Stream Data Classifier with Ensemble Cluster

Ms. Darshana Desai
Department of Computer Engineering
D. J. Sanghavi College of engineering
Mumbai, India

Dr. Abhijit Joshi
Department of Information Technology
D. J. Sanghavi College of engineering
Mumbai, India

Abstract— **Ensemble methods have been one of the powerful methods for improving the robustness and the accuracy of machine learning. Moreover, internet generates huge amounts of data which are continuous, so it is important to identify and consolidate various concepts for improving the performance of any classifier. Numerous studies carried out in the past decade on the problem of combining competing models into a committee. The success of ensemble of models has been observed in many fields including recommendation systems, anomaly detection, stream mining, and web applications. However, the combining process works on the basic principle, i.e., combination of different base models strengthens weak models. Also, merging of different models leads to better performance in comparison to their availability for single task. The generation of clustering technique and classification technique are fixed due to this reasons the selection of optimal decision making is very difficult. In this paper, we are exploring ensemble technique which is based on Ant Optimization Colony (ACO) for selection of clustering and classification technique.**

Keywords **— Ensemble Model, Machine Learning, Classification, ANT Ensemble Classifier and Cluster**

I. 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. Cluster and classification play an important role in data mining and machine learning paradigm.

The process of prototype classification combines two or more methods with same nature. Stream classification is required and needed for online transaction in data. Actually, stream classification saves time of computation and storage area of network. In data mining, to improve and increase the classification and recognition of pattern ensemble classification technique is used. Ensemble classification technique improves the performance of individual classifier with another classifier. The ensemble technique follows a prototype of classification. For classification, decisions of the fused ensemble process are combined, which can be done usually with the cluster scheme [10]. The issues of ensemble cluster over single classifiers in the data stream classification have been proved empirically and theoretically [1, 3]. However, less ensemble methods have been designed and implemented to take into consideration the problem of recurring contexts [6, 7]. Specifically, in complications where concepts invert, models of the ensemble should be continued in memory even if they do not perform well in the latest batch of data. Moreover, every classifier should be functional in a different concept, meaning that it should be trained from data belonging to this concept and used for classifying similar data. In [9, 10], a methodology that identifies concepts by grouping classifiers of similar performance on specific time intervals is described. Clusters are then appointed to classifiers according to act on the latest batch of data. While classifying real-world data set having overlap features from different classes are considered and weighted averaging is used for making predictions. Extreme preparation of the base classifiers will lead to accurate training of the decision border. But it leads to misclassification instances of test data. On the other hand, learning derived boundaries will avoid over fitting but at the cost of always misclassifying some overlapping features. This issue on learning the class boundaries of extending 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 can use clustering to solve this problem. Clusters can contain overlapping features from multiple classes. The cluster ensemble classifier performs an important role in classification technique, but the choose process of cluster is difficult to define. Now this issue is addressed by ant colony optimization. We will discuss it in detail in subsequent section.

The rest of the paper is organised as - in section II we describe related work of ensemble classifier with clustering. Section III addresses the problem occurred while fusing stream data classification with ensemble cluster. In section IV we discuss about our approach for building ensemble model. Finally the paper ends with conclusion and future work.

II. RELATED WORK

The congestion problem of each classifier is reduced by amalgam of ensemble classifier with clustering. Ensemble classifiers on data streams add a generic framework for handling massive volume data streams with concept drifting. The concept of ensemble classifiers is to dividing continuous data streams into small data chunks, from which a count of base classifiers are made and combined together for prediction. For example, [Rich et al., 2004] first proposed a weighted ensemble classifier framework, and demonstrated that their model out performs a single learning model. Inspired by their work, various ensemble models have been proposed, such as ensemble different learning algorithms [3], ensemble active learners [4], to name a few.

[Verma and Rahman, 2012] have used cluster with ensemble classifier which is based on ordinal concepts where cluster boundaries are cultivated by the base classifier and cluster confidences are mapped with the help of fusion classifier to the class decision. They assume that if the patterns are labelled with their cluster number and the base classifiers are trained on the changed data set then base classifier will learn the cluster boundaries. The accuracy is improved of the ensemble classifier clusters by classifying into multiple clusters and is combined into class decided by a fusion classifier.

[Grossi and Sperduti, 2010] has proposed a process of stream data classification by Kernel-Based Selective Ensemble Learning. This is a Kernel based methods that models the structured data in learning algorithms. However, they are computationally difficult. Kernel methods provide a powerful tool for modelling structured objects in learning algorithms but with high computational complexity to be used in streaming environments. Kernel methods can be employed to define an ensemble approach which quickly reacts to concept drifting and this guarantees an efficient kernel computation.

[Ko and Sabourin, 2008] have shown that while combining the outputs of different classifiers the output of single classifier is improved. Even though the clustering diversities might only be able to represent data diversities in Random Subspaces, for Bagging, which only use a chunk of the samples, there is still no adequate measure for their data diversities. Finally, we have to mention that, due to its appropriate ensemble generating mechanism.

[Rodrıguez and Maudes, 2007] has used a method for the building of classifier ensembles called boosting. Boosting is a set of methods for the construction of classifier ensembles. The exceptional feature of these methods is that they allow obtaining a strong classifier from the combination of weak classifiers. Therefore, it is available to use boosting methods with very simple base classifiers. One of the simplest classifiers is decision stumps, decision trees with only single decision node. There also exist a variant of the most well-known boosting method, AdaBoost. It is dependent on seeing as the base classifiers for boosting, not one of the last frail classifier, but a classifier formed by the last r selected frail classifiers (r is a parameter of the method). If the frail classifiers are decision stumps, the combination of r frail classifiers is a decision tree. Given more than one classification methods, one of the most known schemes of obtaining more accurate classifiers is the use of ensembles.

[Fumera and etal., 2009] suggested an analytical framework for the analysis of linearly combined classifiers to ensembles generated by bagging. This detailed model of bagging misclassifies probability as a function of the ensemble size, which is a novel result. Several methods for the construction of classifier ensembles, like bagging, the random subspace scheme, tree randomization and random forests, are based on introducing some sort of randomness into the design process of individual classifiers. One of the famous methods is bagging which gives effectiveness in many real pattern recognition issues.

[Crew and Ksikes, 2010] have suggested a method for constructing ensembles from libraries of thousands of models. Using distinct learning algorithms and parameter settings, model libraries are generated. In this, the maximization in the performance of the ensemble models added with a forward stepwise selection. An ensemble is a collection of models whose predictions are combined by weighted averaging or voting. According to [Dietterich, 2002], "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 diverse and precise." The plain forward model selection action is fast and effective, but sometimes over fits to the hill climbing set, reducing ensemble performance. To minimize the over fitting selection with restoration, stored ensemble initialization and bagged ensemble selection methods are added. [Dudoit and Fridlyand, 2011] came with a bagging application for cluster analysis that is use to improve the accuracy of clustering procedure.

III. Problem with Ensemble Cluster with Classification

Various machine learning algorithms are applied for the purpose of stream data classification, such as clustering, classification, and regression. Looking at the various algorithms, now let us see the problems with these algorithms. The most critical and well generalized problems of data streams are its infinite length and concept-drift. Considering the characteristic of data stream being infinite length and high speed, it is difficult to store and use all the historical data for training [10]. The most discover alternative is an incremental learning technique. Several incremental models have been proposed to address this problem [7, 9]. In addition, concept-drift appears in the stream since it changes timely. A variety of techniques have also been proposed for addressing concept-drift in data stream classification [9, 10]. However, Most of the existing techniques are ignored the two other important characteristics of data streams, i.e., concept evolution and feature evolution. Ensemble classifier with clustering is also used to reduce feature evaluation problem. The selection of optimal number of cluster is also denouncing job in stream data classification. On the analyses process we found some important issues such as infinite length, concept drifting, ever changing data etc. in cluster oriented stream data classification. Our proposed approach addresses these issues.

IV. Our Approach for Optimal Selection of Cluster

Ensemble classifier using clustering is a well known method for stream data classification. The selection of optimal number of cluster improves the performance of cluster oriented ensemble classifier for stream data classification. The selection of optimal cluster is done by heuristic function for which we use ant colony optimization technique. Ant is meta-heuristic function inspired by biological ant. Using Ant Colony Optimization (ACO) we address following issues:

- The selection process of clustering technique and noise removal of boundary base class.
- Selection of optimal number of cluster in ensemble classifier which uses features sub set selection process with the help of ACO.
- Affinity matrix for clustering without alteration of ensemble classifier.

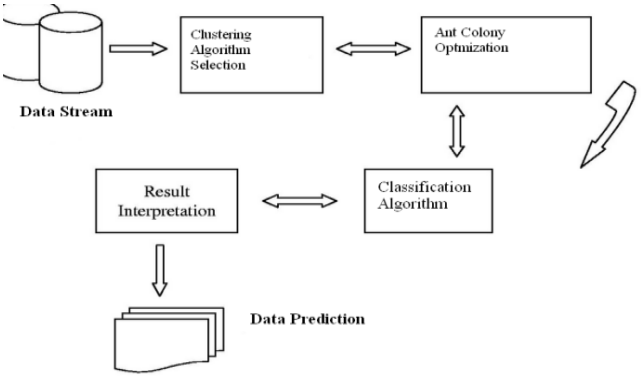

Figure 1: Architecture for selection of cluster for data stream classification

The figure 1 describes the architecture of selection of clusters using ACO for data stream classification. Cluster oriented ensemble classifier is one of the known method for stream data classification. Cluster ensemble with classifier is a replacement to many selection techniques. The selection of maximum number of cluster enhances the ensemble process. The optimality of cluster is decided by Meta function. For this process we use ACO technique. The fitness constrains of ACO is numerous. Using this we maintain the selection process of clustering technique and noise removal of boundary base class. Noise reduction and selection of maximum number of cluster in ensemble classifier used cluster index selection process using ant colony optimization technique. The excepted cluster index selection method based on ACO technique that searches the most similar cluster index for ensemble of classifier. In this method we introduce the continuity of ants for identical features and dissimilar features collect into next node. In this process of ACO, optimal selection of cluster index is made. Every ant of features analyses their property value according to initial features set.

With the help of ACO, we are addressing the following problem as mentioned in section III.

- Classification of data stream/ever changing data.
- Selection of optimal cluster for classification.
- Assortment of feature selection process/sub -set selection
- Boundary value of cluster/infinite length.
- Data outlier/concept drifting.

ACO Algorithm can be implemented with 3 variations explained below [9].

a. **ACO1**: In this method, the ant packets are not allowed to visit a node that it has already visited before, i.e., the ant packets are not allowed to form loops. If a packet reaches a state that it has no other way except to form a loop, the packet is discarded.

b. **ACO2**: In this method, the ant packets are allowed to form loops and visit an already visited node. However, they cannot visit the node last visited by it. In this method, to prevent a packet from going into an infinite loop, if the packet has not reached the destination after a certain interval of time, it should be marked as unsuccessful.

c. **ACO3**: ACO2 is modified with the restriction that the ant packet will not visit the last n nodes already visited by it. A Tabu list [7] is maintained to keep a list of the last n nodes visited.

Depending upon the needs the ACO technique can be selected which accordingly will produce the clusters for data stream classification. The selected ACO would predict which clusters should be formed.

V. Capabilities of Ensemble Model

- Our model should be capable to learn and update with every new data – labelled or unlabeled.
- Will need and deed the knowledge in further learning.
- Will not build on the previously learned ability.
- Will generate a new class/cluster as required and take decisions to merge or divide them as well.
- Will enable the classifier itself to evolve and be dynamic in nature with the changing environment.

VI. 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. We also argue on the enhancement technique of classifier. Such a new ensemble technique uses cluster oriented mechanism for improvement of stream data classification. The selection of clustering and classification technique for ensemble model is a critical task. Ant Colony Optimization technique helps in selecting optimal cluster for data stream classification.

Ant Colony Optimization technique can be used with combination with genetic algorithm for better results.

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

[4] Albert Hung-Ren Ko and Robert Sabourin "The Implication of Data Diversity for a Classifier-free Ensemble Selection in Random Subspaces" in IEEE Transactions.
[5] Juan J. Rodrı´guez and Jesu ´s Maudes "Boosting recombined weak classifiers" in Science Direct, 2007.
[6] Tao, Dacheng, Tang, Xiaoou, Li, Xuelong, Wu and Xindong "Asymmetric bagging and random subspace for support vector 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.
[7] Masaya Yoshikawa and Kazuo Otani, "Ant Colony Optimization Routing Algorithm with Tabu Search", Proceedings of the International Mulitconference of Engineers and Computer Scientists 2010, Volume – III.
[8] Vincent Verstraete, Matthias Strobbe, Erik Van Breusegem, Jan Coppens, Mario Pickavet and Piet Demeester, "AntNet: ACO routing algorithm in practice", Ghent University – IBBT – IMEC, Department of Information Technology.
[9] 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.
[10] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew and Alex Ksikes "Ensemble Selection from Libraries of Models" 21st International Confer-ence on Machine Learning, 2004.
[11] Thomas Dietterich "Ensemble Methods For Machine Learning" Oregon State University retrieved from http://www.cs.orst.edu/`tgd.