International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

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

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

Pune University, Pune, Maharashtra, India

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.***

Keywords: Outlier, concept evaluation, novel class detection, concept drift, feature evaluation.

1. 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 [4], [3]. 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 [2] Masud et al 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 [1], the problem of feature evolution is addressed, this method also address the problem of concept evolution. Since [2] and [1] 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.

2. Background

In this section we briefly describe the existing novel class detection method which is proposed in [2]. In this method the classifier contains L classification models, {N1…NL}.

Initially we see the basic definition of novel class and existing class. ***Existing class and novel class***: *Let N is the ensemble of the classification models. The class C is the existing class if somewhat one of the models N_i belongs to the N has trained with the class C. Else C is novel class.*

Data streams are split up into equal size of chunks. By using the ensemble, the data points are most recent data chunks are classified. The data chunk are used for the training purpose when the data points in a chunk become tagged by the user.

The fundamental steps in the novel class detection and the classification is as follows: The outlier detection module initially examined each incoming instances in the data stream, it check whether it is an outlier or not. If the incoming instance is not an outlier then it is restricted as existing class using the majority voting amongst the classifier in the ensemble. If the incoming instance is an outlier then it is momentarily store in the buffer. When the buffer contained adequate instances, then the novel class model is invoked. The instances of the novel class are labeled consequently if the novel class is originated. Otherwise the instances stored in the buffer are considered as an existing class and classified normally by using the ensemble of models.
The ensemble of models is invoked in the both the outlier detection and novel class detection modules. The process of outlier detection utilizes the decision boundary of the ensemble of models to conclude that the instance is outlier or not. During the phase of training the decision boundary is built. To decide whether a novel class has arrived or not, the process novel class detection evaluates the cohesion between the outlier in the buffer and separation of outliers from the existing classes. In the following section we discussed in detail about the training and classification phase.

1. Training phase
   The training data is trained with the k-NN based classifier. Instead of storing raw training data, by using the semi supervised k means clustering the k clusters are built and the cluster summary of each clusters are saved. The stored summary subsists of the centroid, radius and frequency of the data points belongs to each class. The distance among the centroid and the farthest data point in the cluster is equal to the radius of the pseudopoint. After generating the summary the raw data points are discarded. Since the new model is trained it substitutes one of the existing models in the ensemble. By calculating the each model on the latest training data the candidate for substitution is selected, the model is selected with the worst prediction error. These demonstrate that we have accurately L model in the ensemble at the given time.
2. Classification and Novel Class Detection
   The instances in the most unlabeled chunk is inspected by the ensemble model to observe if is outside the decision boundary of the ensemble, and if it is inside the decision boundary then it is classify usually by using the ensemble of the model. Else the model is considered as the filtered outlier. The important hypothesis behind detecting the novel class is that any of the class has the following properties:

Property1: *A data point should be nearer to the data points of its own class and beyond from the data points of the other class.* From the above property if the novel class present in the stream then the instance belonging to the class will beyond from the existing class and close to other novel class instance. Hence the f-outlier is outside the decision boundary they are away from the existing class instances. So the division property for a novel class is content by the f-outliers.

3. Related Works

For handling the efficiency and concept drift aspect of the classification many existing data stream model are designed [5-9], [4], [3], [10-15]. For tackling the problems of infinite length and concept drift all of these methods follow the incremental learning approached. The single model incremental approached is the first approached in which the single model is animatedly maintained with the new data. For example, in [4] the decision tree is incrementally updated with the incoming data, the technique in [5] micro-cluster in the model is incremented with the new data. One more approached is the hybrid batch incremental approach in which by using the batch learning technique each model is build. When the older techniques are become outdated, the older techniques are replaced by the new one [7-10], [16-18]. Like [9], uses the single model for classifying the unlabeled data, where like [8] uses the ensemble model for classifying the unlabeled data. to update the model hybrid approached need very simple operation it is the advantage of the hybrid approach over the single model incremental approached. The proposed model not only focused on the infinite length and concept drift problem but the approached focus on the concept evolution and feature evolution.

In [21], Spinosa et al, for detecting the novel class in data streams author applied the cluster based method. By defining the hyper sphere surrounding all the clusters of the normal data for building normal model of the data by using the clustering. With the stream progression the model is continuously updated. The novel class is declared when any cluster is formed outside the hyper-sphere satisfied the density constraint. This method considers all class as a novel class and only one class as a normal class. As the method communicate to a one class classifier, since the method is not applicable to the multiclass data stream classification. Additionally the method considers that the topological outline of the normal class instance in the feature space in convex.

In [20], Katakis et al. proposed feature selection methods for data streams which have the dynamic feature space. This method subsists of an incremental feature ranking method and has the incremental learning algorithm. In this method, when a novel document appears belonging to class C, initially it checked there is any new word in the document. [19] Proposed the method called as FAE, this method applies the incremental feature selection but their incremental learner is an ensemble of model. In this paper we extend the by work by tackling feature evolution extending the algorithm for dynamic decision boundary and the algorithm for detecting the multiple novel class.

4. 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.

References

[1] 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.

[2] 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.

[3] 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.

[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] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[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] 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.

[16] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Tracking Recurring Contexts Using Ensemble Classifiers: An Application to Email Filtering," Knowledge and Information Systems, vol. 22, pp. 371-391, 2010.

[17] A. 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.

[18] M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data," Proc. IEEE Eighth Int'ln Conf. Data Mining (ICDM), pp. 929-934, 2008.

[19] B. Wenerstrom and C. Giraud-Carrier, "Temporal Data Mining in Dynamic Feature Spaces," Proc. Sixth Int'l Conf. Data Mining (ICDM), pp. 1141-1145, 2006.

[20] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Dynamic Feature Space and Incremental Feature Selection for the Classification of Textual Data Streams," Proc. Int'l Workshop Knowledge Discovery from Data Streams (ECML/PKDD), pp. 102-116, 2006.

[21] E.J. Spinosa, A.P. de Leon F. de Carvalho, and J. Gama, "Cluster- Based Novel Concept Detection in Data Streams Applied to Intrusion Detection in Computer Networks," Proc. ACM Symp. Applied Computing (SAC), pp. 976-980, 2008.