IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-04, Issue-09 E-ISSN: 01793-1

Optimal Feature Selection in Stream Data Classification Using Improved Ensemble Classifier for High Dimension Data

Gaurav Pandey[1]* and Nitin Mishra[2]

[1,2] SIRT Excellence, Bhopal M P, India

Available online at: www.ijcseonline.org

Received: 16/Aug/2016 Revised: 27/Aug/2016 Accepted: 16/Sept/2016 Published: 30/Sep/2016

***Abstract*—** Dynamic feature evaluation and concept evaluation is major challenging task in the field of stream data classification. The continuity of data induced a new feature during classification process, but the classification process is predefined task for assigning data into class. Stream data comes into multiple feature sub-set format into infinite length. The infinite length not decided the how many class are assigned. Genetic algorithm is well known population based method. The performance of genetic algorithm is better than other optimization technique such as POS and ANT colony optimization. The dynamic nature of genetic algorithm maintains the dynamic feature evaluation. The optimization process goes through multiple stages in terms of selection of feature and optimization of feature. The optimized feature reduces the unclassified region of class during classification. The proposed method for stream data classification is MMCM-GA is implemented in MATLAB 7.8.0. And test the validation process used some reputed data set from UCI machine learning prosperity. These data are corpus, forest and finally used glass dataset. Our empirical evaluation of result shows better feature evaluation and minimization of error rate in comprehension of MCM stream data classification

Keywords— **Stream Data Classification, POS, Ensemble, Optimal Feature, Genetic Algorithm (GA)**

.

I. Introduction

Data stream classification is more challenging than classifying static data because of several unique properties of data streams. First, data streams are assumed to have infinite length, which makes it impractical to store and use all the historical data for training. Therefore, traditional multi-pass learning algorithms are not directly applicable to data streams. Second, data streams observe concept-drift, which occurs when the underlying concept of the data changes over time. In order to address concept drift, a classification model must continuously adapt itself to the most recent concept. Third, data streams also observe concept evolution, which occurs when a novel class appears in the stream. With the advent of advanced data streaming technologies [1], we are able to continuously collect large amounts of data in various application domains, e.g., daily fluctuations of stock market, traces of dynamic processes, credit card transactions, web click stream, network traffic monitoring, position updates of moving objects in location-based services and text streams from news etc [2]. Due to its potential in industry applications, data stream mining has been studied intensively in the past few years. The general approach is to first learn one or multiple classification models from the past records of the evolving data, and then use a selected model that best matches the current data to predict the new data records. All the existing data stream classification techniques assume that at each time stamp there are both large amounts of positive and negative training data available for learning. The goal of data stream classification is to learn a model from past labeled data, and classify future instances using the model. There are many challenges in data stream classification. First, data streams have infinite length, and so, it is impossible to store all the historical data for training [3]. Therefore, traditional learning algorithms that require multiple passes over the whole training data are not directly applicable to data streams. Second, data streams observe concept-drift, which occurs when the underlying concept of the data changes over time. A classification model must adapt itself to the most recent concept in order to cope with concept-drift. Third, novel classes may appear in the stream, which we call concept-evolution. In order to cope with concept-evolution, a classification model must be able to automatically detect novel classes [4]. Data stream classifiers may either be single model incremental approaches, or ensemble techniques, in which the classification output is a function of the predictions of different classifiers. Ensemble techniques have been more popular than their single model counterparts because of their simpler implementation and higher efficiency. Most of these ensemble techniques use a chunk-based approach for learning which they divide the data stream into chunks, and train a model from one chunk [5]. We refer to these approaches as “chunk-based” approaches. An ensemble of chunk-based models is used to classify unlabeled data. These approaches usually keep a fixed-sized ensemble, which is continuously updated by replacing an older model with a newly trained model These approaches usually keep a fixed-sized ensemble, which is continuously updated by replacing an older model with a newly trained model. Some chunk-based techniques, such as, cannot detect novel classes, whereas others can do so. Chunk-based techniques that cannot detect novel classes cannot detect recurrent classes as well. The rest of paper is organized as follows. In Section II Describe the previous work. The Section III state the problem IV discusses proposed methodology. In section V discuss performance evaluation and result analysis followed by a conclusion in Section VI.

II RELATED WORK

In this section discuss the related work of Optimal Feature Selection in Stream Data Classification Using Improved Ensemble Classifier for High Dimension Data. Some technique discuss here.

[1] In this paper author address several open challenges of big data stream classification, including high volume, high velocity, high dimensionality, and high sparsity. Many existing studies in data mining literature solve data stream classification tasks in a batch learning setting, which suffers from poor efficiency and scalability when dealing with big data. To overcome the limitations, this paper investigates an online learning framework for big data stream classification tasks. Unlike some existing online data stream classification techniques that are often based on first order online learning, they propose a framework of Sparse Online Classification (SOC) for data stream classification, which includes some state-of-the-art first-order sparse online learning algorithms as special cases and allows us to derive a new effective second-order online learning algorithm for data stream classification. They conduct an extensive set of experiments, in which encouraging results validate the efficacy of the proposed algorithms in comparison to a family of state of-the-art techniques on a variety of data stream classification tasks.

[2] In this paper author address this issue and propose a data stream classification technique that integrates a novel class detection mechanism into traditional classifiers, enabling automatic detection of novel classes before the true labels of the novel class instances arrive. Novel class detection problem becomes more challenging in the presence of concept-drift, when the underlying data distributions evolve in streams. In order to determine whether an instance belongs to a novel class, the classification model sometimes needs to wait for more test instances to discover similarities among those instances.

[3] In this research work author described about the multi label data stream classification and the details are, these methods, each instance can only be tagged with one label. However, in many realistic applications, each instance should be tagged with more than one label. To address the challenge of classifying multi-label stream in evolving environment, author propose a novel Multi-Label Dynamic Ensemble (MLDE) approach. The proposed MLDE integrates a number of Multi-Label Cluster-based Classifiers (MLCCs). MLDE includes an adaptive ensemble method and an ensemble voting method with two important weights, subset accuracy weight and similarity weight. Experimental results reveal that MLDE achieves better performance than state-of-the-art multi-label stream classification algorithms**.**

[4] In this paper author focus on a novel evolving fuzzy rule-based Classifier and the details are, the proposed classifier addresses the three fundamental issues of data stream learning, viz., computational efficiency in terms of processing time and memory requirements, adaptive to changes, and robustness to noise . Though, there are several online classifiers available, most of them do not take into account all the three issues simultaneously. The newly proposed classifier is inherently adaptive and can attend to any minute changes as it learns the rules in online manner by considering each incoming example. However, it should be emphasized that it can easily distinguish noise from new concepts and automatically handles noise. The performance of the classifier is evaluated using real-life data with evolving characteristic and compared with state-of-the-art adaptive classifiers.

[5] Some online algorithms for linear classification model the uncertainty in their weights over the course of learning. Modeling the full covariance structure of the weights can provide a significant advantage for classification. However, for high-dimensional, large scale data, even though there may be many second-order feature interactions, it is computationally infeasible to maintain this covariance structure. To extend second-order methods to high-dimensional data, authors develop low-rank approximations of the covariance structure. Authors evaluate their approach on both synthetic and real-world data sets using the confidence-weighted online learning framework. They show improvements over diagonal covariance matrices for both low and high-dimensional data.

[6] Here author work with the problem of data stream classification is challenging because of many practical aspects associated with efficient processing and temporal behavior of the stream. Two such well studied aspects are infinite length and concept-drift. Since a data stream may be considered a continuous process, which is theoretically infinite in length, it is impractical to store and use all the historical data for training. Data streams also frequently experience concept-drift as a result of changes in the underlying concepts. However, another important characteristic of data streams, namely, concept-evolution is rarely addressed in the literature. Concept-evolution occurs as a result of new classes evolving in the stream. In this paper author addresses concept-evolution in addition to the existing challenges of infinite-length and concept-drift. In this paper,

the concept-evolution phenomenon is studied, and the insights are used to construct superior novel class detection techniques. First, they propose an adaptive threshold for outlier detection, which is a vital part of novel class detection. Second, they propose a probabilistic approach for novel class detection using discrete Gini Coefficient, and prove its effectiveness both theoretically and empirically. Finally, they address the issue of simultaneous multiple novel class occurrence, and provide an elegant solution to detect more than one novel classes at the same time.

[7] In this paper author defined the data stream challenges and the details are, concept-evolution is one of the major challenges in data stream classification, which occurs when a new class evolves in the stream. This problem remains unaddressed by most state-of-the-art techniques. A recurring class is a special case of concept-evolution. This special case takes place when a class appears in the stream, then disappears for a long time, and again appears. Existing data stream classification techniques that address the concept-evolution problem, wrongly detect the recurring classes as novel class. This creates two main problems. First, much resource is wasted in detecting a recurring class as novel class, because novel class detection is much more computationally- and memory intensive, as compared to simply recognizing an existing class. Second, when a novel class is identified, human experts are involved in collecting and labeling the instances of that class for future modeling. If a recurrent class is reported as novel class, it will be only a waste of human effort to find out whether it is really a novel class. In this paper, author address the recurring issue, and propose a more realistic novel class detection technique, which remembers a class and identifies it as “not novel” when it re-appears after a long disappearance. Our approach has shown significant reduction in classification error over state-of-the-art stream classification techniques on several benchmark data streams.

III PROBLEM STATEMENT

Stream data classification is new area of research due to

dynamic nature of stream data. The utility of stream data classification is very high due to diversity of internet data and environmental data. By virtue of stream data have some common problem based attribute such as infinite length, concept evaluation, feature evaluation and data drift? These entire attribute create some problem and deceases the classification rate of stream data classification. Some common problem mention below.

1. Unbalanced ratio of train and test data [1]
2. Selection of optimal features for ensemble classifier [6], [7]
3. Diversity of feature selection process [9]
4. Boundary value of features [10]
5. Outlier data treat as noise [11].

IV. PROPOSED METHODOLOGY

Feature optimization and data classification in stream data mining is challenging task for researchers for controlling new feature evaluation of incoming data for classification. The new evolution of feature compromised the process of proper class mapping. The mapping of class is not perform the performance of classifier is degraded. For the improving the classification rate of stream data in current decade proposed a multi-class miner method for evaluation of new feature and improvement of classification. For maintaining a new evolve feature attribute of data for classification one problem automatically recall such problem is called data drift. Data drift induced a discontinuity of stream data and suffered a problem of classification. For a lack of data drift and feature evolving concept technique multi-class miner degrade the performance of classifier in mode of single class and multi-class. For fulfillment of researcher gap in stream data classification proposed an optimized pattern selection method for stream data classification. The optimized pattern selection method basically based on genetic algorithm. Genetic algorithm well knows algorithm for purpose of optimal solution.

The multi-class miner algorithm basically consists of ensemble technique of clustering and classification [10]. The main idea in detecting multiple novel classes is to construct a graph, and identify the connected components in the graph. The number of connected components determines the number of novel classes. The basic assumption in determining the multiple novel classes follows property:
A data point should be closer to the data points of its own class (cohesion) and farther apart from the data points of other classes (separation). If there is a novel class in the stream, instances. For example, if there are two novel classes, then the separation among the different novel class instances should be higher than the cohesion among the same-class instances.

The main idea in detecting multiple novel classes is to construct a graph, and identify the connected components in the graph. The number of connected components determines the number of novel classes. The basic assumption in determining the multiple Novel classes follow property. For example, if there are two novel classes, then the separation among the different novel class instances should be higher than the cohesion among the same-class instances.

Algorithm Detect-Multinovel (N list)

Input: N_list: List of novel class instances

```
Output: N_type: predicted class label of the novel instances

1:  G = (V, E) ←empty //initialize graph
2:  NP_list ← K-means (N_list, $K_v$)
3:  Input NP_list X , the clustering number cn , population scale XN , crossover probability cP , mutation probability mP , Pattern probability vP , stop conditions cS ;
4:  Code the chromosome in real number and initialize population A(i),i = 0 at random;
5:  Calculate the fitness of each individual in the current instant;
6:  MCM clustering creates stored pattern for classification, which means find dissimilar feature cluster. Hence the fitness function of algorithm is determined by f(x).
7:  F(x) = {(α +2β)-αi, αi<β+2α
            0, αi≥αi+2β
    I=1,2,……………………………..,N
8:  Judge the termination conditions. If the termination conditions are satisfied, then turn to step 9, otherwise, turn to step 10;
9:  Decode to find and calculate the optimal clustering and pattern matrixes. And set the optimal clustering for classification.
10: Do the parallel crossover and mutation operation on population A(i), then get population B(i), C(i) respectively;
11: Carry out the genetic selection on the instant composed of population A(i), B(i), C(i) and population D(i) is got;
12: Take the MCM optimization on population D(i) and generate the next generation A(i +1) . Then turn to step
13: for h ∈ A(i+1) do
14: h.nn ← Nearest-neighbor (A(i+1)- {h})
15: h.sc ← Compute-SC (h, h.nn)
16: V←V ⋃ {h}
17: V←V ⋃ {h.nn}
18: if h.sc < $th_{sc}$ then
19: E←E ⋃ {(h,h.nn)}
20: End if
21: end for
22: count ← Con-Components (G) for each pair of components (g1,g2) ∈ G do
23: $\mu_1$←mean-dist (g1), $\mu_2$←mean-dist (g2)
24: if $\frac{\mu_1+\mu_2}{2*centroiid\_dist(g1,g2)}$ > 1 then g1←Merge (g1, g2)

end for

25: N_type ← empty
26: for $\mathcal{X}$ ∈ Nlist do
27: h ← Pattern-recallOf ($\mathcal{X}$)
28: N_type ← N_type ⋃ {( $\mathcal{X}$, h.componentno)}
29: end for
```

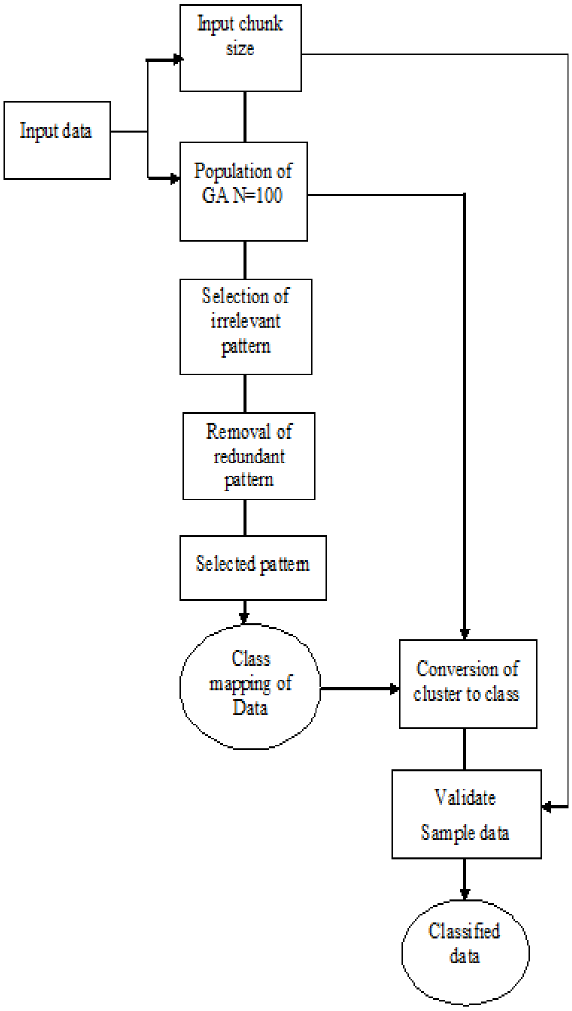


Figure 1: Flow diagram of proposed system.

V EXPERIMENTAL RESULT ANALYSIS AND PERFORMANCE EVALUATION

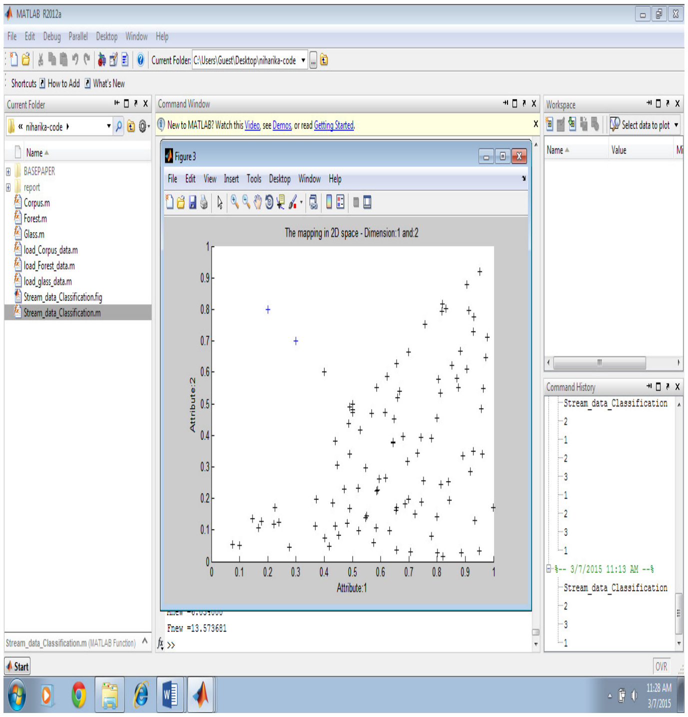
It is simulating on mat lab 7.14.0 and for this work we use Intel 1.4 GHz Machine. MATLAB is a high level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation MATLAB is a software program that allows you to do data manipulation and visualization, calculations, math and programming [12]. For the performance evaluation of Association classification technique and our model used MATLAB software package. MATLAB is a software package for high- performance numerical computation and visualization. It provides an interactive environment with hundreds of built-in function for technical computation, graphics and animation. Best of all, it also provides easy extensibility with its own high- level programming language. The MATLAB stands for matrix laboratory [12]. There are also several optional "toolboxes" available from the developers of MATLAB.

CHUNK SIZE	METHOD NAME	ELAPSED TIME	ERROR	M-NEW	F-NEW
10	ENSEMBLE	158.467	0.445	9.134	82.805
	PROPOSED	21.5949	0.645	6.134	15.203
20	ENSEMBLE	144.485	0.545	7.034	74.576
	PROPOSED	22.5300	0.745	6.034	15.008
30	ENSEMBLE	182.574	0.645	10.93	93.522
	PROPOSED	23.509	0.845	5.930	15.094

Table 1: Shows that the comparative performance evaluation with the ensemble and proposed method using CORPUS dataset for the same and different chunk size.

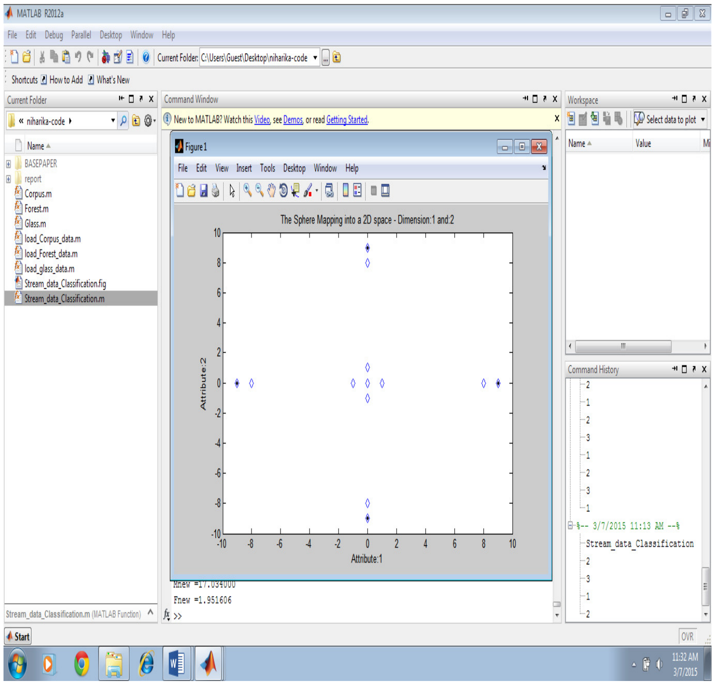
CHUNK SIZE	METHOD NAME	ELAPSED TIME	ERROR	M-NEW	F-NEW
10	ENSEMBLE	0.582	0.4450	16.13	0.8346
	PROPOSED	0.5459	0.6450	17.13	1.7875
20	ENSEMBLE	0.5522	0.5450	16.03	0.8190
	PROPOSED	0.5548	0.7450	17.034	1.8346
30	ENSEMBLE	0.5544	0.6450	15.934	0.8190
	PROPOSED	0.5573	0.8450	16.934	1.8268

Table 2: Shows that the comparative performance evaluation with the ensemble and proposed method using GLASS dataset for the same and different chunk size.


MATLAB R2012a
File Edit Debug Parallel Desktop Window Help
Current Folder: C:\Users\Guest\Desktop\niharika-code
Shortcuts How to Add What's New
Current Folder
Command Window
Workspace
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Figure 3
File Edit View Insert Tools Desktop Window Help
The mapping in 2D space - Dimension:1 and:2
Attribute 2
Attribute:1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
BASEPAPER
report
Corpus.m
Forest.m
Glass.m
load_Corpus_data.m
load_Forest_data.m
load_glass_data.m
Stream_data_Classification.fig
Stream_data_Classification.m
Command History
Stream_data_Classification
%-- 3/7/2015 11:13 AM --%
Start
OVR
11:28 AM
3/7/2015

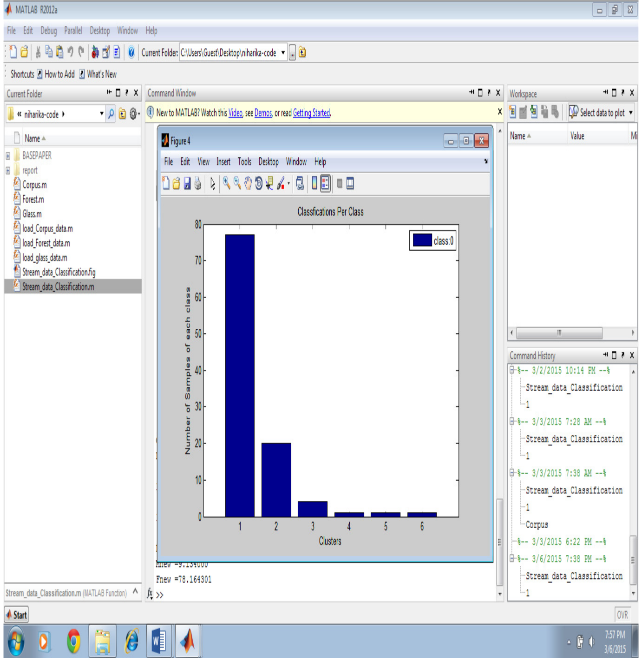
```
Mnew =6.034000
Fnew =13.573681
>>
```

Figure 2: Shows that the method using proposed on with the corpus data set for the chunk size value is 20 for the mapping in 2D space.


MATLAB R2012a
File Edit Debug Parallel Desktop Window Help
Current Folder: C:\Users\Guest\Desktop\niharika-code
Shortcuts How to Add What's New
Current Folder
Command Window
Workspace
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Figure 1
File Edit View Insert Tools Desktop Window Help
The Sphere Mapping into a 2D space - Dimension:1 and:2
Attribute 2
Attribute:1
-10
-8
-6
-4
-2
0
2
4
6
8
10
BASEPAPER
report
Corpus.m
Forest.m
Glass.m
load_Corpus_data.m
load_Forest_data.m
load_glass_data.m
Stream_data_Classification.fig
Stream_data_Classification.m
Command History
%-- 3/7/2015 11:13 AM --%
Stream_data_Classification
Start
OVR
11:32 AM
3/7/2015

```
Mnew =17.034000
Fnew =1.951606
>>
```

Figure 3: Shows that the method using proposed on with the corpus data for the chunk size value is 20 for the 2D space mapping.


MATLAB R2012a
File Edit Debug Parallel Desktop Window Help
Current Folder: C:\Users\Guest\Desktop\niharika-code
Shortcuts How to Add What's New
Current Folder
Command Window
Workspace
New to MATLAB? Watch this Video, see Demos, or read Getting Started.
Select data to plot
Name Value
niharika-code
Name
BASEPAPER
report
Corpus.m
Forest.m
Glass.m
load_Corpus_data.m
load_Forest_data.m
load_glass_data.m
Stream_data_Classification.fig
Stream_data_Classification.m
Stream_data_Classification.m (MATLAB Function)
Figure 4
File Edit View Insert Tools Desktop Window Help
Classfications Per Class
class:0
Number of Samples of each class
0
10
20
30
40
50
60
70
80
1
2
3
4
5
6
Clusters
Command History
Start
OVR
7:57 PM
3/6/2015

```
Fnew =78.164301
fx >>
```

```
%-- 3/2/2015 10:14 PM --%
Stream_data_Classification
1
%-- 3/3/2015 7:28 AM --%
Stream_data_Classification
1
%-- 3/3/2015 7:38 AM --%
Stream_data_Classification
1
Corpus
%-- 3/3/2015 6:22 PM --%
%-- 3/6/2015 7:38 PM --%
Stream_data_Classification
1
```

Figure 4: Shows that the method using ensemble on with the corpus data for the chunk size value is 10 for the data classification as per class.

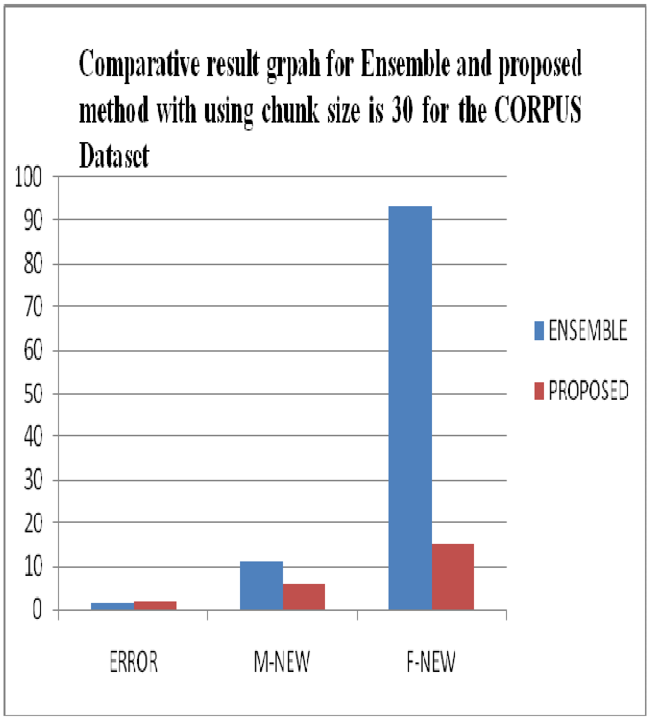


Figure 5: Shows that the comparative result graph for Ensemble and proposed method and find the value of Error, M-New and F-New with using CORPUS Dataset and here the chunk size is 30.

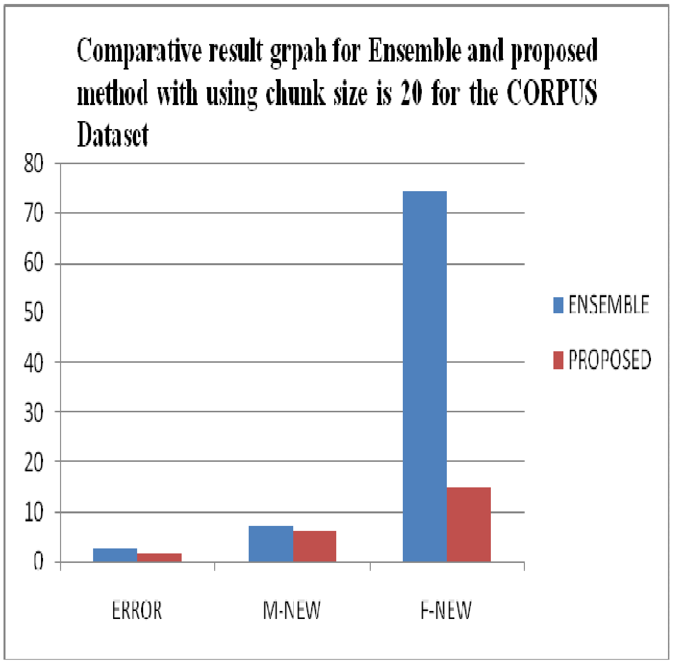


Figure 6: Shows that the comparative result graph for Ensemble and proposed method and find the value of Error, M-New and F-New with using CORPUS Dataset and here the chunk size is 20.

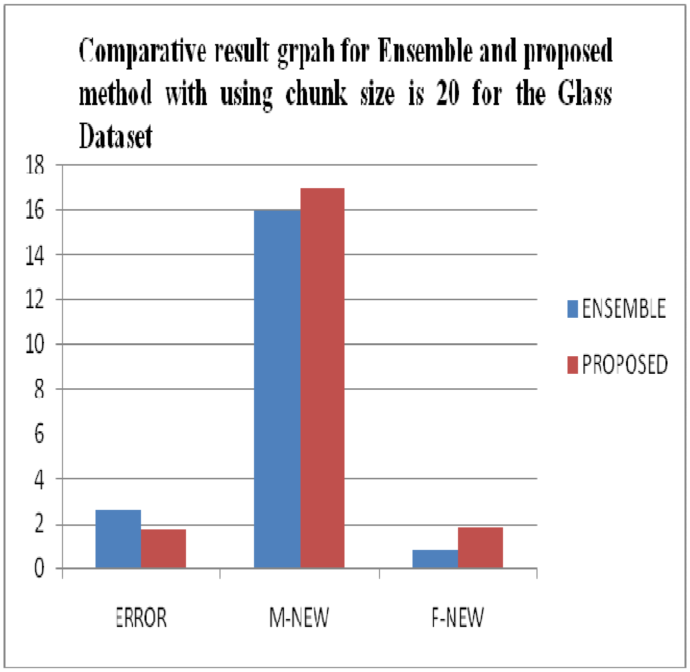


Figure 7: Shows that the comparative result graph for Ensemble and proposed method and find the value of

Error, M-New and F-New with using GLASS Dataset and here the chunk size is 20.

VI CONCLUSION AND FUTURE WORK

Traditional stream classification techniques also make impractical assumptions about the availability of labeled data. Most techniques assume that the true label of a data point can be accessed as soon as it has been classified by the classification model. Thus, according to their postulation, the existing model can be updated without delay using the labeled instance. In reality, one would not be so lucky in obtaining the label of a data instance immediately, since manual labeling of data is time consuming and costly. We claim two major contributions in novel class detection for data streams. First, we propose a dynamic selection of boundary for outlier detection by allowing a slack space outer the decision boundary. Multi-class miner is very efficient data mining tool for stream data classification. Stream data classification is challenging task in the field of classification. Evaluation of new feature creates a problem in feature selection during the classification process of multi-class miner. In this paper we reduces these problems using genetic algorithm, genetic algorithm used to control new feature evolution problem. Genetic algorithm creates a feature prototype for cluster used in classification. The controlled feature evaluation process proposed a modified multi-class miner is called MMCM-GA. The empirical evaluation of modified algorithm is better in compression of MCM algorithm. The error rate of modified algorithm decreases in compression of MCM algorithm. Also improved the rate of F new and M new for evolution of result, after these improvement still some problem is still remain such as infinite length and data drift. Infinite length and data drift problem are not considered here. The proposed method modified multi-class miner solved the problem of feature evaluation and concept evaluation. The controlled feature evaluation process increases the value of F new and M new and reduces the error rate. The genetic prototype cluster faced a problem of right number of cluster, in future used self optimal clustering technique along with genetic algorithm.

REFERENCES

[1] Dayong Wang, Pengcheng Wu, Peilin Zhao, Yue Wu, Chunyan Miao, Steven C.H. Hoi "High-dimensional Data Stream Classification via Sparse Online Learning" IEEE International Conference on Data Mining, 2014. Pp 1007-1012.

[2] Mohammad M. Masud,Jing Gao, Latifur Khan, , Jiawei Han and Bhavani Thurai singham "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. Pp 859-874.

[3] Ge Song, Yunming Ye "A New Ensemble Method for Multi-label Data Stream Classification in Non-stationary Environment" 2014 International Joint Conference on Neural Networks July 6-11, 2014, Beijing, China, Pp 1776-1783.

[4] Rashmi Dutta Baruah, PlamenAngelov, Diganta Baruah "Dynamically Evolving Fuzzy Classifier for Real-time Classification of Data Streams" IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) , July 6-11, 2014, Beijing, China , Pp 383-389.

[5] Justin Ma, Alex Kulesza, Mark Dredze, Koby Crammer, Lawrence K. Saul, Fernando Pereira "Exploiting Feature Covariance in High-Dimensional Online Learning" 2009, Pp 493-500.

[6] Mohammad M. Masud , Qing Chen , Latifur Khan, Charu Aggarwal ,Jing Gao, Jiawei Han and Bhavani Thurai singham "Addressing Concept-Evolution in Concept-Drifting Data Streams" 2009, Pp 124-130

[7] Mohammad M. Masud, Tahseen M. Al-Khateeb, Latifur Khan,Charu Aggarwal, Jing Gao Jiawei Hanand Bhavani Thuraisingham"Detecting Recurring and Novel Classes in Concept-Drifting Data Streams" 2012. Pp 897-902.

[8] Mohammad M. Masud, Qing Chen, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham "Classification and Novel Class Detection of Data Streams in a Dynamic Feature Space" J.L. Balcazar et al. (Eds.): ECML PKDD 2010, Pp. 337–352.

[9] Zhihui Lai, Zhong Jin, Jian Yang, W.K Wong "Sparse Local Discriminant Projections for Face Feature Extraction" IEEE, 2010, Pp1051-1060.

[10] Clay Woolam, Mohammad M. Masud, and Latifur Khan "Lacking Labels in the Stream: Classifying Evolving Stream Data with Few Labels" 2009, LNAI 5722. Pp. 552–562.

[11] Li Su, Hong-yanLiu, Zhen-Hui Song, "A New Classification Algorithm for Data Stream" , I.J.Modern Education and Computer Science, 2011. Pp 32-39

[12] Manjeet Kaur, Manoj agnihotri "A Hybrid technique using Genetic algorithm and ANT colony optimization for improving in cloud datacenter". IJCSE, Volume 04, Issue-08.