A SURVEY ON TECHNIQUES FOR NOVEL CLASS DETECTION IN DATA STREAM

Megha Patel, Neha Gupta
Computer Engineering Department,
Silver Oak College of Engineering & Technology,
Gujarat Technological University, India

***Abstract*- Data stream mining is a process of extracting the knowledge structure from continuously coming rapid data records. Data stream can be viewed as an ordered sequence of instances arrives at any time. Classification of data stream is more difficult task. There are three main problems in data stream classification: infinite length, concept drift and concept evolution or novel class detection. Researches available for first two problem, but novel class detection had not taken more attention. In this paper, different technique for novel class detection and its comparative analysis is presented.**

***Index Terms*- Data stream, Novel class, Ensemble learning, Incremental learning, Decision tree**

I. INTRODUCTION

Data mining is the process of collecting, searching through and analyzing a large amount of data in a database, as to discover the patterns or relationships. Applications - such as network monitoring, security, sensor network, telecommunication data management, web applications, financial applications and others - in which data is generated at very high rates in the form of transient *data streams*. Data stream is a sequence of tuples that appears continuously at any time, doesn't permit to store them permanently into the memory. This generation of continuous stream of information has challenged our storage, communication and computation potential in computing system [1].

The traditional data mining techniques cannot be directly apply to data stream because they require multiple scans of data to extract the information which is unrealistic with data stream. Data stream mining is the process of extracting information form continuous, rapid data records. Mining task includes association mining, classification and clustering. Classification extracts the information and knowledge form continuously coming data instances. It maps data into predefined classes that is supervised learning because classes are determined before examining data that analyzes the training set and builds the model for each class according the feature present in the data. In clustering, classes are not predefined, but defined by data alone, referred to as unsupervised learning. Classifier is used to predict the class value for unseen new instances whose attribute value is known but class value is unknown [2].

There are three major challenges in stream classification [3].

1. Data stream is theoretically infinite in length, so it is not possible to store and use all the historical data for training, since it would require infinite storage and running time.
2. Concept drift occurs as an underlying concept of the data may change over time.
3. Concept evolution occurs when a novel class may appear in the data stream.

In data stream classification most of the existing wok related to first two issues, later one not concentrated. Here we focus on the third issue novel class detection. Existing solution assumes that the total numbers of classes are fixed and traditional classifier only correctly classifies instances of those classes that have been trained. When new class appears in the stream, all the instances belongs to that class are misclassified until new class has been manually identified by some expert and new model is trained with the labeled instances of that class, but in real world, novel classes may arrive at any time from continuous data stream. Data mining classifier should update continuously so that it reveals the most recent concept. There are different approaches to develop the classification model using decision trees, neural networks, rule-set based methods and nearest neighbor methods [7].

Stream classifiers are divided into two categories: single model and ensemble model [2]. Single model

incrementally update a single classifier and effectively responds to arrival of new class so that reflects most recent concept in data stream effectively and efficiently so it is more attractive. It is also called as incremental learning approach. This approach is beneficial to deal with the classification task when datasets are too large or when new examples can arrive at any time [8]. Incremental algorithms can be defined as follow: A learning task is incremental if the training examples used to solve it become available over time, usually one at a time [9].

Ensemble algorithms are sets of single classifiers (components) whose decisions are aggregated by a voting rule. The combined decision of many single classifiers is usually given by a single component. Studies show that to obtain this accuracy boost, it is necessary to diversify ensemble members from each other. Components can differ from each other by the data they have been trained on, the attributes they use, or the base learner they have been created from. For a new example, class predictions are usually established by member voting. Ensemble training is a costly process. It requires at least k times more processing than the training of a single classifier, plus member example selection and weight assignment usually make the process even longer. In massive data streams, single classifier models can perform better because they might not require the time for running and updating an ensemble. Ensemble Model are popular method for learning on stationary data, they are not able to give an explicit interpretation of change made to the model [9].

There are many techniques for novelty detection. MineClass provides the solution for novel class detection. ActMiner expands MineClass and works on less labeled instances. ECSMiner detect novel class automatically even when the classification model is not trained with the novel class instances. SCANR is more realistic technique for novel class detection, it identifies the recurring class and declare it as "not novel" when it reappears after long disappearance. Decision tree classifier work on incremental learning approach for handling novel class detection problem. ID3 and C4.5 learning algorithms used for novelty detection [2].

II. CLASSIFICATION AND NOVEL CLASS DETECTION

Novel class detection in stream data classification is interesting research topic and researches available for concept drift problem but not for novel class detection. This approach fall into two categories: Single model (Incremental approach), Ensemble Model. Data stream classification and novelty detection received increasing attention in many real-world applications, such as spam detection, climate change or intrusion detection, where data distributions inherently change over time [2].

Definition 1 (*Existing class and Novel class*): Let L be the current ensemble of classification models. A class c is an existing class if at least one of the models Li ∈ L has been with the instances of class c. Otherwise, c is a novel class[3][4].

To detect a novel class that has the following essential property:

Property 1: A data point should be closer to the data points of its own class (*cohesion)* and farther apart from the data points of any other classes (*separation*) [3].

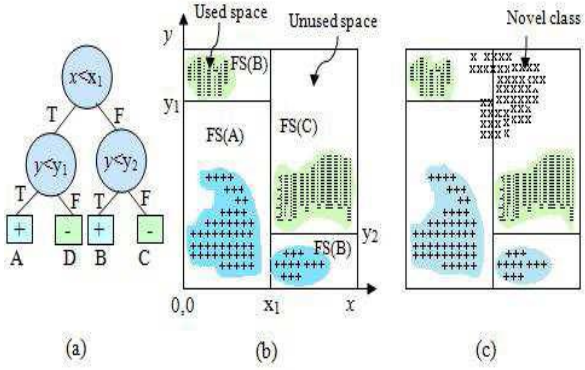

Figure 1: (a) A decision tree, (b) corresponding feature space partitioning where FS(X) denotes the Feature space defined by a leaf node X The shaded areas show the used spaces of each partition. (c) A Novel class (denoted by x) arrives in the unused space [3].

Fig 1. shows the basic idea of novel class detection using decision tree. The notion of used space denotes a feature space occupied by any instance, and unused space to denote a feature space unused by an instance. According to *property 1*(*cohesion*), a novel class must arrive in the unused spaces. There must be strong cohesion among the instances of the novel class. Two basic steps for novel class detection. First,

the classifier is trained such that an inventory of the used spaces is created and saved, done by clustering and saving the cluster summary as "pseudo point". Secondly, these Pseudo points are used to identify outliers in the test data, and declare a novel class if there is strong cohesion among the outliers.

III. RELATED WORK

There are two categories for novel class detection: statistical and neural network based. Statistical approach has two types: parametric and non-parametric. Parametric approaches assume that data distributions are known (e.g. Gaussian), and try to approximate the parameters (e.g. mean and variance) of the distribution. If any test data falls outside the normal parameters of the model, it is declared as novel; whereas non-parametric approach estimates the density of training data and rejects data whose density is beyond a certain threshold.

Traditional stream classification techniques are incapable of detecting the novel class instances until the emergence of novel class is manually identified, and labeled instances of that class are presented to learning algorithm for training. The problem becomes more challenging when the underlying concept of data distribution changes over time. A novel and efficient technique that automatically detect the novel class in the presence of concept drift by quantifying the cohesion among unlabeled test instances and separation of test instances from training instances [3]. Here the algorithm MineClass [3], which stands for Mining novel Classes in data streams, is non-parametric, meaning, it does not assumes any underlying distributions of data. It is based on ensemble learning approach in which decision tree and K-NN classifiers are used for novelty detection. There are two basic steps for novel class detection. First, the classifier is trained such that an inventory used space is created and saved which is done by clustering and saving the cluster summary as "psudopoint". Secondly, these psudopoints are used to detect outliers in the test data, declare as novel class if there is a strong cohesion among the outliers. This approach is able to detect the novel classes in the presence of concept-drift, and even when the model consists of multiple existing classes, but it requires 100% labeled instances.

ActMiner [4], which stand for Active Classifier for Data Streams with novel class Miner, is non-parametric ensemble learning based approach which extends MineClass, and addresses the limited labeled data problem. MineClass requires all the instances in data stream be labeled by human experts and become available for training, which is impractical, since data labeling is time consuming and costly. ActMiner selects only those data points for labeling for which the expected classification error is high thereby reducing the labeling cost. It applies to active learning, but the data selection process is different from others. And it is not applicable to multi-label classification and dynamic feature set problem.

ECSMiner [5], stands for Enhanced Classifier for data Streams with novel class Miner, pronounced as "ExMiner", based on ensemble approach which is different from traditional novelty detection technique. It is a nonparametric approach, not restricted to any specific data distribution. ECSMiner is different from other technique, it is not only considers difference of test instances from training data but also similarities among them. It is "multiclass" novelty detection technique; discover the appearance of a novel class, even if there is a concept-drift. This technique is applied on two different classifiers: decision tree and k-nearest neighbor. When decision tree is used as a classifier, each training data chunk is used to build a decision tree; whereas K-NN strategy would lead to an inefficient classification model, both in terms of memory and running time. Even when the classification model is not trained with the novel class instances, ECSMiner can detects novel class automatically.

SCANR [6], stands for Stream Classifier And Novel and Recurring class detector, ensemble based novelty detection technique which detect both the novel and recurring class. A recurring class is a special case of concept evolution, which take place when a class appears in the stream, then disappears for a long time and again appears. ECSMiner wrongly detect the recurring class as novel class which creates two main problems; first much resource are wasted in detecting recurring class as novel, because novel class detection is computationally and memory intensive, as compared to simply distinguishing an existing class; second it increase the human effort, in cases where the output of classification is used by human analyst. SCANR is more realistic novel class detection technique, which remember a class and declare it as "not novel" when it reappears after long

disappearance and greatly reduce the human effort. Each incoming instances is first examine by the *outlier detection* module of primary ensemble to check whether it is outlier; if it is not outlier, it is classified as an existing class using majority voting among the classifiers of primary ensemble. If is outlier, called *primary outlier*, which is again examined by outlier detection module of auxiliary ensemble; if it is determined as not outlier by auxiliary ensemble then it is recurring class instance and classified by auxiliary ensemble, otherwise it is called *secondary outlier* and stored in buffer for further analysis, and novel class detection module is called when there are enough instances in the buffer. As this technique reduces human effort and false alarm rate, as well as total error rate, it requires extra running time for auxiliary ensemble.

Decision tree [2], classifier is an incremental learning approach for detecting the novel class in concept drifting data stream. This approach builds decision tree from training data points and calculate the threshold value based on the ratio of percentage of data points between each leaf node and training datasets based on similarity of attribute value. If number of data points classify by leaf node in the tree increases than the threshold value, means novel class may arrived. Then compare new points with existing data points based on similarity of attribute value; if it is different from existing data points and new data points doesn't belongs to any cluster, confirms the novel class arrived and rebuild the decision tree which is continuously updates with recent data points so it represents the most recent concept in the data stream. ID3 (Iterative Dichotomiser) technique builds DT using information theory to choose best splitting attributes from a dataset with the highest information gain. The C4.5, successor of ID3 uses the largest GainRatio and improves the performance of building tree using boosting. CART (Classification And Regression Trees) generates the binary tree for decision making and handles missing data and contains pruning strategy. The SPRINT (Scalable Parallelizable Induction of Decision Tree) algorithm uses an impurity function based on gini index to find best splitting attribute. Decision tree classifier is very popular supervised learning algorithm which is easy to implement and require little prior knowledge, but not address this problem under dynamic attribute sets.

IV. COMPARISION OF DIFFERENT TECHNIQUES FOR NOVEL CLASS DETECTION

A comparison of the various algorithms approaches and limitations that have been defined in various research publications have been given in this section.

Table 1 summarizes the comparison of different existing Approaches

No.	Name of Algorithm	Learning Approach	Limitation	Idea of Improvement
1.	MineClass[3]	Ensemble	Requires 100% labeled instances	Automatically detect novel class in the presence concept-drift
2.	ActMiner[4]	Ensemble	Not address dynamic feature set problem and multi-label classification problem	It works on less labeled instances and reduces labeling cost
3.	ECSMiner[5]	Ensemble	Can't identifies recurring class and inefficient in terms of memory and run time	Multiclass novelty detection technique. Detect novel class even when classification model is not trained with novel class instances.
4.	SCANR[6]	Ensemble	Requires extra running time for auxiliary ensemble	Detect recurring class. Save resources, human effort and reduce false alarm rate.
5.	Decision Tree[2]	Incremental	Not work for dynamic feature set	Require little prior knowledge and more efficient because only one classifier is used

V. CONCLUSION

Novel class detection is the challenging task in the stream classification. In this paper we have revised the different approach for novel class detection with Incremental and Ensemble learning approach. Supervised learning algorithm is easy to implement and requires little prior knowledge, so it is very popular. Decision tree classifier represent most recent concept in data stream. Future work focuses on tackling this problem under dynamic feature set.

REFERENCE

[1] Elena Ikonomovska, Suzana Loskovska, and Dejan Gjorgjevik,"A survey of stream data mining," 2007.

[2] Dewan Md. Farid, Amit Biswas and Chowdhury Mofizur Rahman, "A New Decision Tree Learning Approach for Novel Class Detection in Concept Drifting Data Stream Classification,"*Journal of Computer Science and Engineering, Volume 14, Issue 1, July 2012*.

[3] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han and Bhavani Thuraisingham, "Integrating Novel Class Detection with Classification for Concept-Drifting DataStreams," *W. Buntine et al. (Eds.):ECML PKDD 2009,Part II, LNAI 5782, pp. 79-94, Springer-Verlag Berlin Heidelberg 2009*.

[4] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han and Bhavani Thuraisingham, "Classification and Novel Class Detection in Data Streams with Active Mining," *M.J.Zaki et al. (Eds.):PAKDD 2010,Part II, LNAI 6119, pp. 311-324 Springer-Verlag Berlin Heidelberg 2010.*

[5] S.Thangamani, "DYNAMIC FEATURE SET BASED CLASSIFICATION SCHEME UNDER DATA STREAMS," *International Journal of Communications and Engineering Volume 04 – No.4, Issue: 01 March2012*

[6] Mohammad M. Masud, Tahseen M. Al-Khateeb, Latifur Khan, Charu Aggarwal, Jing Gao, Jiawei Han and Bhavani Thuraisingham, "Detecting Recurring and Novel Classes in Concept-Drifting Data Streams," *icdm, pp.1176-1181, 2011 IEEE 11th International Conference on Data Mining, 2011.*

[7] Ahmed Sultan Al-Hegami, "Classical and Incremental Classification in Data Mining Process*," IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.12, December 2007.*

[8] Amit Ganatra, Prerana Gupta and Amit Thakkar, "Comprehensive study on techniques of Incremental learning with decision trees for streamed data," *International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-3, February 2012*.

[9] Bassem Khouzam, "Incremental Decision Trees," *ECD Master Thesis Report*