A Feature Selection process Optimization in multi-class Miner for Stream Data Classification

Manish rai
Department of Computer Science & Engg., LNCT, Bhopal

Rekha Pandit
Department of Computer Science & Engg., LNCT, Bhopal

Vineet Richhariya
Department of Computer Science & Engg., LNCT, Bhopal

ABSTRACT

Multi-class miner resolves the problem of feature evaluation, data drift and concept evaluation of stream data classification. The process of stream data classification in multi-class miner based on ensemble technique of clustering and classification on feature evaluation technique. The process of feature evaluation technique faced a problem of correct point selection of cluster centre for the process of data grouping. For the proper selection of features point we used optimization technique for feature selection process. The feature selection process based on advance genetic algorithm (AGA). The advance genetic algorithm poses a process of feature point for neighbor class detection for finding a correct point in classification. Our proposed algorithm tested on some well knows data set provided by UCI machine learning repository. Our empirical evaluation result shows that better result in comparison of multi-class miner for stream data classification.

KEYWORDS: - Stream Data classification, MCM, AGA and MGM-GA

INTRODUCTION

Data stream is a fast and continuous phenomenon, it is assumed of infinite length. So it is not possible to store the data and use it for training. Infinite length, concept-evolution, feature-evolution and concept-drift are major challenges in data streaming. The infinite length problem is dividing the stream into equal-sized chunks so that every chunk can be stored in memory and processed online. Every chunk is used to train a classification model as soon as all the instances in the chunk are labeled. Concept-drift occurs when the concept of time changes over the time. Concept-evolution occurs when new classes are involved in the data. In practical when types of attacks are labeled as new class in intrusion detection system then it creates problem. Feature-evolution occurs when new features are added and old features faded away. To detect the presence of concept-drift and infinite length ensemble classification technique is used where every classifier is equipped with a novel class detector. In this technique an ensemble of models is used to classify the unlabeled data, and to detect novel classes. Each model in the ensemble is evaluated periodically, and old, obsolete models are discarded often. The data stream is divided into equal size called chunks and the ensemble classifies the data point inside the chunks. Every incoming stream is checked with outlier module and if it is found outlier then it is stored in the buffer but if it is not found outlier then it is classified as an existing class using ensemble technique. Ensemble techniques need relatively simple operations to revise the current concept than their single model counterparts, and also handle concept-drift proficiently. To detect the novel class in a data stream multiclass miner is also used. Multiclass miner is combination of OLINDDA and FAE approach. This combination work with dynamic feature vector and detect novel class. OLINDDA is used to detect the novel class and FAE classifies the data chunks. MCM detects outlier classification and also used in recognizing the novel class instances. MCM is the fastest method in all datasets. MCM is roughly 25% faster than Mine Class. The reason for faster running time is it uses the dynamic threshes holding and Gini coefficient analysis; MCM filters out the majority of the outliers and reduces the cost of novel class detection because it is proportional to the number of outliers. Genetic algorithm is also advantageous to the data streaming. This genetic algorithm is used to rectify the problem of segmenting the data stream. Properly segmented streams can be better arranged and reused. They provide points of access that facilitate browsing and retrieval. Data stream classification has many approaches. These approaches fall into two categories: single model and ensemble classification. Single model classification techniques uphold and incrementally renew a single classification model and effectively react to concept-drift.

The above section discuss introduction of stream data classification. In section II we discuss various related method for stream data classification. In section III discuss MCM. In section IV we discuss proposed method of MCM-GA. In section V we discuss experimental result and finally conclude in section VI.

II. RELATED WORK

In this section we discuss method for stream data classification for minimization and removal a problem such as infinite length, data drift, concept evaluation and feature evaluation. All these method reduce such problem, Yan-Nei Law and Carlo Zanily entitled" An Adaptive Nearest Neighbor Classification Algorithm for Data Streams" [2] describe a process of stream data classification by adaptive nearest classification as the algorithm achieves excellent performance by using small classifier ensembles where approximation error bounds are guaranteed for each ensemble size. The very low update cost of our incremental classifier makes it highly suitable for data stream applications. ANNCAD is very suitable for mining data streams as its update

speed is very quick. Also, the accuracy compares favorably with existing algorithms for mining fact streams. ANNCAD adapts to concept drift efficaciously by the exponential bury approach. However, the very detection of sudden concept drift is of interest in many applications. The ANNCAD framework can also be extended to detect concept drift, for example changes in class label of blocks is a good indicator of possible concept drift.

. Mohammad M. Masud, Latifur Khan, Bhavani Thuraisingham entitled "Classification And Novel Class Detection In Concept-Drifting Data Streams Under Time Constraints" [3] describe a process of stream data classification by novel class detection In Concept-Drifting Data Streams Under Time Constraints as Novel class detection problem becomes more challenging in the presence of concept drift, when the underlying data distributions develop in streams. In order to determine whether an instance belongs to a Novel class, the classification models sometimes require waiting for more test instances to discover similarities among those instances. A maximum allowable wait time Tc is imposed as a time constraint to classify a test instance. Moreover most existing stream classification approaches assume that the true label of a data point can be accessed immediately after the data point is classified. In realness, a time delay Tl is involved in obtaining the true label of a data point since manual labeling is time overwhelming. We show how to make fast and accurate classification decisions under these constraints and apply them to real benchmark data. Comparing with state of the art stream classification techniques proves the superiority of our approach. The concept evolution problem, which has been neglect by most of the existing data stream classification techniques. Existing data stream classification techniques assume that total number of classes in the stream is set. Therefore instances belonging to a novel class are misclassified by the currently techniques. We show how to detect novel classes automatically even when the classification model is not trained with the novel class instances. Novel class detection becomes more challenging in the presence of concept-drift.

Mohammad M. Masud, Qing Chen, Latifur Khan, Charu Aggarwal and Jing Gao , Jiawei Han and Bhavani Thuraisingham [4]"Addressing Concept-Evolution in Concept-Drifting Data Streams **"**in this title author describe a process of stream data classification by Concept-Evolution in Concept-Drifting Data Streams as Concept-evolution occurs as a result of new classes evolving in the stream. This method addresses concept-evolution in addition to the existing challenges of infinite-length and concept-drift. The concept-evolution phenomenon is studied and the insights are used to construct a superior novel class detecting techniques. Firstly, we suggest an adaptive threshold for outlier detection, which is a vital part of novel class detection. Secondly, we suggest a probabilistic approach for novel class detection using discrete gini Coefficient and this prove its effectiveness both theoretically and empirically. Finally, address the issue of simultaneous multiple novel class occurrence and give a refined solution to detect more than one novel class simultaneously. We also consider feature evolution in text data streams which occurs because new features (i.e., words) evolve in the stream data classification. Comparison with state of the art data stream classification techniques establishes the effectiveness of the propose approach we propose an improved technique for outlier detection by defining a dynamic slack space outside the decision boundary of each classification pattern. Secondly we suggest a better alternative for identifying novel class instances using discrete Gini Coefficient. Finally, we propose a graph-based approach for distinguishing among multiple novel classes. We apply our technique on several real data streams that experience concept-drift and concept-evolution, and achieve significant performance improvements over the existing techniques.

Valerio Grossi, Alessandro Sperduti "Kernel-Based Selective Ensemble Learning for Streams of Trees"[5] author proposed a process of stream data classification by Kernel-Based Selective Ensemble Learning as Kernel methods enable the modeling of structured data in learning algorithms, still they are computationally demanding. Both efficacy and efficiency of the proposed approach are assessed for different models by using data sets exhibiting different levels and types of concept drift. Kernel methods provide a powerful tool for modeling structured objects in learning algorithms. Unfortunately, they require a high computational complexity to be used in streaming environments. This work is the first that demonstrates how kernel methods can be employed to define an ensemble approach able to quickly react to concept drifting and guarantees an efficient kernel computation.

Li Su Xi, Hong-yan Liu, Zhen-Hui Song. "A New Classification Algorithm for Data Stream" [6] in this method describes a process of stream data classification by Associative classification (AC) as Associative classification (AC) which is based on association rules has shown great promise over many other classification techniques on static dataset. Meanwhile, a new challenge has been proposed in that the increasing prominence of data streams arising in a wide range of advanced application. This technique describes and evaluates a new associative classification algorithm for data streams which is based on the estimation mechanism of the lossy Counting (LC) and landmark window model. And this technique was applied to mining several datasets obtained from the UCI Machine Learning Repository and the result show that the algorithm is effective and efficient.

Clay Woolam, Mohammad M. Masud, and Latifur Khan "Lacking Labels in the Stream: Classifying Evolving Stream Data with Few Labels" [7]in this title

author describe a process of stream data classification by Evolving Stream Data with Few Labels as It is practical to assume that only a small fraction of instances in the stream are tagged. A more practical assumption would be that the labeled data may not be independently distributed among all train documents. How can we ensure that a good classification model would be built in these scenarios, considering that the data stream also has changing nature? In our previous work we apply semi-supervised clustering to build classification models using limited amount of labeled train data. However, it assumed that the data to be labeled should be chosen randomly.

Mohammad M. Masud, Qing Chen, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham "[8]Classification and Novel Class Detection of Data Streams in a Dynamic Feature Space" here author describe a process of stream data classification by DXMiner as DXMiner, which addresses four major challenges to data stream classification such as infinite length, concept-drift, concept-evolution, and feature evolution. Data streams are assumed to be limitless in length which demands single pass incremental learning techniques. Concept-drift occurs in a data stream when the underlying concept alteration over time. Most presenting data stream classification techniques address only the infinite length and concept drift problems. Still, concept-evolution and feature evolution are also major challenges, and these are neglect by most of the presenting approaches. Concept-evolution occurs in the stream when novel classes arrive, and feature-evolution occurs when new features emerge in the stream and old features fade away. Our previous work addresses the concept evolution problem in addition to addressing the infinite length and concept-drift problems. DXMiner considers the dynamic nature of the feature space and provides an elegant solution for classification and novel class detection when the feature space is dynamic. We show that our approach outperforms state-of-the-art stream classification techniques in classifying and detecting novel classes in real data streams. Most of the existing data stream classification techniques either cannot detect novel class, or does not consider the Dynamic nature of feature spaces.

Charu C. Aggarwal ,Jiawei Han, Jianyong Wang, Philip S. Yu "A Framework for On-Demand Classification of Evolving Data Streams],This model indicate real-life situations effectively, since it is desirable to classify test streams in real time over an evolving training and test stream. The objective here is to make a classification system in which the training model can adapt quickly to the changes of the underlying data stream. In order to achieve this goal, we propose an on-demand classification process which can dynamically select the appropriate window of past training data to build the classifier. The empirical results show that the system maintains high classification accuracy in a developing data stream, while providing an efficient solution to the classification task. The stream classification framework proposed in this study has the following fundamental differences from the previous stream classification work in design philosophy. First, due to the dynamic nature of evolving data streams.

III MCM

In this section we discuss multi-class miner for feature evaluation, infinite length problem for data stream classification. The multi-class miner algorithm basically consists of ensemble technique of clustering and classification [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.

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) //clustering
3: for h ∈ N P_list do
h.nn ← Nearest-neighbor (NP_list - {h})
4: h.sc ← Compute-SC (h, h.nn) //silhouette coefficient
5: V←V ∪ {h} //add these nodes
6: V←V ∪ {h.nn}
7: if h.sc < th_{sc} then //relatively closer to the nearest neighbor
8: E←E ∪ {(h,h.nn)} //add this directed edge
9: end if
10: end for
11: count ← Con-Components (G) //find connected components

// merging phase

12: for each pair of components (g1,g2) ∈ G do
13: μ_1←mean-dist (g1), μ_2←mean-dist (g2)
14: if $\frac{\mu_1+\mu_2}{2*centroid\_dist(g1,g2)} > 1$ then g1←Merge (g1, g2)
15: end for

// Now assign the class labels

16: N_type ← empty
17: for x ∈ Nlist do

18: h ← PseudopointOf (x) //find the corresponding pseudopoint
19: N_type ← N_type ∪ {(x, h.componentno)}
20: end for

IV. PROPOSED METHOD MODIFIED MCM-GA

In this section we discuss the modification of multi-class miner algorithm with genetic algorithm. Genetic algorithm is heuristic function; the nature of genetic algorithm is single objective for optimization of given problem. In multi-class miner genetic algorithm play a role of seed selection of better generation of cluster radius for grouping of new feature data in ensemble process. In the process of MCM the graph points of number of feature point selection executed by genetic algorithm.

Modified Steps of MCM-GA algorithm

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 Kv , population scale XN , crossover probability cP , mutation probability mP , vaccination 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 need optimization of cluster center, which means loss of data of waiting 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 centers and fuzzy partition matrixes. And set the optimal clustering partition according to maximum membership principle and output the results.

MCM and MCM-GA

MCM

Data Chunk Size	Error Rate	Mnew	Fnew	Elapse Time In Seconds
10	0.789	7.234	88.108	85.141486
20	0.689	5.234	77.063697	74.886859
30	0.589	6.589	88.124168	85.426146
40	0.489	12.234	103.24067	101.24276
50	0.389	14.234	94.489009	92.242545
60	0.289	14.234	138.84009	137.0451
70	0.189	15.234	114.2075	112.5605

10: Do the parallel crossover and mutation operation on population A(i), then we can 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: μ_1←mean-dist (g1), μ_2←mean-dist (g2)
24: if $\frac{\mu_1+\mu_2}{2*centroid\_dist(g1,g2)} > 1$ then g1←Merge (g1, g2)
25: end for

// Now assign the class labels

26: N_type ← empty
27: for x ∈ Nlist do
28: h ← PseudopointOf (x) //find the corresponding pseudopoint
29: N_type ← N_type ∪ {(x, h.componentno)}
30: end for

V EXPERIMENTAL RESULT ANALYSIS

For the evaluation of performance of MCM and MCM-GA,

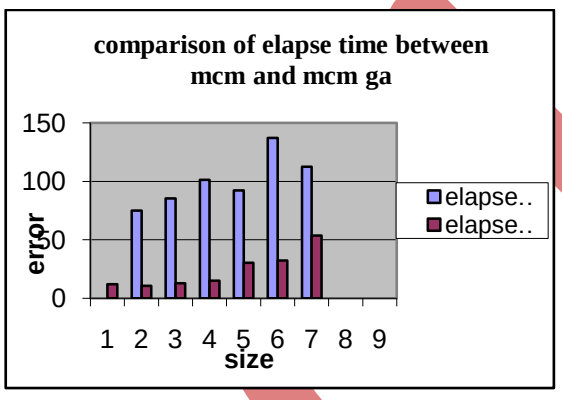
We implement our algorithm in matlab 7.8.0 and for tested of result used UCI machine repository data set. Here we used three data set glass data set crops data set and finally used forest fire data. The result measurement parameter is Fnew, Mnew and Error rate of classification. Here shows the evaluation table of result.

For crops data set both method computed result

MCM-GA

Size	Error rate	Mnew	Fnew	Elapse Time In Sec
10	0.478	4.234	14.501694	12.112107
20	0.378	3.234	13.128885	10.809094
30	0.278	4.234	15.219298	12.79616
40	0.178	4.234	17.4813	15.0065
50	0.078	5.234	32.863011	30.513
60	0.022	5.234	34.407421	32.2463
70	0.122	7.234	55.857559	53.68276

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



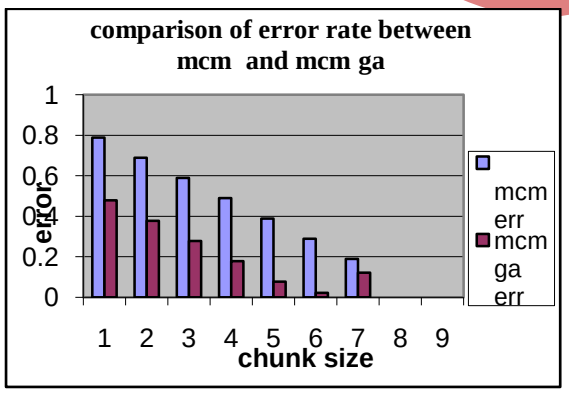
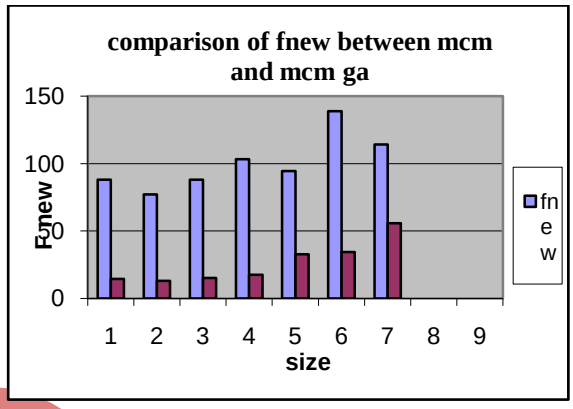


Figure 1 shows that the comparison graph between time complexity and Error rate of both method MCM and MCM-GA for crops data set



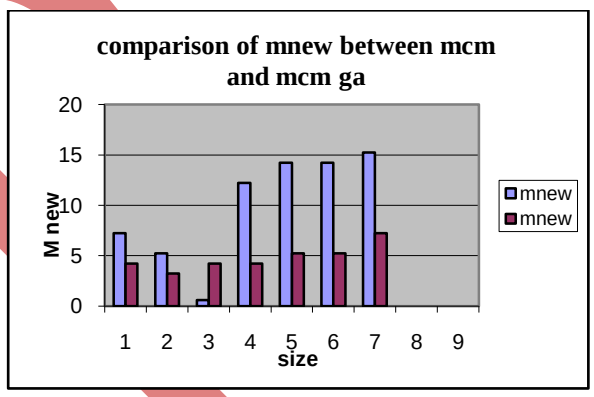


Figure 2 shows that the comparison graph between Fnew and Mnew rate of both methods MCM and MCM-GA for crops data set.

VI CONCLUSION

In this paper we modified a multi-class miner algorithm using genetic algorithm. The empirical evaluation of modified algorithm is better in compression of MCM algorithm. The error rate of modified algorithm decreases in comparison of MCM algorithm. Also improved the rate of Fnew and Mnew for evolution of result. The error rate reduced 20% in comparison of mcm .instate of that classification of data are improved.

REFERENCES:-

[1]Urvesh Bhowan, Mark Johnston, Mengjie Zhang and Xin Yao "Evolving Diverse Ensembles using Genetic Programming for Classification with Unbalanced Data" in IEEE Tansaction2010.

[2]Yan-Nei Law and Carlo Zanily entitled" An Adaptive Nearest Neighbor Classification Algorithm for Data Streams" in PKDD 2005, LNAI 3721, pp. 108–120, 2005.

[3]Mohammad M. Masud, Latifur Khan, Bhavani Thuraisingham entilied "Classification And Novel Class Detection In Concept-Drifting Data Streams Under Time Constraints" in IEEE TRANSACTIONS ON

KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 6, JUNE 2011

[4]Mohammad M. Masud, Qing Chen, Latifur Khan, Charu Aggarwal and Jing Gao , Jiawei Han and Bhavani Thuraisingham "Addressing Concept-Evolution in Concept-Drifting Data Streams " in IEEE Transaction 2010.

[5]Valerio Grossi, Alessandro Sperduti "Kernel-Based Selective Ensemble Learning for Streams of Trees" in Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence 2010.

[6]Li Su Xi, Hong-yan Liu, Zhen-Hui Song. "A New Classification Algorithm for Data Stream"

[7]Clay Woolam, Mohammad M. Masud, and Latifur Khan "Lacking Labels in the Stream: Classifying Evolving Stream Data with Few Labels" in *I.J.Modern Education and Computer Science*, 2011

[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" in ISMIS 2009, LNAI 5722, pp. 552

[9]Charu C. Aggarwal ,Jiawei Han, Jianyong Wang, Philip S. Yu "A Framework for On-Demand Classification of Evolving Data Streams" in ECML PKDD 2010, Part II, LNAI 6322, pp. 337–352,

[10]Mohammad M. Masud,Qing Chen,Latifur Khan,Charu C. Aggarwal,Jing Gao, Jiawei Han, Ashok Srivastava and Nikunj C. Oza" Classification and Adaptive Novel ClassDetection of Feature-Evolving Data Streams" in IEEE TRANSACTION -2012.

[11]A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavald. New ensemble methods for evolving data streams. In *Proc.SIGKDD*, pages 139–148, 2009.

[12]S. Chen, H. Wang, S. Zhou, and P. Yu. Stop chasing trends:Discovering high order models in evolving data. In *Proc. ICDE*,pages 923–932, 2008

[13] S. Hashemi, Y. Yang, Z. Mirzamomen, and M. Kangavari. Adapted one-versus-all decision trees for data stream classification. *IEEE Trans. Knowl. Data Eng*, 21(5):624–637, 2009.

[14] G. Hulten, L. Spencer, and P. Domingos. Mining time changing data streams. In *Proc. SIGKDD*, pages 97–106, 2001.

[15] I. Katakis, G. Tsoumakas, and I. Vlahavas. Dynamic feature space and incremental feature selection for the classificationof textual data streams. In *Proc. ECML PKDD*, pages 102–116,

2006.

[16] I. Katakis, G. Tsoumakas, and I. Vlahavas. Tracking recurring contexts using ensemble classifiers: an application to email filtering. *Knowledge and Information Systems*, 22:371–391, 2010