknown example and prediction is made on it by using each of the Ct classifiers. The final prediction is made by selecting the most common prediction from each of the classifier's Ct. The finial classification of the test example made from the target classifiers is called a voting scheme where the prediction of each target classifier is a "vote" towards the final prediction. Bagging works because if a learning algorithm (i.e. decision tress) is unstable a small change in training set causes large changes in the learned classifier and Bagging always improves performance. A pictorial view of bagging is shown in Fig.2.

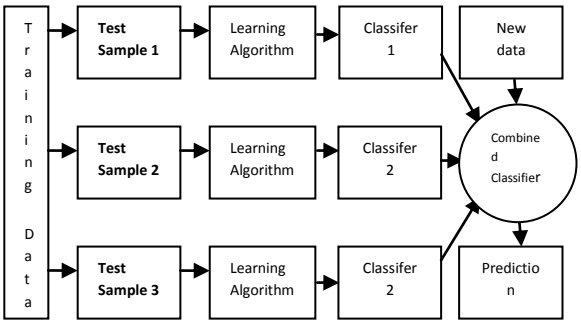

Fig.2 Bagging.

4.1.1 Bagging Algorithm

Let Si the given data set, at each iteration i, a training set Si is sampled with replacement from S (i.e. bootstrapping) and a classifier Ci is learned for each Si.

1. for i = 1 to m // m ... number of iterations

a) draw (with replacement) a bootstrap sample Si of the data

b) learn a classifier Ci from Si

2. for each test example

a) try all classifiers Ci

b) predict the class that receives the highest number of votes

4.2 Boosting

Boosting is a technique for combining multiple base classifiers whose combined performance is significantly better than that of any of the base classifiers. Sequential training of weak learners Each base classifier is trained on data that is weighted based on the performance of the previous classifiers e.g. after a classifier Ci is learned, the weights are adjusted to allow the subsequent classifier Ci+1 to "pay more attention" to the tuples that were misclassified by Mi. Finally each classifier votes to obtain a final outcome. In this technique multiple models are developed in sequence by assigning higher weights (boosting) for those which are difficult to classify. One of the major problem come across in implementing this concept of boosting is improper distribution of data and method may require a large training data set. The basic idea behind boosting is sown in fig.3. One of the popular boosting algorithm is Fruend and Schapire's AdaBoostM1 which implements the basic idea of boosting for more than two classes [7].

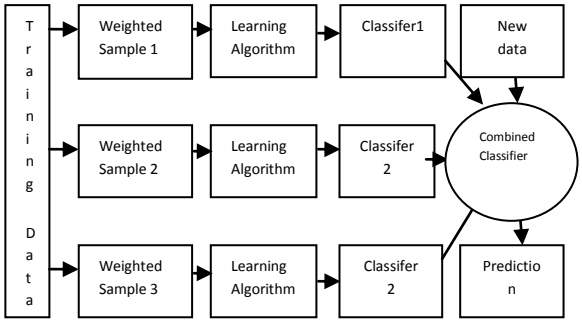

Fig.3: Boosting

4.2.1 AdaBoostM1 Algorithm

1. Set w [1, i] =1/N for case i=1, 2,….N
2. For each t=1,2…….,T
   - Find classifier C(t) using w[t,*]
   - Set E[t] =sum/i{w[t,i]|C[t] misclassifies case i}
   - If E[t]=0, stop
   - If E[t]=0.5 ,set T=t+1 and stop
   - Otherwise ,set w[t+1,i]=
     w [t,i]/2 E[t] (if C[t] misclassifies case i)
     otherwise set to w[t,i]/2(1-E[t]
2. To classify a case:
   - chose class k to maximize
     sum/t {log((1-E[t])/E[t])|C[t] predicts class k}

4.3 Random Forest

A random forest is an ensemble (i.e., a collection) of unpruned decision trees. The algorithm was developed by Leo Breiman and Adele Cutler and "Random Forest" is their trademark. The main principle behind ensemble methods is that a group of "weak learners" can come together to form a "strong learner". Random forests are often used when very large training datasets are given along with very large number of input variables (hundreds or even thousands of input variables). A random forest model is a classifier that consists of many decision trees and outputs the class that is the node of the class output by individual trees [6]. The visualization of random forest is sown in fig.3.

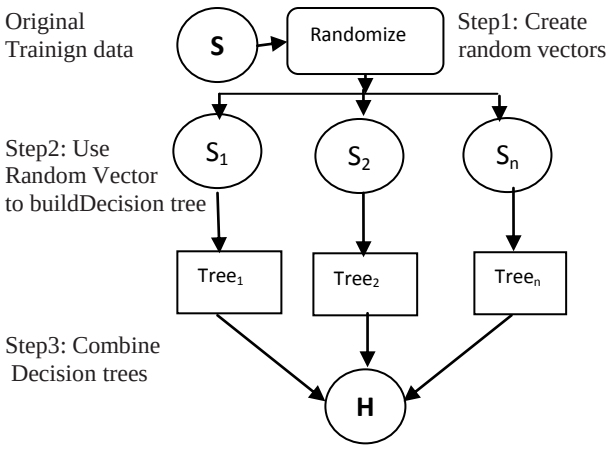

Fig.4: Random Forest

4.3.1.1 Random Forest Algorithm

Given a training set S

For i=1 to n do:

Build subset Si by sampling with replacement from S

Learn tree Ti from Si

At each node:

Choose best split from random subset of F features.

Each tree grows to the largest extend, and no pruning

Make predictions according to majority vote of the set of n trees.

4.3.1.2 Advantages of Random Forest Method

.

- For many data sets, it produces a highly accurate classifier.
- It handles a very large number of input variables.
- It estimates the importance of variables in determining classification.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.
- It includes a good method for estimating missing data and maintains accuracy when proportion of the data is missing.
- It can balance error in the class population of unbalanced data sets.

5. EXPERIMENT AND RESULTS

Performance evolution of the discussed methods has been made by using breast cancer dataset from UCI Machine Learning Repository [22]. This dataset is commonly used among researchers working with machine learning methods for cancer diagnosis. The dataset contains 286 examples, 10 attributes and two classes: (a) non-reoccurrence-events and (b) reoccurrence-events. These data has been tested with Weka 3.6.9 machine learning software [21]. The empirical evaluations of method discussed in this paper are summarized as follows:

Table 1: Comparative result analysis of Bagging, Boosting and Random Forest Algorithm

	Parameter	Bagging	Boosting (Adaboost M1)	Random Forest
1	Correct classification	197	198	199
2	Incorrect Classification	89	88	87
3	Root Mean Sqr. error	0.45	0.46	0.45
4	Accuracy percentage	68.88	69.23	69.58

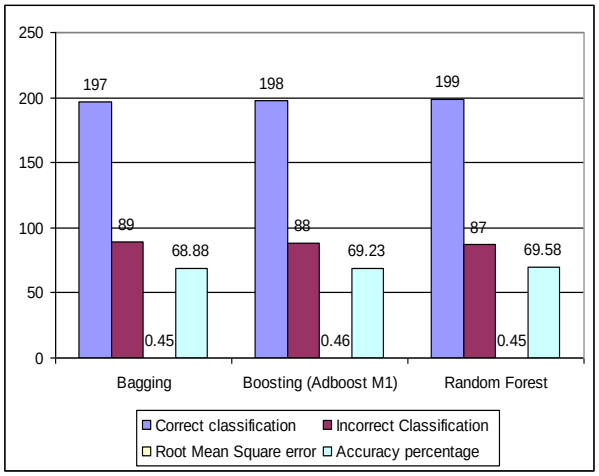

Fig.5: Comparative result analysis of Bagging, Boosting and Random Forest Algorithms

6. CONCLUSION AND FUTURE WORK

The discussed goal in class prediction in human cancer treatment is a precise classification of cancerous malignancies at an early stage, allowing for directed and more successful therapies. Important for this task are classification algorithms that can deal with the high dimensionality of gene expression data, and that exploit as much of the available information as possible. Ensemble classifier play an important role in that as the individual classifier may not be powerful enough to classify all the data or the classification may be biased toward particular features of data set. The three popular ensemble methods, Bagging, Boosting and Random Forest have been discussed here. By evaluating performance of all the discussed methods on UCI cancer data set, it has been shown that random forest lowers the misclassification error as compared to boosting and bagging. Bagging is technique which combines output from decision of models generated from bootstrap samples of training dataset while boosting is an iterative process of weighting more heavily cases classified incorrectly by the classifier model and then combining all the models generated during the process.

Although ensemble based classifier are slow as compared to fast learner like decisions trees, experiments have shown that the prediction accuracy is significantly increased for ensemble classifiers, and random forest has shown better performance among discussed ensemble classifiers. Considering the relative simplicity in regards to implementation and the predictive power improvements, bagging and boosting provide an excellent means of improving performance to an existing machine learning algorithm implementation. Bagging and boosting have the opportunity to both increase and decrease the error on example predictions on case to case basis, while random forest has capacity to deal with various types and large datasets. Ensemble classifiers have better performance than individual classifier by combining individual learning models, however determining which individual models best combination from training result is difficult. Simply selecting the best individual models do not necessarily lead to an improvement in result.. Neural networks can be used for approximation of final prediction to be made by and ensemble classifier. Radial Bias function (RBF) network is one of the options for this purpose which used Gaussian function for approximation based on training data. Further ensemble methods may be cascaded with neural networks for further improvement in results.

7. REFERENCES

[1] L.Breiman ,Bagging predictors, Machine Learning, 24(2):123-140, 1996.

[2] Henrik Bostrom,Ronnie Johansson and Alexadder Karlsson, On evidential combination rules for ensemble classifiers, Modelling and Simulation, 18(2), 112-116, 1998.

[3] T.K.Ho, The random subspace method of constructing decision forest , IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8):832-844, 1998.

[4] M Ozaki, Y. Adachi, Y. Iwahori, and N. Ishii, Application of fuzzy theory to writer recognition of Chinese characters, International Journal of Modelling and Simulation, 18(2), 112-116, 1998.

[5] E.Baur and R.Kohavi,An empirical comparison of voting classification algorithms: Bagging, boosting and variants, Machine earning, 36(1-2):105-139, 1999.

[6] L. Brieman, Random Forests, Machine Learning, vol.5, 2001.

[7] Robert E. Schapire, The Boosting Approach to Machine Learning An Overview, Nonlinear Estimation and Classification,Springer, 2003

[8] Tuba Kiyan,Tuba Yildirim, Breast Cancer Diagnosis using statistical neural networks, Journel of Electrical and Eelectronics Engineering, Istanbul Universit, 1149-1153, 2004.

[9] R.J. Roiger,M.W, Data Ming : A Tutorial-Based Primer, Addison Wesly, 2004

[10] I.H. Witten. E.Frank, Data Mining, 2nd ed, Morgan Kaufmann, 2005.

[11] Juan J. Rodriguez,Jesus M.Maudes, Carlos J. Alonso, Rotation-based Ensembles of RBF networks, European Symposium on Artificial Neural Networks, 2-930307-06-4, 2006.

[12] Feng Chu and Lipo Wang ,Applying RBF neural networks to cancer classification based on gene expression Intentional Joint Conference on Neural Networks, Vancouver,BC,Canada, 2006.

[13] Hyontai Sug, International Journal of Mathematics and Computers in Sumulation, 2010.

[14] G.Sophia Reena,P.Rajeswarei, A survey of human cancer classification using micro array data, International Journal o Computer technology and Applications), 1523-1533, 2011.

[15] Kobalan Moorthy and Mohd Saberi Mohammad, Random forest for gee selection and microarray data classification,Bioinformation, 0973-2063, 2011.

[16] Dr A. Chirta and S Uma, An Ensemble Model of Multiple Classifiers, International Journal of Computer Theory and Engineering, 1793-8201,Jun-2012.

[17] Evathia E. Tripoliti,Dimitrios I.,Fotiadis,George Manis, Automated diagnosis of diseases based on classification : dynamic determination of the number of tress in random forest algorithm, IEEE Transactions of Information Technology in Biomedicine, Jul-2012

[18] Eun-Hye Jang,Byoung-Jun Park,Sang-Hyeob Kim and Jin-Hun Sohn, Emotion classification based on physiological signals induced by negative emotions, IEEE, 978-1-4673-0390-3/12, 2012.

[19] Gouda I. Salama, M.B.Abdelhalim, and Magdy Abd-elghany Zeid, Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers, International Journal of Computer and Information Technology, (2277 – 0764), Sep-2012.

[20] Cuong Nguyen,Yong Wang,, Ha Nam Nguyen, Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic , Journal of Biomedical Science and Enginnerring, 6, 551-560, 2013.

[21] WEKA, The University of Waikato, www.cs.waikato.ac.nz/ml/weka

[22] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/.