www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 4 April, 2014 Page No. 5314-5318

An Effective Classification and Novel Class Detection of Data Streams

G.Divya[1], MR.D.BrightAnand[2]

[1]Computer science and Engineering, RVS College of Engineering and Technology,Dindigul,Tamilnadu,India. email: ittechdhivya@gmail.com,

[2] Professor-CSE Dept, RVS College of Engineering and Technology, RVS Nagar, Dindigul, Tamilnadu,India. email: brightanand@reddiffmail.com

***Abstract*— Data stream classification suffered from the problem of infinite length, concept drift, concept-evolution and feature-evolution in data mining community. Usually data streams are infinite in length and it makes difficult to store and use all the historical data for training. Several research is eliminating the concept-evolution and feature-evolution concepts. Recently, many researchers have been focused on data streams as an important approach against huge database mining instead of mine the entire database. In this paper, an efficient approach is proposed for classification and novel class detection of data streams. The proposed method uses the outlier detection method to remove the unwanted data present on the data streams. Also, this approach uses the Nearest Neighbor algorithm and the Naive Bayes classifier concepts for novel class detection. The performance is evaluated with respect to error rates, final word count, speed and time. The result shows that the proposed method for classification and novel class detection provides better results than the existing techniques.**

***Index Terms*— Concept drift, Concept-evolution, Data stream, Feature-evolution, Infinite Length, Nearest Neighbor, and Naive Bayes classifier**

I. INTRODUCTION

The major challenge of data mining community has been to mine the ever-growing stream data. Usually there are three major problems occurred in this domain. First, it is impractical to store and access all the historical data for training. Because, it needs infinite storage and running time to train the historical data. Second, there may be concept-drift in the data i.e. the underlying concept of the data changes over time. Third, novel classes may evolve in the data stream. Most of the existing techniques are not capable of detecting the novel classes in the data stream. Also, there are two major characteristics of data streams exist. They are concept-evolution and feature-evolution. But these concepts are ignored by most of the existing techniques. The infinite length problem is dividing the stream into equal-sized chunks, so that every chunk can be stored in memory and processed online. To cope up with the concept-drift, a classifier must continuously update with the recent concepts.

Concept-evolution occurs when new classes evolve in the data. It means the statistical property of the target variable; the model is trying to predict it over time. The problems may arise, because the predictions become less accurate as time passes. The problem of concept-evolution is addressed in a limited number of the existing systems. The proposed system addresses the concept and feature-evolution problem in data streams, like text streams. When a new feature emerges, the old one fades out and the new feature occupies. Outlier detection techniques are used to remove the unwanted data on the data streams to obtain the accurate results. To generate a novel class a machine learning techniques like supervised and unsupervised learning techniques can be applied. In supervised learning, training data includes both the input and the desired results. For example, the correct targets are known and are given as the input to the model during the learning process. In unsupervised learning, the model is not provided with the correct results during the training.

In this paper, a hybrid method is proposed to detect and classify the novel class in the feature-evolving data streams. The proposed system discussed about the four major challenges occurred in the classification and novel class detection of the feature-evolving data streams. The challenges are infinite length, concept-drift, concept-evolution and feature-evolution. The first step of the proposed technique is to remove the unwanted data and space occurred in the data chunks. Second step is to remove the spaces on the chunks. Then, outliers are detected and removed from the dataset. Finally it is classified based on the Nearest Neighbor algorithm and Naive Bayes classifier technique. The problem of concept-evolution is addressed in limited approaches by the currently available data stream classification techniques. So, in this proposed approach the problem of concept-evolution is addressed with the improved solutions. Also the feature-

evolution problem is also addressed in data streams such as text streams.

The rest of the paper is organized as follows. Section II presents a description about the previous research which is relevant to the classification and the novel class detection of data streams. Section III involves the detailed description about the proposed method. Section IV presents the performance analysis. This paper concludes in Section V.

II. Related Work

This section deals with the works related to the classification detection of a novel class of data streams. *Masud et al* proposed a data stream classification approach that integrates a class detection mechanism into traditional classifiers. It enables automatic detection of novel classes before the true labels of the novel class instance arrive. The classification model was used to determine the class instances [1]. *Aggarwal et al* proposed a cluster histogram concepts, which provides an efficient way to estimate and summarize the most important data distribution profiles over different stream segments. The profiles can be constructed in a supervised or unsupervised way depending upon the nature of the underlying application. These profiles can also be used for change detection, anomaly detection, segmental nearest neighbor search or supervised stream segment classification. These techniques can also be used to model other kinds of data such as text and categorical data [2].

Bifet et al proposed a data stream context by building an ensemble of Heoffging trees. Those were restricted to a small subset of attributes. Each tree was restricted to model interactions between attributes. A mechanism was proposed for setting the perceptrons learning rate using the ADWIN change detection method for data streams and also ADWIN used to reset ensemble members [3]. *Chen et al* proposed a multiple selective recursive approach (MUSERA). This approach deals with the problem of learning from imbalanced data streams. An ensemble was maintained, which consist of hypotheses built upon the coming training data chunks. MUSERA learnt the target concept of the imbalanced data streams [4]. *Ho et al* proposed a martingale approach to detect the changes in the data streams by testing the exchangeability property of the observed data. The martingale approach is a nonparametric, one-pass algorithm on the classification, cluster and regression data generating models. Also, an adaptive Support Vector Machine (SVM) was proposed, which utilizes the martingale methodology [5].

Masud et al proposed a DXMiner approach, which addresses four challenges to data stream classification. They were infinite length, concept-drift, concept-evolution and feature-evolution. The single-pass incremental learning technique was used. Concept-drift occurs in a data stream when the underlying concept changes over time [6]. *Rai et al* proposed a process of feature selection optimization in multiclass miner for stream data classification. The optimization technique was used for feature selection process. The feature selection process was based on advance genetic algorithm (AGA) [7]. Masud et al studied the concept-evolution and the insights were used to construct the superior novel class detcetion techniques. Also an adaptive threshold was introduced for outlier detection, which was a vital part of novel class detection. A probabilistic approach was proposed for novel class detection using discrete Gini Coefficient [8].

Miao et al introduced a class instance detection technique to deal with the mixed attribute data. The VFDTC was adopted as base classifier to speed up the process and reduced the model size [9]. *Upadhyay et al* provided the study about various data stream clustering techniques and various dimension reduction techniques with their characteristics to enhance the quality of clustering. Also, Fuzzy c-means algorithm was applied to improve the clustering quality [10]. *Chen et al* proposed a selectively recursive approach (SERA) to deal with the problem of learning from the nonstationary imbalanced data streams. SERA can alleviate the difficulty confronted by the conventional stream data mining methods when they have to learn from the nonstationary imbalanced data streams [11].

Khalilian et al presented a different problem definitions related to data stream clustering and the specific difficulties encountered in this field [12]. *Rai et al* presented a various method for reducing the problems occurred in stream data classification. A machine learning technique for feature evaluation process for generation of novel class [13]. *Farid et al* proposed an approach for detecting novel class in data stream mining. The decision tree classifier was used to determine whether an unseen or new instance belongs to a novel class. This approach built a decision tree model from training dataset which continuously updates [14]. *Parker et al* proposed a Hierarchical Stream Miner (HSMiner) takes a hierarchical decomposition approach to the ensemble classifier concept [15].

III. Proposed Methodology

The proposed method is used to classify and detect the novel class for data streams. The following section describes the complete structure of the entire system.

A. Data Classification

The classification process can permit searches based on the file size. In this process, the document is selected and divides the document into equal sized chunks. Initially, the data point in the most recent chunk is classified using the ensemble. When the data points in a chunk becomes labeled, then it is used for training.

The basic steps in classification and novel class detection are performed based on the following steps:

1. Each incoming data stream i.e input document is examined by a outlier detection process to check whether the document contains any outliers
2. If the document doesn't have any outlier, then
    a. Classified as an existing class using mass voting among the classifier
3. Else
    a. Temporarily stored in a buffer
4. When there are enough occurences in the buffer
    a. Novel class detection process is invoked

    If a novel class is found
        1. The occurences of the novel class are tagged

    Else
        2. The buffer is considered as an existing class
        3. Classified normally using the ensemble

B. Detection of novel class

Each occurrence in the most recent unlabeled chunk is examined by the ensemble of models to check if it is outside the decision boundary of the ensemble. If it is within the decision boundary, then classified based on the mass vote. Else, fix it as cleaned-outliers. As, the cleaned-outliers are outside the decision boundary, they are absent from the existing class instances. The cleaned-outliers are possible novel class instances and it will be stored in a temporary buffer *temp_buf*. The *temp_buf* is periodically investigated to see whether there are enough cleaned-outliers that are close together. It is computed based on the nearest neighbor rule.

1) Nearest Neighborhood Rule

Assume that the instance belongs to a class *c*, which are generated from the underlying generative model and the instances in each class. The instances which are close together under the distance metric are supposed to be generated by the same class.

Algorithm

1. Store the output values of the k nearest neighbors to query scenario *q* in vector v = { v^1,… v^k} by repeating the following loop k times
   a. Go to the next step n^i in the data set, where i is the recent iteration within the domain (1,..,S}.
   b. If q is not set or $q < d(q, n^i): q \leftarrow d(q, n^i): , t \leftarrow o^i$
   c. Loop until the end of the dataset comes *(i = P)*
   d. Store *q* into vector *a* and *t* into vector
2. Compute the arithmetic mean output across *b*
   a. $\bar{b} = \frac{1}{k}\sum_{i=1}^{k} b_i$
3. Revisit *b* as the output value for the query set-up *q*

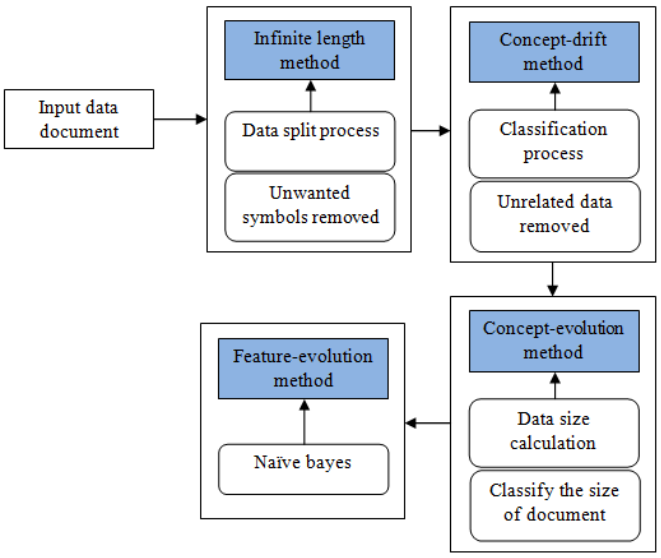

Fig.1. Flow of the proposed methodology

C. Novel Class Detection

The discrete Gini Coefficient G(p) is used to detect the novel classes. For a random sample of m_i as follows

$$G(p) = \frac{1}{k}(k + 1 - 2(\frac{\sum_{i=1}^{k}(k+1-i)m_i}{\sum_{i=1}^{k} m_i})) \quad (1)$$

A compound measure for each cleaned-outlier is measured called Novelty score (Nov-score).

Consider the following three cases and observe the behavior of G(p):

1. If G(p)=0, then the cleaned-outlier actually belongs to the existing class (all the Nov-score(z) are very low).
2. If G(p) = $\frac{k-1}{k}$, then all the cleaned-outlier belongs to the novel class (all the Nov-score(z) are very high).
3. Nov-score(z) is evenly distributed across all the intervals, m_i= i/k . Hence, G(p) becomes $\frac{k-1}{3k}$.

D. Feature Space Conversion

It is clear that the data streams do not have any fixed feature space will have a diverse feature spaces for diverse models in the ensemble. Because, the various sets of features are usually selected from different chunks. The three possible conversions are Lossy-fixed conversion, Lossy-local conversion and Lossless homogenizing conversion.

1) Lossy-fixed conversion

The same feature set is used for the entire process i.e. the same feature set is elected for the first data chunk or the first *n* data chunks. This makes the feature set is predetermined and maps all the training and test occurrences to this feature set.

2) Lossy-local conversion

Using the feature extraction and selection technique, each training chunk and the model built from the chunk have its own feature set. When classifying the test instance, it is expected to the feature set. Both the above conversion approaches lose some of the major features due to the conversion [6]. As a result, the Lossless homogeneous conversion is proposed.

3) Lossless-homogenous conversion

This conversion is otherwise called as lossless conversion. When a test instance *x* is to be classified, then both the model and the instance expand the resultant feature set to the union of their feature sets. Hence, no useful features are misplaced due to lossless conversion. This conversion is more suitable to detect the novel classes.

E. Multiple Novel Class Detection

The major idea behind the multiple novel class detection is to build a graph and recognize the connected components in the graph. The number of connected components regulates the number of novel classes. Suppose, two novel classes are obtained, then the separation between the various novel class instances must be higher than the cohesion among the same-class instances.

Algorithm

1. Initialize the graph G : (V, E)
2. Perform k-means clustering
3. If *a* ϵ List S
   a. *a.nn* ←nearest neighbor
   b. *a.sil*←Compute silhouette coefficient
   c. V ← V ∪ {*a*}
   d. V ← V ∪ {*a.nn*}
   e. If *a.sil* < *th*
      i. *E* ←*E* ∪ {*a, a.nn*}
4. Count ← Connected-component
5. For each pair of components (g_1,g_2) ϵ G
6. M_1 ← mean(g_1) , M_2 ← mean(g_2)
7. If $\frac{M1+M2}{2*centroid(g1,g2)}$ >1
8. g_1 ←merge (g_1, g_2)
9. For y ϵ list S
10. a ←Pesudopoint of (y)
11. Z ← Z ∪{y, a.componentnumber}

Here Z is the predicted class label of the novel instance; List S is the list of novel class instances. The silhouette coefficient using the following equation:

$$a.sil = \frac{dist(a,a.nn) - a.M}{\max(dist(a,a.nn),a.M)} \quad (2)$$

Where $dist(a, a.nn)$ is the distance between the centroids of a and $a.nn$, $a.M$ is the mean distance from the centroid of a. The feature selection is based on the naive bayes classifier [16].

IV. Performance Analysis

The proposed system is implemented based on the following datasets: Twitter, KDD, Forest and ASRS. The twitter dataset contains the twitter messages, Forest dataset contains the geospatial descriptions for different kinds of forests, Knowledge Discovered in Databases, and ASRS contains the NASA Aviation Safety Reporting Systems text documents. The performance is tested based on the words count, time, speed and error rate for the proposed system to detect and classify the novel classes with the existing systems [17].

A. Word count

The word count is the number of words used in a document or passage of text. The proposed system for classification and novel class detection (C&CD) results the reduced number of word counts than the existing W-OP and W-OS.

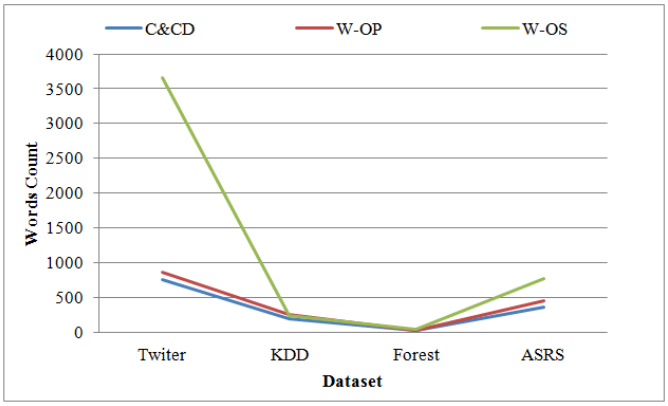

Fig.2. Word Count for C&CD (proposed) and W-OP, W-OS (existing)

B. Speed

Fig.3 shows the speed comparison between the proposed system C&CD and the existing systems like W-OS and W-OP. It shows that our proposed system performs better than the existing methods.

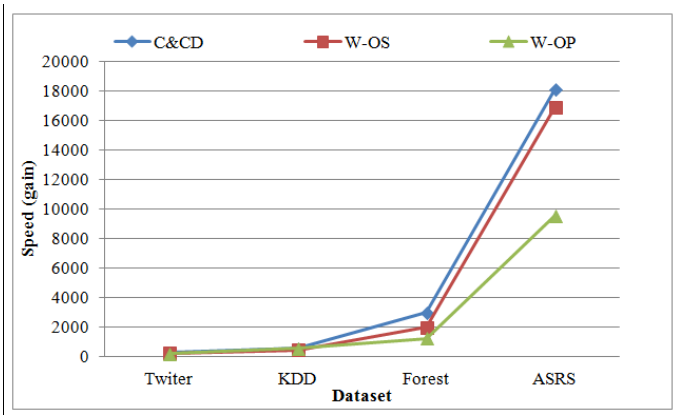

Fig.3. Speed analysis for C&CD (proposed) and W-OP, W-OS (existing)

C. Execution Time

Execution time is the amount of time needed to execute the proposed process. Fig.4 shows the comparison between the proposed and the existing systems. The result shows that the proposed the classification and novel class detection takes less time to compute the results than the existing systems.

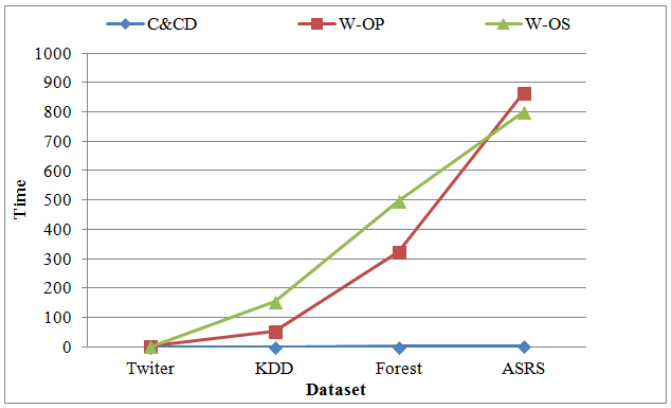

Fig.4.Time analysis for C&CD (proposed) and W-OP, W-OS (existing)

D. Error rates

The overall error is estimated based on the following equation:

$$Error.rate = \frac{(100)(A+B+C)}{N} \quad (3)$$

Where A is the total novel class instances that are misclassified as existing class, B is the total existing class instances misclassified as the novel class and C is the total existing class instances misclassified as another existing class and N is the total number of instances.

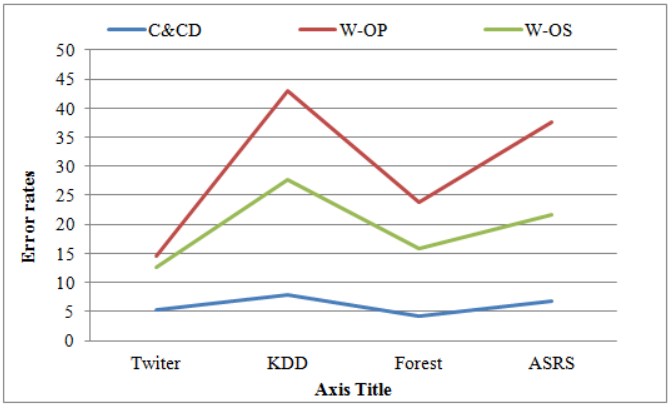

Fig.5.Error rates between C&CD, W-OP and W-OS

Fig.5 shows the comparison error rate between the proposed method and the existing method. It shows that the proposed method results lesser error rate when compared with the existing methods.

V. Conclusion

An efficient approach is implemented for classifying and detecting the novel classes in the data streams. The existing novel class detection approaches could not address the feature-evolution problems. It provides improved solutions for the issues occurred in the data mining domain. Also, the Nearest Neighbor algorithms and the Naive Bayes classifier results better performance. The implementation results show that the proposed system utilizes less time, error rate and results reduced word counts with higher speed.

In future, the work is extended with the recent classifiers to

detect the novel classes in the data streams.

REFERENCES

[1] 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," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, pp. 859-874, 2011.

[2] C. C. Aggarwal, "A segment-based framework for modeling and mining data streams," *Knowledge and information systems*, vol. 30, pp. 1-29, 2012.

[3] A. Bifet, E. Frank, G. Holmes, and B. Pfahringer, "Accurate Ensembles for Data Streams: Combining Restricted Hoeffding Trees using Stacking," *Journal of Machine Learning Research-Proceedings Track*, vol. 13, pp. 225-240, 2010.

[4] S. Chen, H. He, K. Li, and S. Desai, "Musera: multiple selectively recursive approach towards imbalanced stream data mining," in *Neural Networks (IJCNN), The 2010 International Joint Conference on*, 2010, pp. 1-8.

[5] S.-S. Ho and H. Wechsler, "A martingale framework for detecting changes in data streams by testing exchangeability," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, pp. 2113-2127, 2010.

[6] M. M. Masud, Q. Chen, J. Gao, L. Khan, J. Han, and B. Thuraisingham, "Classification and novel class detection of data streams in a dynamic feature space," in *Machine Learning and Knowledge Discovery in Databases*, ed: Springer, 2010, pp. 337-352.

[7] M. Rai and V. Richhariya, "A Feature Selection process Optimization in multi-class Miner for Stream Data Classification," *INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY*, vol. 3, pp. 359-364, 2012.

[8] M. M. Masud, Q. Chen, L. Khan, C. Aggarwal, J. Gao, J. Han, *et al.*, "Addressing concept-evolution in concept-drifting data streams," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, 2010, pp. 929-934.

[9] Y. Miao, L. Qiu, H. Chen, J. Zhang, and Y. Wen, "Novel Class Detection within Classification for Data Streams," in *Advances in Neural Networks – ISNN 2013*. vol. 7952, C. Guo, Z.-G. Hou, and Z. Zeng, Eds., ed: Springer Berlin Heidelberg, 2013, pp. 413-420.

[10] D. Upadhyay, S. Jain, and A. Jain, "Comparative Analysis of Various Data Stream Mining Procedures and Various Dimension Reduction Techniques," *Control Theory and Informatics*, vol. 3, pp. 60-64, 2013.

[11] S. Chen and H. He, "Sera: selectively recursive approach towards nonstationary imbalanced stream data mining," in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, 2009, pp. 522-529.

[12] M. Khalilian and N. Mustapha, "Data stream clustering: Challenges and issues," *arXiv preprint arXiv:1006.5261*, 2010.

[13] M. Rai and R. Pandit, "A Review of Classification and Novel Class Detection Technique of Data Streams," *International Journal of Computers & Technology*, vol. 3, pp. 314-316, 2012.

[14] D. M. Farid and C. M. Rahman, "Novel class detection in concept-drifting data stream mining employing decision tree," in *Electrical & Computer Engineering (ICECE), 2012 7th International Conference on*, 2012, pp. 630-633.

[15] B. Parker, A. M. Mustafa, and L. Khan, "Novel Class Detection and Feature via a Tiered Ensemble Approach for Stream Mining," in *Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th International Conference on*, 2012, pp. 1171-1178.

[16] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," *Expert Systems with Applications*, vol. 36, pp. 5432-5435, 4// 2009.

[17] M. M. Masud, J. Gao, L. Khan, J. Han, and B. Thuraisingham, "Integrating novel class detection with classification for concept-drifting data streams," in *Machine Learning and Knowledge Discovery in Databases*, ed: Springer, 2009, pp. 79-94.