Categorization and Detection of Adaptive Novel Class of Feature Evolving Data Streams

Ms. Chaitrali T. Chavan[1], Prof. Vinod S. Wadne[2]
[1,2]Pune University, Pune

Abstract— *Classification in the data stream is the challenging fact for the data mining community. In this paper, we tackle four major challenges which are infinite length, concept drift, concept evolution, and feature evolution. As we know that the data streams are huge in amount, so practically it is not possible to store the data and used it for the training purpose. The results of changes in the underlying concepts are occurred because of concept drift, which is the general observable fact in the data streams. The result of new classes surfacing in the data streams occurs because of concept evaluation. The feature evaluation generally occurs in many streams like text streams, in text streams new features emerge as stream advancement. Many existing methods of the data stream classification tackle only first two challenges and ignore last two challenges. Here in this paper we proposed an ensemble classification skeleton, in which each classifier is prepared with a novel class detector to tackle the concept drift and concept evolution. We also proposed the feature set homogenization methods for feature evaluation. We improve the component of novel class detection by making it more adaptive to the evolving stream, and enable it to notice more than one novel class at a time. As comparing with the existing methods of the novel class detector method the efficiency of the proposed method is more than the existing one.*

I. INTRODUCTION

In the recent years the concept of data stream classification has been widely studied research problem. As we know that the data stream has the dynamic and evolving nature, for this nature data stream required efficient and effectual methods which are considerably distinct from static data classification methods. Infinite length and concept drift are two most challenging and well studied characteristics of data streams. As we know that the data stream is a fast and continuous phenomenon therefore the data stream is assume to have infinite in length. Therefore practically it is impossible to store and use all the data for the training. The optional for this is the incremental learning techniques. Number of incremental learners had been proposed for solving this problem of data stream by [4] Hulten et al. (2001) and (W. Fan, 2004) [16]. Additionally the concept drift appear in the stream when the concept of underlying of the stream change over time.

Though, the concept evolution and feature evolution are the other two significant characteristics of data streams, these two characteristics are generally ignored in most of the existing methods. When the new classes evolve in the data the concept evolution occurred. Suppose the example of intrusion detection in a network traffic streams and the example of case of text data stream which occurs in the social networking sites like facebook. In the case of second example new classes are recurrently materialize in the underlying stream of the text messages. The problem regarding the concept evolution is noticed in very inadequate way by the presently presented data stream classification methods. In this paper we examine the problem of concept evolution and proposed improved solution. We also focus on the feature evolution problem occurs in the data streams.

In [8] Masud et al (2004) proposed the novel class detection problem in the presence of concept drift and infinite length. In this method for classifying the unlabeled data and for detecting the novel class the method used the ensemble models. The processes of novel class detection method consist of three steps; in the first step at the time of training decision boundary is built. In the second steps the test points which are falling outside the decision boundary is stated as a outlier. And in the third steps the outliers are examined to see if there is sufficient cohesion between them and separation from the existing class instances. However, the author did not address the feature evolution problem. In [9] Masud et al (2009), the problem of feature evolution is addressed, this method also address the problem of concept evolution. Since [8] and [9] have two deficiencies, first the false alarm rate is high for some data sets. Second the method is fails to distinguish between the two novel classes.

Therefore we proposed a superior technique for the both outlier detection and novel class detection for reducing the false alarm rate and for increasing the detection rate. The proposed framework is successfully able to distinguish among the two novel classes.

In the proposed work we assert the four major contributions in the novel class detection for the data streams. First, we proposed a flexible decision boundary for outlier detection by permitting the slack space outside the decision boundary. The allotted space is proscribed by the threshold and the threshold is adapted continuously to decrease the risk of alarm rate and missed novel classes. Second by using the discrete Gini Coefficient we apply the probabilistic approach for detecting the novel class instances. By using this approach, it is possible to distinguish the different causes for the appearances of the outliers which are noise, concept drift and concept evaluation. Third, for detecting the appearance of more than one novel class we apply the graph based approached. Finally, we addressed the feature evaluation problem on the top of the enhancements as discussed above.

II. RELATED WORK

To handle the efficiency and concept drift aspect of the classification process, we discussed some of the existing methods. In [3], was proposed by C.C. Aggarwal, J. Han, J. Wang, and P.S. YU (2006), in this method model is proposed for data stream classification from the point of view of a dynamic approached in which the simultaneously the training and testing stream are used for dynamic classification of data sets. This model replicate real life situation effectively therefore it is enviable to classify test streams in real time over an evolving training and testing streams. In [4] (C.C. Aggarwal, 2009), here they discuss the details of an online voice recognition system, for this purpose micro clustering algorithm which design concise signatures of the target features. One of the shocking and perceptive annotations from our experiences with such a system is that while it was formerly designed only for effectiveness, we later discovered that it was also more accurate than the widely used Gaussian Mixture Model.

In [5] Wang et al. (2003), we proposed a general framework for mining concept drifting data streams by using the weighted ensemble classifier. Train the ensemble classification model like C4.5 , RIPER, naïve Beyesian, etc. from the sequential chunks of the data streams.

In [15] Yang et al. (2005), uses a part of theoretical equivalence to organizing the data history into a history of approach. Onward the journey of concept change, it identifies new concepts as well as reappearing ones, and learns transition patterns among concepts to help prediction. Distinct from conventional methodology that inactively waits until the concept changes, this method included proactive and reactive predictions.

In [7] (J. Kolter and M. Maloof, 2005), we represent the additive skilled ensemble algorithm Add Exp, which is a novel general technique for using any online learner for concept drift. We adjust techniques for examining expert prediction algorithms to demonstrate mistake and loss bounds for a discrete and a continuous version of Add Exp. Finally, we present pruning methods and empirical results for data sets with concept drift.

III. NOVEL CLASS DETECTION: PROPOSED APPROACH

1. Outlier Detection Using Adaptive Threshold

We permit a slack space away from the surface of each hypersphere. The class is considered as the existing class if any test instances falls within this allocated slack space. The slack space is defined by the threshold which is referred as OUTTH. Since, if the threshold value set to be too small then the false alarm rate will go up, and vice versa. Therefore for adjusting the false alarm rate we applied the flexible methods.

2. By using Gini Coefficient the detection of novel class

We evaluate the Gini coefficient G(s), for a random sample of yi is as follows:

$$G(S) = \frac{1}{n}\left(n + 1 - 2\left(\frac{\sum_{i=1}^{n}(n+1-i)yi}{\sum_{i=1}^{n} yi}\right)\right)$$

By examining the following three cases we can come with the threshold for Gini Coefficients to identify the novel class, the three cases are as follows:

- If G(s) greater than $\frac{n-1}{3n}$ then the class declare as the novel class and the tag F-outliers as novel class instances.
- If G(s) is equal to the zero then classify the F-outliers as the existing class instances.

- If G(s) belongs to (0, $\frac{n-1}{3n}$) then filter out the F-outlier falling in the first interval and considered rest of the outliers as the novel class.

3. Simultaneously Multiple Novel Class Detection

The idea behind detecting the multiple novel classes simultaneously is to construct a graph and recognize the connected components in the graph. The numbers of connected components described the total number of novel classes. The basic hypothesis for determining the novel class detection method is as follows: A data point should be nearer to the data points of its own class and farther apart from the data points of the other classes. For example, suppose there are two novel classes, then the separation between the two different novel class instances is higher than the cohesion between the same class instances.

IV. CONCLUSION

We proposed a classification and novel class detection method for the concept drift data streams which tackle four challenges which are infinite length, concept drift, concept evaluation and feature evaluation. Existing class detection method for data streams do not address the problem of feature evaluation or experienced from high false alarm rate and false detection rate in many scenarios. In this paper we converse about the feature space conversion techniques for addressing the feature evaluation problem. After that we recognize two key mechanisms of the novel class detection methods which are outlier detection and identifying the novel class examples which is the prime reason of high error rates in the existing approaches. To overcome this problem we proposed an enhanced method for outlier detection by defining a slack space outside the decision boundary of each classification model, and adaptively changing the slack space based on the uniqueness of the evolving data. We also proposed improved optional approached for identifying novel class examples by using the distinct Gini coefficient and theoretically set up its helpfulness. Finally we proposed a graph based approached for distinguish between multiple novel classes. We apply our technique on several real data streams that experience concept-drift and concept-evolution and achieve much better performance than existing techniques.

REFERENCE

[1] Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda`, "New Ensemble Methods for Evolving Data Streams," Proc. ACM SIGKDD 15th Int'l Conf. Knowledge Discovery and Data Mining, pp. 139-148, 2009.

[2] C.C. Aggarwal, "On Classification and Segmentation of Massive Audio Data Streams," Knowledge and Information System, vol. 20, pp. 137-156, July 2009.

[3] C.C. Aggarwal, J. Han, J. Wang, and P.S. Yu, "A Framework for On-Demand Classification of Evolving Data Streams," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 5, pp. 577-589, May 2006.

[4] G. Hulten, L. Spencer, and P. Domingos, "Mining Time-Changing Data Streams," Proc. ACM SIGKDD Seventh Int'l Conf. Knowledge Discovery and Data Mining, pp. 97-106, 2001.

[5] H. Wang, W. Fan, P.S. Yu, and J. Han, "Mining Concept-Drifting Data Streams Using Ensemble Classifiers," Proc. ACM SIGKDD Ninth Int'l Conf. Knowledge Discovery and Data Mining, pp. 226-235, 2003.

[6] J. Gao, W. Fan, and J. Han, "On Appropriate Assumptions to Mine Data Streams," Proc. IEEE Seventh Int'l Conf. Data Mining (ICDM), pp. 143-152, 2007.

[7] J. Kolter and M. Maloof, "Using Additive Expert Ensembles to Cope with Concept Drift," Proc. 22nd Int'l Conf. Machine Learning (ICML), pp. 449-456, 2005.

[8] M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "Integrating Novel Class Detection with Classification for Concept- Drifting Data Streams," Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML PKDD), pp. 79-94, 2009.

[9] M.M. Masud, Q. Chen, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "Classification and Novel Class Detection of Data Streams in a Dynamic Feature Space," Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML PKDD), pp. 337-352, 2010.

[10] P. Wang, H. Wang, X. Wu, W. Wang, and B. Shi, "A Low- Granularity Classifier for Data Streams with Concept Drifts and Biased Class Distribution," IEEE

Trans. Knowledge and Data Eng., vol. 19, no. 9, pp. 1202-1213, Sept. 2007.

[11] P. Zhang, X. Zhu, and L. Guo, "Mining Data Streams with Labeled and Unlabeled Training Examples," Proc. IEEE Ninth Int'l Conf. Data Mining (ICDM), pp. 627-636, 2009.

[12] S. Chen, H. Wang, S. Zhou, and P. Yu, "Stop Chasing Trends: Discovering High Order Models in Evolving Data," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 923-932, 2008.

[13] S. Hashemi, Y. Yang, Z. Mirzamomen, and M. Kangavari, "Adapted One-versus-All Decision Trees for Data Stream Classification," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 5, pp. 624-637, May 2009

[14] X. Li, P.S. Yu, B. Liu, and S.-K. Ng, "Positive Unlabeled Learning for Data Stream Classification," Proc. Ninth SIAM Int'l Conf. Data Mining (SDM), pp. 257-268, 2009.

[15] Y. Yang, X. Wu, and X. Zhu, "Combining Proactive and Reactive Predictions for Data Streams," Proc. ACM SIGKDD 11th Int'l Conf. Knowledge Discovery in Data Mining, pp. 710-715, 2005.

[16] W. Fan, "Systematic Data Selection to Mine Concept-Drifting Data Streams," Proc. ACM SIGKDD 10th Int'l Conf. Knowledge Discovery and Data Mining, pp. 128-137, 2004.