MJRET
Open Access
Multidisciplinary Journal of Research in Engineering and Technology
Journal homepage: www.mjret.in
ISSN:2348 - 6953

A SURVEY ON VARIOUS CLASSIFICATION AND NOVEL CLASS DETECTION APPROACHES FOR FEATURE EVOLVING DATA STREAM

Ms. Najneen Momin[1], Prof. Nitin Hambir [2]
Computer Dept.
Dr. D. Y. Patil School Of Engineering, (Affiliated to Savitribai Phule Pune University)
Pune, India
[1]najmomin@gmail.com, [2]nitin.hambir@dypic.in

Abstract:*The classification of data stream is challenging task for data mining community. Dynamic changing nature of data stream has some difficulties such as feature evolution, concept evolution, concept drift and infinite length. As we know that the data streams are huge in amount, it is impractical to store and use all the data for training. Concept drift occurs when underlying concept changes. Concept-evolution occurs as a result of new classes evolving in the stream. Another important characteristic of data streams, namely, feature evolution, in data stream new features emerge as stream advancement. In this paper we discuss the stream data classification processes and method of these classification techniques. Different authors used different method such as data miner and tree based approach for reduced such types of issues. Ensemble of classifier is used to detect novel classes for feature evolving data stream.*

Keywords*: data stream, novel class, classification technique, outlier.*

1. INTRODUCTION

Data stream classification is more difficult due to its dynamic changing nature as compared to stationary data. First data streams are in infinite length so it is not feasible to use all the historical data for training. For this issue multi-pass learning algorithms are not applicable. Incremental learning approach is well suited for this problem. Second concept drift, when underlying concept changes over time, concept drift occurs. Various techniques have been proposed to address this problem. In order to deal with concept drift, classification model must be updated with recent data. Another characteristic of data stream is concept evolution, when new classes evolve in data concept evolution occurs. In order to deal with this problem

classification model must be able to detect novel classes when they appear. For example intrusion detection in a network traffic. Most important characteristic is feature evolution in which new features (words) emerge and old features fade away.

Ensemble techniques have been more popular than single model [1]. In this technique more than one classifier is used for classification with higher efficiency. Each classifier in the classification model is trained on different data chunks. With the help of advanced data streaming technologies [2], we are now able to collect large volume of data for different application domains. For example credit card transaction, network traffic monitoring etc. the presence of irrelevant and redundant data slows down the learning algorithms [3] [4]. By removing or ignoring irrelevant and redundant feature, prediction performance and computational efficiency can be improved. Multiclass miner works with dynamic feature vector and detects novel classes. It is a combination of OLINDDA and FAE approach. OLINDDA and FAE are used to detect novel classes and to classify data chunks respectively. MCM detects outliers as well as recognize novel class instances. It uses dynamic thresholding and Gini coefficient analysis. It is considered as fastest method for classification of dynamic feature data stream. It filters out majority of outliers and reduces the cost of finding a novel class because cost is proportional to the number of outliers. These approaches fall into two categories: single model classification and ensemble of model classification.

The remainder of the paper is organized as follows: Section II describe related work of stream data classification. Section III describes data stream classification and finally concludes our approach in section IV.

2. RELATED WORK

The challenges of data stream classification are addressed by different researchers in different ways.

2.1 Incremental Learning Approach

Most of the existing data stream classification is addressed to handle concept drift and infinite length problem [5], [6], [7], [8]. Each of these techniques follows some sort of incremental technique. There are two variations of this technique: Single-model incremental approach and hybrid batch incremental approach. In single model incremental approach, model is dynamically maintained with new data. For example, incrementally modifies a decision tree with new data [9]. Second approach is hybrid batch incremental approach, in which batch learning technique is used to build each model. When older models become obsolete, they are replaced by newer models [10], [11]. The hybrid approaches require much simpler operations to update a model as compared with single-model.

2.2 Cluster Based Approach

Another category of data-stream classification technique is cluster based approach which addresses the problem of concept evolution in addition to infinite length and concept drift to detect novel classes in data streams [12]. It defines hyper-sphere for all clusters and continuously updated with stream progression. If any cluster found out of this hyper-sphere and if that cluster have some density then novel class is declared. This approach assumes only one 'normal' class and all other classes as 'novel', so it is not useful for multiclass data stream classification.

2.3 Feature Selection For Data Streams Having Dynamic Feature Space

It consists of an incremental feature ranking method in which whenever new document arrives, first check for new word, if found it is added to library and according to its frequency statistics are updated. Based on these statistics new ranking of words are computed and top N words are selected to update classifier (Navie Bayes or kNN). [13] FAE, which applies incremental feature selection. Classification is done by voting among different models in ensemble. This approach has better performance than above approaches. But this approach uses Lossy-L conversion and doesn't detect novel class.

2.4 Multiclass Classifier And Novel Class Detector

To classify unlabeled data, it uses ensemble of models. During training phase decision boundaries are build. If any instance of test found outside the boundary considered as outliers. If enough outliers are found and if they satisfy cohesion and separation property then considered as novel class. But this approach does not consider feature evolution problem. [14] If more than one novel class found at the same time, it cannot distinguish among them.

3. APPROACHES OF DATA STREAM CLASSIFICATION

This section describes the comparative study of different approaches. Table 1 shows various approaches, methods used in that approach and its demerits.

Methods	Techniques	Demerits
SVSTREAM (Support Vector stream clustering)	In this method cluster labeling approach is used.	Boundary values and outlier noise degrades the performance of clustering.
O-F Miner [12], [13]	It is a combination of OLINDDA and FAE.	It assumes only one normal class so it is not useful for

	OLINDDA works as novel class detector and FAE does classification.	multi-class classification.
DX Miner [15]	It considered dynamic nature of feature space.	Feature space generated can be ambiguous so little data conversion loss occurred.
MineClass [16]	Proposed a solution for concept evolution problem. Detect novel classes automatically.	It does not address the classification problem under dynamic feature sets.
MCM (MultiClass Miner)[17]	It applies as adaptive threshold for outlier detection with ensemble of classification model.	It addresses concept drift as well as concept evolution problem. It gives satisfactory result.

Table 1: Comparison of classification Techniques

4. CONCLUSION AND FUTURE SCOPE

There are various approaches for data stream classification and to address its problems such as infinite length, concept drift, concept evolution and feature evolution. We describe advantages and disadvantages of various existing approaches to address these four problems. Existing classification techniques do not address feature evolution problem very well or experienced from high false alarm rate and false detection rate in many scenario. In the future we would like to extend these techniques to reduce the risk of false alarm and missed novel classes.

ACKNOWLEDGEMENT

I wish to express my sincere thanks to the guide Prof. Nitin Hambir and Head of Department, Prof. Soumitra Das, also Grateful thanks to our PG Coordinator Prof. Pankaj Agarkar and last but not least, the departmental staff members for their support.

REFERENCES

[1]. Mohammad M. Masud,Qing Chen,Latifur Khan,Charu C. Aggarwal "Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams" TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE 2011. pp. 1-14.

[2]. Mohammad M. Masud, Qing Chen, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham "Classification and Novel Class Detection of Data Streams in a Dynamic Feature Space" springer 2010. pp. 337-352.

[3]. Xin Xu, Wei Wang, Guilin Zhang, Yongsheng Yu "An Adaptive Feature Selection Method for Multi-class Classification" IEEE 2011. pp. 225-233.

[4]. Salvador Garcıa, Joaquın Derrac "Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 34, 2012. pp. 417-435.

[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]. 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.

[8]. 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.

[9]. 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.

[10]. 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.

[11]. 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.

[12]. 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.

[13]. 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.

[14]. 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.

[15]. 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.

[16]. M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints," IEEE Trans. Knowledge and Data Eng., vol. 23, no. 6, pp. 859-874, June 2011.

[17]. M.M. Masud, Q. Chen, L. Khan, C. Aggarwal, J. Gao, J. Han, and B.M. Thuraisingham, "Addressing Concept-Evolution in Concept- Drifting Data Streams," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 929-934, 2010.