A REVIEW OF CLUSTER BASED CLASSIFICATION TECHNIQUE

Shivam Mishra\textsuperscript{1}, Anurag Jain\textsuperscript{2}

shivammishrajhansi@gmail.com\textsuperscript{1}, anurag.akjain@gmail.com\textsuperscript{2}

Department of Computer Science & Engineering\textsuperscript{1, 2}

RITS, Bhopal, India

ABSTRACT

Fusion and ensemble is important technique of machine learning. Fusion fused the feature attribute of different classifier and improved the classification of binary classifier. Instead of that ensemble technique provide the facility of merge two individual classifier and improve the performance of both classifiers. The ensemble technique of classifier depends on number of nearer point of classifier. For the selection of nearest point of classifier various authors used clustering technique for ensemble process. The cluster based ensemble technique suffered a problem of optimal selection of cluster during ensemble process. The generation of clustering technique is fixed due to this reasons the selection of optimal cluster is very difficult. In this paper gives the information about ensemble technique based on clustering and other technique.

Keywords: - Classification, ANT Ensemble Classifier and Cluster

INTRODUCTION

Data classification and pattern recognition are important research area in data mining and computer vision. For the purpose of data classification and pattern recognition various machine learning algorithm are used, such as clustering techniques, classification, and neural network. For the improvement and increasing the classification and recognition of pattern in data mining used ensemble classification technique. Ensemble classification technique improves the performance of individual classifier with another classifier. The technique of ensemble method follows a prototype of classification. When a decrease in performance is practical, new classifier are fused into the ensemble while aged and horrific the stage fusion process are neutral. For classification, decisions of fusion in the ensemble are combined, usually with a cluster scheme [10]. The advantage of cluster ensembles over single classifiers in the data classification problem has been proved empirically and theoretically [1, 3]. However, few ensemble methods have been designed to take into consideration the problem of recurring contexts [6, 7]. Specifically, in problems where concepts resume, models of the ensemble should be maintained in memory even if they do not perform well in the latest batch of data. Moreover, each classifier should be specific in a unique concept, meaning that it should be trained from data belonging to this concept and used for classifying similar data. In [9, 12], a methodology that identifies concepts by grouping classifiers of similar performance on specific time intervals is described. Clusters are then assigned to classifiers according to performance on the latest batch of data. Predictions are made by using weighted averaging. Although this strategy fits very well with the recurring contexts difficulty, it has an offline step for the finding of concepts that is not suitable for data streams. In particular, this framework will possibly inaccurate with concepts that did not appear in the training set. To classified real-world data set with overlapping features from different classes. The training of class borders between overlap class features in such cases is a hard crisis. Extreme preparation of the base classifiers will lead to accurate training of the decision border but resulting in over fitting thus mis-classifying instances of experimental data[16,19]. On the other hand, learning global outlier will avoid over fitting but at the cost of always misclassifying some overlapping features. This type of problem on learning the class boundaries of overlapping features remains inherent in all the base classifiers and is propagated to the decision fusion stage as well even though the base classifier errors are uncorrelated. Clustering is the process of unsupervised a data set into multiple groups where each group contains data points that are very close in Euclidean distance. The iteration have well defined and easy to learn outlier data. Let’s assume that the features are labeled with their cluster number [18]. Now if the base classifiers are trained on the modified data set they will learn the cluster outlier. As the clusters have well defined easy to learn outlier the base classifiers can learn them with high accuracy. Clusters can hold overlapping features from different class. A fused classifier can be trained to vote the class of a pattern from the dedicated cluster. The above
section discuss introduction of stream data classification and ensemble classification. In section II we describe related work of ensemble classifier. In section III problem of stream data classification used ensemble cluster. In section IV discuss our approach for ensemble cluster classification and finally conclude in section V.

II. RELATED WORK

In this section describe method for ensemble classifier for data classification using clustering technique and other method for ensemble classifier. The method of cluster ensemble classifier and fusion of ensemble classifier reduces the bottleneck problem of individual classifier. Clustering and other data grouping technique provide flexibility for classification fusion in different domain of data.

cluster oriented ensemble classifier is based on original concepts where cluster boundaries are learned by the base classifier and cluster confidences are mapped with the help of fusion classifier to the class decision. According to this paper an ensemble classifier is constructed using a set of base classifier which learns the class boundaries separately over the pattern. Clustering is the method of separating an item set into multiple item sets group. Clustering assumed that if the patterns are labeled with their cluster number and the base classifiers are trained on the modified data set then base classifier will learn the cluster boundaries [1]. To gain improved and better accuracy of the ensemble classifier clusters are classified into multiple clusters and cluster decisions produced by the base classifier are combined into class decision by a fusion classifier.

Ensembles are designed in such a way that each classifier is trained independently and the decision in pattern classification, multiple classifier systems are often use a practical and effective solution for difficult recognition problems fusion is performed as a post-process module. In some cases, the experimental observations of the performance of specialized classifiers justify the use of multiple classifiers [2]. In other cases, the implementation of multiple classifiers stems from the problem decomposition such as the need to employ a variety of sensor types, or the need to avoid making commitments to arbitrary initial conditions and parameters. There are many methods to use more than one classifier in a recognition problem.

A method for generating multiple version of a predictor and using these to get an aggregated predictor [3]. The aggregation averages over the description when predicting a numerical outcome and does a plurality vote when predicting a class. Number of constraints is formed by making bootstrap replication of the learning set and using these as new learning group. Tests on experimental datasets using classification and regression technique and feature selection in linear regression show that bagging gives better result prediction. Bagging is one of the oldest, simpler, and better known methodology for creating an ensemble of classifiers. A number of other randomization-based ensemble techniques have been introduced. Some of the include boosting random sub spaces, random forests [20].

Analysis of Bagging as a Linear Combination of Classifiers as applying an analytical framework for the analysis of linearly combined classifiers to ensembles generated by bagging [5]. This provides an analytical model of bagging misclassification probability as a function of the ensemble size, which is a novel result. This allows us to derive a novel and theoretically grounded guideline for choosing bagging ensemble size. The technique of ensemble classifier are bagging, boosting and random forest tree, are based on introducing some kind of randomness into the design process of single classification technique. The ensemble process of linear technique is going on boosting process. Author applied an analytical framework for linear combiners developed in, and to the particular case of linearly combined classifiers generated by bagging.

Several methods for the construction of classifier ensembles, like bagging, random subspace technique, tree randomization and random forests technique, these methods are based on introducing some kind of randomness into the design process of individual classifiers. Bagging is perhaps the most admired method, and its efficiency has been empirically shown in number of real pattern recognition problems. Author applied a systematic framework for linear combiners developed in, and to the particular case of linearly combined classifiers generated by bagging.

Ensemble of Classifiers (EoC) has been shown effective in improving the performance of single classifiers by combining their outputs [6]. Even though the clustering diversities might only be able to represent data diversities in random Subspaces, for Bagging method, which only use a part of the samples, there is still no adequate measure for their data diversities. It will be big interest to figure out how to calculate the data diversities in Bagging. Finally, we have to point out that, due to its special ensemble generating methods, which ares not likely to be related in Boosting.

A bagging can push a good but unstable procedure a major step towards optimality. On the other hand, it can slightly degrade the performance of established steps. There has been latest work in the literature with some of the flavor of bagging [4].

The goal of ensemble learning methods is to construct a collection (an ensemble) of individual classifiers that are diverse and yet accurate. The highly accurate classification decisions can be obtained by voting the decisions of the
individual classifiers in the ensemble. Two of the most popular techniques for constructing ensembles are bootstrap aggregation and the Adaboost family of algorithms. Both of these methods operate by taking a base learning algorithm and invoking it many times with different training sets. In bagging, each training set is constructed by forming a bootstrap replicate of the original training set[22]. Ensemble learning methods have become an active research topic within the computational intelligence community. Over the past decade, many theoretical analyses, practical algorithms, and empirical studies have been proposed in this field. Ensemble training techniques also have been widely applied in many real-world applications, including Web mining, financial engineering, geosciences and remote Sensing, biomedical data analysis. Bootstrap aggregating (bagging) is an ensemble learning method based on the idea of developing multiple hypotheses by bootstrap sampling (with replacement) of the available training instances. In the bagging method, the probability sampling function is uniformly distributed across all the training instances. In order to dynamically adjust the weights for different data instances according to their distributions, various boosting algorithms have been developed [11].

Ensemble methods make predictions by combining the predictions from a set of individual classifiers. To achieve high prediction accuracy, traditionally it is believed that ensemble methods should have accurate and diverse individual classifiers. “Accurate classifiers” means the prediction accuracy of each classifier should be better than random, that is, larger than 0.5 for a binary classifier. “Diverse classifiers” means each classifier should make prediction independently, so that a combination of these predictions will result in high prediction accuracy for ensemble methods [14].

Boosting is a set of methods for the construction of classifier ensembles [8]. The differential feature of these methods is that they allow obtaining a strong classifier from the combination of minor classifiers. Therefore, it is possible to use boosting methods with very simple base classifiers. [9] The simple classifier as decision tress is one decision node. This method is a alternative of the boosting method. It is based on considering, as the base classifiers for boosting, but a classifier formed by the last selected weak classifiers. If the weak classifiers are decision tree, the combination of weak classifiers is a decision tree.

[17] Author describe a process of stream data classification by Kernel-Based Selective Ensemble Learning as Kernel methods enable the modeling of structured data in learning algorithms. Kernel methods provide a dominant technique for modeling structured objects in instant based technique, they require a high computational complexity to be used in streaming environments. This method is the first that demonstrates how kernel methods can be employed to define an ensemble approach able to quickly react to concept drifting and guarantees an efficient kernel computation.

[21] There are several applications for Machine Learning (ML), the most significant of which is data mining. People are often prone to creating errors during analyses or probably, when trying to creating relationships between various dataset. This makes it difficult for them to find solutions to some certain problems. Machine learning can be successfully applied to these problems. Every instance in any feature set used by machine learning algorithms is represented using the same data sets. The features may be continuous, categorical or binary, particular, this work is concerned with classification problems in which the output of instances admits only discrete, unordered values.[18] The library of machine learning algorithm used tools and technique of kernel function. To maximize the recital of the ensemble models a forward process selection is joined. An ensemble is a group of models who’s voting are combined by weighted averaging value of classifier. A necessary and ample condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are good and bad.

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. An essential and sufficient condition for an ensemble of classifiers to be more specific than any of its individual members is if the classifiers are precise and diverse. The area of ‘diversity’ has been a favorites buzzword in the multiple classifier systems community for long time. Various diversity measures have been proposed, measured and maximized and all with the goal to increase ensemble performance by balancing “individual accuracy” against “diversity”. It is therefore ironic that after so much time and attempt, we still have no distinctively agreed definition for “diversity” [15].

The simple forward model selection procedure is fast and effective, but sometimes over fits to the hill climbing set, reducing ensemble execution process. To decrease the over fitting selection with alternate, stored ensemble initialization and bagged ensemble selection methods are used.

III PROBLEM OF DATA CLASSIFICATION IN CLUSTER BASED ENSEMBLE

Cluster based ensemble classifier compromised with selection of optimal cluster selection. The selection of optimal cluster
faced a problem of fixed number of cluster generation using k-means clustering. The fixed number of cluster technique induces a problem of data unbalancing in classification and pattern recognition. The unbalance data ratio issue aird a problem of minority and majority voting principle. The selection of cluster in classification is bottleneck. An alternative approach to producing diverse binary classifiers involving random sampling over the feature space was proposed in [12]. In this work each base classifier was generated with a randomly selected subset of features. The final ensemble with a combining voting technique was able to improve performance in comparison with the binary base classifiers. A similar method based on this concept was presented by [13]. In this work each binary model was built by applying a different subset of features. Each feature had a weight assigned to reflect its relevance to the problem being considered. A weighted Euclidean distance metric was applied while selecting neighbors. To make the final decision a voting scheme was applied among the instances selected by all the binary in the ensemble. Following evaluations it was found that the proposed approach based on sampling over the feature space provided a significant improvement in comparison with a single binary classifier. Some current problem found in review process.

1. Unbalanced ratio of train and test data[17]
2. Selection of optimal cluster for ensemble classifier[1,3,6]
3. Diversity of feature selection process.[12] 
4. Boundary value of cluster[9,13]
5. Outlier data treat as noise [24].

IV OUR APPROACH OPTIMAL CLUSTER SELECTION

Cluster oriented ensemble classifier is well know method for stream data classification. In cluster oriented ensemble classifier is suffered from a selection of optimal number of cluster selection technique. The selection of optimal number of cluster enhanced the process of cluster oriented ensemble classification for data classification. The optimality of cluster is selected by Meta function. For this process we used ACO technique. Ant is meta-heuristic function inspired by real ANT. The fitness constrains of ACO is multiple. Using ant colony optimization we maintain the selection process of clustering technique and noise removal of boundary base class. Noise reduction and selection of optimal number of cluster in ensemble classifier used cluster index selection process using ant colony optimization technique. We introduce a new feature sub set selection method for finding similarity matrix for clustering without alteration of ensemble classifier. The proposed cluster index selection method based on ant colony optimization, ACO technique find the most similar cluster index for ensemble of classifier. In this method we introduced continuity of ants for similar features and dissimilar features collect into next node. In that process ACO find optimal selection of cluster index. Suppose ants find features of similarity in continuous root. Every ant of features compares their property value according to initial features set.

V CONCLUSION AND FUTURE WORK

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 a new ensemble technique is used in cluster oriented mechanism for improvement of stream data classification performance. The selection of optimal number in ensemble classifier is significant task. All authors’ method suffered from this type of problem. The selection of ensemble classifier mainly based on bagging, boosting and random forest technique. These techniques are not deals in the area of data diversity and suffered stream data classification. For the development of data diversity and boundary class training used clustering technique for ensemble classifier. For the survey problem, I will resolve this issue using ant colony optimization technique for selection of optimal cluster and base boundary value.

REFERENCES:


Giorgio Fumera, Fabio Roli and Alessandra Serrau “A Theoretical Analysis of Bagging as a Linear Combination of Classifiers” in IEEE Transactions.


Xueyi Wang “A New Model for Measuring the Accuracies of” in IEEE World Congress on Computational Intelligence, 2012.

Gavin Brown, and Ludmila I. Kuncheva “‘Good’ and ‘Bad’ Diversity in Majority Vote Ensembles” in IEEE Transaction.


Thomas G. Dietterich “Ensemble Methods in Machine Learning” in IEEE TRANSACTIONS.

