A Review of Classification and Novel Class Detection Technique of Data Streams

Manish rai[1]
Department of Computer Science & Engineering, LNCT, Bhopal

Rekha Pandit[2]
Department of Computer Science & Engineering, LNCT, Bhopal

ABSTRACT

Stream data classification suffered from a problem of infinite length, concept evaluation, feature evaluation and data drift. Data stream labeling is more challenging than label static data because of several unique properties of data streams. Data streams are suppose to have infinite length, which makes it difficult to store and use all the historical data for training. Earlier multi-pass machine learning technique is not directly applied to data streams. Data streams discover concept-drift, which occurs when the discontinue concept of the data changes over time. In order to address concept drift, a classification model must endlessly adapt itself to the most recent concept. Various authors reduce these problem using machine learning approach and feature optimization technique. In this paper we present various method for reducing such problem occurred in stream data classification. Here we also discuss a machine learning technique for feature evaluation process for generation of novel class.

Keywords: Stream Data classification, data drift, novel class

I. INTRODUCTION

Stream data classification plays important role in the field of data mining. The need and requirement of online transaction of data is stream classification, due to stream classification save time of computation and storage area of network. For the purpose of stream data classification various machine learning algorithm are applied, such as clustering, classification, and regression. Two of the most critical and well generalized problems of data streams are its infinite length and concept-drift. Since a data stream is a fast and continuous event, it is assumed to have infinite length. Therefore, it is difficult to store and use all the historical data for training. The most discover alternative is an incremental learning technique. Several incremental learners have been proposed to address this problem [8], [5]. 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 [2], [6], [7] in data stream classification. However, there are two other significant characteristics of data streams, such as concept evolution and feature evolution that are ignored by most of the existing techniques. 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 happen when a completely new kind of attack take place in the traffic. Another example is the case of a text data stream, such as that occurring in a social network such as Twitter. In the classification process of an developing data stream, either the short-term or long-term behavior of the stream may be more important, or it often cannot be known a priori as to which one is more significant. How do we determine the window or horizon of the training data to use so as to obtain the best classification accuracy? While techniques namely decision trees are useful for one-pass mining of data streams, these cannot be easily used in the context of an on-demand classifier in an developing environment. This is because such a classifier needed rapid variation in the horizon selection process due to data stream evolution. In this respect, nearest-neighbor classifiers tend to be more amenable to quick horizon adjustments because of their easiness. However, it is still too costly to keep track of the entire history of the data in its original precise granularity. Therefore, the on-demand classification process still requires the appropriate machinery for efficient and adjustable statistical data collection in order to perform the corresponding operations in an efficient way. The above section discuss introduction of stream data classification. In section II we discuss various proposed method for stream data classification. in section III conclude the paper.

II. Method for Stream Data classification

In this section we discuss method for stream data classification for minimization and removal a problem such as infinite length, data drift, concept evaluation and feature evaluation. All these method reduce such problem, Yan-Nei Law and Carlo Zanily entitled" An Adaptive Nearest Neighbor Classification Algorithm for Data Streams" [2] describe a process of stream data classification by adaptive nearest classification as the algorithm achieves excellent performance by using small classifier ensembles where approximation error bounds are guaranteed for each ensemble size. The very low update cost of our incremental classifier makes it highly suitable for data stream applications. ANNCAD is very suitable for mining data streams as its update speed is very quick. Also, the accuracy compares favorably with existing algorithms for mining fact streams. ANNCAD adapts to concept drift efficaciously by the exponential bury approach. However, the very detection of sudden concept drift is of interest in many applications. The ANNCAD framework can also be extended to detect concept drift,for example changes in class label of blocks is a good indicator of possible concept drift.

Mohammad M. Masud, Latifur Khan, Bhavani Thuraisingham entilied "Classification And Novel Class Detection In Concept-Drifting Data Streams Under Time Constraints"[3] describe a process of stream data classification by novel class detection In Concept-Drifting Data Streams Under Time Constraints as Novel class detection problem becomes more challenging in the presence of concept drift,when the underlying data distributions develop in streams. In order to determine whether an instance belongs to a Novel class, the classification models sometimes require to wait for more test instances to discover similarities among those instances. A maximum allowable wait time Tc is imposed as a time constraint to classify a test instance. Moreover most existing stream classification approaches assume that the true label of a data point can be accessed immediately after the data point is classified. In realness, a time delay Tl is involved in obtaining the true label of a data point since manual labeling is time overwhelming. We show how to make fast and accurate classification decisions under these constraints and apply them to real benchmark data. Comparing with state of the art stream classification techniques proves the superiority of our approach. The concept evolution problem, which has been neglect by most of the existing data stream classification techniques. Existing data stream classification techniques assume that total number of classes in the stream is set. Therefore instances belonging to a novel class are misclassified by the currently techniques. We show how to detect novel classes automatically even when the classification model is not trained with the novel class instances. Novel class detection becomes more challenging in the presence of concept-drift.

Mohammad M. Masud, Qing Chen, Latifur Khan, Charu Aggarwal and Jing Gao , Jiawei Han and Bhavani Thuraisingham [4]"Addressing Concept-Evolution in Concept-Drifting Data Streams **"**in this title author describe a process of stream data classification by Concept-Evolution in Concept-Drifting Data Streams as Concept-evolution occurs as a result of new classes evolving in the stream. This method addresses concept-evolution in addition to the existing challenges of infinite-length and concept-drift. The concept-evolution phenomenon is studied and the insights are used to construct a superior novel class detecting techniques. Firstly, we suggest an adaptive threshold for outlier detection, which is a vital part of novel class detection. Secondly, we suggest a probabilistic approach for novel class detection using discrete gini Coefficient and this prove its effectiveness both theoretically and empirically. Finally, address the issue of simultaneous multiple novel class occurrence and give an refined solution to detect more than one novel class simultaneously. We also consider feature evolution in text data streams which occurs because new features (i.e., words) evolve in the stream data classification. Comparison with state of the art data stream classification techniques establishes the effectiveness of the propose approach we propose an improved technique for outlier detection by defining a dynamic slack space outside the decision boundary of each classification pattern. Secondly we suggest 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.

Valerio Grossi, Alessandro Sperduti "Kernel-Based Selective Ensemble Learning for Streams of Trees"[5] author proposed a process of stream data classification by Kernel-Based Selective Ensemble Learning as Kernel methods enable the modeling of structured data in learning algorithms, still they are computationally demanding. Both efficacy and efficiency of the proposed approach are assessed for different models by using data sets exhibiting different levels and types of concept drift. Kernel methods provide a powerful tool for modeling structured objects in learning algorithms. Unfortunately, they require a high computational complexity to be used in streaming environments. This work is the first that demonstrates how kernel methods can be employed to define an ensemble approach able to quickly react to concept drifting and guarantees an efficient kernel computation.

Li Su Xi, Hong-yan Liu, Zhen-Hui Song. "A New Classification Algorithm for Data Stream"[6] in this method describe a process of stream data classification by Associative classification (AC) as Associative classification (AC) which is based on association rules has shown great promise over many other classification techniques on static dataset. Meanwhile, a new challenge has been proposed in that the increasing prominence of data streams arising in a wide range of advanced application. This technique describes and evaluates a new associative classification algorithm for data streams which is based on the estimation mechanism of the lossy Counting (LC) and landmark window model. And this technique was applied to mining several datasets obtained from the UCI Machine Learning Repository and the result show that the algorithm is effective and efficient.

Clay Woolam, Mohammad M. Masud, and Latifur Khan "Lacking Labels in the Stream: Classifying Evolving Stream Data with Few Labels" [7]in this title author describe a process of stream data classification by Evolving Stream Data with Few Labels as It is practical to assume that only a small fraction of instances in the stream are tagged. A more practical assumption would be that the labeled data may not be independently distributed among all train documents. How can we ensure that a good classification model would be built in these scenarios, considering that the data stream also has changing nature? In our previous work we apply semi-supervised clustering to build classification models using limited amount of labeled train data. However, it assumed that the data to be labeled should be chosen randomly.

Mohammad M. Masud, Qing Chen, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham "[8]Classification and Novel Class Detection of Data Streams in a Dynamic Feature Space" here author describe a process of stream data classification by DXMiner as DXMiner, which addresses four major challenges to data stream classification such as infinite length, concept-drift, concept-evolution, and feature evolution. Data streams are assumed to be limitless in length which demands single pass incremental learning techniques. Concept-drift occurs in a data stream when the underlying concept alteration over time. Most presenting data stream classification techniques address only the infinite length and concept drift problems. still, concept-evolution and feature evolution are also major challenges, and these are neglect by most of the presenting approaches. Concept-evolution occurs in the stream when novel classes arrive, and feature-evolution occurs when new features emerge in the stream and old features fade away. Our previous work addresses the concept evolution problem in addition to addressing the infinite length and concept-drift problems. DXMiner considers the dynamic nature of the feature space and provides an elegant solution for classification and novel class detection when the feature space is dynamic. We show that our approach outperforms state-of-the-art stream classification techniques in classifying and detecting novel classes in real data streams. Most of the existing data stream classification techniques either cannot detect novel class, or does not consider the Dynamic nature of feature spaces.

Charu C. Aggarwal ,Jiawei Han, Jianyong Wang, Philip S. Yu "A Framework for On-Demand Classification of Evolving Data

Streams],This model indicate real-life situations effectively, since it is desirable to classify test streams in real time over an evolving training and test stream. The objective here is to make a classification system in which the training model can adapt quickly to the changes of the underlying data stream. In order to achieve this goal, we propose an on-demand classification process which can dynamically select the appropriate window of past training data to build the classifier. The empirical results show that the system maintains high classification accuracy in an developing data stream, while providing an efficient solution to the classification task. The stream classification framework proposed in this study has the following fundamental differences from the previous stream classification work in design philosophy. First, due to the dynamic nature of evolving data streams.

III. CONCLUSION

In this paper we review a various method of stream data classification and handling a problem during stream classification, such as infinite length of data, feature evaluation, concept evaluation and storage efficiency of stream data. The method of stream data classification generates a drift in case of stream. The garneted drift discovers a problem of computational efficiency and rate of classification. The method such as general purpose programming and problastic reduced the infinite length and drift problem. Also all these methods take a process of optimization for the resolution of feature evaluation problem. Mining data streams is still in its early state. Addressed along with open issues in classification data stream mining are discussed in this paper. Furthermore evolution would be realised over the next few years to address these problems. Having these systems that address the above research issues create, that would speed up the science discovery in physical and astronomical applications in addition to business and financial ones that would improve the real-time decision making process.

References:-

[1]Urvesh Bhowan, Mark Johnston, Mengjie Zhang and Xin Yao "Evolving Diverse Ensembles using Genetic Programming for Classification with Unbalanced Data" in IEEE Tansaction2010.

[2]Yan-Nei Law and Carlo Zanily entitled" An Adaptive Nearest Neighbor Classification Algorithm for Data Streams" in PKDD 2005, LNAI 3721, pp. 108–120, 2005.

[3]Mohammad M. Masud, Latifur Khan, Bhavani Thuraisingham entilied "Classification And Novel Class Detection In Concept-Drifting Data Streams Under Time Constraints" in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 6, JUNE 2011

[4]Mohammad M. Masud, Qing Chen, Latifur Khan, Charu Aggarwal and Jing Gao , Jiawei Han and Bhavani Thuraisingham "Addressing Concept-Evolution in Concept-Drifting Data Streams " in IEEE Transaction 2010.

[5]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.

[6]Li Su Xi, Hong-yan Liu, Zhen-Hui Song. "A New Classification Algorithm for Data Stream"

[7]Clay Woolam, Mohammad M. Masud, and Latifur Khan "Lacking Labels in the Stream: Classifying Evolving Stream Data with Few Labels" in *I.J.Modern Education and Computer Science engg,* 2011

[8]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" in ISMIS 2009, LNAI 5722, pp. 552

[9]Charu C. Aggarwal ,Jiawei Han, Jianyong Wang, Philip S. Yu "A Framework for On-Demand Classification of Evolving Data Streams" in ECML PKDD 2010, Part II, LNAI 6322, pp. 337–352,