Multi Novel Class Classification of Feature Evolving Data Streams with J48

Punam D. Dhande
ME IInd Year Student, Dept. of Computer Engineering
PVPIT, Bavdhan
Pune, India

A.M. Dixit, PhD
Computer Engineering Department, PVPIT, Bavdhan,Pune
Department of Technology, Savitribai Phule Pune University, Shivajinagar, Pune, India

ABSTRACT

In the Data stream classification main issues are infinite length, concept drift, concept development, and feature development. Hypothetically data stream is infinite in length; it is impossible for storing and use all the traditional for training. In the existing system of data stream method researcher tackle on the only two issues i.e. concept drift and concept evolution problem of classification. In the existing system for tackling the issue of feature evolution feature set homogeneous technique was developed and also focus on the novel class detection technique for detecting the novel class at a time, but this method required more time for detecting novel and multi novel class detection. Therefore we used the method for detecting the novel class method for data stream classification, we used J48 classification algorithm for detecting the novel class and reducing the time for detecting the novel class. Finally we compared our result with the existing novel class detection method.

Keywords

Classification; Data Stream Classification; J48 classifier; novel class; features evaluation

1. INTRODUCTION

With the substantial number of transactions which are recorded and accumulate by numerous associations, the significance of being able to examine trends in fast data streams has become very important. Commonly, such databases are made by unbroken activity over long period of time and are therefore databases which develop without limit. For instance, even basic transactions of regular life, like, paying by credit card or utilizing the phone are recorded in an automated way by using the current hardware technology. The volume of such transactions might effectively run in the millions on a daily basis. Frequently, the information may indicate important changes in the patterns over the long because of basic alter in the underlying phenomena. This methodology is referred to as information advancement. By understanding the nature of such changes, a client may have the capacity to gather important experiences into rising patterns in the basic transactional alternately spatial movement. Subsequently, it is helpful to create tools and techniques which would give a visual and symptomatic overview of the key qualities in the information which have changed over the long run in a fast and easy to understand way.

Now days, data stream classification is the focus area for researchers. Data stream has the active and increasing nature of data streams need effectual and flourishing methods that are not quite the similar as still data classification technique. Infinite length and concept drift are two most tricky and usually measured attributes for information streams. Data stream is a rapid and reliable event, it is considered to have infinite length. It is impossible to store the tremendous amount of data for the training process. Incremental learning procedure is one of the most preferred options. Incremental learners have been proposed for preventing the problem of classifications. For preventing the concept drift problem number of methods have been proposed. However, there are two other grave qualities of information stream, concept evolution and feature evolution which are ignored by the number of the current system of novel class detection method.

When the new class arrived in the data then the problem of concept evolution are proposed. For example, the problem of intrusion detection. The concept evolution takes place, if every sort of attack is measured as a class name, when a totally new sort of attack takes place in the traffic. An alternative solution is the circumstances of a text data stream. Because of this situation, novel classes might frequently increase in the core stream of text messages. The issue of concept development is tackled in number of existing data stream classification systems.

Initially in [13] studied the novel class detection problem in the neighbourhood of concept-drift and infinite length. An assembly of model is utilized for categorizing the unlabeled data, and identifying the novel classes. Novel class detection contains the three stages which are: first one limit of decision is made among the training. Second one is test points are announced as anomaly which is fall outside the decision limit. Third are the anomalies which are examined to check whether there is sufficient unity among themselves and division from the current class examples. The problem of feature evolution is not addressed in this study. In [14], the problem of feature-evolution is studied; additionally concentrate on the problem of concept-evolution issue. On the other hand, both [13] and [14] have two limitations. i) The false alarm rate that is recognition of existing classes as novel is more for few data sets. ii) If there is present more than one novel class, they are not able to recognize among them.

System focuses on the issues shown up during the process of classification. Here they focus on the problem of concept drift, concept evolution, infinite length and feature evolution. They also proposed the method for novel class detection; we will use J48 classification method which detect multi novel class detection. They proposed Density based clustering algorithm for chunking the data streams and compare the result with the existing k-means algorithm. The proposed method increases the efficiency of clustering algorithm. We will going to compare the efficiency and computational time for novel class detection algorithm.

Remaining paper will be organized as follows: in section II we discussed about the related work done for the data stream classification and discussed the literatures regarding the data stream classification. In Section III we discussed about proposed system, i.e. in this section we will show whole system with system representation, implementation details, mathematical models and proposed algorithm. In section IV we will discussed about results obtain from the proposed system. In section V we will discussed the conclusion and future scope. And finally shows the references used for the paper.

2. RELATED WORK

In this section we discussed some of the existing data stream classification methods designed for resolving the issues of concept drift, concept evolution, feature evolution, etc. In this section discussed literatures who work on the data stream classification method.

Fan [1] proposed cross validation decision tree ensemble technique, where initially algorithm senses all features with information gain. Later on builds multiple decision trees by randomly choosing from these features with information gain and disregard the irrelevant features. C. Agrawal [2] converse the concept of velocity density estimation, technology used to understand, visualize and determine tendency in the evolution of fast data streams. Babcock, B., Babu, S., Datar, M., Motawani, R., and Widom, J.[3] stimulate the necessitate for and research issues arising from a new model of data processing. In their work they do not take data in the form of persistent relations, rather arrives in multiple, continuous, rapid, time varying data streams. J. Gao, W. Fan, and J. Han [4] reveals that the robustness of a model averaging and simple voting based framework for data stream.J. Gao, W. Fan, J. Han, and P. Yu [5] introduced a novel method for mining data stream by estimating dependable posterior probabilities using an group of model for matching the distribution over under samples of negative and repeated sample of positives. W. Fan, P. S. Yu, and H. Wang [6] introduced the methodical technique for mining very skewed distribution in very large volume of data.

Ali, P Chia, K. Ong [7] represents a data stream summary which can answer point queries with the threshold accuracy and shows the space needed. The method demon-strate that the significant skew is present in both textual and telecommunication data. G. Hulten, L. Spencer, and P. Domingos introduced [8] a competent algorithm to mine decision tree from incessantly changing data streams on the basis of VFDT decision tree learner which is a ultra fast decision tree learner.I. Katakis, G. Tsoumakas, and I. Vlahavas highlight [9] the requirement for a dynamic feature space and usefulness of addition feature selection in streaming text classification tasks. Additionally they show a computationally straightforward incremental learning architecture which could serve as the baseline in the field. Finally they introduce a novel method drifting dataset which could help other researchers in the development of novel methodologies. T. Fawcett [10] shows the uniqueness which make it rich and demanding domain for the data mining and dispute that the researchers follow in vivo spam filtering as an accessible domain for examining them.F. Ferrer-Troyano, J. S. Aguilar-Ruiz, and J. C. Riquelme [11] demonstrate the classification technique which is based on the incremental learning approach. Here they demonstrate a classification system on the basis of decision rules which may store up to date examples of up to date for avoiding needless revisions when the virtual drifts are exist in the data. Additionally system offers an understood forgetting heuristic so that the positive and negative instances are removed from a rule when they are not near one another's. Kolter and M. Maloof [12] represent the additive expert ensemble algorithm, which is a novel method for using any online learner for drifting concept. They acclimatize methods for examining expert prediction algorithm for proving the error and loss bounds for a discrete and a continuous version of addexp.

3. IMPLEMENTATION DETAILS

A. *Propose System*

The proposed System is as follows:

1) Initially user uploads the adult dataset for feature selection, detecting novel and multi novel class detection.

2) After that the process of feature selection is done. In feature selection module important features from the dataset are selected. In feature selection module by implementing the feature selection algorithm important attributes of the dataset are selected. The output of this stage or important features is saved in the txt file for the further process.

3) After selecting the features, outlier detection using DBscan is done. In data mining outliers are the exits data object which does not comply with the general behaviour or model of the data. Such data objects are revoltingly distinct from the remaining set of data. We pass two parameters for detecting outliers which are minimum points and radius using DBscan algorithm.

4) After detecting outliers they are group into clusters by using k-means clustering. Then novel classes are detected by Gini coefficient.

5) Multi novel class detected from the buffer by using the J48 classifier. In the following section we will discussed the J48 algorithm in detail.

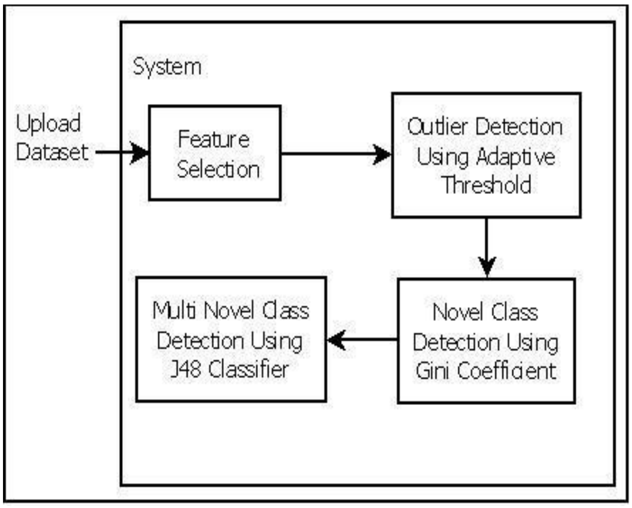

System
Upload Dataset
Feature Selection
Outlier Detection Using Adaptive Threshold
Novel Class Detection Using Gini Coefficient
Multi Novel Class Detection Using J48 Classifier

Fig.1: System Architecture

B. Algorithm

J48 classifier is the classification algorithm used for detecting the novel and multi novel class. For the problem to the classification the methodology of decision tree is used. For modelling the classification process tree is build. While the tree is generated it is connected with each column of the database and results in classification for that column.

Algorithm 1 Novel Class Detection Using Gini Coefficient

1: The F outliers detected during the outlier detection phase may occur because of three different reasons: noise, concept drift, or concept evolution.
2: These three cases may occur simultaneously too.
3: To distinguish the F outliers that occur because of concept evolution only, we compute a metric called discrete Gini Coefficient of the F-outlier instances. Gini Coefficient is usually used to measure statistical dispersion.
4: The value of Gini Coefficient is within the range [0, 1].
5: Ideally, the higher the dispersion, the higher the value of the Gini Coefficient.

Table Notation

DS	Dataset
DT	Decision Tree

Algorithm 2 J48 algorithm:

```
Input: Training data DS
Output: Decision Tree DT
DSTBUILD (*DS)
f
DT=';
DT=Generate root node and label with splitting attribute;
DT=Add arch to root node for each splitting predicate and label;
DS=By applying split predicate to DS database is created;
If stopping point reached for this path, then;
$DT^0$ =generate leaf node and label with the appropriate class;
$DT^0$ =DSTBUILD(*DS);
Else
$DT^0$ =DSTBUILD (DS);
DT=add $DT^0$ to arc;
g
```

The J48 classifier for establish the tree does not require any code. While constructing a tree, J48 rejects the missing qualities i.e. the quality for those things can be anticipated focused around which is the thought about characteristics qualities for the other record.

C. Experimental Setup

The system is built using Java framework (version jdk 8) on Windows platform. The Netbeans (version 8.0) is used as a development tool. The system doesn't require any specific hardware to run; any standard machine is capable of running the application.

4. RESULTS AND DISCUSSION

A. Dataset Used

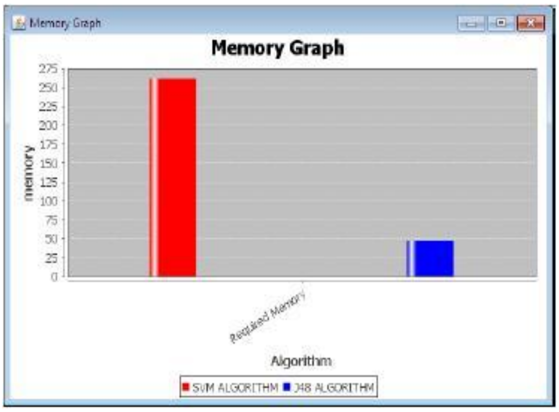
Here, we used two real data sets they are ASRS dataset and Forest Cover data set.

B. Result

In the following graph we compare the result of the proposed method with existing method.i.e.J48 method with SVM method.

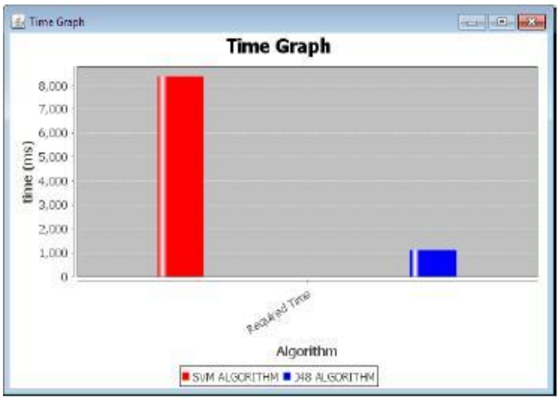
1. Memory Graph

In the comparison we discussed about memory required for proposed algorithm and existing algorithm



Graph 1: Memory Graph

2. Time Graph
In the comparison we discussed about time required for proposed algorithm and existing algorithm

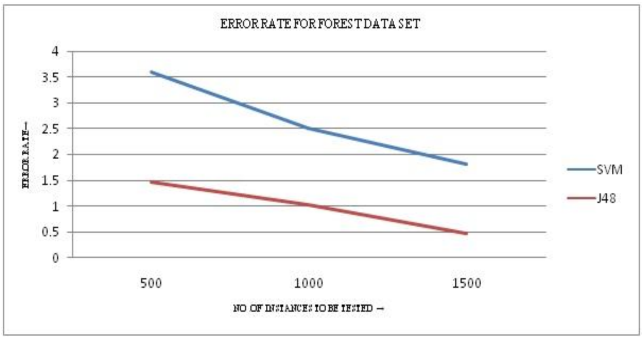


Graph 2: Time Graph

In the following graph we compare the error rate of proposed novel class detection method with the existing method. In X axis represent no. of instances for comparison and Y axis the error rate.

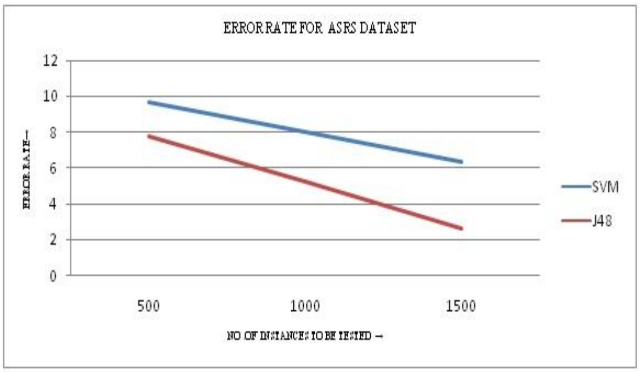
3. Error rate Graph

A. For Forest_Outlier.arff



Graph 3: Error rate for forest dataset

B. ASRS dataset

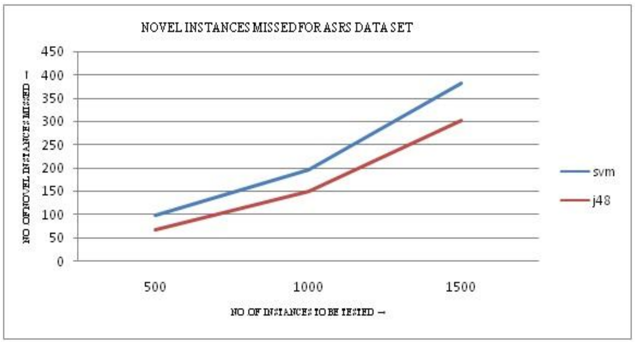


Graph 4:Error rate for ASRS dataset

4. Novel instance missed

In the following graphs we will show how many novel instance missed in Forest and asrs dataset.

A. Asrs.arff

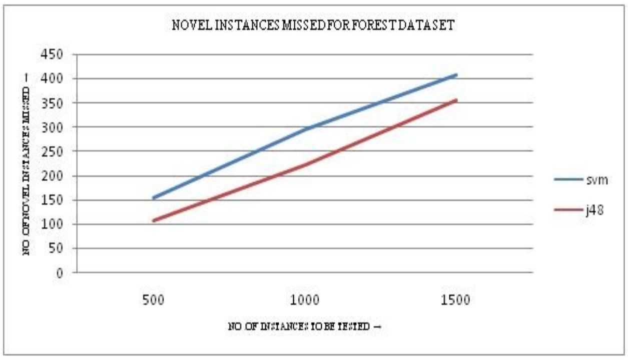


Graph 5: Novel instance missed in ASRS dataset

Y axis: Number of Novel Instances Missed

X axis: - Number of Instances tested

B. Forest Outlier Data



Graph 6: Novel instance missed in forest dataset

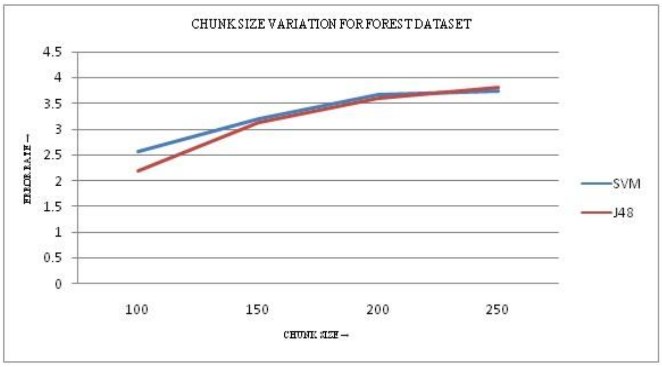
Y axis: Number of Novel Instances Missed

X axis: - Number of Instances tested

5. Error rate with chunk size variation

In the following graphs we show the effect of different chunk size with changing error rates for both dataset.

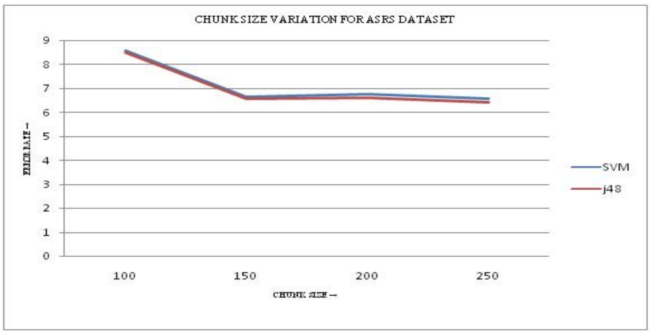
A. Forest Outlier Data



Graph 7: Error rate with chunk size variation in forest dataset

Y axis: Error rate
X axis: Chunk Size

B. ASRS dataset



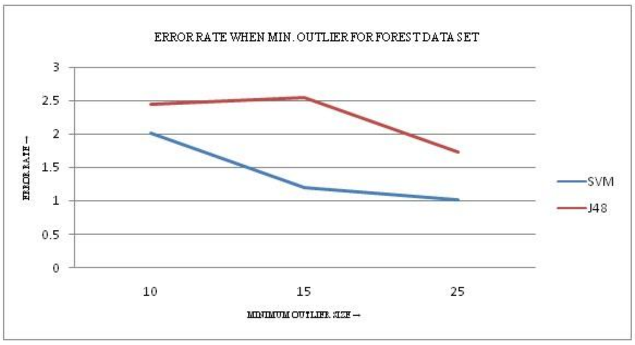
Graph 8: Error rate with chunk size variation in ASRS dataset

Y axis: Error rat

X axis: Chunk Size

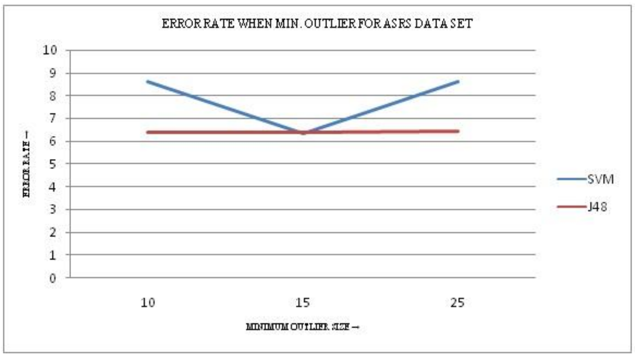
6. Error Rate when we vary minimum outlier required for novel class detection

A. Forest dataset



Graph 9: Error Rate when we vary minimum outlier required for novel class detection

B. ASRS



Graph 10: Error Rate when we vary minimum outlier required for novel class detection

X axis Minimum outliers for novel class detection

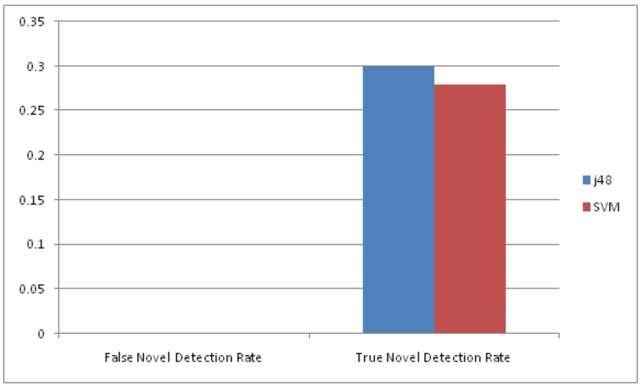
Y axis – err rate

7. ROC curve

This is Receiver Operating Characteristic curve used to plot true novel detection rate against false novel detection rate.

On the data set of the 1500 instances

In following graph, False Novel Detection Rate come 0 for both algorithms and True Novel class detection rate comes 0.3 and 0.28 for j48 and SVM respectively.



Graph 11: ROC curve

X axis –False/True Novel Detection Rate

5. CONCLUSION

Data stream classification faces many problems such as infinite length, concept drift, concept evolution and feature evolution. In this system we used the method for classification based novel class detection method by considering the four problems of classification like infinite length, concept drift, concept assessment and feature assessment. We introduced and used J48 classification algorithm for detecting the multi novel class detection. We used density based clustering algorithm for the process of clustering the data into chunks. Finally we compare the results of existing and proposed system, and we conclude that the proposed method for detecting the multi novel class detection by using the J48 classification is efficient than the existing methods of detecting novel class. ROC rate that calculates true novel verses false novel class detection. False Novel Detection Rate come 0 for both algorithms and True Novel class detection rate comes 0.3 and 0.28 for j48 and SVM respectively.

An interesting future work would be to recognize the special case more exactly to differentiate from the actual influx of a novel class.

6. REFERENCES

[1] M.M. Masud, Q. Chen, L. Khan, C. Aggarwal, J. Gao, J. Han and Nikunj C. Oza " Classification and adaptive novel class detection of Feature-Evolving Data stream" IEEE Trans. Knowledge and Data Eng., vol. 25, no. 7, July2013.

[2] W. Fan, "Systematic Data Selection to Mine Concept-Drifting Data Streams," Proc. ACM SIGKDD 10th Int'l Conf. Knowledge Discovery and Data Mining, pp. 128-137, 2004.

[3] Aggarwal, C. C. (2003). A framework for diagnosing changes in evolving data streams. In Proceedings of ACM SIGMOD 2003, pages 575–586.

[4] Babcock, B., Babu, S., Datar, M., Motawani, R., and Widom, J. (2002). Models and issues in data stream systems. In ACM Symposium on Principles of Database Systems (PODS).

[5] J. Gao, W. Fan, and J. Han, "On Appropriate Assumptions to Mine Data Streams," Proc. IEEE Seventh Int'l Conf. Data Mining (ICDM), pp. 143-152, 2007.

[6] J. Gao, W. Fan, J. Han, and P. Yu. A general framework for mining concept-drifting data streams with skewed distributions. In Proc. SDM'07.

[7] W. Fan, P. S. Yu, and H. Wang. Mining extremely skewed trading anomalies. In Proc. of EDBT'04

[8] F. Korn, S. Muthukrishnan, and Y. Wu. Modeling skew in data streams. In Proc. of SIGMOD '06

[9] G. Hulten, L. Spencer, and P. Domingos, "Mining Time-Changing Data Streams," Proc. ACM SIGKDD Seventh Int'l Conf. Knowledge Discovery and Data Mining, pp. 97-106, 2001.

[10] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Dynamic Feature Space and Incremental Feature Selection for the Classification of Textual Data Streams," Proc. Int'l Workshop Knowledge Discovery from Data Streams (ECML/PKDD), pp. 102-116, 2006.

[11] T. Fawcett. "in vivo" spam filtering: A challenge problem for data mining. KDD Explorations, 5(2), December 2003

[12] F. Ferrer-Troyano, J. S. Aguilar-Ruiz, and J. C. Riquelme. Incremental rule learning based on example nearness from numerical data streams. In SAC '05: Proceedings of the 2005 ACM symposium on Applied computing, pages 568–572, New York, NY, USA, 2005. ACM Press.

[13] J. Kolter and M. Maloof, "Using Additive Expert Ensembles to Cope with Concept Drift," Proc. 22nd Int'l Conf. Machine Learning (ICML), pp. 449-456, 2005.

[14] M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "Integrating Novel Class Detection with Classification for Concept-Drifting Data Streams," Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML PKDD), pp. 79-94, 2009.

[15] 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 (ECML PKDD), pp. 337-352, 2010.