International Journal of Engineering Research and General Science Volume 4, Issue 3, May-June, 2016
ISSN 2091-2730

Classification And Adaptive Novel Class Detection Of Feature – Evolving Data Streams

[1] Ms. Pranali R Gajbhiye , [2] Prof. Swapnil G Vaidya

2 ME Student , Department of Computer Science & Engineering , BAMU University, SYCET, Aurangabad.

[2] Assistant Professor, Department of Computer Science & Engineering, BAMU University, SYCET, Aurangabad

Abstract*:* —Data stream Classification poses many challenges to the data mining community. Most existing data mining classifiers cannot detect and classify the novel class instances in real-time data stream mining problems like weather conditions, economical changes, and intrusion detection etc., until the classification models are trained with the labelled instances of the novel class. In this thesis, a new approach for detecting "Multiple Novel Class" in data stream mining using "Decision Tree Classifier" that can determine whether an unseen or new instance belongs to a novel class and detection of more than one novel class at a time are proposed. Arrival of a novel class in "Concept-Drift "occurs in data stream mining when new data introduce the new concept classes or remove the old Ones. We have proposed a general Framework for mining concept-drifting data streams using weighted ensemble classifiers is used for this study. We compute Classification model is trained using these weighted ensemble classifiers. Deals with Classification technique for Feature-Evolving Data Stream, thereby helping the users to construct more secure information system. We have proposed immense quantities of high-dimensional data renew the challenges to the state-of-the-art data mining techniques. Analysis of Large (Big) Data and Data mining has become an important study in the field of E-commerce. Data Stream Classification may pose many challenges in the area of Data Mining community. Multiple Novel Class and Decision Tree Classifier approaches to mine feedback comments for Classification.

Keywords: Data stream, concept-evolution, novel class, Outlier, Decision Tree Classifier, Concept drift, data stream mining, Simultaneous multiple novel class.

Introduction

We have proposed the dynamic and evolving nature of data streams requires efficient and effective techniques that are significantly different from static data classification techniques. Two of the most challenging and well-studied characteristics of data streams. Since a data stream is a fast and continuous phenomenon, it is assumed to have infinite length. Therefore, it is impractical to store and use all the historical data for training. The most obvious alternative is an incremental learning technique. Several incremental learners have been proposed to address this problem. 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 in data stream classification.

The contributions of this thesis are:

- We propose to use Concept-evolution occurs when new classes evolve in the data. For example, consider the problem of intrusion detection in a network traffic stream. If we consider each type of attack as a class label, then concept-evolution occurs when a completely new kind of attack occurs in the traffic.
- We propose a superior technique for both outlier detection and novel class detection to reduce both false alarm rate and increase detection rate. Our framework also allows for methods to distinguish among two or more novel classes. We claim four major contributions algorithm to identify dimension rating expresses from feedback comments by applying Data Mining techniques in combination with the Naive Bayes classifier.

We tackle the four research questions by two approaches:

6. Data stream classification techniques address only the first two challenges namely infinite length and concept-drift. The dynamic and evolving nature of data streams requires efficient and effective techniques that are significantly different from

static data classification techniques. Important of data streams are infinite length and concept-drift. It is impractical to store and use all the historical data for training. Each classifier is equipped with a novel class detector, to address concept-drift and concept-evolution. A decision boundary is built during training then test points falling outside the decision boundary are declared as outliers.

7. Data streams, namely, concept-evolution and feature-evolution that are ignored by most of the traditional method technique. Concept-evolution occurs when new classes evolve in the data. The problem of concept-evolution is addressed in only a very limited way by the currently available data stream classification techniques. Address the novel class detection problem in the presence of concept-drift and infinite length. An ensemble of models is used to classify the unlabelled data, and detect novel classes. The new classes will be detected as novel, because our SVM class detection technique detects. If it is found in the unused feature spaces. Our new SVM method including to effectively retrieving data from input data content. In this method using effectively and less timely to retrieve the document.

We propose Superior technique for both outlier detection and novel class detection to reduce both false alarm rate and increase detection rate. A flexible decision boundary for outlier detection by allowing a slack space outside the decision boundary. Improved technique for outlier detection by defining a slack space outside the decision boundary of each classification model. Enabling it to detect more than one novel class at a time. Each classifier is equipped with a novel class detector, to address concept-drift and concept-evolution. We propose a graph-based approach for distinguishing among multiple novel classes. We apply our technique on several real data streams that experience concept-drift and concept-evolution, and achieve significant performance improvements over the existing techniques.

2. Survey CLASSIFICATION

The dynamic and evolving nature of data streams pose special challenges to the development of effective and efficient algorithms. Two of the most challenging characteristics of data streams are its infinite length and concept-drift. Since a data stream is a high volume phenomenon, which can be considered infinite in length, it is impractical to store and use all the historical data for training. Several incremental learners have been proposed to address this problem. In addition, concept-drift occurs in the stream when the underlying concepts of the stream change over time.

a) Addressing Concept-Evolution in Concept-Drifting Data Streams

To have an existing classification and novel class detection technique. First, we pro-pose an improved technique for outlier detection by defining a dynamic slack s pace outside the decision boundary of each classification model. Second, we propose a better alternative for identifying novel class instances using discrete Gini Coefficient. Finally, we propose a graph-based approach for distinguishing among multiple novel classes. We apply our technique on several real data streams that experience concept-drift and concept-evolution, and achieve significant performance improvements over the existing techniques. In the future, we would like to extend our technique to text and multi-label stream classification problems.

b) **A Framework for O n-Demand Classification of Evolving Data Streams**

An interesting frame-work for online classification of dynamically evolving data streams. The new framework has been designed carefully based on our analysis and reasoning and has been tested based on our experiments on a real intrusion detection data set. As evidenced by the empirical results, the system developed here is able to provide significantly better results than a static classification model on classification accuracy. In addition, it is efficient and scalable at handling large data stream.

C) **Classification an d Novel C lass Detection of Data Streams in a Dynamic Feature Space**

A novel technique to detect new classes in concept-drifting data streams having dynamic feature space. Most of the existing data stream classification techniques either cannot detect novel class, or does not consider the dynamic nature of feature spaces. We have analytically demonstrated the effectiveness of our approach, and empirically shown that our approach outperforms the state-of-the art data stream classification techniques in both classification accuracy and processing speed. In the future, we would like to address the multi-label classification problem in data streams.

Related work divided into three main areas:

1) Data Classification

2) Outlier detection and filtering

3) Enhance data Detection

1) Data Classification

In literature [8]-[10], The classification process can enable searches based on values that go much further than standard attributes such as filename or creation date. Information can aid in compliance or business intelligence searches. As proposed in [10], In our process select document to classification .Classification is based on document size wise classified. In this method data classification two or three split document .Split process document size calculation after split. Model is provided in [11] to create a document structure create process. The comprehensive overview of trust model is provided in [11]. Individual level trust models aims to compute the reliability of peers and assist buyers in their work of decision making [12]–[14]. To regulate the behaviour of peers, avoid fraudsters and ensure system security was the system level models aim [11].

2) Data verify and Detection

[10], [15], [16], [17] examined analysing feedback the historical data for training since it would require infinite storage and running time. In our data document increasing line and also unwanted symbols are detecting for this process. The data stream is divided into equal sized. The latest chunk, which is unlabelled, is provided to the algorithm as input. Unwanted symbols search method split out all data searching symbols .In this process using to decrees document size. This modules main object for symbols remove to data size automatically decreases. The main focus of [10] and [16] was sentiment classification of feedback comments. It is proved that feedback comments feedback comments. It is proved that feedback comments are noisy and hence analysing them is a challenge. [10] States that the missing aspect comments are deemed negative. Models built from aspect ratings are used to classify comments into positive or negative. [16] Proposed a technique for summarizing feedback. It aims at to filter out courteous comments that do not provide real feedback.

8. Model Analysis

We view feedback comments to use the same feature set for the entire stream, which had been selected for the first data chunk .This will make the feature set fixed, and therefore all the instances in the stream, whether training or testing, will be mapped to this feature set. We call this a loss conversion because future models and instances may lose important features due to this conversion.

Attest instance x is to be classified using a model MI, both the model and the instance will convert their feature sets to the union of their feature sets. We call this conversion “loss-less homogenizing” since both the model and the test instance preserve their dimensions, and the converted feature space becomes homogeneous for both the model and the test instance. Therefore, no useful features are lost as a result of the conversion.

Our experimental results, which show that Loss-L conversion misclassifies most of the novel class instances as existing class. It might appear to the reader that increasing the dimension of the models and the test instances may have an undesirable side effect due to curse of dimensionality. However, it is reasonable to assume that the feature set of the test instances is not dramatically different from the feature sets of the classification models because the models usually represent the most recent concept. Therefore, the converted dimension of the feature space should be almost the same as the original feature spaces. Furthermore, this type of conversion has been proved to be successful in other popular classification techniques such as Support Vector Machines.

Dataset	Method	ERR	M_{new}	F_{new}	AUC	FP	FN
Twitter	DXMiner	4.2	30.5	0.8	**0.887**	-	-
	Lossy-F	32.5	0.0	32.6	0.834	-	-
	Lossy-L	1.6	82.0	0.0	0.764	-	-
	O-F	3.4	96.7	1.6	0.557	-	-
ASRS	DXMiner	**0.02**	-	-	**0.996**	**0.00**	**0.1**
	DXMiner(info-gain)	1.4	-	-	0.967	0.04	10.3
	O-F	3.4	-	-	0.876	0.00	24.7
Forest	DXMiner	**3.6**	**8.4**	1.3	**0.973**	-	-
	O-F	5.9	20.6	1.1	0.743	-	-
KDD	DXMiner	**1.2**	**5.9**	**0.9**	**0.986**	-	-
	O-F	4.7	9.6	4.4	0.967	-	-

Most of the existing data stream classification techniques either cannot detect novel class, or does not consider the dynamic nature of feature spaces. We have analytically demonstrated the effectiveness of our approach, and empirically shown that our approach outperforms the state-of-the art data stream classification techniques in both classification accuracy and processing speed. In the future, we would like to address the multi-label classification problem in data streams. We make use of two types of lexical knowledge to \supervise" grouping dimension expressions to dimensions so roduningful clusters.

- Comments are short and therefore co-occurrence of head terms in comments is not very informative. We instead use the co-occurrence of dimension expressions with respect to a same modifier across comments, which potentially can provide more meaningful contexts

Support vector machines are considered a must try it offers one of the most robust and accurate methods among all well-known algorithms. In addition efficient methods for training SVM are also being developed at a fast pace. The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data.

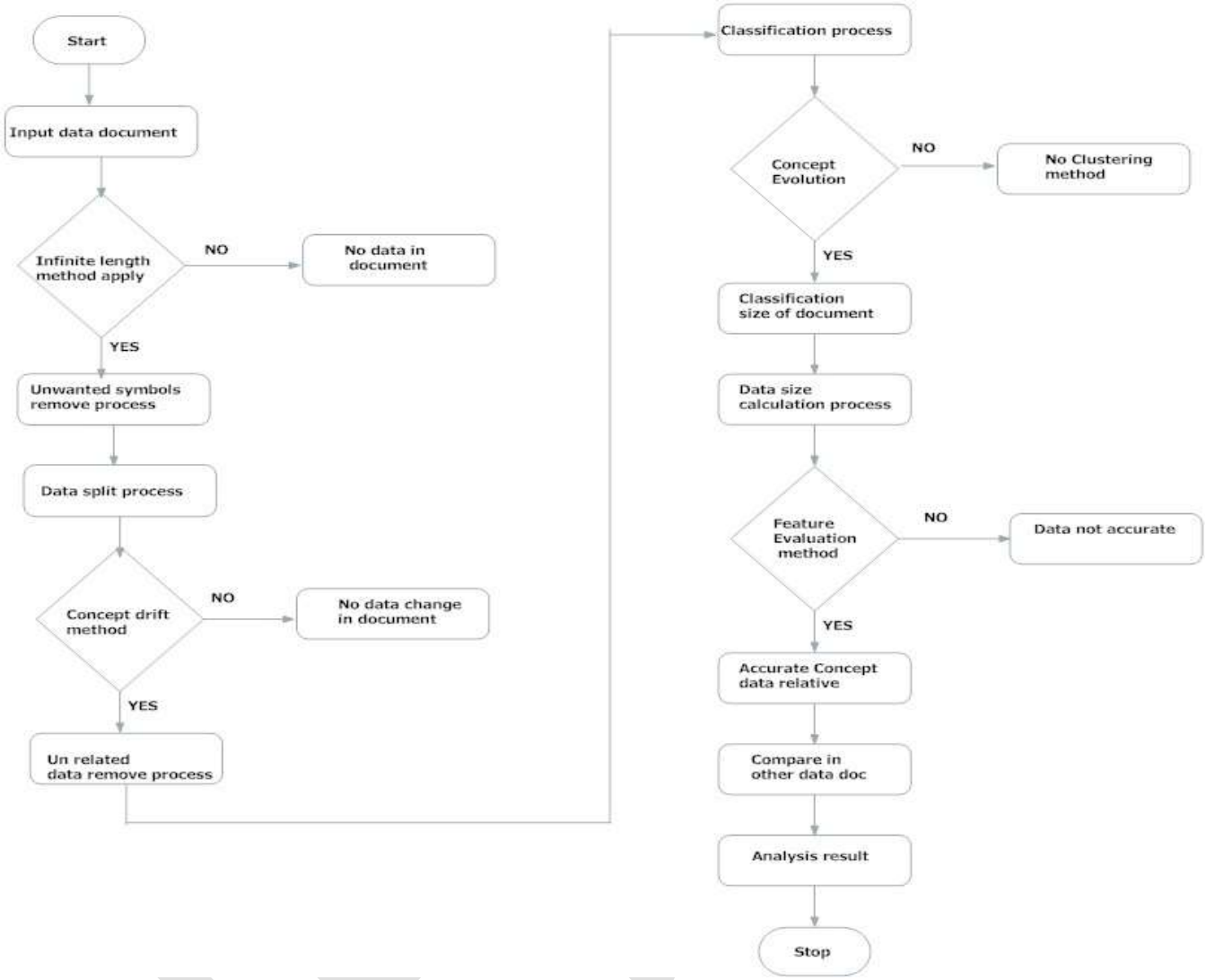

Figure 27: Classification Framework

The metric for the concept of the best classification function can be realized geometrically. A linear classification function corresponds to a separating hyper plane that passes through the middle of the two classes, separating the two.

4) **Outlier detection**

A test instance is identified as an F -outlier if it is outside the

Radius of all the pseudo points in the ensemble of models. Therefore, if a test instance is outside the hyper sphere of a pseudo point, but very close to its surface, it will still be an outlier. However, this case might be frequent due to concept-drift or noise; existing class instances may be outside and near to the surface of the hyper sphere. As a result, the false alarm rate detecting existing classes as novel would be high. In order to solve this problem, we follow an adaptive approach for detecting

The outliers. We allow a slack space beyond the surface of each hyper sphere. If any test instance falls within this slack space, it is considered as existing class instance. This slack space is defined by a threshold, OUTTH. We apply an adaptive technique to adjust the threshold. First, we explain how the threshold is used.

4. Outlier Detection

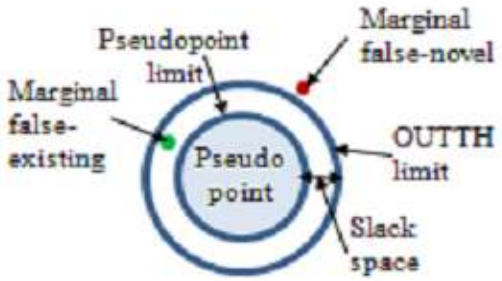

Figure 2: Outlier Phase

The F -outliers detected during the outlier detection phase may occur because of one or more of the three different reasons: noise, concept-drift, or concept-evolution. In order to distinguish the F -outliers that occur because of concept-evolution only, we compute a metric called discrete Gini Coefficient of the F -outlier instances. We show that if the Gina Coefficient is higher than a particular threshold, then we can be confident of the concept-evolution scenario.

It is possible that more than one novel class may arrive at the same time in the same chunk. This is a common scenario in text streams, such as Twitter messages. The number of connected components determines the number of novel classes. The basic assumption in determining the multiple novel classes follows from the cohesion and separation property. For example, if there are two novel classes, then the separation among the different novel class instances should be higher than the cohesion among the same-class instances.

5. Conclusion

We have proposed a novel technique to detect new classes in concept-drifting data streams having dynamic feature space. Most of the existing data stream classification techniques either cannot detect novel class, or does not consider the dynamic nature of feature spaces. We have analytically demonstrated the effectiveness of our approach, and empirically shown that our approach outperforms the state-of-the art data stream classification techniques in both classification accuracy and processing. Data streams, namely, concept-evolution and feature-evolution that are ignored by most of the traditional method technique. Concept-evolution occurs when new classes evolve in the data. The problem of concept-evolution is addressed in only a very limited way by the currently available data stream classification techniques. Address the novel class detection problem in the presence of concept-drift and infinite length. An ensemble of models is used to classify the unlabeled data, and detect novel classes. The new classes will be detected as novel, because our SVM class detection technique detects. If it is found in the unused feature spaces. Our new SVM method including to effectively retrieving data from input data content. In this method using effectively and less timely to retrieve the document.

REFERENCES:

[1] C.C. Aggarwal, "On Classification and Segmentation of Massive Audio Data Streams," Knowledge and Information System".

[2] 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.

[3] 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.

[4] 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.

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

[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] 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] G. Hulten, L. Spencer, and P. Domingos, "Mining Time-Changing Data Streams," Proc. ACM SIGKDD Seventh Int'l Conf. Knowl.

[9] 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.

[10] 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.

[11] 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