International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 10, October-2015

Analysis of Classification and Clustering based Novel Class Detection Techniques for Stream Data Mining

Kamini Tandel
Shri S'ad Vidya Mandal Institute of Technology,
Bharuch,Gujarat, India.

Jignasa N. Patel
Shri S'ad Vidya Mandal Institute of Technology,
Bharuch, Gujarat, India.

***Abstract*— Data stream is continuous and always change in nature. Data stream mining is the process of extracting knowledge form continuous data. Due to its dynamic changing nature it has some major challenges like infinite length, novel class detection and concept-drift. Data stream is infinite in length and we cannot store it for historical purpose. Concept drift means data changes rapidly over time and novel class define as new class appear in continuous data stream. Classification is the challenging task in data stream and existing data mining classifier cannot detect novel class until the classification models are not trained. Different classification and clustering based techniques are used to detect novel class in data stream. In this paper we have discuss the different techniques of novel class detection and its comparative analysis.**

***Keywords*— Concept drift, Data stream mining, Learing approach, Novel class detection.**

I. INTRODUCTION

Data mining is the process of mining useful hidden information from large amount of database. Data mining generally have two major task classifications and clustering. Data mining classification is used to predict group membership for data instances. Data mining classifiers that predict the class value of a new or previously unseen instance, whose attribute values are known but the class value we need to predict. Using supervised learning and unsupervised learning techniques we classify our data. In supervised learning techniques we are referring the training dataset, classes are determined based on that training data and model is prepared for future prediction. In unsupervised learning techniques we have not any training instances or any predefined group but use individual data, like clustering techniques.

Data stream classification is most important research topic in recent years. Due to dynamic nature of data stream we required effective techniques that can handle its continuous changing nature and it also differ from static classification techniques. Data stream has three most challenging characteristics that is infinite length, concept drift and concept evolution. Data stream is infinite in length and we cannot store it for historical purpose and use all historical data as training dataset. Concept drift that means data always changes over time and concept evolution means new class arrived in data stream. For example intrusion detection in network traffic stream when completely new attack arrived in traffic stream that time novel class detection occur in stream [9].

II. MAJOR CHALLENGES OF DATA STREAM

There are three major challenges of data stream: Infinite length, Concept drift and Concept evolution. Most of existing work done on infinite length and concept drift. Here we discuss more about the novel class detection.

A. Infinite Length

Data stream is infinite in length and it requires infinite storage and also more running time. We cannot store it for historical purpose and cannot use all data stream as training dataset. Only the small amount of summary information of data stream is used and store and other are discarded. Some of the summarization techniques like sampling, sketching, wavelet, histograms and sliding window are used to reduce data stream size and speed up analysis.

B. Concept Drift

Concept drift means data always changes over time that means classifier boundary and clustering centers are changed as time elapsed. Because of this past prediction model become irrelevant for current data, which decrease the prediction accuracy. To handle this problem model should be updated with the recent concepts. Many of the techniques are used in existing work to handle concept drift like Weighed Window, FISH, ADWIN, DDM, EDDM, CVFDT, DDD, and Streaming Ensemble.

We have two fundamental approaches to handle the concept drift: Incremental learning approach and Ensemble learning approach. In Incremental learning approach one single model is used and if a new concept is appearing updates the model accordingly. In incremental learning approach assume that they have not sufficient training data at the stating learning process and data comes over time elapsed. In Ensemble learning approach multiple models are combining and create one composite model. In Ensemble learning approach, we are dividing the large data stream into smaller data chunks and each chunk train with single classifier model and create one composite model. The most important difference between ensemble and increment approach is that ensemble approach may discard training model which is outdated but incremental approach may not. Ensemble learning approach have many advantages: Each data chunk is smaller in size so cost of training classifier is relatively small. Memory requirement is very less because it store trained classifier not the training instance of data chunk. So ensemble approach easily handle

increasing length of data streams and concept drift problem in data stream mining. Compare to incremental approach ensemble learning approach is very flexible to concept drift. In ensemble learning approach we can set the data chunk size according to different changes occurs in concept drift: First, if there is sudden concept drift data chunk size should be small and for smooth concept drift data chunk size should be large. Second, it can assign different weighting values to different base classifiers to satisfy various concept drift [11]. Third, we can change outdated classifier as per requirement. Incremental algorithm is better and faster anti-noise capacity than ensemble algorithm [11], but it has more restriction and not all classification algorithm use with incremental approach. And generally speaking, where smooth concept drift and simple data stream is present incremental approach is better choice and for huge concept drift and complicated or unknown distribution of data stream ensemble approach is better choice.

C. Novel Class Detection

In major challenges of data stream classification concept evolution not taken much more attention. Concept evolution refers to the appearance of new class. Most existing data stream classifier assumes numbers of classes are fixed however in data stream new classes may often appear. For example a different kind of attack appears in network traffic, which is not previously seen or a new category of text appears in social text stream such as Twitter [4]. When new class appears, traditional classifier misclassified those instances. So it is important to perfectly classify novel class in data stream. Novel class detection is closely related to outlier detection techniques. Outlier detection is the intermediate part of novel class detection techniques. Most of the novel class detection techniques follow two approaches [5]: parametric and nonparametric. Parametric approaches assume a particular distribution of data and estimate parameters of the distribution from the normal data, and nonparametric approach not restricted to any specific data distribution [5]. A recurring class is also common concept in concept evolution. It occurs when a class reappears after long disappearance from the stream. If recurring class misclassify they create several undesirable effect. They increase false alarm rate because when they reappear they falsely identified as novel class and also increase human effort.

In [1], [5] author gives the definition of the existing class and Novel class.

Definition (*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 [1].

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*) [1].

In [1] show the basic idea of novel class detection using decision tree in Fig 1. Figure described the occupied used space and free unused space. Novel class that arrived in unused space, and according to property 1, strong cohesion present between instances of novel class. Keeping track of the used spaces of each leaf node in a decision tree and finding strong cohesion among the test instances that fall into the unused spaces [1].

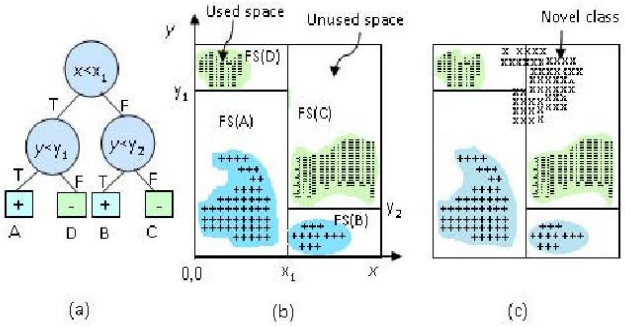

Fig. 1. (a) A decision tree and (b) Corresponding feature space partitioning. FS(X) denotes the feature space defined by a leaf node X. The shaded areas show the used spaces of each partition. (c) A Novel class (denoted by x) arrives in the unused space [1].

III. NOVEL CLASS DETECTION TECHNIQUES

MineClass: In [1] author describe "MineClass", which stands for Mining novel Classes in data streams with base learner K-NN (K - Nearest Neighbor) and decision tree. This technique follows the non parametric approach and ensemble multiple classifier for classify unlabeled data. This technique is different from traditional novelty (or anomaly/outlier) detection techniques in several ways. Traditional novel class detection techniques build model on normal data and identify that point that deviated from normal data and find strong cohesion among the multiple data. Traditional novel class detector find only one class model, which simply differentiates normal data and anomalies data but it cannot differentiates two different kinds of anomalies. But MineClass work on a "multiclass" model; it can differentiate among different classes of data and at the same time detect a novel class also. Existing techniques are not work if there are more than one class are consider as existing class but "MineClass" work for any number of existing classes.

ActMiner: ActMiner, define as Active Classifier for Data Streams with Novel Class Miner that performs classification and novel class detection in data streams while requiring small amount of labeled data for training [2]. Existing techniques work on infinite length, concept drift and concept evolution but not concentrate on limited labeled data. ActMiner extends MineClass and address limited labeled data problem, if we have less labeled data that cannot generate accurate classification model. ActMiner only choose those data points which have higher classification error, only those data points are labeled [2]. ActMiner only choose limited instances for labeling that reduce labeling cost and also time than traditional approach required.

DETECTNOD: DETECTNOD [3] (DiscrETE Cosine Transform based Novelty and Drift detection) is single class classification techniques that detect novel class and handle concept drift also. This approach outperform in terms of discriminating normal concept from novelty and concept drift by using deviation instead of distance [3], and considering cluster energy instead of just as center [3]. In first phase of this technique it generate initial model using cluster based on normal data. Apply data interpolation and DCT to build generative model. This generative model differentiates normal data with novel classes and concept drift. In experiment compare the result with other novel class detection techniques on different dataset. Experiment result show that in normal detection phase DETECTNOD gives better result than other and minimum error rate. In novelty and concept drift detection DETECTNOD outperformed OLINDDA. DETECTNOD storage space and computational complexity is also low as compare to OLINDDA.

SCANR: In [4] author describes "SCANR", which stand for Stream Classifier And Novel and Recurring class detector. Recurring class is a most important part of novel class detection in data stream. If any class appear in the stream and again reappear after long period of time than its called recurring class. If recurring class present in data stream than it increase false alarm rate because when they reappear it falsely identify as novel class. It also increases human effort to again identify the class as novel class. So author designed approach as multi-class classifier for concept-drifting data stream that identify novel class and distinguish recurring classes from novel classes. When data stream arrive, the outlier detection module of primary ensemble check that outlier present or not. If it is not outlier than it classify as existing class otherwise it again examined by outlier detection module of the auxiliary ensemble. Auxiliary ensemble checks that it is outlier or recurring class. If it is not outlier than detect as recurring class by auxiliary ensemble. This technique reduces false alarm rate, as well as overall error rates.

ESCMiner: In [5] author described ESCMiner novel class detection techniques. This technique is different from other traditional "one class" novelty detection techniques. This approach offers a "multiclass" framework that finds novel class as well as differentiates between the classes of data. This approach follows the ensemble technique and non parametric approach, which is not restricted to any specific data distribution. ECSMiner is different from existing techniques in three aspects: (I) It not only detects different instances but strong cohesion among the instances that indicate the arrival of novel class. (II) It is "multiclass" novelty detection technique differ from single class that only differentiate between normal and novel. (III) It detects novel class even if concept drift occurs in existing classes. This approach also considers time constraints, which impose several restrictions on classification algorithm, making classification more challenging. With the time constraint classification model have to wait some time to show similarity among instances. Decision tree and KNN classifier are used as base learner.

Decision Tree: Decision tree that is built on ID3 and detect multiple novel class. In [6] author build decision tree from training data points and calculate the percentage of number of data points in each leaf node in the tree with respect to data point in training dataset. And cluster the data points based on similarity attribute values for each leaf node in the tree. When classifying the data streams in real-time, if number of data points classify by a leaf node increases than the percentage calculated before it means a novel class arrived. Then we check in which cluster the new data point belongs based on the similarity of attribute values, if this new data point does not belongs to any cluster, which confirms a novel class arrived. Then we add the new data point into training dataset and rebuild or updated the decision tree model. The decision tree classifier continuously updated so that it represents the most recent concept in the data stream. This technique follows the incremental approach and tested on different dataset and also proved that approach efficiently detects novel class and improves the classification accuracy.

CLAM: In CLAM [7] author proposes a class based ensemble technique that overcomes the drawback of chunk based ensemble techniques [7]. In chunk based ensemble technique divide data into chunks, and train a model from one chunk. An ensemble of chunk based models is used to classify unlabeled data. Chunk based ensemble usually keep fixes sized ensemble, and update old model with new trained model and that detect recurrent class as novel class [7]. We choose *K*-means clustering because of its lower time complexity compared to those alternatives. Essentially, fast running is vital for mining data streams. The second reason for clustering is to reduce space complexity. By storing only the cluster summaries and discarding the raw data, the space complexity per model M^b is reduced from $O(S)$ to constant (i.e., K), where S is the chunk size. This work reduces false alarm rate and overall classification error compare to ESCMiner and detects recurrent class as existing class that increases classification accuracy.

DTNC: DTNC [8], which stand for Detecting Novel Classes in Data Streams. Author uses VFDT (Very Fast Decision Tree) as base classifier and developed *k*-prototypes++ algorithm to deal with the mixed attribute data and initial centers [8]. Because the original *k*-means deals with numerical attribute only. It is sensitive to initial centers and noise, which method may lead to lower accuracy of novel detection and classification. *K*-means++ proposed a way to initialize *k*-means by choosing starting centers. *K*-prototypes presented a measure similar to the squared distances to deal with both numeric and categorical attributes [8]. Author combine the two technique, the new algorithm is named as *k*-prototypes++. The classification accuracy of DTNC is better than MineClass and MCM. It is the result of the better base learner and the more precise novel detection.

MCM: Multi Class Miner [9] proposed an ensemble classification framework to detect novel class. It uses K-NN classifier and K-means clustering techniques to detect novel class. Novel class detection process consists of three steps. In first steps, a decision boundary is built during training. In

second steps, test points falling outside the decision boundary are declared as outliers. And in final steps, the outliers are analyzed to see if there is enough cohesion among themselves and separation from the existing class instances. Here author first use the flexible decision boundary for outlier detection by allowing a slack space outside the decision boundary [9]. This space is controlled by threshold, and the threshold is adapted continuously to reduce the risk of false alarms and missed novel classes [9]. Second author apply a probabilistic approach to detect novel class instances using the discrete Gini Coefficient [9]. With this approach, we are able to distinguish noise, concept drift or concept evolution. Third author use graph based approach to detect the appearance of more than one novel classes simultaneously and separate the instances of novel class from others. Author also considers feature evolution in data stream and use lossless homogenizing conversion for feature evolution. Experiment conducted on different dataset and compare with other approaches like ESCMiner and O-F (OLINDDA for novel class and FAE for classification). Author proposed approach give better performance than other existing techniques. And it only worked with fixed chunk size of the data stream.

SCND: In [10] author handles the problems using the string comparison operations. In that author gathering all words present in test instances and compare with existing set of words. They use an ensemble of models. First they detect the outliers from each model and then they find the final outliers. Then they work to separate instances based upon the cause of their occurrence *i.e.* concept-evolution, concept drift or noise. And then they update the model with the new features. This technique is non-parametric and model is able to detect multiple novel classes present in the data and is a multi class classifier. And it not works for dynamic chunk size. Experiment result show that no novel class instance classified as existing class instance so false alarm rate is quite low and is negligible.

HOTDC: HOTDC [12] means Hoeffding Option Tree with detection of Novel Class. It is a regular Hoeffding tree containing option nodes. HOT is a well known supervised learning classifier that used some little knowledge and it follow incremental approach. In HOTDC author improve the classification accuracy and performance of HOT in concept drift data stream classification. HOTDC continuously update with new data points so most recent concept it represents and efficiently find novel class.

HOTND: HOTND [13] that is different from HOTDC. HOTND is a new voting method for novel class detection using hoeffding option tree. Hoeffding option tree is better classification method than HOT because options are present to deal with equally discriminative attribute. In HOTND author first decide the classes [13], when data comes than generate the tree. If any single instance misclassify than they do not store that instance but update the trained model and create the new class. Here, not required to store misclassify instance so memory requirement is low and not create any cluster so CPU overhead is also low.

IV. COMPARATIVE ANALYSIS OF NOVEL CLASS DETECTION TECHNIQUES

The Table 1 below describes comparative analysis between different techniques of Novel class detection based on Learning Approach, Type of Classifier, Advantage and Disadvantage.

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

Algorithm	Learning Approach	Classifier	Advantage	Disadvantage
Mine Class [1]	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.
Algorithm	Learning Approach	Classifier	Advantage	Disadvantage
ACT Miner [2]	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.
DETECTNOD (Discrete Cosine Transform) [3]	Ensemble	K Means clustering	- DCT coefficient efficiently detects novel class & concept drift. -Memory usage is very low.	- Not work on multi class.
SCANR [4]	Ensemble	Multiclass classifier	- Remember a class and detect it as recurring class if it appears after a long time. - Reduce false alarm rate and overall error rate.	- Auxiliary ensemble is used so take extra time to detect the class.

ECSMiner [5]	Ensemble	Classical classifier work with K-NN and decision tree.	- Non parametric. - Does not require data in convex shape.	- Inefficient in terms of memory and running time. - Detect recurring class as novel class.
Decision Tree [6]	Incremental	Decision tree based classifier	-Improve classification accuracy and error rate. - Detect arrival of new class and update tree with new recent concept.	- Not work for dynamic attribute sets.
CLAM (Class based ensemble) [7]	Ensemble	K means clustering approach	- Perform better than chunk based approach. - Detect recurring class efficiently.	- Storage space increase.
DTNC[8]	Ensemble	VFDT classification Modified clustering method – K prototype ++	-Work on numeric categorical attributes. - Low error rate - Classification accuracy better than existing.	- Complexity is too high.

Algorithm	Learning Approach	Classifier	Advantage	Disadvantage
MCM (Multiclass miner) [9]	Ensemble	K-NN based classifier	- Adjust decision boundary of model using slack space concept that reduce the false alarm rate and missed novel class. -Gini Coefficient identifies causes of outlier. -Multiple novel class detection	- Approach not work for dynamic chunk size of data stream.
SCND [10]	Ensemble	K Medoid clustering	- It uses string matching instead of distance. - False alarm rate is negligible.	-Not work for multi label instance. -Not handle future evolution effectively. - Not work for dynamic chunk size of data stream.
HOTDC [12]	Incremental	Hoeffding Option Tree classifier	- Better classification accuracy than HOT because of option tree. - Ambiguity removes between instances.	-Compare to HOT take more time.
HOTND [13]	Incremental	Hoeffding Option Tree classifier	Memory requirement is low. -Clustering is not required so CPU overhead is less.	- Time requirement slightly increase compare to HOT.

V. CONCLUSION

Novel class detection is the most challenging task in data stream. In this paper we have studied many classification and clustering based techniques, that provide solution for novel class detection with incremental and ensemble approach. Some techniques provide good solution for recurring class that appears after long time in the stream. Supervised learning techniques are more popular than unsupervised learning because they are simple and easy to implement just required little prior knowledge. But, where smooth concept drift and simple data stream is present incremental approach is better choice and for huge concept drift and complicated or unknown distribution of data stream ensemble approach is better choice.

VI. REFERENCES

[1] 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.

[2] 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.

[3] Morteza Zi Hayat, Mahmoud Reza Hashemi," A DCT Based Approach for Detecting Novelty and Concept Drift in Data Streams" 978-1-4244-7896-5/10/$26.00 c 2010 IEEE.

[4] 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" 1550-4786/11 $26.00 © 2011 IEEE DOI 10.1109/ICDM.2011.49.

[5] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham, "Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints" 1041-4347/11/$26.00 2011 IEEE Published by the IEEE Computer Society.

[6] Dewan Md. Farid and Chowdhury Mofizur Rahman "Novel Class Detection in Concept-Drifting Data Stream Mining Employing Decision Tree" 978-1-4673-1436-7/12/$31.00 ©2012 IEEE.

[7] Tahseen Al-Khateeb, Mohammad M. Masud, Latifur Khan, Charu Aggarwa, Jiawei Han and Bhavani Thuraisingham, " Stream Classification with Recurring and Novel Class Detection using Class-Based Ensemble" 1550-4786/12 $26.00 © 2012 IEEE DOI 10.1109/ICDM.2012.125.

[8] Yuqing Miao, Liangpei Qiu, Hong Chen, Jingxin Zhang, and Yimin Wen "Novel Class Detection within Classification for Data Streams" C. Guo, Z.-G. Hou, and Z. Zeng (Eds.): ISNN 2013, Part II, LNCS 7952, pp. 413–420, 2013. © Springer-Verlag Berlin Heidelberg 2013.

[9] Mohammad M. Masud, Qing Chen, Latifur Khan, Charu C. Aggarwal, Jing Gao, Jiawei Han, Ashok Srivastava and Nikunj C. Oza, "Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams" 1041-4347/13/$31.00 2013 IEEE Published by the IEEE Computer Society.

[10] Singh, R. and Chandak, M.B. (2015) "Classification and Novel Class Detection in Data Streams UsingStrings" *Open Access Library Journal*, 2.

[11] Wenyu Zang1, Peng Zhang, Chuan Zhou and Li Guo, "Comparative study between incremental and ensemble learning on data streams: Case study", © 2014 Zang et al.; licensee Springer.

[12] JIGNASA N. PATEL, SHEETAL MEHTA, "Detection of Novel Class with Incremental Learning for Data Streams", (IJRMEET) ISSN: 2320-6586, Vol. 1, Issue: 3, April-2013.

[13] Darshana Parikh , Priyanka Tirkha, A New Voting Method to Novel Class Detection Using Hoeffding Option Tree", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013.