Novel Class Detection for Feature Evolving Data Streams

Harshada Wagaskar[1], Prof. Gayatri Bhandari[2]

[1] JSPM's BSIOTR, Savitribai Phule Pune University, Gate No.720/1/2 Wagholi, Pune Maharashtra, India

[2] JSPM's BSIOTR, Savitribai Phule Pune University, Gate No.720/1/2 Wagholi, Pune Maharashtra, India

Abstract: ***Data Stream Mining is the method of deriving knowledge from constant and quickly developing records of information. A data stream is an ordered sequence of occurrences. These occurrences can be read just once or a less number of times utilizing restricted using limited storage capabilities and computing. Examples of such data streams include ATM exchanges, sensor information, telephone discussions and so forth. Data stream characterization has many difficulties in the information mining field. The four major challenges in the field of Data stream classification which are infinite length, concept-drift and concept-evolution are proposed here. A data stream is never-ending in length, hence it is not practical to store and utilize all the historical data for training purpose. Concept-drift occurs as a result of changes in the fundamental concepts. Concept-evolution happens when new classes develop in the information data. An example of concept-evolution is Twitter, where new themes develop routinely in the stream of instant messages. Feature-evolution is a regularly happening process in data streams, where new features evolve and old features vanish. This problem is investigated during this paper, and improved solutions are proposed. The current work additionally addresses the recurring class problem in data streams.***

Keywords: Data stream, concept-evolution, novel class, outlier

1. Introduction

Data mining derives its name from the similarities between checking out valuable business info during a massive info - for instance, _finding linked product in gigabytes of store scanner information - and mining a mountain for a vein of valuable ore. Each process needs either winnowing through associate huge quantity of fabric, or showing intelligence inquisitor to seek out precisely wherever the worth resides. Given databases of sufficient size and quality, data processing technology will generate new business opportunities by providing these capabilities. Data mining automates the method of finding predictive info in massive databases. Queries that historically needed in depth active analysis will currently be answered directly from the data quickly. A typical example of a predictive drawback is targeted selling. Data processing uses information on past promotional mailings to spot the targets possibly to maximize come back on investment in future mailings. Different predictive issues embody statement bankruptcy and different varieties of default, and characteristic segments of a population seemingly to retort equally to given events.

Data mining tools make databases and determine antecedent hidden patterns in one step. An example of pattern discovery is that the analyses of retail sales information to spot apparently unrelated merchandise that are usually purchased together. Different pattern discovery issues include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors. Data mining techniques will yield the platforms, and might be enforced on new systems as existing platforms are upgraded and new product developed. Once data processing tools are unit enforced on high performance data processing systems, they'll analyze large databases in minutes. Quicker process implies that users will mechanically experiment with additional models to know complicated knowledge. High speed makes it sensible for users to investigate vast quantities of information. Larger databases, in turn, yield improved predictions.

1.1 Concept Drift

In concept drift the data stream changes regarding every product is checked by calculating the threshold of the same.

1.2 Data Stream

In our project we consider amazon as our data stream, where we search many products and after data extraction we select one particular product.

1.3 Concept Evaluation

Here the data browsing is done out of threshold then a new class is generated called as a novel class.

1.4 Outlier

Outlier is a value which lies outside the threshold.

2. Problem Definition and Scope

2.1 Problem Definition

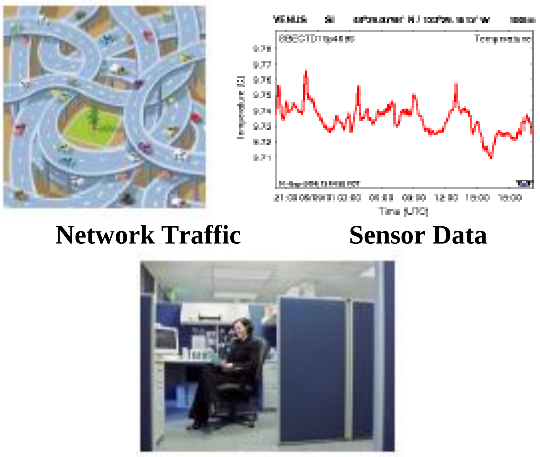
The existing novel class detection techniques for data streams either do not address the feature-evolution problem or suffer from high false alarm rate and false detection rates in many scenarios. To Overcome the Problems found in existing system, we propose a classification and novel class detection technique for concept-drifting data streams that addresses four major challenges namely, infinite length, concept-drift, concept-evolution, and recurring class. Our proposed approach applies ensemble classification framework, where each classifier is equipped with a novel class detector, to address concept-drift and concept-evolution. The problem of concept-evolution is addressed in only a very limited way by the currently available data stream classification techniques. We investigate this problem in this paper, and propose improved solutions. The current work also addresses the recurring class issue problem in data streams, such as text streams, where new features (words) emerge and old features fade away. Masud et al. address the novel class detection problem in the presence of concept-drift and infinite length. In this technique, an ensemble of models is used to classify the unlabeled data, and detect novel classes. The novel class detection process consists of three steps. First, a decision boundary is built during training. Second, test points falling outside the decision boundary are declared as outliers.

Finally, the outliers are analyzed to see if there is enough cohesion among themselves (i.e., among the outliers) and separation from the existing class instances.

2.2 Project Scope

To Present novel class detection technique for concept-drifting data streams that addresses four major challenges, namely, infinite length, concept-drift, concept-evolution, and feature-evolution. To enhance the novel class detection module by making it more adaptive to the evolving stream, and enabling it to detect more than one novel class at a time. An ensemble classification framework, where each classifier is equipped with a novel class detector, to address concept-drift and concept evolution. The scope of this paper is for Clustering and Informative selection purpose.

Data Stream means continuous flow of information. Example of information stream includes computer network traffic, phone communication, ATM transactions, and internet Searches and Sensor data.



Network Traffic **Sensor Data**

Call Center Records

Figure 2.1: Examples of Data Streams

Data Stream Mining could be a method of extracting data structure from continuous, fast information records. It behaves as a subfield of data mining. Information Stream Will be classified into on-line streams and offline streams. Online data stream mining utilized in variety of world applications, including internet work traffic watching, intrusion detection and credit card fraud detection. And offline information stream mining utilized in like generating report based on blog streams. Data Stream Classification uses past labelled data to build classification model. It predicts the labels of future instances using the model and helps in decision making.

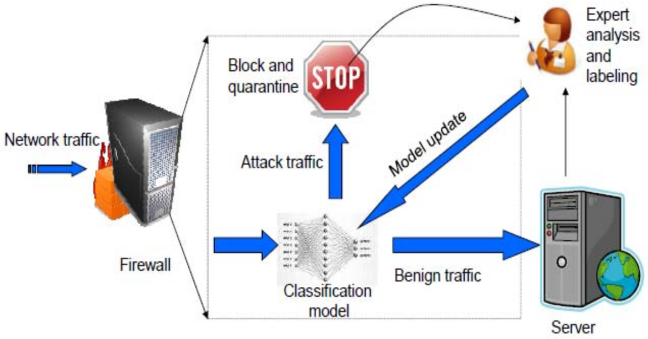


Figure 2.2: Data Stream Classification

Characteristics of data stream are continuous flow of information. Data size is extraordinarily giant and infinite. It is not possible to store all the data. Data Stream classification faces three major challenges Concept-Drift, Concept Evolution and Feature Evolution. Concept-drift occurs in the stream when the fundamental ideas of the stream change over time. A mixture of methods have been proposed in the writing for solving addressing concept drift in data stream classification. However, there are two other important qualities of data stream which are concept-evolution and feature evolution that are disregarded by a large portion of the current systems. Concept-evolution takes place when new classes develop in the data. For instance, consider the issue of intrusion detection in a network traffic stream. Here, if every attack is considered as Class label, then concept-evolution occurs when completely new type of attack happens in the traffic. An alternate illustration is the situation of a content information stream. For example, that happening in an informal organization, for example, Twitter. For this situation, new points (classes) might often develop in the basic stream of instant messages. The issue of concept-evolution is handled in a very limited way by the presently available data stream classification techniques. Finally, the feature space that represents a data point in the stream may change over time. For instance, consider a data stream where every information point is a record, and each one expression is a gimmick. Since it is difficult to know which words will show up later on, the complete peculiarity space is obscure. Plus, it is standard to utilize just a subset of the words as the list of capabilities in light of the fact that the vast majority of the words are liable to be repetitive for order. In this manner at any given time, the peculiarity space is characterized by the helpful words (i.e., peculiarities) chose utilizing some choice criteria. Since later on, new words may get to be valuable and old helpful words may get to be repetitive, the gimmick space changes alterably. Concept-drift occurs in the stream when the fundamental ideas of the stream change over time. Mixtures of methods have been proposed in the writing for solving addressing concept-drift. New approach proposed here addresses the feature-evolution problem in DataStream, such as text streams, where new features (words) emerge and old features fade away. Detection process consists of three steps. First, a decision boundary is built during training. Second, test points falling outside the decision boundary are declared as outliers .Finally, the outliers are analyzed to see if there is enough cohesion among themselves (i.e., among the outliers) and separation from the existing class instances. The proposed technique is applied on various data streams like Twitter data set, Forest cover data set, etc.

3. Proposed System

3.1 System Architecture

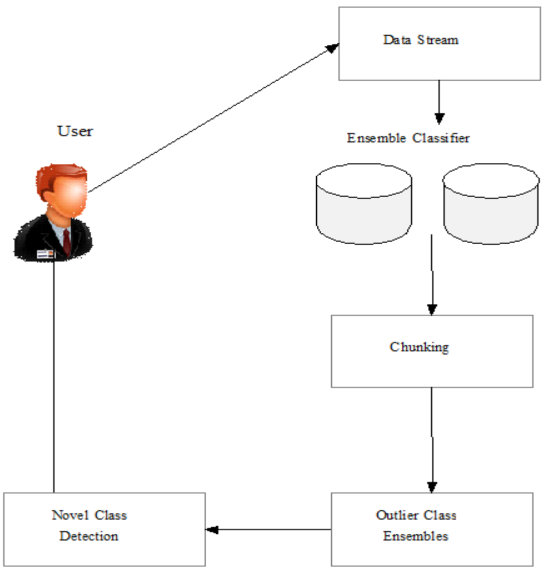


Figure 3.1: Business logic and architecture

The architecture represents the entire flow of the Novel Class Detection of Concept Evolving Data Streams: Stream is the main input given to the system, date stream acts as a container of multiple text documents. Each incoming instances in the data stream is passed through and examined by the outlier detection module of the primary ensemble for checking whether there is an outlier or not. It is classified as an existing class if it is not an outlier using majority voting by the primary ensemble and if is an outlier, it is then passed through the outlier detection module of the auxiliary ensemble and it is called as a primary outlier and if it is not an outlier it is considered as an recurring class instance and classified by the auxiliary ensemble and if it is an outlier that is detected by the auxiliary ensemble it is considered as an secondary outlier which is temporarily stored in the buffer for further analysis for recovery of the outlier.

3.2 Project Modules

1. Registration

In this module, an user have to register first, then only he/she has to access the data base. User registers with his own mobile number through his phone. The registration helps keeping the track of the usage and search history of the user. The registered user can see his past searches and make a decision more efficiently the next time.

2. Login:

In this module, any of the above mentioned person have to login, they should login by giving their username and password or can keep the mob logged in as being using from His own mobile. The person using his own mobile does not need to login as his application is already logged in with his no. The user can log out and log in as and when required.

3. Category

In this module Admin registers the category. Also he/she additional data regarding to group .This data helps the user search**.**

4. Sub Category

In this module Admin registers the sub category. Also he/she additional data regarding BSIOTR, Department of computer engineering (2014-15) 29 Novel class detection for feature evolving data streams to group .This data helps the user search.

5. Download code

In this module code present in the web page related to the category gets fetched and made available to user.

6. Parse

In this module code related to the subcategory gets fetched and made available to user.

7. Align

In this module the niche code regarding to the specific element sorted out and made available.

8. Threshold

Here we are using edge point for searching the user key.

9.Add Product:

This module will be suitable for offline system and will add products to the database.

10. Offers:

This module will provide offers to the less frequently searched products.

11. Add/Remove Offer

The Add module will apply offer on less searched products and Remove module will remove the offer if the product is being is searched and will be applying to the next least searched product.

12. Search techniques

Key Search:

Means that the user can give the key in which category and subcategory items should displayed.

4. Algorithm Strategy

4.1 Detect-Novel class using Threshold

```
1. cnt = 0;
2. total count = 0;
3. main total count = 0;
```

```
4. total added count = 0;
5. ProgressBarSimilarity.Value = 0;
6. ProgressBarSimilarity.Maximum = listTitle.Count - 1;
7. if listTitle.Count > 0
8. for each mainIndex < listTitle.Count
9. oat sum = 0;
10. cnt = 0;
11. for each subIndex < listTitle.Count
12. String value = null;
13. String value2 = null;
14. value = listTitle[mainIndex].ToString();
15. value2 = listTitle[subIndex].ToString();
16. MatchsMaker(value, value2);
17. total count++;
18. oat average = sum / (cnt - 1);
19. mainSum = mainSum + average;
20. ProgressBarSimilarity.Value = mainIndex;
21. mainAvg = (mainSum / total count);
22. THRESHOLD VALUE SIMILARITY = mainAvg;
23. discarded = 0;
24. for each mainIndex < ListTitleTemp.Count
25. if(ListTitleTemp[mainIndex].avg > THRESHOLD
VALUE SIMILARITY);
26. total added count++;
```

4.2 Clustering Algorithm

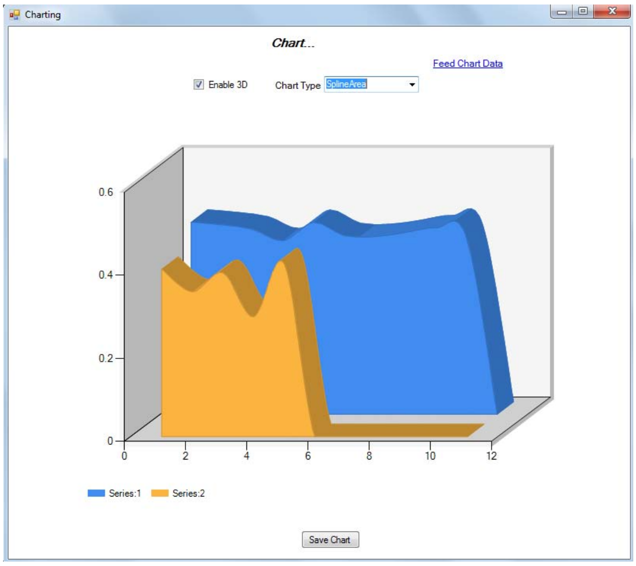
The reason for using K-means clustering is two-fold. The first reason is to create the decision boundary. While there are other alternatives such as finding the convex-hull or using density-based clustering to find the clusters that encompass the training data ; The K-means clustering is chosen because 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 Mb is reduced from O(S) to constant (i.e., K),where S is the chunk size.

4.2.1 Fuzzy K-means:

```
1. Initial means m1, m2,,,, mk
2. Where m1,m2,,,mk = 0;
3. Degree of means U (j,i) of xjin each cluster j
4. U(j,i) = a(j,i) / sum - j a(j,i)
5. For i=1 to k
6. Replace mi with mean cluster of j
7. Mi= U(j-i)2 xj /U(j-1)2
8. End for
9. End if
```

5. Evaluation And Results

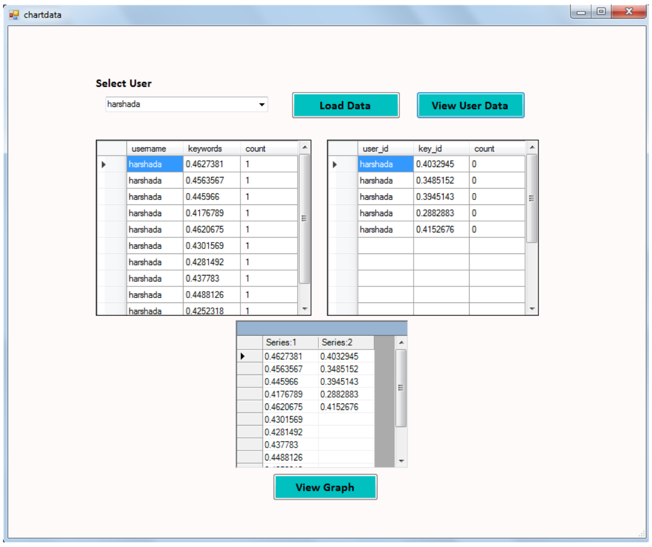
Here, we have explained a concept drift and concept evolution along with design aspects. The below graph explains the concept of Concept-Evolution. Series 1 indicates the values which lie inside the threshold. Series 2 indicates the values which lie outside the threshold which means a new concept has evolved. This is called as Concept Evolution.



☑ Enable 3D

Chart Type: SplineArea

Graph 5.1: Graph for Concept –Evolution



Select User: harshada

	username	keywords	count
▶	harshada	0.4627381	1
	harshada	0.4563567	1
	harshada	0.445966	1
	harshada	0.4176789	1
	harshada	0.4620675	1
	harshada	0.4301569	1
	harshada	0.4281492	1
	harshada	0.437783	1
	harshada	0.4488126	1
	harshada	0.4252318	1

	user_id	key_id	count
▶	harshada	0.4032945	0
	harshada	0.3485152	0
	harshada	0.3945143	0
	harshada	0.2882883	0
	harshada	0.4152676	0

	Series:1	Series:2
▶	0.4627381	0.4032945
	0.4563567	0.3485152
	0.445966	0.3945143
	0.4176789	0.2882883
	0.4620675	0.4152676
	0.4301569	
	0.4281492	
	0.437783	
	0.4488126	

Figure 5.2: Concept Drift

6. Conclusion and Future Work

A novel ensemble technique, which is superior to other data stream classification techniques because of its ability to detect novel class, and distinguish a recurring class from novel class is proposed. A recurring class is a class that disappears from the stream for a long time and reappears. Existing data stream classification techniques misclassifies a recurring class as another class, or identifies it as a novel class, they forget the class during its absence. The proposed approach can be considered a class-based ensemble, as Opposed to the chunk-based ensemble that is more popular in data stream classification. As future work, designing a new solution to identifying the Novel Class and also enhancing the drift changes by extracting the various techniques will be considered. The issue of Novel Class identification by extending the classifiers of new classes to detect the novel classes in the data stream changing drift may be the added reason for the data classification in the stream of data. The system will utilize less time, error rate and results reduced word count with Higher Speed.

References

[1] H. Wang, W. Fan, P.S. Yu, and J. Han, "Mining Concept-Drifting Data Streams Using Ensemble Classifiers," Proc. ACM SIGKDD Ninth Int'l Conf. Knowledge Discovery and Data Mining, pp. 226-235,2003.

[2] 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.

[3] Bennett, C. H., Bessette, F., Brassard, G., Salvail, L., & Smolin, J. (1992). Experimental quantum cryptography. Journal of Cryptology , 5 (1), 3-28.

[4] 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 (ECMLPKDD), pp. 79-94,2009.

[5] M.M. Masud, Q. Chen, J. Gao, L. Khan, J. Han, and B.M.Thuraisingham, Classification and Novel Class Detection of DataStreams in a Dynamic Feature Space,"

[6] Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML PKDD), pp. 337-352, 2010.

[7] 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.

[8] E.J. Spinosa, A.P. de Leon F. de Carvalho, and J. Gama, "Cluster-Based Novel Concept Detection in Data Streams Applied to Intrusion Detection in Computer Networks," Proc. ACM Symp. Applied Computing (SAC), pp. 976-980, and 2008.

[9] M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham,"Integrating Novel

[10] 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.

[11] 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.