International Journal of Computer Applications Technology and Research
Volume 1– Issue 3, 70-74, 2012

Comparative Analysis of Different Techniques for Novel Class Detection

Patel Jignasa N.
Parul institute of Engineering & Technology,
Waghodia, Vadodara,
Gujarat, India

Sheetal Mehta
Parul institute of Engineering & Technology,
Waghodia, Vadodara,
Gujarat, India

Abstract: Data stream mining is the process of extracting knowledge from continuous data. Data stream can be viewed as a sequence of relational touples arrives continuously at time varying. Classification of data stream is more challenging task due to three major problems in data stream mining: Infinite length, Concept-drift, Arrival of novel class. Novel class detection in stream data classification is interesting research topic and researches available for concept drift problem but not attention on the Novel class detection. In this paper we have discussed various techniques of the novel class detection. And have also covered comparative analysis of various techniques for the same.

Keywords: Data stream, Novel class, Incremental learning, Ensemble Technique, Decision tree.

1. INTRODUCTION

Data mining is the process of extracting hidden useful information from large volume of database. A data stream is an ordered sequence of instances that arrive at any time does not permit to permanently store them in memory. Data mining process has two major functions: classification and clustering. Data stream classification is the process of extracting knowledge and information from continuous data instances. The goal of data mining classifiers is to predict the class value of a new or unseen instance, whose attribute values are known but the class value is unknown [1].Classification maps data into predefined that is referred to a supervised learning because the classes are determined before examining the data and that analyses a given training set and develops a model for each class according to the features present in the data. In clustering class or groups are not predefined, but rather defined by the data alone. It is referred to as unsupervised learning.

There are three major problems related to stream data classification [2].

1. It is impractical to store and use all the historical data for training
2. There may be concept-drift in the data, meaning, the underlying concept of the data may change over time.
3. Novel classes may evolve in the stream.

In data stream classification most of the existing work related to infinite length and concept drift here we focus on the novel class detection. Most of the existing solutions assume that the total number of classes in the data stream is fixed but in real-world data stream classification problems, such as intrusion detection, text classification, fault detection, novel classes may arrive at any time in the continuous stream. There are many approaches to develop the classification model including decision trees, neural networks, nearest neighbor methods and rough set-based methods [4].The data stream classifiers are divided into two categories: single model and ensemble model [1]. Single model incrementally update a single classifier and effectively respond to concept drifting so that reflects most recent concept in data stream. Ensemble model use a combination of classifiers with the aim of creating an improved composite model, and also handle concept drifting efficiently. The traditional tree induction algorithm is that they do not consider the time in which the data arrived. The incremental classifier that reflects the changing data trends effective and efficient so it is more attractive. Incremental learning is an approach to deal with the classification task when datasets are too large or when new examples can arrive at any time [5]. Incremental learning most important in applications where data arrives over long periods of time and storage capacities are very limited. In [7] author Defines incremental tasks and incremental algorithms as follows:

Definition 1: A learning task is incremental if the training examples used to solve it become available over time, usually one at a time.

Definition 2: A learning algorithm is incremental if, for any given training sample *e1... en*, it produces a sequence of hypotheses h_0, $h_1, \ldots, h_n$ such that h_{n+1} depends only on *hi* and the current example e_i.

As per [8] the learning to be one that is: Capable to learn and update with every new data (labeled or unlabeled), Will use and exploit the knowledge in further learning, Will not rely on the previously learned knowledge, Will generate a new class as required and take decisions to merge or divide them as well

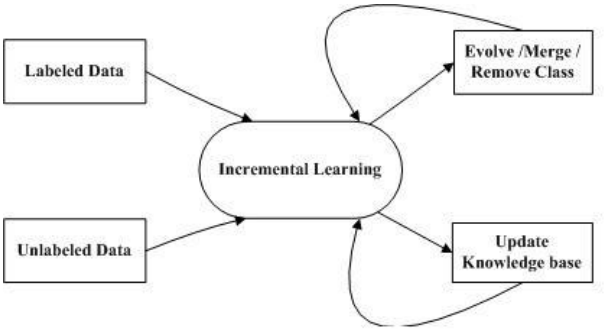

Figure 1: Working of an Incremental learning

Will enable the classifier itself to evolve and be dynamic in nature with the changing environment.

Decision tree that provide the solution for handling novel class detection problem. ID3 is very useful learning algorithm for decision tree.C5.0 algorithm improves the performance of tree using boosting. MineClass that provide solution for Novel Class. ActMiner extends MineClass, and addresses the limited labeled data problem. ECSMiner which stands for Enhanced Classifier for Data Streams with novel class Miner. The stream classification model is enhanced to handle dynamic feature sets. SCANR, which stands for Stream Classifier And Novel and Recurring class detector that address the recurring issue, and propose a more realistic novel class detection technique, which remembers a class and identifies it as "not novel" when it reappears after a long Period of time.

2. NOVEL CLASS DETECTION

Novel class detection in stream data classification is interesting research topic and researches available for concept drift problem but not attention on the Novel class detection. This approach fall into two categories : Single model (Incremental approach), Ensemble Model. Data stream classification and novelty detection recently received increasing attention in many practical real-world applications, such as spam, climate change or intrusion detection, where data distributions inherently change over time[6]. Ensemble techniques maintain an combination of models, and use ensemble voting to classify unlabeled instances. As per [6] In 2011, Masud et al. proposed a novelty detection and data stream classification technique, which integrates a novel class detection mechanism into traditional mining classifiers that enabling automatic detection of novel classes before the true labels of the novel class instances arrive, also In 2011, R. Elwell and R. Polikar introduced an ensemble of classifiers-based approach named Learn++. NSE for incremental learning of concept-drift, characterized by nonstationary environments.

In [9], [10] author gives the definition of the existing class and Novel class.

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 trained with the instances of class c. Otherwise, c is a novel class.

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*).

In [10] show the basic idea of novel class detection using decision tree in Fig 2. That introduces the notion of used space to denote 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. Besides, there must be strong cohesion (e.g. *closeness*) 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. This is done by clustering and saving the cluster summary as "pseudo point" (to be explained shortly). Secondly, these Pseudo points are used to detect outliers in the test data, and declare a novel class if there is strong Cohesion among the outliers

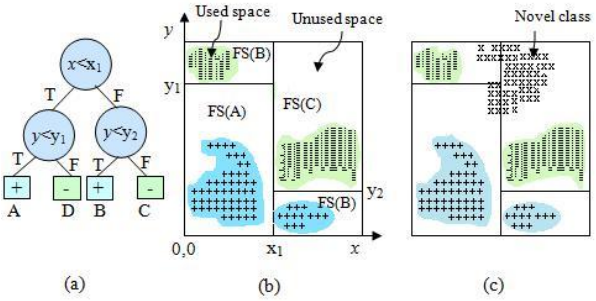

Figure 2: (a) A decision tree, (b) corresponding feature space partitioning where FS(X) denotes the Feature space defined by a leaf node X The shaded areas show the used spaces of each partition. (c) A Novel class (denoted by x) arrives in the unused space.

3. RELATED WORK

Novelty detection techniques into two categories: statistical and neural network based. Statistical approach has two types: parametric, and non-parametric. Some approaches assume that data distributions are known (e.g. Gaussian), and try to estimate the parameters (e.g. mean and variance) of the distribution called Parametric approach. If any test data that outside the normal parameter that detect as Novel.

In [10] author describe "MineClass", which stands for Mining novel Classes in data streams with base learner K-NN (K-nearest neighbor) and decision tree. K-NN based approaches for novelty detection is also non-parametric. Novelty detection is also closely related to outlier/anomaly detection techniques. There are many outlier detection techniques available, some of them are also applicable to data streams However, and the main difference with this outlier detection is

that here primary objective is novel class detection, not outlier detection. Outliers are the by-product of intermediate computational steps in Novel class detection algorithm. Recent work in data stream mining domain describes a clustering approach that can detect both concept-drift and novel classa and assumes that there is only one 'normal' class and all other classes are novel. Thus, it may not work well if more than one class is to be considered as 'normal' or 'nonnovel'. Mine class can detect novel classes in the presence of concept-drift, and proposed model is capable of detecting novel classes even when the model consists of multiple "existing" classes.

In [9] ActMiner applies an ensemble classification technique by addressing the limited labeled data problem. ActMiner extends MineClass, and addresses the Limited labeled data problem in addition to addressing the other three Problems thereby reducing the labeling cost. It also applies active learning, but its data selection process is different from the others. An unsupervised novel concept detection technique for data streams is proposed, but it is not applicable to multi-class classification. As per previously mention work MineClass addresses the concept evolution problem on a multi-class classification framework. MineClass does not address the limited labeled data problem, and requires that all instances in the stream be labeled and available for training.

In [11] author describes ECSMiner for Novel class detection. Novel class detection using ECSMiner is different from traditional one class detection technique. This approach offers a "multiclass" framework for the novelty detection problem that can distinguish between different classes of data and discover the emergence of a novel class. This technique is a nonparametric approach, and therefore, it is not restricted to any specific data distribution. ECSMiner is different from other technique in three aspects: (I) It not only considers difference of test instance from training data but also similarities among them. Technique discovers novelty collectively among several coherent test points to detect the presence of a novel class. (II) It is "multiclass" novelty detection technique, and also discover emergence of a novel class. (III) Approach can detect novel classes even if concept-drift occurs in the existing classes. "ECSMiner" (pronounced like ExMiner).This technique 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. K-NN strategy would lead to an inefficient classification model, both in terms of memory and running time. ECSMiner detect novel classes automatically even when the classification model is not trained with the novel class instances.

In [12] author proposed a *recurring class* is a special case of concept-evolution. A *recurring class* is a special and more common case of concept-evolution in data streams.It occurs when a class reappears after long disappearance from the stream. ECSMiner identifies recurring classes as novel class. Each incoming instance of data stream is first check by primary ensemble if it is outlier called it primary outlier (P-outlier) than again check through auxiliary ensemble if it is outlier than called *secondary outlier(S-outlier)*, and it is temporarily stored in a buffer for further analysis. When there are enough instances in the buffer, the *novel class detection* module is invoked. In this technique compute a unified measure of cohesion and separation for an S-outlier *x*, called *q*-NSC (neighborhood silhouette coefficient), range of q-NSC is [-1, +1]. The *q*-NSC(*x*) value of an S-outliers *x* is computed separately for each classifier $Mi \in M$. A *novel class* is declared if there are S-outliers having positive *q*-NSC for all classifiers $Mi \in M$. Recurring class instance, they should be P-outliers but not S-outliers because the primary ensemble does not contain that class, but secondary ensembles shall contain that class. The instances that are classified by the auxiliary ensembles are not outliers. The technique for Classification with novel and recurring class is called SCANR (Stream Classifier and Novel and Recurring class detector).

ERR is calculated using the following equation:

$$M_{new} = \frac{F_n * 100}{N_c} \quad (1)$$

$$F_{new} = \frac{F_p * 100}{N - N_c} \quad (2)$$

$$ERR = \frac{(F_p + F_n + F_e) * 100}{N} \quad (3)$$

Where,

Fn = Total novel class instances misclassified as existing class, *Fp* = Total existing class instances misclassified as novel class, *Fe* = Total existing class instances misclassified (other than *Fp*), *Nc* = total novel class instances in the stream, *N* = total instances the stream, *Mnew* = % of novel class instances Misclassified as existing class, *Fnew* = % of existing class instances falsely identiified as novel class, *ERR* = Total misclassification error (%) (Including *Mnew* and *Fnew).*

In [12] using (3), authors have demonstrated that OW (*OLINDDA-WCE)* has highest *ERR* rate followed by EM (ECSMiner). The main source of error for OW is *Mnew*, since it fails to detect most of the novel class instances. Therefore, the Fnew rates of OW are also low. The main source of higher error for EM compared to SC (SCANR) can be contributed to the higher Fnew rates of EM, which occurs because EM misclassifies all recurring class instances as novel ("false novel" error). Since SC can correctly identify most of the recurring class instances, the Fnew rates are low. Here describe that *ERR* rate of EM increase with increasing number of recurring classes. This is because EM identifies the recurring classes as novel. Therefore, more recurring class increases its *Fnew* rate, and in turn increases *ERR* rate. For

SC, the *Fnew* rate increases when drift increases, resulting in increased *ERR* rate. The *Fnew* rate (and *ERR*) of EM is almost independent of drift, i.e., whether drift occurs or not, it misclassifies all the recurrent class instances. However, the *Fnew* rate of SC is always less than that of EM. *Fnew* rate increases in OW because the drift causes the internal novelty detection mechanism to misclassify shifted existing class instances as novel However, for EM, here that describe *ERR* increases with increasing chunk size. The reason is that *Fnew* increases with increasing chunk size For OW, on the contrary, the main contributor to *ERR* is the *Mnew* rate. It also increases with the chunk size because of a similar reason, i.e., increased delay between ensembles update SCANR Need Extra running time because of auxiliary ensemble.

In [1] authors have proposed New decision tree learning approach for detection of Novel class. In this approach calculate the threshold value based on the ratio of percentage of data points between each leaf node in the tree and the training dataset t and also cluster the data points of training dataset based on the similarity of attribute values. If number of the data points classify by a leaf node of the tree increases than the threshold value that calculated before, which means a novel class arrived. IN [6] paper describe the decision tree learning algorithm The ID3 (Iterative Dichotomiser) technique builds decision tree using information theory. The C5.0 algorithm improves the performance of building trees using boosting, which is an approach to combining different classifiers. CART (classification and regression trees) is a process of generating a binary tree for decision making. CART handles missing data and contains a pruning strategy. The SPRINT (Scalable Parallelizable Induction of Decision Trees) algorithm uses an impurity function called gini index to find the best split .In this they introduce decision tree classifier based novel class detection in concept drifting data stream classification, which builds a decision tree from data stream. The decision tree continuously updates with new data points so that the most recent tree represents the most recent concept in data stream. Using (3), Compare the traditional decision tree and new decision tree learning approach and demonstrated the efficacy of New approach with less *ERR* rate.

4. COMPARATIVE ANALYSIS FOR NOVEL CLASS DETECTION.

The Table 1 below describes comparative analysis between different techniques of Novel class detection based on Learning Approach, type of classifier, advantages and disadvantages or limitation.

Table 1: Comparative Analysis of Various Techniques for Novel Class Detection

Algorithm	Learning Approach	Classifier	Advantage	Disadvantage
ACT Miner [9]	Ensemble	Active classifier work with K-NN and decision tree.	Work on the less label instance. It saves 90% or more labeling time and cost.	Not directly applicable to multiclass. Not work for the multi label classification.
Mine Class [9][10]	Ensemble	Decision tree and K-NN (Train and create inventory baseline techniques.)	Nonparametric. Does not require data in convex shape.	That requires 100% label instance.
ECS miner [11][12][13]	Ensemble	Classical classifier Work with K-NN and decision tree.	Non parametric Does not require data in convex shape	Not efficient in terms of memory and run time. It Identifies recurring class as Novel class.
SCANR [12]	Ensemble	Multiclass classifier	Remembers a class and identifies it as “not novel” when it reappears after a long disappearance.(Detect Recurring class)	Auxiliary ensemble is used so running time is more than other detection method
Decision tree[1][6]	Incremental	Decision tree based classifier	Detect the arrival of new class and update the tree with new recent concept	Does not work for dynamic attribute sets

5. CHALLENGES

- Concept drift and Arrival of Novel class is the challenging task for stream data mining
- Multiclass classification is challenging problem in stream data mining. [9]
- Work with less label instances and detection of recurring class is the challenging for stream data mining [10], [11].

6. CONCLUSION

Novel class detection is the more challenging task in data stream classification. In this paper we have studied the different approach that provides the solution for novel class detection with Incremental learning and Ensemble Technique. Supervised learning algorithm that has several advantages such as it is easy to implement and requires little prior knowledge, so it is very popular. Incremental approach in decision tree classifier that represent most recent concept in data stream.

7. REFERENCES

[1] Amit Biswas, Dewan Md. Farid 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.

[2] Mohammad M. Masud, Jing Gao, Latifur Khan Integrating Novel Class Detection with Classification for Concept-Drifting Data W. Buntine et al. (Eds.):ECML PKDD 2009,Part II, LNAI 5782, pp. 79-94, Springer-Verlag Berlin Heidelberg 2009.

[3] S.PRASANNALAKSHMI, S.SASIREKHA INTEGRATING NOVEL CLASS DETECTION WITH CONCEPT DRIFTING DATA STREAMS *International Journal of communications and Engineering Volume 03– No.3, Issue: 04 March2012.*

[4] 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.

[5] Prerana Gupta, Amit Thakkar, Amit Ganatra 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.

[6] Dewan Md. Farid, Chowdhury Mofizur Rahman Novel Class Detection in Concept-Drifting Data Stream Mining Employing Decision Tree.

[7] Bassem Khouzam ECD Master Thesis Report INCREMENTAL DECISION TREES.

[8] Prachi Joshi, Dr. Parag Kulkarni Incremental Learning: Areas and Methods – A Survey International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012.

[9] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, 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.

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

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

[12] Mohammad M. Masud, Tahseen M. Al-Khateeb, Latifur Khan, Charu Aggarwal, Jing Gao, Jiawei Han, 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.

[13] S.PRASANNALAKSHMI, S.SASIREKHA INTEGRATING NOVEL CLASS DETECTION WITH CONCEPT DRIFTING DATA STREAMS *International Journal of Communications and Engineering Volume 03, No.3, Issue: 04 March2012.*