International Journal of Master of Engineering Research and Technology, 2394-6172-(O)
Volume-5, Issue-4, April 2018

Stream Data Classification Based on Feature Optimization and SOM Neural Network Model

Mamata Mishra
Department CSE
SVCST, Bhopal (MP)

Mr. Amit Thakur
Department CSE
SVCST, Bhopal (MP)

Abstract

In current decade the stream data classification is major issue. The property of stream data creates the problem of data categorization such as infinite length, data drift and feature evaluation process. The classification process faced a problem of new feature evaluation. The concept of new feature evaluation cannot understand by classification algorithm. the classification algorithms generate error during find the new features set. For the selection of new features set used feature optimization technique. the feature optimization technique minimized the process of features for the mapping of class. in this paper proposed SOM based stream data classification algorithm. for the optimization of features of stream data used glow worm optimization algorithm. the glow worm optimization algorithm works on the principle of lubrification. The proposed algorithm implemented in MATLAB software. For the validation of algorithm used UCI data set glass, forest and croups.

Keywords: - stream data, GSO, SOM, classification.

INTRODUCTION

Data stream mining, involves extraction of useful knowledge structures from rapid data records and continuous data. As huge amount of information from Sensors, Internet, Twitter, Facebook, Online Banking and ATM Transactions, data Stream Mining plays a vital role and gains more practical significance. The dynamic and evolving nature of data streams require techniques which are effective and efficient and that are highly different from classification techniques using static data. The rate at which the data arrives has a great impact on accuracy of classification[2-4]. While analyzing the incoming stream of elements, there are many difficulties to be faced, therefore traditional data mining techniques cannot be applied. Data streams need to be classified effectively and the classification algorithm should be enhanced based on the incoming data stream[1].

The above section discusses introduction of stream data classification and feature optimization. In section II we describe multi-class miner and particle of swarm optimization. In section III discuss proposed method. In section IV experimental process and finally conclude in section V.

II. MULTI-CLASS MINER AND GSO

In this section describe two algorithms one is multi-class miner and other one is glow worm optimization. the first phase of algorithm discusses multi-class miner algorithm and second phase discuss glow worm optimization algorithm[14].

Phase-I

The multi-class miner algorithm is a combination of clustering and classification technique such technique also called ensemble process[3]. 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[12, 14].

Input: N_list: List of novel class instances

Paper ID:- 2018/04/IJMERT/5/322

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 ) //clustering
3:      for h ∈ N P_list do
4:        h.nn ← Nearest-neighbor (NP_list - {h})
5:        h.sc ← Compute-SC (h, h.nn) //silhouette coefficient
6:        V←V ∪ {h}                   //add these nodes
7:        V←V ∪ {h.nn}
8:        if h.sc < th_sc then        //relatively closer to the nearest neighbor
9:          E←E ∪ {(h,h.nn)}          //add this directed edge
10:       endif
11:     end for
12:   count ← Con-Components (G) //find connected components
            // Merging phase
13:     for each pair of components (g1,g2) ∈ G do
14:       μ1 ←mean-dist (g1), μ2 ←mean-dist (g2)
15:       if (μ1+μ2)/(2*centroid_dist(g1,g2)) > 1 then g1←Merge (g1, g2)
16:     end for
        // Now assign the class labels
17:   N_type ← empty
18:     for x ∈ Nlist do
19:       h ← PseudopointOf (x)       //find the corresponding pseudopoint
20:       N_type ← N_type ∪ {( x , h.componentno)}
21:     end for
```

Phase-II

The stream data feature passes through glowworm algorithm. the feature map into glowworm search space. Each glowworm i encode the object function value J(xi(t)) at its current location xi(t) into α luciferin value li and broadcasts the same within it neighborhood. The set of neighbor (Ni(t)) of glowworm i consist of those glowworm that have relatively higher luciferin value that are located within a dynamic decision domain and updating by formula 4.1 at each iteration [5-9].

Local decision range update is given by equation 1

$$r_d^i(t+1) = min\left\{rs, max\{0, r_d^i(t) + \beta(nt - |Ni(t)|)\}\right\} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

And $r_d^i(t+1)$ is glowworm is local decision range at the t+1 iteration, rs is the sensor range, nt is the neighborhood range. The number of glow in local decision range is given by equation (2)

$$N_{i(t)}=\left\{j:\|xi(t)-xi(t)\|<r_d^i ; li(t)<li(t)\right\} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

And xi(t) is the glowworm I position at the t iteration(t) is the glowworm i luciferin at the t iteration the set of neighbor of glowworm i consist of those glowworm that have relatively higher luciferin value and that are located within dynamic decision domain whose range r_d^i is bounded above by a circular sensor range[6-9].

Each glowworm is given in equation (3)

$$p_{ij(t)}=\frac{l_{i(t)-l_{i(t)}}}{\sum_{k\in Ni(t)} lk(t)-li(t)} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3)$$

Movement update is given in equation (4)

$$x_{i(t+1)=xi(t)+s\left(\frac{sj(t)-xi(t)}{\|xj(t)-xi(t)\|}\right)} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (4)$$

Luciferin update is given in equation (5)

$$l_{i(t)=(1-\rho)li(t-1)+\gamma j\left(xi(t)\right)} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (5)$$

And li(t) is a luciferin value of glowworm i at the t iteration, P belong (0,1) lead to the reflection of the cumulative kindness of the path followed by the glowworm in their current luciferin values the parameter Y only scale the function values, J(xi(t)) is the value of test function. Finally gets the optimal feature. The optimal feature passes through SOM model.

III PROPOSED ALGORITHM

In this section discuss the proposed algorithm for stream data clustering. For clustering used SOM neural network model and for the minimization and optimization of feature contains used glowworm optimization algorithm. The glowworm optimization algorithm works on the principle of nearest neighbor. For the extraction of features used partial feature

extraction process. The partial feature extraction process extracted the feature of stream data in terms of partial boundary value. The stream data features are extracted from the stream data using energy function. SOM acts as a clustering mechanism that projects N-dimensional features from the energy function into an M-dimensional feature space. The resulting vectors are fed into an SOM that categorizes them onto one of the relearned n classes. The transformed feature vectors are fed into the SOM, which classifies them. We call the feature space generated from the Energy function output as primary feature space and M-dimensional feature space from SOM output as secondary feature space. The vectors from the secondary feature space are called secondary feature vectors. The concept behind the use of SOM as an intermediate stage is that it can perform and enhanced it. Topology preserving feature mapping from its input space to output space, and these mapped features, which are of reduced dimension, can represent the necessary information in the input features [20]. Thus, the training and segmentation of the upper stage (SOM) can be done in a reduced dimension compared to the higher dimension of the primary feature space.

Step1. Initially input stream data passes through MCM

Step2.the extracted feature passes through glowworm optimization algorithm.

Step3. The glowworm optimized the feature of input stream data

Step4. In phase of feature mapping in feature space of SOM network create a fixed cluster according to objective function.

Step5. Here show steps of processing of SOM network

1) Define weight of each node value.

2) Select a random cluster from training data and present it to the SOM.

3) All nodes satisfied the minimization of objective function.

4) The radius of the area around the objective is calculated. The size of the area decreases with each iteration.

5) Each node in the objective function area has its weights adjusted to become more like the objective. Nodes closest to the objective are altered more than the nodes furthest away in the neighborhood.

6) Repeat from step 2 for enough iteration for convergence.

7) Calculating the objective is done according to the Euclidean distance among the node's weights (W_1, W_2, … ,W_n) and the input vector's values (V_1, V_2, … , V_n).

    1) This gives a good measurement of how similar the two sets of data are to each other.

8) The new weight for a node is the old weight, plus a fraction (L) of the difference between the old weight and the input vector… adjusted (theta) based on distance from the BMU.

9) The learning rate, L, is also an exponential *decay* function.

    1) This ensures that the SOM will converge.

10) The lambda represents a time constant, and t is the time step

Step6. After processing of SOM network out data of stream data is segmented is done

Step7. Finally gets segmented stream data and estimate the value of global consistency error rate.

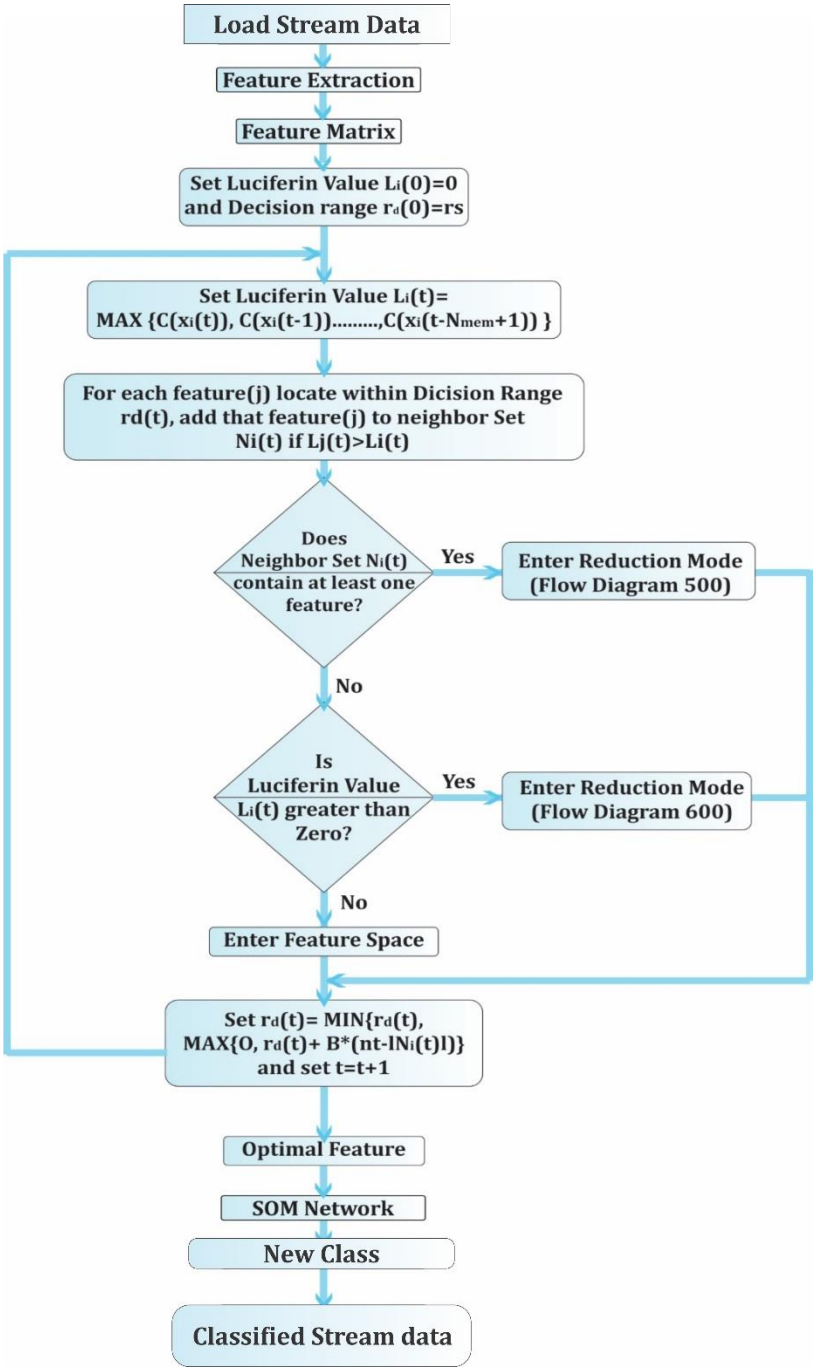

Figure 1: Proposed model of stream data classification.

IV. EXPERIMENTAL PROCESS

For the evaluation of performance of MCM and PROPOSED, we implement our algorithm in matlab 7.8.0 and for tested of result used UCI machine respority data set. Here we used three data set glass data set croups data set and finally used forest fire data.

The result measurement paprmeter is Fnew, Mnew and Error rate of classification. Here shows the evaluation table of result.

For crops data set both method computed result MCM and PROPOSED

Chunk size	Error rate	Mnew	Fnew	Time in seconds
10	0.569	13.5 34	0.4373	0.328215
20	0.769	12.786	0.5216	0.453076
30	0.989	11.368	0.3426	0.631141
40	0.536	15.865	0.3688	0.743038
50	0.434	16.458	0.7586	0.562971
60	0.379	17.336	0.6587	0.860288
70	0.289	15.123	0.5689	0.730222

Table 1 show that the calculated result of feature selection of MCM and PROPOSED for crops data set

Chunk size	Error rate	Mnew	Fnew	Time in seconds
10	0.458	5.164	12.601694	11.212107
20	0.428	4.563	14.228785	12.609094
30	0.364	5.224	14.329298	13.659516
40	0.338	4.144	18.324813	14.230065
50	0.278	5.298	28.463011	28.624513
60	0.122	6.132	31.367421	30.632463
70	0.102	5.456	51.797559	48.618276

Table 2 shows that the for crops data set both method computed result MCM and PROPOSED.

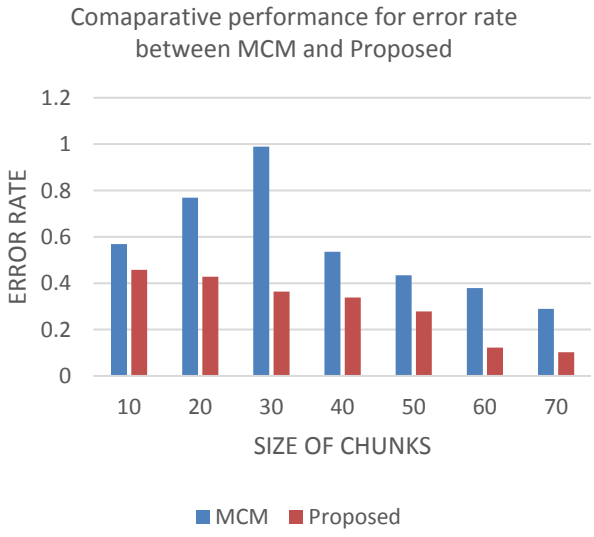

Figure 2: Show that the Error rate performance.

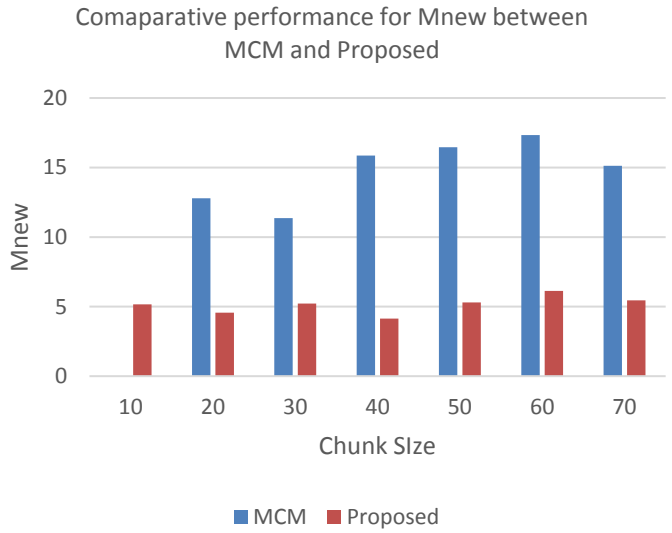

Figure 3: Show that the Mnew performance.

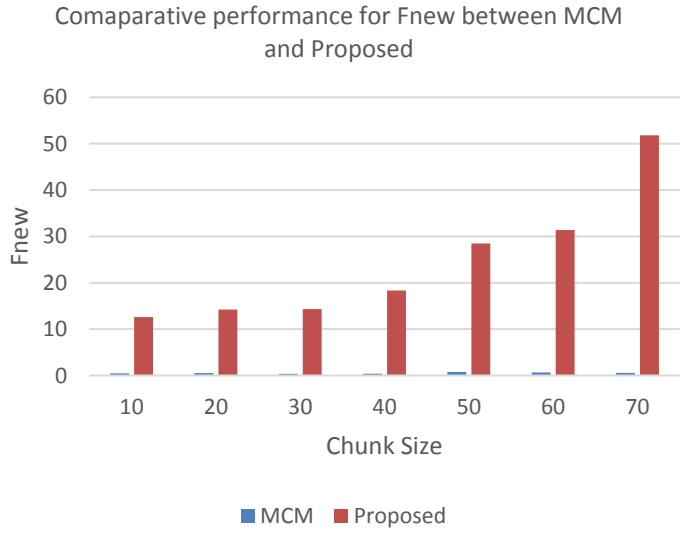

Figure 4: Show that the Fnew performance.

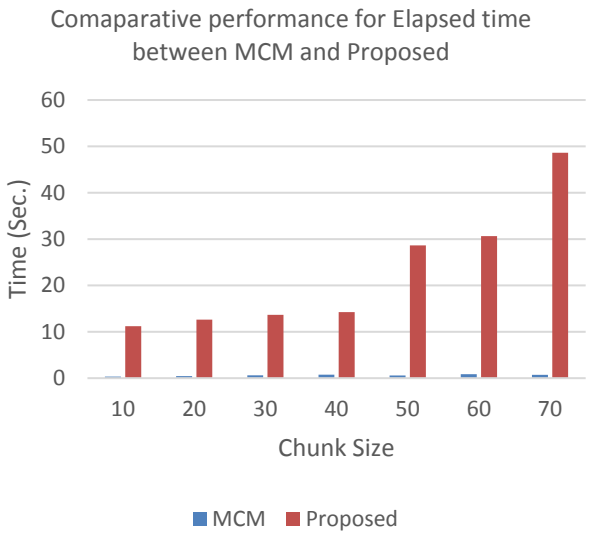

Figure 5: Show that the elapsed time performance.

V. CONCLUSION

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 improvements still some problem is remaining such as infinite length and data drift. Infinite length and data drift problem are not considered in this paper. 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 GSO of swarm prototype cluster faced a problem of right number of cluster, in future used self-optimal clustering technique along with glow worm optimization.

References

[1] Srilakshmi Annapoorna P.V and Mirnalinee T.T "Streaming Data Classification", Trends in Information Technology, 2016, Pp 1-7.

[2] Mahardhika Pratama, Sreenatha G.Anavatti, Meng-joo Er and Edwin Lughofer "pClass : An Effective Classifier for Streaming Examples", IEEE, 2015, Pp 1-19.

[3] Poonam Sonar and Udhav Bhosle and Chandrajit Choudhury "Mammography Classification Using Modified Hybrid SVM-KNN", International Conference on Signal Processing and Communication, 2017, Pp 305-311.

[4] Michal Wozniak,Pawel Ksieniewicz, Boguslaw Cyganek, Andrzej Kasprzak and Krzysztof Walkowiak "Active Learning Classification of Drifted Streaming Data", Procedia Computer Science, 2016, Pp 1724–1733.

[5] Jonathan A. Cox, Conrad D. James and James B. Aimone "A Signal Processing Approach for Cyber Data Classification with Deep Neural Networks", Procedia Computer Science, 2015, Pp 349 – 354.

[6] Heng Wang and Zubin Abraham "Concept Drift Detection for Streaming Data", arXiv, 2015, Pp 1-9.

[7] Nemanja Djuric, Hao Wu, Vladan Radosavljevic, Mihajlo Grbovic and Narayan Bhamidipati "Hierarchical Neural Language Models for Joint Representation of Streaming Documents and their Content", arXiv, 2016, Pp 1-8.

[8] Sara del Rıo, Victoria Lopez, Jose Manuel Benıtez and Francisco Herrera "A MapReduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules", International Journal of Computational Intelligence Systems, 2015, Pp 422-437.

[9] Simon Fong, Raymond Wong and Athanasios V. Vasilakos "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data", IEEE, 2015, Pp 1-14.

[10] Arati Kale and M.D. Ingle "SVM based Feature Extraction for Novel Class Detection from Streaming Data", International Journal of Computer Applications, 2015, Pp 1-3.

[11] Fan Zhang, Junwei Cao, Samee U. Khan, Keqin Li and Kai Hwang "A task-level adaptive MapReduce framework for real-time streaming data in healthcare applications", Future Generation Computer Systems, 2015, Pp 149–160.

[12] Nawel Yala, Belkacem Fergani and Anthony Fleury "Feature extraction for human activity recognition on streaming data", IEEE, 2015, Pp 1-6.

[13] Indre˙ Žliobaite, Albert Bifet, Jesse Read, Bernhard Pfahringer and Geoff Holmes "Evaluation methods and decision theory for classification of

streaming data with temporal dependence", Springer, 2015, Pp 1-28.

[14] Shifei Dinga,b, Yuexuan An, Xiekai Zhanga, Fulin Wu and Yu Xue "Wavelet twin support vector machines based on glowworm swarm optimization", Neurocomputing, 2017, Pp 157–16

3.