Available online at www.ijiere.com

International Journal of Innovative and Emerging Research in Engineering

e-ISSN: 2394 – 3343 p-ISSN: 2394 - 5394

Study on the Different Technique of Concept Drift and Novel Class Detection in Data Stream

Jay Gandhi [a], Amit Thakkar [b], Purvi Prajapati [c]
[a]Charotar University of Science and Technology Changa, India, jaygandhi7591@gmail.com
[b]Asso. Prof, Charotar University of Science and Technology Changa, India, amitthakkar.it@ecchanga.ac.in
[c]Assi. Prof., Charotar University of Science and Technology Changa, India, purviprajapati.it@charusat.ac.in

ABSTRACT:

Data streams mining has become interesting research topic and growing interest in knowledge discovery process. Because of the high speed and huge size of data and mining is processed with limited computing power and limited memory storage capabilities. Therefore our traditional classification technique are not directly applicable. Classification of data stream is more challenging task due to four major problems which is addresses by data stream mining: Infinite length, Concept-drift, Arrival of novel class and limited labeled data. In recent years great amount of work has been done to efficiently solve this problems. In this paper we discusses various technique which efficiently solve the problem of concept drift and novel class detection. Also we have present comparative analysis of this techniques.

Keywords:

Data stream, mining, Classification, Concept drift, Novel class, Limited labeled data

I. INTRODUCTION

Data Stream Mining is the process to extract the knowledge structures from continuous and rapid data records. A data stream is an ordered sequence of instances that in many applications of data stream mining can be read only once (or a small number of times) using limited computing power and limited memory storage capabilities.

There are four major problems related to stream data classification [1].

1. It is impractical to store and use all the historical data for training.
2. There may be concept-drift in the data, it means that underlying concept of the data may change over time.
3. Novel classes may arrive at any time in the stream.
4. High speed data streams always suffer from the insufficient labelling of data.

A stream classification algorithm must meet with several requirements to work with the assumptions and be learning from the data streams. [2]

Requirement 1: Process an example at a time, and inspect it only once.
Requirement 2: Use a limited amount of memory.
Requirement 3: Work in a limited amount of time.
Requirement 4: Be ready to predict at any point

Data streams nowadays can be found in numerous domains, in order to find interesting patterns and Knowledge in data streams, different data mining techniques like Classification, clustering, frequent item-set mining, and outlier detection are introduced. In term of using these techniques for the analysis of static data sets, mining data streams poses new different challenges that have been addressed recently.

In today’s world growing number of applications generate streams of data. Such application are

- Performance measurements in network monitoring and traffic management of continuous data.
- Handling the Call detail records in telecommunications like in customer care to improve the services.
- Transactions in retail chains, ATM operations in banks.
- Log records generated by Web Servers.
- Sensor network data. Continuous monitoring of data from sensor like in weather prediction.

At times, the system performing the analysis task might be busy with other (data mining) tasks, therefore it may possible only little CPU-time and memory available for each individual task, whereas at other times the available resources might be full. There is high rate at which new stream objects are generated, and so can the properties of the objects in the stream. Also the properties of the data stream may vary over time as well, and thus the focus of the stream mining algorithm has to change accordingly. It is therefore desirable that stream mining algorithms are dynamic to all the changes that are happened due to data stream. So there is requirement of new additional methods and techniques for this.
The rest of the paper is organized as follows. Section 2 discusses Concept Drift problem. Section 3 discuss related work of concept drift in detail. Section 4 then describes Novel Class Detection technique. Section 5 discuss its related work. Section 6 discusses comparative study of novel class detection technique. Section 7 conclude both the technique.

II. CONCEPT DRIFT

Concept refers to the target variable, which the learned model is trying to predict. Concept change is the change of the concept as per time. In today, mining and classification with the concept drifting data is been challenging task. Many researcher have done their work to efficiently detect the concept drift and learn model with the latest concept. Example of such type of data are customer's buying habit, whether prediction etc. In such type of application concept which we are going to predict is change over time. So, result of that is poor classification of given data and mining is not accurate.

There are different pattern of which data changes like Sudden/Abrupt, Incremental, Gradual etc. [3]. There are different algorithm which detect concept drift and second the algorithm which have the capability of learning with the concept drifting data. The model is learn from new attribute as well as forgetting the older attribute. So it can efficiently handle the memory management with learning latest concept.

For handling concept drift we have to learn model over time. It means incremental learning of model with new attributes has to be done. Incremental learning is an approach which deal with the classification task when datasets are very large or new examples can be arrive at any point of time.

An algorithm has incremental learning capabilities, if it meets the several criteria given below [4]:
1. Ability to acquire additional knowledge when new datasets are introduced.
2. Ability to retain previously learned information.
3. Ability to learn new classes if introduced by new data.

Based on the number of classifier we can divide the classifier algorithm in two categories.
Single Classifier Approach 2) Ensemble Approach.
As the name suggest single classifier approach handle concept drift using one classifier like Decision Tree, Naïve Bayesian, Hoeffding Tree, Adaptive Hoeffding tree, ADWIN, DDM, Flora, CVFDT etc. Whereas Ensemble Approach uses two or more classifier to handle concept drift like Streaming Ensemble Algorithm, Accuracy Weighted Ensembles, Hoeffding Option Trees etc.

III. RELATED WORK

A. Windowing techniques

This is one of the popular approach to deal with the time changing data using the concept of Sliding Window. Window allow to limiting the example to be learned as well as forgetting the old data points and adding new data points, so that model learns the latest concept. The basic windowing algorithm is very simple. First each example updates the window than the classifier is updated by that window.

Weighted Windows

A one way of forgetting older example is providing the window with a decay function, it assigns the weight to examples in window. In window older examples gets smaller weights and it's importance is lesser than new example. So they are treated as less important by the classifier. Cohen and Strauss gives the use of some decay functions for calculating data stream aggregates. [5]

FISH

FISH algorithm is proposed by Zliobaite, in that uses time and space similarities between examples to dynamically creating a window. The author proposed that the selection of training examples is based on distance measure Dij , which is defined as follows:

$$Dij = a1d(s)ij + a2d(t)ij$$

Where, d(s) indicates distance in attribute space,

d(t) indicates distance in time, and a1, a2 are the weight coefficients.
To manage the balance between the time and space distances, d(s) and d(t) need to be normalized. [6]

ADWIN
ADWIN is sliding window algorithm suitable for data stream in which sudden drift is occurred. It consist sliding window of size w with recently read examples. The main idea behind ADWIN is: whenever two “large enough” subwindows of W are “distinct enough” averages, we can conclude that expected values are different, and the older portion of the window is dropped. pseudo-code checks average in both subwindows differs by more than threshold ϵcut. This threshold is calculated using the Hoeffding bound, thus gives guarantees of the classifiers performance. [7][8]

B. Drift detectors

This is a group of algorithms allowing to adapt mostly any learner to evolving data streams are *drift detectors*. The main activity of this algorithm is to detect concept drift and alert the learner that it should be now rebuilt or update the existing model.

DDM
DDM stands for Drift Detection Method in which each iteration of the online classifier predict the decision class which is either true or it may be false. So for the set of examples error is calculated from Bernoulli trials. For every example in data stream we update two register to keep track of error rate first is p_{min} and Secondly the s_{min} . This two are used to calculate warning level condition and alarm level condition. Whenever the warning level reach examples are remember in the separate window. And if alarm level reach the previously learned classifier is dropped and new is adopted from the example stored in separate warning window. [9]

EDDM
EDDM is the modification of DDM proposed by Baena-Garcia et al. In this algorithm use the same warning-alarm mechanism which is proposed in DDM, but it uses the distance error rate instead of classifier’s error rate. EDDM performs better in the case of slow gradual drift but it is more sensitive to noise. [10]

C. Hoeffding trees

Hoeffding trees are became popular because they represent current state of-the-art for classifying high speed data streams. It was introduced by Domingos and Hulten in their paper “Mining High-Speed Data Streams”. Hoeffding Tree is the theoretical algorithm while VFDT is introduced for practical implementation.

The Hoeffding bound states that with probability 1-δ, the true mean of a random variable of range R will not differ from the estimated mean after n independent observations by more than:

$$\epsilon = \sqrt{\frac{R^2 \ln(\frac{1}{\delta})}{2n}}$$

For selecting best splitting attribute after viewing all the example it uses hoeffding bound to select number of example necessary to select the right split-node with probability 1-δ. [11]

CVFDT
CVFDT is the improved algorithm of VFDT which adapt time changing data. This algorithm is proposed by Hulten et al. in the paper “Mining Time-Changing Data Streams” which solve the problem of classifying time changing data streams using Hoeffding trees. CVFDT uses fixed-size window to determine which nodes are old and may need updating. For Hoeffding tree that become old and less accurate than new node, alternative subtrees are grown and replace the outdated nodes. Outdated examples are removed from window and adapt newly arrived examples. [12]

D. Ensemble approaches

Ensemble classifier algorithm are combination of single classifiers in which decision is made by voting rule. The decision made by voting rule is more accurate than the single classifier. Ensemble classification is costly process. For huge data stream single classifier is better than ensemble classifier because there is not time to update and running the ensemble classifier.

DDD
The algorithm DDD proposed a new online ensemble learning approach which called as Diversity for Dealing with Drifts (DDD). A recent study says that different diversity levels in learning are required for maintaining the both old and new example from data stream. DDD maintains ensembles which has different diversity levels and it is able to maintain good accuracy than

other ensemble approach. Also, it is very robust, performs well to handle concept drift in terms of accuracy in case of higher false positive drift. [13]

Streaming Ensemble Algorithm

Streaming Ensemble Algorithm (SEA) proposed by Street and Kim which is an ensemble method that changes its structure to react to changes. The authors proposed replacement strategy for weakest classifier based on two approach: accuracy and diversity. The incoming data stream is divided into chunk. Each chunk is used to train new classifier. Based on parameter discussed weaker classifier is removed and new one is adopted in ensemble. There are few more ensemble algorithm named Accuracy Weighted Ensemble and Accuracy Diversified Ensemble. [14]

IV. NOVEL CLASS DETECTION

Novel Class Detection is major concept of concept evolution. In today's world novel class is became interesting issue and novel research area for researcher. In data stream classification we can't assume fixed number of class because new class may arrives at any time in real environment. When new class evolve, most existing data classification technique ignore the important aspect of novel class arrival [15].

Example

Classification rules:
R1. If (x > x1 and y < y2) or (x < x1 and y < y1) then class = +
R2. If (x > x1 and y > y2) or (x < x1 and y > y1) then class = -
Existing classification models misclassify novel class instances

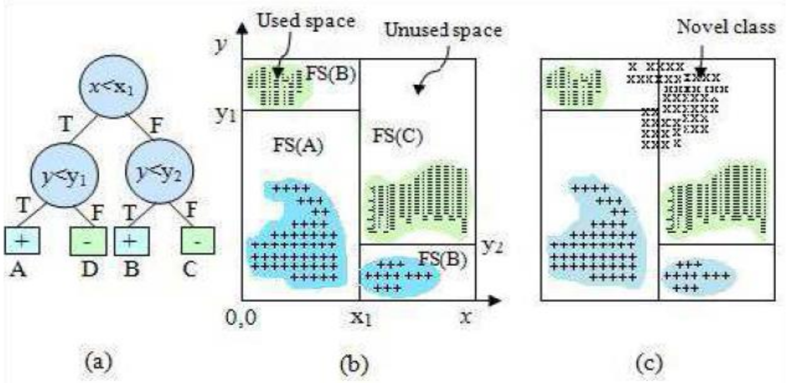

Figure 1: (a) Decision Tree (b) used space in each partition. (c) Novel class (x) arrives in unused space. [16]

In [16] 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*).
So, the traditional data stream classification techniques are not capable of detecting the novel class instances or novel class until it is manually identified by user, and labeled instances of that class are presented to the learning algorithm for training. The problem becomes even critical when there is concept drift, when the underlying data distribution changes over time. For example, in the case of intrusion detection technique, whenever a new kind of intrusion occurs, we should not only be able to detect that it, but also knows that it is a new kind of attack. Therefore our technique should automatically recognize novel instance whenever it comes and find that it forms novel class or not.

V. RELATED WORK

A. OLINDDA

OLINDDA, stands for OnLIne Novelty and Drift Detection Algorithm, OLINDDA technique based on k-mean clustering algorithm aimed at detecting concept drift as well as novel class situations in a single learning strategy. This technique is related to statistical outlier detection. It is one class or single class classification technique. In this technique normal concept is built based on the k-means algorithm. K-mean gives k cluster as well as position of centroid. Now for concept drift and novelty detection if clusters are closer to the boundaries of the model are happens due to a concept drift in the normal concept. And validated cluster appearing far from the normal concept is due to a novel concept. To establish this limit, we calculate dmax. [17]

B. MineClass

MineClass algorithm is stands for mining novel class in streaming data which uses two base learner Decision Tree approach and K-NN (K-nearest neighbor) based approach. MineClass is related to statistical approach because it uses K-NN which is non-parametric. MineClass provides the multiclass framework. Novel class detection technique is mainly related to outlier/anomaly detection. MineClass uses ensemble classifier which is collection of decision trees. MineClass introduce the notion of used space and unused space and according to property (separation) of novel class is will arrive is unused space. In this approach are: for decision tree keep track of the used spaces of each leaf node, and find strong cohesion among the test instances and for K-NN keep track of all training instances for used space. In this cluster summary is saved as “Pseudopoints” to reduce memory requirement. If many instance beyond threshold have a strong cohesion amongst them then we can say novel class is arrived. [16]

C. ECSMiner

ECSMiner is differ from some traditional technique which is one class novel class detection. This approach also offers a “multiclass” framework like MineClass for the novelty detection problem that can differentiate between different classes of data discover novel class. It is Non-Parametric approach. ECSMiner uses two base learner Decision Tree approach and K-NN (K-nearest neighbor) based approach. This technique considers time constraints. To show similarity among instances sometimes the classification model has to wait. A maximum allowable wait time Tc is imposed as a time constraint to classify a test instance. Moreover we can’t assume that the true label of a data point can be available instantly after the data point is classified, most existing technique does so. In reality, a time delay Tl is involved in obtaining the true label of a data point because manual label the instance is time consuming. [18]

D. ActMiner

ActMiner addresses main four challenges of stream classification which is Infinite Length, Concept Drift, Concept Evolution (Novel Class Detection), Limited Labeled Data. Most of the existing data stream classification algorithms address the infinite length and concept-drift problems. Above given two algorithm MineClass and ECSMiner addresses the concept-evolution problem with infinite length and concept-drift problems. The scarcity of labeled data is result in poorly build classifier. ActMiner actively selects only those data points for which has the higher classification error, only those data point are labeled. ActMiner selects only a few instances for labeling, so that it saves lots of labeling time and cost approx. 90% or more, than the traditional approaches require for all instances to be labeled. [15]

E. SCANR

SCANR is stands for “Stream Classifier And Novel class And Recurring Class detector”. Recurring Class is a special case of concept evolution which arrives many times in stream classification. It happens when a class arrives in the stream, then after disappears for a long time, and then again appears. Existing techniques denotes this concept-evolution problem, wrongly detecting as novel is case of the recurring. Problem comes with that is to detect novel class is much more costly operation than recurring class due to that many resource involved in this process. Secondly it increase false alarm. SCANR algorithm solve this problem of recurring class. Each incoming instance is first check by primary ensemble if it is outlier then it is called primary outlier (P-outlier). Then again instance check by auxiliary ensemble if it is outlier in that also than called secondary outlier(S-outlier), and it is temporarily stored in a buffer. When enough instances in the buffer, the novel class detection module is invoked. [19]

F. Decision Tree

In this approach Decision Tree is used to detect multiple novel class. The basic decision tree algorithm ID3 (Iterative Dichotomiser) builds decision tree. In this technique, first build a decision tree from training data points and calculate the percentage of number of data points in each leaf node with respect to data points in training dataset. Now apply cluster on each leaf node of tree based on similarity. In real time classification novel class is arrived if number of data point in leaf node of tree is increase than percentage calculated before. The idea of detecting multiple novel class is to construct graph, novel class obtained is plotted on graph. After constructing the graph identify connected component, the number of connected component determines the number of novel class arrives. [20]

G. MINAS

To deal with novelty detection new technique arrives called MINAS. MINAS is stands for *MultI-class learNing Algorithm for data Streams*. This new algorithm addresses multi class problem of novelty detection in data stream. MINAS has two phases: offline phase and online phase. In the first *offline* phase learning of decision model with the known concept of the problem, which is executed only once. Secondly, the *online* phase done the work of receiving new example as well as classify them in known or unknown classes. Also this algorithm allows to forget the old example and adapt new ones to effectively handle concept drift. [21]

VI. COMPARATIVE ANALYSIS OF NOVEL CLASS DETECTION

Table 1 gives comparative analysis of various algorithms of Novel Class Detection based on Learning approach, Type of classifier, Review, Advantages.

Table 1: Comparative Analysis of different Algorithm of Novel Class Detection

Algorithm	Learning Approach	Classifier	Review	Advantages
OLINDDA	Cluster based	k-mean Clustering	Single Class Approach Detect Concept Drift and Novelty	Efficient in terms of memory and run time.
MineClass	Ensemble	Decision Tree and K-NN	Detect Novel Class With Concept Drift	Does not require data in convex shape.
ECSMiner	Ensemble	Classical Classifier with Decision Tree and K-NN	Detect Novel Class with Time Constraint	Does not require data in convex shape.
ActMiner	Ensemble	Active classifier with Decision Tree and K-NN	Detect Novel Class with limited Labeled Data.	Work with limited labeled data so saves 90% labeling cost.
SCANR	Ensemble	MultiClass Classifier	Detect Recurring and novel class in concept drifting data.	Identify reappearing class as “not novel”.it saves time and cost.
Decision Tree	Incremental	Decision Tree based	Detect Multiple Novel Class using Connected Graph	Simple and efficient for detecting multiple novel class.
MINAS	Cluster based	*k*-mean and Clustream	Detect Novel Class using Online and Offline Phase	Low computation Cost.

VII. CONCLUSION

This paper summarizes the current techniques in data stream mining for concept drift and novel class detection. First it introduce with the concept drift and novel class and it’s importance in today’s world and in real application. Than it summarizes some of concept drift technique like Weighed Window, FISH, ADWIN, DDM, EDDM, CVFDT, DDD, Streaming Ensemble Algorithm. It also summarizes different novel class detection technique like OLINDDA, MineClass, ECSMiner, ActMiner, SCANR, Decision Tree, MINAS. And also gives the comparative analysis of the techniques.

VIII. REFERENCE

[1] 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.
[2] A book on “stream data mining”. the university of waikato.
[3] Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., Bouchachia, A.”A survey on concept drift adaptation” ACM Computing Surveys 46(4) (2014)

[4] Jeffrey t. byorick, assesing classification confidence using a weighted exponential based technique with the learn++ incremental learning algorithm, lecture notes in computer science volume: 2714, 2002, pages 181-188

[5] Edith Cohen and Martin J. Strauss. Maintaining timedecaying stream aggregates. J. Algorithms, 59(1):19–36, 2006.

[6] Indr˙e ˇZliobait˙e. Adaptive training set formation. PhD thesis, Vilnius University, 2010.

[7] Albert Bifet and Ricard Gavald`a. Kalman filters and adaptive windows for learning in data streams. In Ljupco Todorovski, Nada Lavrac, and Klaus P. Jantke, editors, Discovery Science, volume 4265 of Lecture Notes in Computer Science, pages 29–40. Springer, 2006.

[8] Albert Bifet and Ricard Gavald`a. Learning from timechanging data with adaptive windowing. In SDM. SIAM, 2007.

[9] J. Gama, P. Medas, G. Castillo, and P. Rodrigues. Learning with drift detection. In SBIA Brazilian Symposium on Artificial Intelligence, page 286–295, 2004.

[10] Manuel Baena-Garc´ia, Jos´e del Campo- ´Avila, Ra´ul Fidalgo, Albert Bifet, Ricard Gavald´a, and Rafael Morales-Bueno. Early drift detection method. In In Fourth International Workshop on Knowledge Discovery from Data Streams, 2006.

[11] Pedro Domingos and Geoff Hulten. Mining high-speed data streams. In KDD, pages 71–80, 2000.

[12] Geoff Hulten, Laurie Spencer, and Pedro Domingos. Mining time-changing data streams. In KDD, pages 97–106, 2001.

[13] leandro l. minku, member, ieee, and xin yao, fellow, ieee ddd: A new ensemble approach for dealing with concept drift ieee transactions on knowledge and data engineering, vol. 24, no. 4, april, 2012

[14] W. Nick Street and YongSeog Kim. A streaming ensemble algorithm (sea) for large-scale classification. In KDD, pages 377–382, 2001.

[15] Mohammad m. masud, jing gao, latifur khan, jiahan, bhavani thuraisingham classification and novel class detection in data stream with active mining m.j.zaki etal.(eds.): pakdd 2010, part ii,lnai 6119, pp.311-324 springer- verlag berlin heidelberg 2010.

[16] 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.

[17] Spinosa, E.J., de Leon, A.P., de Carvalho, F., Gama, J.: Olindda: a cluster-based approach for detecting novelty and concept drift in data streams. In: Proc. 2007 ACM symposium on Applied computing, pp. 448–452 (2007)

[18] Mohammad m. masud, jing gao, latifur khan, jiawei han, bhavani thuraisingham classification and novel class detection in concept-drifting data streams under time constraints: ieee transactions on knowledge and data engineering, vol. 23, no. 6, june 2011

[19] Mohammad M Masud, Tahseen M, Al-khateeb, Latifur Khan, Charu Aggrawal, Jing Gao, Jiawei Han and Bhawani Thuraisinghum Detecting Recurring and Novel classes in Concept Drift Data Streams icdm, pp. 1176- 1181, 2011 IEEE 11th International Conference On Data Mining.

[20] Amit Biswas, Dewan Md. Farid and Chowdhary Mofizur Rahman A New Decision Tree Learning Approch For novel Class Detection In Cocept Drifting Data Stream Classification journal of computer science and engineering, volume 14, issue 1, july 2012.

[21] Elaine R. Faria, João Gama, André C. P. L. F. Carvalho Novelty Detection Algorithm for Data Streams Multi-Class Problems ACM SAC '13, March 18-22, 2013