ISSN 22773061

A Review of Cluster Oriented Ensemble Classifier for Improving Performance of Stream Data Classification

Richa Gupta[1], Hitesh Gupta[2]
gudiyagour@yahoo.co.in[1], hitesh034@gmail.com[2]

Department of Computer Science & Engineering[1, 2]
PCST, Bhopal, India

Abstract

Ensemble classification technique is great advantage over conventional classifier such as statistical, binary and neural network classifier. Ensemble technique improves the performance of other classifier with some valid constraints. In the improvement of ensemble classifier cluster play important role for data grouping before classification. Cluster oriented ensemble classifier maintain and control the diversity of data during classification process. The diversity of cluster oriented ensemble classifier implied in stream data classification. Stream data classification is critical task due to diversity of data. The critical task of diversity such as infinite population, data drift and feature evaluation technique is overcome through ensemble classifier. Various author proposed a method for stream data classification using ensemble technique. In this paper we give the review of ensemble technique used in stream data classification with clustering technique.

Keywords: - Stream Data, Ensemble Classifier and Cluster

Council for Innovative Research
Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol 8, No 3
editor@cirworld.com
www.cirworld.com, member.cirworld.com

INTRODUCTION

Stream data classification is demand of current storage of internet data. The need and requirement of online transaction of data is stream classification, due to stream classification save time of computation and storage area of network. For the purpose of stream data classification various machine learning algorithm are used, such as clustering, classification, and neural network. In the classification process of a growing data stream, also the temporary or long-standing activities of the stream may be more significant, or it often cannot be known a priori as to which one is more important. We decide the window or horizon of the training data to use so as to obtain the best classification accuracy. For the proper selection of window and horizon used ensemble classifier with support of clustering technique. In ensemble methods, the main strategy is to maintain a dynamic set of classifiers. When a decrease in performance is practical, new classifier are fused into the ensemble while aged and horrific the stage fusion process are disinterested. 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 stream 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 re-occur, models of the ensemble should be maintained in memory even if they do not perform well in the latest batch of data. Moreover, every classifier should be specialized 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 problem, it has an offline step for the discovery of concepts that is not suitable for data streams. In particular, this framework will probably be 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 test data. On the other hand, learning generalized boundaries will avoid over fitting but at the cost of always misclassifying some overlapping features. This 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. We opt to bring in clustering at this point. Clustering is the process of partitioning a data set into multiple groups where each group contains data points that are very close in Euclidean space. The clusters have well defined and easy to learn boundaries. 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 boundaries. As the clusters have well defined easy to learn boundaries the base classifiers can learn them with high accuracy. Clusters can contain overlapping features from multiple classes. A fusion classifier can be trained to predict the class of a pattern from the predicted 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 we discuss method for ensemble classifier for stream data classification using support of clustering technique. The method of cluster ensemble classifier has great overcome of binary and normal classifier, here we describe method of ensemble classifier used for stream data classification.

Brijesh Verma and Ashfaqur Rahman entitled "Cluster-Oriented Ensemble Classifier: Impact of Multi-cluster Characterization on Ensemble Classifier Learning" a novel cluster-oriented ensemble classifier is presented [1]. This 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 process of separating an item set into multiple item sets group. It is 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. To gain better and improved 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.

Nayer M.Wanas, Rozita A. Dara and Mohamed S. Kamel entitled an investigation of Adaptive fusion and co-operative training for classifier ensembles [2] as ensembles are designed in such a way that each classifier is trained independently and the decision In pattern classification, multiple classifier systems are often considered a practical and effective solution for difficult recognition problems fusion is performed as a post-process module. In some cases, the empirical observations of the performance of specialized classifiers justify the use of multiple classifiers. In other cases, the adoption 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 ways to utilize more than one classifier in a recognition problem.

Leo Breiman entitled a Bagging predictor [3] as a method for generating multiple version of a predictor and using these to get an aggregated predictor. The aggregation averages over the version when predicting a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replication of the learning set and using these as new learning sets. Tests on real and simulated data sets using classification and regression trees and subset selection in linear regression show that bagging can give substantial gains in accuracy.

Giorgio Fumera, Fabio Roli and Alessandra Serrau entitled Analysis of Bagging [5] as a Linear Combination of Classifiers as applying an analytical framework for the analysis of linearly combined classifiers to ensembles generated by bagging. 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. Several methods for the construction of classifier ensembles, like bagging, the random subspace method, tree randomization and random forests, are based on introducing some kind of randomness into the design process of individual classifiers. Bagging is perhaps the most popular method, and its effectiveness has been empirically shown in many real pattern recognition problems. Author applied an analytical framework for linear combiners developed in, and to the particular case of linearly combined classifiers generated by bagging.

Albert Hung-Ren Ko and Robert Sabourin entitled an Implication of Data Diversity for a Classifier-free Ensemble Selection in Random Subspaces [6] as Ensemble of Classifiers (EoC) has been shown effective in improving the performance of single classifiers by combining their outputs. Even though the clustering diversities might only be able to represent data diversities in Random Subspaces, for Bagging, which only use a part of the samples, there is still no adequate measure for their data diversities. It will be of great interest to figure out how to measure the data diversities in Bagging. Finally, we have to mention that, due to its special ensemble generating mechanism, the scheme is not likely to be applicable in Boosting.

Juan J. Rodrı guez and Jesu s Maudes entitled a method for the construction of classifier ensembles called boosting [8] as Boosting is a set of methods for the construction of classifier ensembles. The differential feature of these methods is that they allow obtaining a strong classifier from the combination of weak classifiers. Therefore, it is possible to use boosting methods with very simple base classifiers. One of the most simple classifiers are decision stumps, decision trees with only one decision node. This work proposes a variant of the most well-known boosting method, AdaBoost. It is based on considering, as the base classifiers for boosting, not only the last weak classifier, but a classifier formed by the last r selected weak classifiers (r is a parameter of the method). If the weak classifiers are decision stumps, the combination of r weak classifiers is a decision tree. Given one or more classification methods, one of the most natural ways of obtaining more accurate classifiers is the use of ensembles.

Valerio Grossi, Alessandro Sperduti "Kernel-Based Selective Ensemble Learning for Streams of Trees"[17] author proposed a process of stream data classification by Kernel-Based Selective Ensemble Learning as Kernel methods enable the modeling of structured data in learning algorithms, however they are computationally demanding. Both efficacy and efficiency of the proposed approach are assessed for different models by using data sets exhibiting different levels and types of concept drift. Kernel methods provide a powerful tool for modeling structured objects in learning algorithms. Unfortunately, they require a high computational complexity to be used in streaming environments. This work 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.

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 [18]. 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 over fits to the hill climbing set, reducing ensemble performance. To reduce the over fitting 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 [17].

III PROBLEM OF STREAM DATA CLASSIFICATION USED ENSEMBLE CLUSTER

For the purpose of stream data classification various machine learning algorithm are applied, such as clustering, classification, and regression. Two of the most critical and well generalized problems of data streams are its infinite length and concept-drift. Since a data stream is a fast and continuous event, it is assumed to have infinite length. Therefore, it is difficult to store and use all the historical data for training. The most discover alternative is an incremental learning technique. Several incremental learners have been proposed to address this problem [18], [17]. In addition, concept-drift occurs in the stream when the underlying concepts of the stream change over time. A variety of techniques have also been proposed in the literature for addressing concept-drift [12], [16], [17] in data stream classification. However, there are two other significant characteristics of data streams, such as concept evolution and feature evolution that are ignored by most of the existing techniques. Concept-evolution occurs when new classes evolve in the data. Cluster oriented ensemble classifier used to reduce feature evaluation problem in stream data classification. The selection of optimal number of cluster is also challenging job in stream data classification. On the review process we found some important problem in cluster oriented stream data classification. These problems are given below.

Stream data classification suffered from infinite length, concept evaluation ,feature evaluation and data drift[17]

Selection of optimal cluster for ensemble classifier[1,3,6]

Diversity of feature selection process.[12]

Boundary value of cluster[9,13]

Outlier data treat as noise [24, 25].

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 for ensemble. The selection of optimal number of cluster improves the performance of cluster oriented ensemble classifier for stream data classification. The optimality of cluster is selected by heuristic function. For this process we used ant colony optimization technique. Ant is meta-heuristic function inspired by biological ants. The objective of ant colony optimization 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 features sub set 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 features sub set selection method based on ant colony optimization, ant colony optimization is very famous meta-heuristic function for searching for finding similarity of data. 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 features sub set. Suppose ants find features of similarity in continuous root. Every ant of features compares their property value according to initial features set. When deciding data is noise and outlier should consider the two factors: importance degree and easiness degree of noise and outliers. While walking ants secrete phenomenon on the ground according to importance of the outlier and follow, in probability pheromone previously laid by other ants and the easiness degree of the noise.

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 new ensemble technique is used cluster oriented mechanism for improvement of stream data classification. The selection of optimal number in ensemble classifier is important task. All authors' method suffered from this problem. The selection of ensemble classifier basically based on bagging, boosting and random forest technique. These techniques are not deals in the field of data diversity and suffered stream data classification. For the improvement of data diversity and boundary class training used clustering technique for ensemble classifier. For the survey problem I will solve using ant colony optimization technique for selection of optimal cluster and base boundary value.

REFERENCES:-

[1] Brijesh Verma and Ashfaqur Rahman "Cluster-Oriented Ensemble Classifier: Impact of Multicluster Characterization on Ensemble Classifier Learning" in IEEE Transactions on knowledge and data engineering, 2012.

[2] Anne-Laure Bianne-Bernard, Fare`s Menasri, Rami Al-Hajj Mohamad, Chafic Mokbel, Christopher Kermorvant and Laurence Likforman-Sulem "Dynamic and Contextual Information in HMM Modeling for Handwritten Word Recognition" in IEEE transactions on pattern analysis and machine intelligence, 2011.

[3] Nayer M.Wanas, Rozita A. Dara and Mohamed S. Kamel "Adaptive fusion and co-operative training for classifier ensembles" in Pattern Analysis and Machine Intelligence Lab, University of Waterloo, 2006.

[4] LEO BBEIMAN "Bagging Predictors" in Kluwer Academic Publishers, 1996.

[5] Yoshua Bengio "Learning Deep Architectures for AI" in Foundations and Trends in Machine Learning, 2009.

[6] Giorgio Fumera, Fabio Roli and Alessandra Serrau "A Theoretical Analysis of Bagging as a Linear Combination of Classifiers" in IEEE Transactions.

[7] Albert Hung-Ren Ko and Robert Sabourin "The Implication of Data Diversity for a Classifier-free Ensemble Selection in Random Subspaces" in IEEE Transactions.

[8] Zhihui Lai, Zhong Jin, Jian Yang and W.K Wong "Sparse Local Discriminant Projections for Face Feature Extractio" International Conference on Pattern Recognition, 2010.

[9] Juan J. Rodrı´guez and Jesu ´s Maudes "Boosting recombined weak classifiers" in ScienceDirect, 2007.

[10] Oriol Pujol and David Masip "Geometry-Based Ensembles: Toward a Structural Characterization of the Classification Boundary" in IEEE Transactions, 2009.

[11] Haibo He and Yuan Cao "SSC: A Classifier Combination Method Based on Signal Strength" in IEEE Transactions on neural networks and learning systems, 2012.

[12] Nandita Tripathi, Stefan Wermter, Chihli Hung and Michael Oakes "Semantic Subspace Learning with Conditional Significance Vectors" in IEEE Transactions.

[13] Terry Windeatt "Accuracy/Diversity and Ensemble MLP Classifier Design" in IEEE Transactions.

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

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

[16] 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.

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

[18] Valerio Grossi, Alessandro Sperduti "Kernel-Based Selective Ensemble Learning for Streams of Trees" in Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence 2010.

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

[20] 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.
[21] Leo Breiman “Bagging Predictors” in Kluwer Academic Publishers, 2006.
[22] Thomas G. Dietterich “Ensemble Methods in Machine Learning” in IEEE TRANSACTIONS.
[23] S. B. Kotsiantis “Supervised Machine Learning: A Review of Classification Techniques” in Informatica 30, 2007.
[24] Thomas G. Dietterich “An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization” in Kluwer Academic Publishers, 1999.
[25] Guoqiang Peter Zhang entitled “Neural Networks for Classification: A Survey” in IEEE TRANSACTIONS, 2000.