Volume 3, Issue 1, January 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review of Ensemble Technique for Improving Majority Voting for Classifier

Sarwesh Site
M.Tech Scholer LNCT Bhopal
India

Dr. Sadhna K. Mishra
Prof LNCT Bhopal
India

Abstract— ***Data classification plays important role in the field of data mining. The increasing rate of data diversity and size decrease the performance and efficiency of classifier. The decreasing performance of classifier compromised with unvoted data of classifier. Now the merging of two or more classifier for better prediction and voting of data are used, such techniques are called Ensemble classifier. Initially the resembling of classifier used bogging and boosting technique and later on used random Forest technique. The process of classifier improved the performance and efficiency of data classification. But feature selection process of ensemble technique has important part of classifier. In this paper we present various technique of ensemble classifier for binary classification as well as multi-class classification.***

Keywords*—* ***ensemble classifier, Bogging, boosting, Random forest, majority Voting***

I. Introduction

Machine learning has many applications and is used most significantly in data mining. People often made mistakes when analyse data or try to outline some relationship among multiple features of data. These mistakes make difficult to produce the correct output. Machine learning is applied to these problems to improve the efficiency of system and design of machine[1]. When instances with known label are given the learning is called supervised learning and if instances are unlabeled the learning is called unsupervised learning. But unsupervised learning provides useful classes of items which is called clusters. Clusters are groups of similar types of objects. Theses groups are formed with classification methods[3]. These classifications are done by classifiers. But when an object need to be classified into predefined group or class on the basis of number of observed attributes related to that object a classification problem is occurred. Another type of learning is reinforcement learning where information's are provide by the environment in the form of scalar reinforcement signal which constitutes a measurement of system operation i.e. how well system is operating. Meta-learning uses set of attributes called meta-attributes to represent the characteristics of learning tasks[4,5]. So it is not a good way to utilize one method or algorithm to solve a particular problem because every algorithm has strength with some limitations. So the best idea is use strengths of one method over the limitations of another algorithm. So techniques of applying algorithms in such way are called ensemble of classifiers. COB (core, outlier, and boundary) method quantitatively measures the accuracies of majority voting ensembles for binary classification.[8]. Good ensemble methods are that in which each individual classifiers are accurate and diverse. But ensemble methods are combination of predictions made by a set of individual classifiers. Accurate classifier is meant to be produce accurate prediction than the random classifier and Diverse classifier is meant to be produce prediction independently. For experimental purpose of COB three different ensemble methods bagging, random forests, and a randomized ensemble, two different numbers of individual classifiers and three different machine learning algorithms decision trees, k-nearest neighbours, and support vector machines are used. The COB model results that the accuracy of ensemble method is worse with the present of nonempty core subset than the accuracy of the binomial method. The COB model is an enhancement to the binomial model with addition of two subsets core and outlier. The majority votes are decomposed into three terms an average individual accuracy, good diversity and bad diversity[9.10]. Diversity can be defined as a consequence of two decisions (1) the choice of error function and (2) the choice of combiner function in the design of ensemble problem in the machine learning. While illustrating the majority votes two special case pattern of success and pattern of failure were introduced in this paper. Probability distributions over all possible combinations of correct/incorrect votes are defined to improve the individual accuracy p. Each combination where exactly (T+1)/2votes are correct, appears with probability α. In this pattern no votes are wasted[11,12]. The above section discuss introduction of stream data classification. In section II we discuss various proposed method for stream data classification. In section III conclude the paper.

II. Method for Ensemble classification

In this section we discuss method for ensemble classifier for improving majority voting of classifier and improved the accuracy of classification technique. Xueyi Wang entitled "A New Model for Measuring the Accuracies of Majority Voting Ensembles "a new model called COB (core, outlier, and boundary) which quantitatively measures the accuracies of majority voting ensembles for binary classification [1]. Good ensemble methods are that in which each individual classifiers are accurate and diverse. But ensemble methods are combination of predictions made by a set of individual classifiers. Accurate classifier is meant to be produce accurate prediction than the random classifier and Diverse classifier is meant to be produce prediction independently. For experimental purpose of COB three different ensemble methods bagging, random forests, and a randomized ensemble, two different numbers of individual classifiers and three different machine learning algorithms decision trees, k-nearest neighbours, and support vector machines are used. The COB model results that the accuracy of ensemble method is worse with the present of nonempty core subset than the accuracy of the binomial method. The COB model is an enhancement to the binomial model with addition of two subsets core and outlier. Gavin Brown, and Ludmila I. Kuncheva entitled ""Good" and "Bad" Diversity in Majority Vote Ensembles" accuracy is not straight forward with the desired diversity in classifier ensembles is proposed [2]. The majority votes are decomposed into three terms an average individual accuracy, good diversity and bad diversity. Diversity can be defined as a consequences of two decision (1) the choice of error function and (2) the choice of combiner function in the design of ensemble problem in the machine learning. While illustrating the majority votes two special case pattern of success and pattern of failure were introduced in this paper. A probability distribution over all possible combinations of correct/incorrect votes is defined to improve the individual accuracy p. Each combination where exactly (T+1)/2votes are correct, appears with probability α. In this pattern no votes are wasted. Tao, Dacheng, Tang, Xiaoou, Li, Xuelong, Wu and Xindong entitled "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval" a new asymmetric bagging and random subspace mechanism is designed [3]. Relevance feedback schemes based on support vector machines (SVM) have been widely used in content-based image retrieval (CBIR). However, the performance of SVM-based relevance feedback is often poor when the number of labeled positive feedback samples is small. This is mainly due to three reasons: 1) an SVM classifier is unstable on a small-sized training set, 2) SVM's optimal hyper plane may be biased when the positive feedback samples are much less than the negative feedback samples, and 3) over fitting happens because the number of feature dimensions is much higher than the size of the training set. The proposed method addressed all these three problems. In a relevance feedback process, the user first labels a number of relevant retrieval results as positive feedback samples and some irrelevant retrieval results as negative feedback samples. Then, a CBIR system refines all retrieval results based on these feedback samples. These two steps are carried out iteratively to improve the performance of the image retrieval system by gradually learning the user's preferences. Many relevance feedback methods like discriminate learning, heuristic method the density estimation method have been developed in recent years. They either adjust the weights of various features to adapt to the user's preferences or estimate the density of the positive feedback examples. Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew and Alex Ksikes entitled "Ensemble Selection from Libraries of Models" a method for constructing ensembles from libraries of thousands of models is presented [4]. Using distinct learning algorithms and parameter settings, model libraries are generated. To maximize the performance of the ensemble models a forward stepwise selection is added. An ensemble is a collection of models whose predictions are combined by weighted averaging or voting. According to Dietterich "A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate and diverse." The simple forward model selection procedure is fast and effective, but sometimes overfits to the hillclimbing set, reducing ensemble performance. To reduce the overfitting selection with replacement, stored ensemble initialization and bagged ensemble selection methods are added.Sandrine Dudoit and Jane Fridlyand entitled "Bagging to improve the accuracy of a clustering procedure" an application of bagging to cluster analysis is proposed [5]. Bagging can substantially improve clustering accuracy and yields information on the accuracy of cluster assignments for individual observations. In addition, bagged clustering procedures are more robust to the variable selection scheme, i.e. their ac-curacy is less sensitive to the number and type of variables used in the clustering. Improving and assessing the accuracy of a given clustering procedure using a resembling method is known as bagging. In supervised learning bagging is used to generate and aggregate multiple clustering's. In this paper two new sampling methods BagClust1 and BagClust2 are proposed to improve and assess the accuracy of a given clustering procedure. In BagClust1 the clustering procedure is repeatedly applied to each bootstrap sample and the final partition is obtained by plurality voting. The BagClust2 method forms a new dissimilarity matrix by recording for each pair of observations the proportion of time they were clustered together in the bootstrap clusters. Nikunj C. Oza and Kagan Tumer entitled "Classifier Ensembles: Select Real-World Applications" classifier ensembles and ensemble applications are presented [6]. Ensuring that the particular classification algorithm matches the properties of the data is crucial in providing results that meet the needs of the particular application domain. One way in which the impact of this algorithm/application match can be alleviated is by using ensembles of classifiers, where a variety of classifiers are pooled before a final classification decision is made. Classifier ensembles provide an extra degree of freedom in the classical bias/variance tradeoff, allowing solutions that would be difficult to reach with only a single classifier. Many learning algorithms generate a single classifier that can be used to make predictions for new examples. The way in which multiple classifiers are combined are simple averaging, weighed averaging, stacking, bagging and boosting. Robert E. Banfield, Lawrence O. Hall, Kevin W. Bowyer and W.P. Kegelmeyer entitled "A Comparison of Decision Tree Ensemble Creation Techniques" Randomization-Based technique for creating an ensemble of classifiers is proposed [7]. BAGGING is one of the older, simpler, and better known techniques for creating an ensemble of classifiers. Bagging creates an ensemble of classifiers by sampling with replacement from the set of training data to create new training sets called "bags". A number of other randomization-based ensemble techniques boosting, random subspaces, random forests, and randomized C4.5 have been introduced. In bagging, only a subset of examples typically appears in the bag which will be used in training the classifier. Out-of-bag error provides an estimate of the true error by testing on those examples which did not appear in the training set. Authors have developed an algorithm which appears to provide a reasonable solution to the problem of deciding when enough classifiers have been created for an ensemble. It works by first smoothing the out-of-bag error graph with a sliding window in order to reduce the variance. Leo Breiman entitled "Bagging Predictors" a method for generating multiple versions of a predictor and using these to get an aggregated predictor is proposed [8]. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets. Thomas G. Dietterich entitled "Ensemble Methods in Machine Learning" methods for constructing ensembles [9]. Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions. The original ensemble method is Bayesian averaging, but more recent algorithms include error-correcting output coding, Bagging, and boosting. An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way (typically by weighted or un-weighted voting) to classify new examples. A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate and diverse. An accurate classifier is one that has an error rate of better than random guessing on new x values. Author also explained various methods for manipulating the training data for constructing ensembles such as bagging, bootstrap replicates, cross validated committee, inject randomness and ADA Boost algorithm. According to author experiment it has been concluded that the ADA Boost gives best result. Bagging and randomized trees give similar performance, although randomization is able to do better in some cases than Bagging on very large datasets. S. B. Kotsiantis entitled "Supervised Machine Learning: A Review of Classification Techniques" a best-known supervised techniques in relative detail is proposed [10]. Supervised machine learning is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances. In other words, the goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown. Every instance in any dataset used by machine learning algorithms is represented using the same set of features. The features may be continuous, categorical or binary. If instances are given with known labels then the learning is called supervised in contrast to unsupervised learning, where instances are unlabeled. A common method for comparing supervised ML algorithms is to perform statistical comparisons of the accuracies of trained classifiers on specific datasets. If we have sufficient supply of data, we can sample a number of training sets of size N, run the two learning algorithms on each of them, and estimate the difference in accuracy for each pair of classifiers on a large test set. Our next step is to perform paired t-test to check the null hypothesis that the mean difference between the classifiers is zero. The average of these differences is an estimate of the expected difference in generalization error across all possible training sets of size N, and their variance is an estimate of the variance of the classifier in the total set. Thomas G. Dietterich entitled "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization" an alternative method for constructing good ensembles that does not rely on instability is explored [11]. Randomize the internal decisions of the learning algorithm. Specifically, author implemented a modified version of the C4.5 learning algorithm in which the decision about which split to introduce at each internal node of the tree is randomized. The author has compared three methods for constructing ensemble classifiers using C4.5: Randomizing, Bagging, and Boosting. Author observed that Boosting gives the best results in most cases. Randomizing and Bagging give quite similar results-there is some evidence that Randomizing is slightly better than Bagging in low noise settings and concluded Bagging is clearly the best method. Guoqiang Peter Zhang entitled "Neural Networks for Classification: A Survey" some of the most important developments in neural network classification research is surveyed [12]. Classification is one of the most frequently encountered decision making tasks of human activity. A classification problem occurs when an object needs to be assigned into a predefined group or class based on a number of observed attributes related to that object. Classification is the most researched topic of neural networks. Author focused review of several important issues and recent developments of neural networks for classification problems. These include the posterior probability estimation, the link between neural and conventional classifiers, the relationship between learning and generalization in neural net-work classification, and issues to improve neural classifier performance. Neural networks have been demonstrated to be a competitive alternative to traditional classifiers for many practical classification problems. Numerous insights have also been gained into the neural networks in performing classification as well as other tasks. All selection criteria and search procedures in feature selection with neural networks are heuristic in nature and lack of rigorous statistical tests to justify the removal or addition of features. Statistical properties of the saliency measures as well as the search algorithms must be established in order to have more general and systematic feature selection procedures. More theoretical developments and experimental investigations are needed in the field of feature selection.

III. Conclusion

In this paper we review a various method of ensemble classifier and discuss the problem of ensemble classifier for large data. And also discuss the enhancement technique of classifier. Such new ensemble technique is COB is great performance of classification, but it also compromised with noise and outlier data of classification. Also discuss the hybrid technique for ensemble of classifier and its suffered from learning rate of classifier and generate negative rate of classification. Finally, we have concluded that ensemble-based algorithms are worthwhile, improving the results that are obtained by the usage of data pre-processing techniques and training a single classifier. The use of more classifiers makes them more complex, but this growth is justified by the better results that can be assessed. We have to remark the good performance of approaches such as RUS Boost or Under Bagging, which despite being simple approaches; achieve higher performances than many other more complex algorithms.

REFERENCES

[1] [1]Xueyi Wang "A New Model for Measuring the Accuracies of" in IEEE World Congress on Computational Intelligence, 2012.

[2] [2]Gavin Brown, and Ludmila I. Kuncheva ""Good" and "Bad" Diversity in Majority Vote Ensembles" in IEEE Transaction.

[3] [3]Tao, Dacheng, Tang, Xiaoou, Li, Xuelong, Wu and Xindong "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval" in IEEE Transactions, 2006.

[4] [4]Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew and Alex Ksikes "Ensemble Selection from Libraries of Models" 21st International Confer-ence on Machine Learning, 2004.

[5] [5]Sandrine Dudoit and Jane Fridlyand "Bagging to improve the accuracy of a clustering procedure" in IEEE Transcation, 2002.

[6] [6]Nikunj C. Oza and Kagan Tumer "Classifier Ensembles: Select Real-World Applications" in Elsevier, 2007.

[7] [7]Robert E. Banfield, Lawrence O. Hall, Kevin W. Bowyer and W.P. Kegelmeyer "A Comparison of Decision Tree Ensemble Creation Techniques" in IEEE TRANSACTIONS, 2007.

[8] [8]Leo Breiman "Bagging Predictors" in Kluwer Academic Publishers, 2006.

[9] [9]Thomas G. Dietterich "Ensemble Methods in Machine Learning" in IEEE TRANSACTIONS.

[10] [10]S. B. Kotsiantis "Supervised Machine Learning: A Review of Classification Techniques" in Informatica 30, 2007.

[11] [11]Thomas G. Dietterich "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization" in Kluwer Academic Publishers, 1999.

[12] [12]Guoqiang Peter Zhang entitled "Neural Networks for Classification: A Survey" in IEEE TRANSACTIONS, 2000.