# A Survey On Various Strategies For Classification And Novel Class Detection Of Data Streams

Rimjhim Singh[1] Dr. M. B. Chandak[2]
RamdeoBaba College of Engg. & Management.
Nagpur, India
[1]rimjhimsingh1012@gmail.com
[2]chandakmb@gmail.com

**Abstract - Data Stream: A continuous stream of raw data has to be converted into some intelligible form to extract meaningful information from it. Only then the data can be put to use. Handling real time data streams in data mining isn't easy and poses various challenges to researchers. It poses four main challenges namely, Infinite length, Concept-drift, Concept-evolution and Feature evolution. Various researchers have proposed various techniques for overcoming these difficulties. But major work has been done to handle infinite length and concept-drift in data streams only. The other two problems have not been tackled that efficiently. In this paper, we make an effort to list and summarize various strategies that have been proposed to overcome the above stated problems for efficiently making use of stream data.**

**Index term-: Data Stream; Concept-Drift; Concept-Evolution; Outliers; Novel Class.**

## I. INTRODUCTION:

Now days the remarkable work in the field of storage of data has enabled the researchers and developers to store the real-time data. Such type of data keeps on growing continuously without any boundaries. The data is dynamic in nature and is referred to as stream data. This data is always in some raw and unknowledgeable form and needs to be treated to gain meaningful information from it. Being dynamic in nature it poses various challenges to the researchers. Hence the strategies that work with static data don't work properly with stream data. Efficient and effective strategies are required to be developed for stream data.

The four main characteristics of stream data are: infinite length, concept-drift, concept evolution and feature evolution. As stream data is fast and continuously growing, it has infinite length and it is difficult to store and use all the available data for classification. Hence we can use incremental learning approach to handle it and many approaches have already been proposed. Concept-drift arises because the underlying concept of the stream changes with the duration of time. Hence the classification model must keep itself updated with the latest ideas evolving gradually in the stream data. Various approaches have been proposed to handle this efficiently. Third is the concept-evolution. According to this, novel classes emerge in the stream data. Many classifiers assume that number of classes is fixed in a data stream. But in real world problem this is not the case. In many application like intrusion detection, credit card fraud detection, new classes keep on evolving that are not defined by the classifier model. In network intrusion detection, each type of attack is a class label. Here concept-evolution occurs when a new type of attack is detected and this attack is not identified by the classifier. Traditional classifiers that work with labeled data will not be able to detect novel classes when they arrive until they are trained with the labeled instances of the novel class. Hence in order to detect novel classes the classifier model must be able to mechanically detect the novel classes before it is trained with the instances of novel class. Last challenge posed is that of feature-evolution. Because of the dynamic nature of stream data, new features appear in the data and the old ones fade away. Hence in order to make our model more efficient we need to keep track of new features arriving in the data set.

First two problems have been efficiently solved by various researchers but the not much work has been done yet to handle the other two. Like various incremental approaches have been proposed. These approaches are basically of two types. Single model incremental approach (like in [9]) in which only one model is used and it is updated dynamically time to time. Other one is the Hybrid-batch incremental approach (proposed in [10], [2]). It uses an ensemble of models and each model is generated using batch learning technique from recent data the older and obsolete ones are simply discarded based on their efficiency. The advantage of using hybrid approach is that updating a model is easy and simple. Outliers (explained below) occur in a data stream due to several reasons like noise, concept-drift or concept-evolution.

We need to distinguish between the reasons behind occurrence of an outlier. Otherwise an instance that belongs to an existing class, due to concept-drift, may be misclassified as an outlier and false alarm rate will be high( misclassifying existing class instances as novel). Spinosa et al (in [11]) addresses concept evolution problem using a one-class classifier. It assumes only one class as normal and takes all other classes as novel. Hence cannot be used for multi class data. It also assumes that the feature space of data is convex in shape but this is not the case with real time data. Majority of the models use parametric approaches for classification and novel class detection. In such approaches certain parameters are calculated and decision is made on the basis of these parameters. In this paper we will throw light on some major contribution of researchers in handling data stream.

## II. SOME BASIC CONCEPTS:

As we are dealing with stream data, we need to understand certain facts about the properties of stream data. Firstly, data stream contains two types of classes namely, an existing class and a novel class. Say L is an ensemble of models and the models $\{M_1, M_2, \ldots\ldots Mn\}$ belong to the ensemble L.

*Definition 1: Existing Class:* The class C that is defined and addressed by any of the model $M_i$ belonging to the ensemble is called an existing classes. That is, at least one model in M is trained on class C.

*Definition 2: Novel Class:* The class N that is not defined by any the models $M_i$ belonging to M and is unknown to the model is called a novel class. No model in the ensemble is trained on novel class before it occurs in the stream.

*Definition 3: Outliers:* if x is a test instance and if it does not fall within the range of the feature vector of the ensemble of models, then it is called an outlier. Outliers do not belong to any of the class on which the models in an ensemble are trained.

*Property of cohesion and separation:* data point should be closer to the data points of its own class (cohesion) and farther apart from the data points of other classes (separation). All of the strategies discussed below are based on this property.

## III. APPROACHES FOR CLASSIFICTION AND NOVEL CLASS DETECTION:

### *A. Olindda:*

OLINDDA (An Online Novelty and Drift Detection Algorithm) is a clustered based approach for detecting novelty and concept drift in data streams. It is a one class classifier model. It considers only one class as normal and all the other classes as novel classes. This approach makes use of k-means clustering and forms k clusters from the chunk of data. It stores the summary of these clusters (centroid, radius etc) and uses it to identify, whether, there is a concept drift or concept-evolution. It calculates two types of centroids. First the centroid of all clusters individually and the second is the centroid of global cluster. When the test instance arrives we differentiate between the reason (concept-drift / novelty) behind that in a following way. Calculate distance between the instance and global centroid. If distance is smaller than it's a concept- drift but if distance is larger then it's a novel instance. It cannot be used normally when multiple classes are considered. In that case we need to generate one separate OLINDDA model for each class. It is not used for classification. Its detailed explanation can be found in [7].

### *B. X-Miner:*

X-Miner (discussed in [1]) is a non-parametric approach for classification and novel class detection and does not assume any proper distribution of data. It requires much labeled data to act efficiently. It addresses infinite-length, concept-drift and concept-evolution problem using an ensemble of models. It first divides the data into chunks. The clusters are then created from these chunks and the models are generated and trained using these chunks (clusters). It uses a threshold for outlier detection and once outliers are detected they are tested for cohesion among themselves and separation from the existing class instances. If the cohesion among them is high then they are said to belong to a novel class, a new model is trained then with recent data and novel class instances to work efficiently in future. This is how the ensemble of models is updated. Problem with X-Miner is that it does not address feature evolution problem and it fails to work efficiently when multiple novel classes appear in data. Moreover, if the unlabeled data is more than labeled data then it doesn't work properly. This strategy can be studied in [1].

### *C. Act-Miner:*

ActMiner is an extension of X-Miner. It addresses the problem of limited labeled data by identifying only few labeled instances. It identifies the instances for which the model has high expected error without knowing its true label. By doing so it saves 90% of the cost and labeling time required. It uses a hybrid batch incremental approach to address infinite length and concept drift and addresses concept-evolution by updating the ensemble of models on recent. It doesn't address feature evolution problem and sometimes do

not work efficiently when multiple novel classes arrive. The detailed explanation of ActMiner can be studied in [2].

*D. DX-Miner*

DX-Miner addresses all the four issues of data stream namely, infinite length, concept-drift, concept-evolution and feature-evolution. The main key feature of this approach is that also considered the dynamic feature space of the data stream for the first time. It used fixed number of models to address infinite length, updates the ensemble continuously for concept-drift. It uses threshold to detect outliers and the concept of high cohesion among instances of same class and high separation among the instances of separate classes to detect novel classes. In order to handle dynamic and evolving feature space it combines the heterogeneous feature sets of both test instance and the models without losing any feature. For dynamic feature set it applies one of the two strategies for feature selection. Predictive Feature selection: here we predict the feature set for test instances not by obtaining their information but by analyzing the feature set of the previously classified chunks. Second one is the informative feature selection: It uses the test chunk to select the features. It selects all that main features and then selects R best features based on some criteria, like we can choose the features with highest frequency. The problem with this algorithm is that it has high false alarm rate and it doesn't work properly in presence of multiple novel classes. The detailed strategy can be studied in [3].

*E. MCM 1(Multi class Miner):*

Multi-Class Miner1 is an enhanced version of previous miners and addresses the problem of infinite-length, concept-drift and concept-evolution efficiently. It again addresses the infinite-length problem of data stream by using an ensemble of fixed number of models and handles concept-drift by updating the ensemble on recent data. Its major contribution is that is reduces the false alarm rate efficiently. For doing so, it uses the concept of Adaptive thresh holding. When an instances that actually belongs to an existing class, but due to some noise is classified as novel (Marginal False-novel), at that time the threshold radius is increased to make it an existing class instance. On the other hand when the Novel class instance is classified as existing and is quite close to outer boundary (Marginal false-existing), at that time the radius is decreased to exclude that instance. It uses the probabilistic approach for detecting the novel class instances by calculating the Gini-coefficient. Now for the separation of two novel classes it proposes a graph theoretic technique that

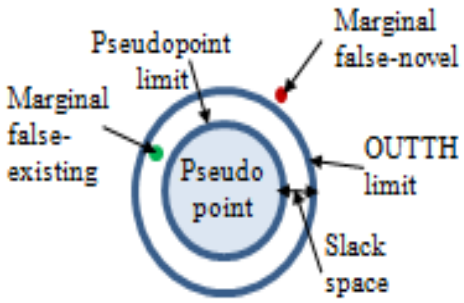


Figure1: Outlier Detection using Adaptive Threshold

makes use of silhouette coefficient to estimate the cohesion and separation between the instances of novel classes. As it is an enhanced technique, so it reduces the false alarm rates but it does not address the problem of feature-evolution. The detailed explanation can be studied in [4].

*F. ECS-Miner:*

ECS-Miner is a multi-class classification approach that handles the problem of infinite length, concept-drift and concept-evolution under particular timing constraints. It says that the test instance once appears must be classified with in time $T_c$ and it must be labeled within time $T_l$, otherwise it should be discarded. That is the outliers that are store in a buffer to see whether they belong to existing class or novel class must be classified under proper time constraints. Previous approaches assumed that the labeling of data required no time, but this is not the case with real-time data. It also uses the adaptive threshold for outlier detection, probabilistic approach using Gini-coefficient for novel class detection and Silhouette coefficient for measuring the concept of cohesion and separation among the instances and Silhouette coefficient for measuring the concept of cohesion and separation among the instances. Basically, it uses the nearest neighbor rule for classification and novel class detection. As it uses adaptive threshold, it has comparatively lower false alarm rate but it again

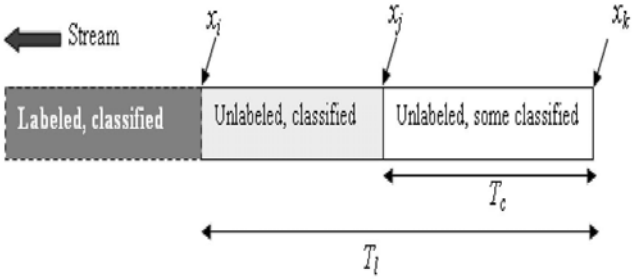


Figure 2: Time Constraints on Labeling of Data.

doesn't address the problem of feature evolution. The detailed approach can be studied in paper [5].

*G. Multi-Class Miner 2(MCM 2):*

This multi class miner is the extension of MCM 1. It addresses all the four problems namely; infinite length, concept-drift, concept-evolution and feature-evolution.MCM 1didnot address feature-evolution. It used the fixed ensemble of models to address the infinite length, handled concept-drift by continuously updating the ensemble. It used the nearest neighbor approach to detect novel classes by applying a probabilistic approach of calculating Gini-coefficient for novel class detection. It used silhouette coefficient to differentiate between multiple novel classes. It handled the feature evolution problem by applying the lossless feature space conversion technique. It has the reduced false alarm rate and addresses all the challenges of data streams. This is the best known strategy so far. This strategy can be studied from [6].

Following table summarizes the Error rate of different miners for Forest data set.

**Table1: Summary of Error Rates.**

| APPROACH | ERROR RATE |
|---|---|
| X-Miner | 7.3% |
| ActMiner | 7.1% |
| DX-Miner | 3.6% |
| MCM(Multi class miner)1 | 3.1% |
| ECS Miner. | 3.6% |
| MCM(Multi class Miner)2 | 3.1% |

## IV. PROPOSED APPROACH:

*A. Data Set*

*"4 University Data Set":* While we were studying the concepts we tried to use the "4 University Data Set" for this purpose. The data set required a lot of preprocessing as it was given in the form of HTML files in which each file represented a separate person. We were supposed to parse those files and convert them into some intelligible form. Because the HTML files did not have any proper predefined structure, we were not able to convert them into intelligible and usable form. Due to which we were forced to stop working on that data set.

*4.2 NASA Aviation Safety Reporting System:* From the NASA ASRS website we downloaded instances of the dataset. Each of which represent an accident and the possible reasons and outcomes related to them. Here each event has an anomaly related to it. Each of these event anomalies can be considered as a separate class, like Aircraft problem: critical, Aircraft problem: less severe etc. Data cannot be used directly, preprocessing was required. We first removed the columns that were having some missed entries. Then some attributes were multi valued. Such attributes also need to be removed from the data set.

Strategy:

While studying various approaches we found that most of them used K-means algorithm for clustering purpose. We also found that k-medoids algorithm works more efficiently than k-means when there is presence of outliers in the dataset. Hence, we followed the k-medoid algorithm for our work. We plan to handle all the four challenges efficiently. Proper algorithms to be followed are yet to be decided.

## V. CONCLUSION

In the above sections we presented and discussed various approaches that have been developed to handle the stream data and that handle the challenges namely, infinite-length, and concept-drift, concept-evolution ad feature-evolution, posed by the stream data. We saw that major contributions in research have been done to typically solve the infinite-length problem and concept-drift problem and efficient algorithms have been generated. But much work is not done to solve the problem of concept-evolution and feature-evolution. The MCM 2 is the only known possible approach that works effectively for stream data and that handle all four challenges. We can work the direction of devising some better algorithms that may work in a more enhanced and efficient manner.

## VI. REFERENCES:

[1] M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "Integrating Novel Class Detection with Classification for Concept-Drifting Data Streams," IEEE Trans. Knowledge andData Eng , vol. 25, no. 7, July 2009.

[2] M.M. Masud, J. Ago, L. Khan, J. Han, and B.M. Thuraisingham, "Classification and Novel Class Detection in Data Streams with Active Mining,".

[3] M.M. Masud, Q. Chen, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "Classification and Novel Class Detection of Data Streams in a Dynamic Feature Space," Proc. European Conf.Machine Learning and Knowledge Discovery in Databases (ECMLPKDD), pp. 337-352, 2010.

[4] M.M. Masud, Q. Chen, L. Khan, C. Aggarwal, J. Gao, J. Han, and B.M. Thuraisingham, "Addressing Concept-Evolution in Concept-Drifting Data Streams,"

Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 929-934, 2010.

[5] M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints," IEEE Trans. Knowledge andData Eng,vol. 23, no. 6, pp. 859-874, June 2011.

[6] M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "Classification and Novel Class Detection in Feature Based Stream Data," IEEE Trans. Knowledge andData Eng , vol. 25, no. 7, July 2013.

[7] OLINDDA: A Cluster Based Approach for Detecting Novelty and Concept-Drift in Data Stream by Eduardo J. Spinosa, Andr´e Ponce de Leon F.de Carvalho, Jo ao Gama in ACM Symposium of Applied Computing SAC'07

[8] A.Bopche, M. Nagle and H. Gupta. A Review of Method of Stream data classification through Optimized Feature Evolution Process, International Journal Of Engineering And Computer Science ISSN:2319-7242Volume 3 Issue 1 January, 2014 Page No.3778-3783.

[9] 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.

[10] Y. Yang, X. Wu, and X. Zhu, "Combining Proactive and Reactive Predictions for Data Streams," Proc. ACM SIGKDD 11th Int'l Conf. Knowledge Discovery in Data Mining, pp. 710-715, 2005.

[11] E. J. Spinosa, A. P. de Leon F. de Carvalho, and J. Gama.Cluster-based novel concept detection in data streams appliedto intrusion detection in computer networks. In *ACM SAC*,pages 976-980, 2008.