A Survey on Ensemble Methods for High Dimensional Data Classification in Biomedicine Field

Shweta B. Meshram
Student, Computer Engg.,
JSCOE, Pune, India

Sharmila M. Shinde
Computer Engineering Dept.
JSCOE, Pune, India.

ABSTRACT
The volume of the information is growing exceptionally large. So, there is growing interest to help people to categorize, handle and control these resources. In last few years, the data mining is applied widely to discover the knowledge from information system. Classification is one of the tools which are used for data mining. Ensemble methods proved to be superior to individual classification method for high dimensional datasets. Hence, this paper surveys many ensemble methods along with bagging, boosting and random forest. It gives the idea about the earlier proposed categories of ensemble methods. It also discusses about the performance of ensemble methods.

Keywords
Classification, Ensemble method, High dimensional data.

1. INTRODUCTION

The information is increasing continuously with the growth of internet, with organizations and with evolving research. Many researchers are taking interest to provide some resources that can handle such huge information. So that new knowledge can be discovered from datasets. Classification algorithm is presented with a set of records, where each record is described with a predefined number of features. These features are described with a class and that label also represents its target. The accuracy of the classification surely degrades if we directly perform the classification task without preprocessing the data. The accuracy is affected by imbalance dataset, high noise, missing values, redundant data. We can notice that classification has wide applications in biomedicine field, image processing. So, it motivates us to develop learning method to classify high-dimensional data in ensemble way.

Growing problem of data dimensionality makes a various challenges for supervised learning. Generally used classification methods are naïve bayes, decision tree, k nearest neighbor, neural network and support vector machines were difficult to be directly applied on high-dimensional datasets [1]. In general, the ensemble classifiers provide better classification accuracy than individual classifier produced. Hence ensemble classifier methods are used to achieve better accuracy. Individual classification can not handle the noise and imbalance data in the high dimensional datasets. There are many ensemble methods proposed till now but they are not much accurate on bioinformatics datasets. So, there is a need to improve the ensemble methods in terms of learning performance, scalability, training and testing time.

The paper consists of two parts. The first part describe the different ensemble methods in brief while second part gives summary of the describe methods.

2. LITERATURE SURVEY

2.1 About Ensemble Method

Ensemble methods, also known as classifier combiner. It generates a group of base classifiers from learning data. Lastly classification is performed by combining the results of each base classifier. However, the scalability and learning performance of classifier are greatly affected by increasing dimensionality of datasets. The classification ability of a single classifier is limited [1]. To improve the accuracy of ensemble methods, each base classifier of the method should be distinct and independent. Accuracy of the ensemble classifier is mainly depend on the two main factors i.e., diversity and independency of features.

The methods for generating ensemble classifier can be broadly categorized into four groups [2] that are based on i) manipulation of the training set, ii) manipulation of the input features, iii) manipulation of the class labels, iv) manipulation of the learning algorithm. But methods are classified into five groups when homogeneous base classifiers are used. These groups are (i) manipulation of the training parameters, (ii) manipulation of the error function, (iii) manipulation of the feature space, (iv) manipulation of the output labels, and (v) manipulation of the training patterns [2].

All these methods aim to achieve diversity among the base classifiers. Diversity can be achieved by manipulating the training parameters of the base classifiers in an ensemble. In another group of ensemble classifiers diversity among the base classifiers is achieved by manipulating the input feature space. The random subspace ensemble classifiers perform relatively inferior to other ensemble classifiers. The largest set of ensembles generates ensemble classifiers by manipulating the training patterns where the base classifiers are trained on different subsets of the training patterns.

Mohammad Ali Bagheri, Qigang Gao, Sergio Escalera presented a combing structure for multiple classifier systems that conceptually merges a large variety of ensemble classification methods [3]. They propose four general approaches of ensemble strategies. Three approaches are based on manipulation of training data. Three approaches are subsample approach, subspace method, subclass approach. Fourth approach is based on manipulation of learner. The combination based approach combines the above four different approaches i.e., combination of subsample and subspace approach, combination of subsample and subclass approach, combination of subspace and subclass approach, combination of data manipulation approach and learner manipulation. In subsample approach, different learning samples are used to build the base classifiers and each subset is drawn from dataset. Hence, each subset has the same features as original data. In subspace method, each classifier with original dataset is trained using different feature subsets. It is most beneficial for the high dimensional datasets. In

subclass approach, classifiers are trained with samples which belong to different target classes. Learner manipulation approach can be achieved by using different algorithms for the base classification or by using different parameters of the base classifiers.

2.2 Examples of Ensemble Methods

There are many proposed ensemble methods that deals with high dimensional data. High dimensional data are the data that are characterized by few dozen to many thousands of dimensions i.e. nothing but features. Most of ecological data, data on health status of patients, movie rating data, climate data, bioinformatics data are examples of the high dimensional data.

Bagging is one of the ensemble methods. It is also called as bootstrap aggregation [4]. Each classifier returns its class guess. In bagging, each prediction is considered as one vote. Maximum votes to class are considered for final classification. Bagging performs the aggregation of the votes and considers only maximum number of votes. Good classifiers are transformed into optimal one in this method. The best bagged predictor is always best subset predictor. It is preferable with trees. Also it is simple and having interpretable structure. Bagging is more robust to noisy data. It can handle the unstable procedures more efficiently. Also it is simple and having interpretable structure. The performance of the bagging degrades when it deals with the stable procedures.

Boosting is another ensemble method. In boosting, weights are assigned to each training tuple and these tuples are learned iteratively using classifiers. The weights of tuples are updated in each iteration. If the tuple is classified correctly then weights of the tuples are decreased. The weight of the misclassified tuple is increased, so that more attention can be paid to it. The error rate of the model is calculated to update the weights of correctly classified tuples. There was a need to improve the performance of boosting because of the two reasons [5]. They are as i) it generates an assumption such that it can merge training sets with large error and results in those with small error ii) divergence reduction. Hence, two versions of boosting algorithms are proposed i.e., AdaBoost.M1 and AdaBoost.M2. In AdaBoost.M1, it calls the algorithm that minimizes the error during learning. If the tuple is not classified correctly then it doesn't change its weights. It reduces the error to zero if any hypothesis has constantly error slightly more than ½. Hence AdaBoost.M1 is unable to handle weak hypothesis with error greater than 1/2. AdaBoost.M2 performs better than AdaBoost.M1 because it considers a set of probable labels instead of any one feature. Pseudo loss is used to measure the error in Adaboost.M2. It is calculated using distribution of set of all pairs of samples and misclassified labels. Hence it can pay more attention to the misclassified labels. The minimized pseudo loss is required for accurate classification. The performance of the boosting for large datasets can be improved if it is used with complex algorithm. It also minimizes the error of classifier. But it is highly sensitive to noise and outlier.

Breiman proposed the "random forest". It constructs many decision trees for classification using input vector [6]. The decision tree with the maximum votes is used for final classification task. The random forest improves the performance significantly when it is compared with single decision tree. Random forests are simple, work very fast and most accurate classifier. It is difficult to analyze. It also overfits data that are noisy. Random forest can not predict data beyond the range of training data.

Y. Piao, H. W. Park, C. H. Ji, K. H. Ryu proposed the ensemble method that uses the Fast Correlation- Based Filter method (FCBF) to generate multiple feature subsets based on the correlations among features and learns the model from each feature subset using support vector machine as base classifier. The results of each classifier are combined by majority voting [1]. In the first step, different sets of features are created by partitioning redundant features to achieve the diversity. The Classifier is trained on the diverse subsets instead of choosing a best subset. This approach shows the good accuracy when compared with bagging, boosting and random forest. It has the highest level of dimensionality reduction. It can handle the data of different feature type [11]. But FCBF is not much accurate method for feature selection. It may create instable behavior of predictive algorithms.

S.B. Cho, H. Won proposed ensemble of neural networks for cancer classification with multiple significant gene subsets [7]. It consists of three steps i.e., significant gene subset selection, neural Network used as base classifier and bayesian approach to combine the results. In the first step, gene vector is constructed for each class. This vector distinguishes between class j and other classes using similarity measure. Depending on the similarity with representative gene vector, the significant gene subsets are selected. The maximum similarity is considered for constructing significant gene subsets. Three layered perceptrons are used for learning this model. Final classification is based on probability calculated of the results of base classifiers. It maintains originality of the features. But the performance of classification doesn't improve much.

H. I. Elshazly, A.M. Elkorany, A. E.Hassanien proposed ensemble method for prostate cancer diagnosis [8]. In this method each decision tree is trained on rotated feature space. Rotation forest algorithm is proposed here. Bootstrap samples are extracted from original datasets. New subsets are projected into new feature space with principal component analysis. These new subsets are trained using decision tree. It achieves good accuracy on prostate cancer. But accuracy achieved with loss of originality. It leads to overfitting of data i.e., accuracy of the method increases with the increase in number of selected features.

Hualong Yu and Jun Ni proposed ensemble learning method for imbalanced biomedicine data [9].

It addresses the imbalance classification. The method proposed the feature subspace. It combines clustering to remove redundant features and feature selection to remove noisy data. It first selects the groups of similar features. Features are selected using feature selection method using signal to noise ratio. It is also called feature space. The features are randomly projected into subsets. Asymmetric bagging is applied on these subsets i.e., bootstrap executed on majority class samples and random undersample minority class samples. It gives training subsets which are trained using support vector machine as the base classifier. Lastly the majority voting is used for final classification. It increases accuracy. But this method does not suit for low dimensional tasks. The complexity is again higher than its predecessor.

3. SUMMARY

This paper discusses seven ensemble methods for classification of high dimensional data in biomedicine area.

The table below gives the limitation of each method and also presents datasets used in each method.

Ensemble Method	Datasets used	Limitation
Bagging	Waveform, heart, breast cancer, ionosphere, diabetes, soybean, glass	Its performance degrades when deals with stable procedure.
Boosting	soybean-small, labor, promoters, iris, hepatitis, sonar, glass, audiology.stand, cleve, soybean-large, ionosphere, house-votes, votes1, crx, breast-cancer, pima-indians-di, vehicle, vowel, german, segmentation, hypothyroid, sick-euthyroid, splice, kr-vs-kp, satimage, agaricus-lepiot, letter-recognition	It is highly sensitive to noise and outlier
Random forest	Glass, breast cancer, diabetes, sonar, vowel, ionosphere, vehicle, soybean, German credit, image, ecoli, votes, liver, letters, sat-images, zip-code, Waveform, twonorm, threenorm, Ringnorm	It is difficult to analyze. It overfits data that are noisy. It can not predict beyond the range of training data.
Y. Piao, H. W. Park, C. H. Ji, K. H. Ryu	Leukemia, Prostate cancer.	The FCBF is not much accurate. It may create instable behavior of predictive algorithms.
S.B. Cho, H. Won	Leukemia, Colon, Lymphoma	The performance of the classification doesn't improve much
H. I. Elshazly, A.M. Elkorany, A. E. Hassanien	Prostate cancer.	Features are transformed into new ones hence loss of originality. It leads to overfitting of data.
Hualong Yu and Jun Ni	Colon, Lung, Ovarian I, Ovarian II.	It does not suit for low dimensional tasks.

4. CONCLUSION

This paper surveys the different ensemble methods based on different feature selection criteria with same base classifier. But they are not much accurate. We have seen the advantages and limitations of the some of ensemble methods. It is not uniform when it is applied on different datasets. The performance of these methods is varying because of the characteristics of the features in the datasets and approach of subset generation with classifier. For the classification task, one has to decide the priority between accuracy and efficiency. This work will help to analyze the existing methods and to propose the new ensemble method. It will also help to overcome limitation of existing methods.

5. REFERENCES

[1] Yongjun Piao, Hyun Woo Park, Cheng Hao Ji, Keun Ho Ryu,"Ensemble Method for Classification of High-Dimensional Data", 978-1-4799-3919-0/14/ IEEE Big Comp.

[2] A. Rahman, B. Verma,"Ensemble Classifier Generation using Non uniform Layered Clustering and Genetic Algorithm, Knowledge-Based System, 2013, in press.

[3] Mohammad Ali Bagheri, Qigang Gao, Sergio Escalera ,"A Framework towards the Unification of Ensemble Classification Methods", 2013 12th International Conference on Machine Learning and Applications.

[4] L. Breiman,"Bagging predictors", Mach. Learning. 24, 1996, pp.123- 140.

[5] Y. Freund, R.E. Schapire,"Experiments with a new boosting algorithm", International Conference on Machine Learning, 1996, pp.148-156.

[6] Leo Breiman," Random Forests", Statistics Department, University of California, Berkeley, CA 94720.

[7] Sung-Bae Cho Hong-Hee Won,"Cancer classification using ensemble of neural networks with multiple significant gene subsets", Published online: 12 November 2006 Springer Science and Business Media, LLC 2007.

[8] Hanaa Ismail Elshazly, Abeer Mohamed Elkorany, Aboul Ella Hassanien,"Ensemble-based classifiers for prostate cancer diagnosis", Scientific Research Group in Egypt (SRGE).

[9] Hualong Yu and Jun Ni,"An Improved Ensemble Learning Method for Classifying High-Dimensional and Imbalanced Biomedicine Data", IEEE/ACM transactions on computational biology and bioinformatics vol.11, no.04, July/August 2014.