VOL. 10, NO. 8, MAY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

INCREMENTAL AGGREGATION MODEL FOR DATA STREAM CLASSIFICATION

S. Jayanthi[1] and B. Karthikeyan[2]
[1]Department of Computer Science and Engineering, Karpagam University, Coimbatore, India
[2]Dhanalakshmi Srinivsan Institute of Research and Technology, Siruvachur, Perambalur, India
E-Mail: nigilakash@gmail.com

ABSTRACT

In online data stream processing, data stream classification task confronts several challenges such as, concept drift, concept evolution and partial labeling due to the dynamic nature of data streams. Amid these issues, concept drift is on the top concern that degrades the accuracy of data stream classification task, immediately upon its occurrence. However, concept evolution and partial labeling are also equally notable plights that are not focused by most of the existing approaches. Ensemble learning is a widely accepted prominent method that attempts to reconcile the issues encountering in the data stream classification. Our previous work addresses only the different types of concept drifts. This paper expounds a Novel Incremental Aggregation Model (IAM) which makes use of Adaptive Probabilistic Neural Network (APNN), Aggregate Weighted Ensemble Model (AWEM) and Ensemble Cloning that makes the system impeccable by combating against all the above said issues. The performance of the proposed algorithm has been experimentally tested with few synthetic data sets. Experimental results show that our model outperforms the existing ensemble approaches in terms of accuracy.

Keywords: data stream processing, data stream classification, concept drift, concept evolution, partial labelling, ensemble learning;

1. INTRODUCTION

In this technological era, streams of incremental data are being generated in almost all digitalized organizations. Extracting knowledge from such data streams is the key for the success of these organizations. In general, data streams are massive in size, dynamic in nature, infinite in length. Hence, conventional data mining techniques which work well only on stationary data become unsuitable for handing dynamic data streams. Moreover, Data stream processing techniques have several resource constraints such as, single scanning of data streams, limited memory size and processing time. Indeed, these constraints might not be met with conventional classification techniques [12], [15], [16], [17].

Among several task of data stream processing, data stream classification is a most prominent supervised task that predicts and classifies the upcoming data streams in ever-changing data distribution center [21]. While data stream processing, data stream classification task confronts several challenges, such as, concept drift, concept evolution, partial labeling, outlier and real time analysis. Concept drift is one of the most common phenomenons in data stream processing that occurs due to any of the changes in data distribution center, interest besides the target concept and the rules underlying the classification task [2], [8], [11]. In general, incremental learning data set is subject to concept drift. Due to this, almost all incremental learning algorithms have constant look out on concept drift. Concept evolution and partial labeling occurs due to the emergence of novel classes and unlabelled instances respectively.

This paper is segmented into six sections. Section two discusses some of the research issues stood behind the data stream classification task. Section three addresses some of the most cited related work of the proposed approach. Section four expounds the proposed novel incremental aggregation model and its architecture. Section five illustrates the experimental results and the last section is concluded by instilling the tactics for the enhancement of the proposed work.

2. RESEARCH ISSUES

The following research issues are the motivation behind the proposed research work.

A. It is found that many of the data stream classification algorithms in the literature do not forecast about concept drift and works well only with stable data distribution centre.
B. Few algorithms are good in confronting different types of concept drifts altogether. That is the algorithm that confronts gradual concept drift efficiently fails to be good in other type of concept drifts, and vice versa.
C. Even the methods good in handling concept drift are not good in other issues such as novel class occurrence and partial labeling. For example, our previous model named Aggregate Weighted Ensemble Model exclusively focused on combating against different types of concept drifts. However, this model did not forecast about the concept evolution and partial labeling [22], [23].
D. Few approaches attempted to expound the method for handling novel class occurrence and partial labeling [23].
E. Finally, no one method is good in confronting all the above said issues altogether in dynamic data streams.
F. In light of these challenges, the proposed work is focused on defending against all the above said issues.

3. RELATED WORK

In this section, we provide a thorough analysis of unattended data stream classification issues, namely, the concept evolution, concept drift, and partial labeling. A growing number of researches are going on to resolve the problems encountering in data stream classification. Most of these researches follows supervised learning, where the complete label of incoming instances are known and can be used for predicting the class labels of upcoming mysterious data streams. From this, it is implied that supervised learning algorithms are good only in handling completely labeled data streams [4], [6], [13].

Hence the contribution of unsupervised learning algorithm is highly imperative to handle concept evolution and partial labeling. Since the data stream distribution centers are highly subject to fluctuation. Both the supervised and unsupervised classifiers are needed to classify labeled and unlabeled instances (or novel classes). Moreover, it is very common for the misconception of the recurrent classes or outlier as novel classes.

In literature, there are several approaches used to handle the concept drift in different mode. In this section, we analyze the related work in two dimensions. First dimension is related to batch learning and the other dimension is related to incremental learning of dynamic data streams.

That is, these categories are framed based on the way it process the streams, namely, batch processing and incremental learning approaches [1], [5]. Batch processing approaches, however, processes the data streams; it produces deprived results in case of abrupt drift occurrence in data streams. Moreover, batch processing approaches are based on supervised learning method, which are able to classify only the class labels on which it is trained in advance.

In incremental learning, the classifiers learn from dynamic training data stream by incrementally revising the model, either by using single classifier learning approach or ensemble learning approach [10]. This approach achieves classification on dynamic data stream with less or no access to the previously used training data while preserving the knowledge about historical data and also it have the ability to learn novel classes. Interested readers shall construe with the detailed survey of data stream classification techniques [15].

4. PROPOSED WORK

In this section, we briefly discusses about the most prominent methods, namely, Adaptive Probabilistic Neural Network (APNN), Aggregate Weighted Ensemble Model and Ensemble Cloning which are used as the building components in our research experiments.

A. Adaptive PNN

In our proposed work, a variation of Probabilistic Neural Network (PNN) called the Adaptive PNN (APNN) is used to cope up with concept evolution in incremental learning. APNN works based on the competitive learning principle, “Winner takes all attitude”. To the best of our knowledge, till now no other existing research work in data stream classification implemented APNN. In this research work, Adaptive probabilistic neural network is used for novel class detection. Novel class or concept evolution means that the arrival of data in data streams are subject to form new classes when it significantly differs from the data stream used to train the classifier. APNN with some pre processing is used to address novel classes emerging from incremental data streams. Transformation and normalization are the two widely used preprocessing methods. Transformation manipulates input data stream to create a single input, while normalization is a transformation performed on an input data stream to scale it into an acceptable range. Since the neural network does not perceive alpha numeric data, first the data are transformed into neural network understandable format. At first, when a novel concept emerges from data streams, it can be considered as outlier. As time evolves, if growing number of similar concepts emerges in the data stream, and also high cohesion is found between them, then that concept could be enunciated as a novel class.

B. Aggregate weighted ensemble model

In our previous work, Aggregate weighted ensemble model has been expounded, which achieves better performance in data stream classification over all kinds of concept drifts such as gradual, incremental, sudden and recurrent concept drift. It achieves so by conferring more priority for the classifiers which is trained on recent data chunks and also tracking of the classifiers that performed well in the historical data chunks, The AWEM achieves better performance in data stream classification over all types of concept drifts such as gradual concept drift, incremental concept drift, sudden concept drift and recurrent concept drift [15], [16].

C. Ensemble cloning

In case of concept drift and novel class occurrence, the classifiers in the ensemble need to be updated and trained to learn new data. In our experimental work, ensemble cloning is implemented to achieve generality. Cloning of the ensemble is created only to learn new data chunks emerging in the data stream. In case of stable data stream, the current ensemble classifies the data chunks without any further learning. In case of successive novel class arrivals, one of the two ensembles can be chosen by using a selection procedure.

5. SYSTEM ARCHITECTURE

Initially, data streams are segmented into equal sized data chunks in order to cope up with infinite length. Then aggregate weighted ensemble model is applied where each classifier is trained on different data chunks separately. A synopsis on the status of the ensemble model is maintained in order to avoid the delay in data stream classification.

Each time when a new data chunk arrives, it is tested to confirm whether the data chunk contains labeled instances or unlabeled instances. If it contains only labeled instances, it is implicit that the system can able to classify the labeled instances. Labeled instances are the instances

by which the model is trained on. Unlabeled instances are alien to the model, hence the model needs to be trained to classify upon its arrival.

Concept drift occurs only when there is a change in target concept, rules used for labeling and data distribution center. If the data chunk contains unlabelled instance as well then it is tested for confirming outlier. If it is found as outlier, all the outliers are buffered for a while to find the correlation between them. If strong correlation is found among the outliers, they can be formed as a novel class [24], [25]. Else, unlabeled instances can be treated as noise.

In case of concept drift these systems effectively cope up with several types of concept drift using aggregate weighted ensemble model. When concept drift occurs the ensemble model is cloned and updated to cope up with concept drift and concept evolution. If there is no concept drift, the ensemble model is kept intact.

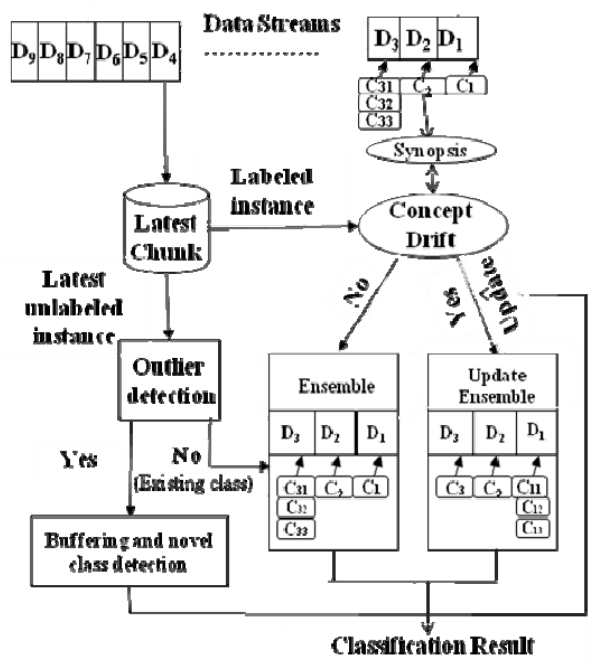

Figure-1. Incremental aggregation model.

Incremental Aggregate Model (M, Di)

Input: M: Ensemble of Classifiers, Di: Data Stream Instances

Output: C: Class Labels of Data Stream Instances

1. Partition D into Chunks, Cnks;
2. Train Classifiers using instances in Chunks
3. n1 = Cnks; //n is the size of ensemble
4. Scan Instances to discriminate labeled instances, X and unlabeled instances, Y
5. For X=1 to n1 do //n1 is the number of labelled instances
6. If Check _Outlier() = true
7. Remove ()
8. Else
9. Check for concept drift with synopsis
10. If Synopsis!=Cnks
11. Use M to learn chunks // SVM, Decision Tree, NaiveBase
12. Else
13. Use cloned ensemble M' to learn chunks
14. Endif
15. Endif
16. Calculate Avg on M, //avg: Average weighting
17. If Avg > Threshold_Value;
18. Classify Data Chunk using Classifier having majority vote; // Majority Voting();
19. Clone Ensemble=(Update_ Ensemble_Model());
20. Else
21. Classify Cnks with Ensemble_Model();
22. Endif
23. Endfor
24. End

Update _Ensemble Model (m, wi)

Input: M: Ensemble Model, Wi: Weight of classifiers

Output: C_{an}: best ensemble classifier

1) Choose best ensemble classifier, C_{an}, from aggregate ensemble; // Majority Voting
2) Current= C_{an}
3) For C1= 1 to M-1 do
4) For c2=1 to N-1 do
5) Choose best classifier, Cbn, from each ensemble ;
6) // Majority Voting
7) Train Cbn with novel classes or concept drift
8) Endfor
9) Endfor
10) End

6. EXPERIMENTS

We implemented the Incremental Aggregation Model using Java language. The code for naïve bayes classifier, decision tree, has been adapted from the Weka machine learning open source repository http: //www.cs.waikato.ac.nz/ml/weka/. The experiments were run on an Intel P-V machine with 4 GB memory and 3 GHz dual processor CPU. To investigate the accuracy of the proposed incremental aggregate ensemble model, we conducted experiments on two UCI datasets, namely, NSL-KDD data, Soybean data. Data its test dataset descriptions are displayed in Table-1.

Table-1. Test dataset descriptions.

Dataset	No. of att.	Att. types	Instances classes
NSL-KDD data	41	Real and Nominal	25192 23
Soybean data	35	Nominal	68319

Mean Square Error (MSE) evaluates the performance of an underlying classifier. The mean square error is also useful to sustain the concepts of bias,

precision, and accuracy. Mean square error rate can be calculated by calculating the sum of the variance and squared bias of the predictor.

$$MSE(\overline{\theta}) = Var(\overline{\theta}) + (Bias(\theta, \overline{\theta}))^2 \quad (1)$$

Mean Square Error Rate is slated on the following snapshot of the experiment, Error Rate Estimation (Table-1).

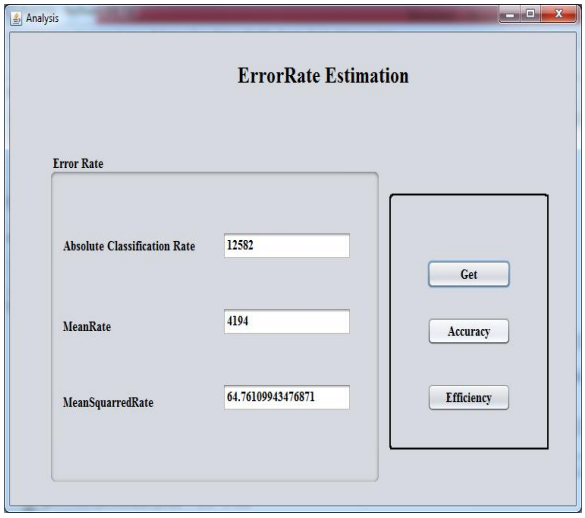

Figure-2. Error rate estimation.

Finally, we experimentally compared IAM with three online ensembles approaches, the Adaptive Classifier Ensemble (ACE), Dynamic Weighted Majority (DWM), and Leveraging Bagging (LB). The obtained result shows that IAM offers high classification accuracy in dynamic data stream environments. The average accuracy on three data sets of four algorithms is shown in Figure-3.

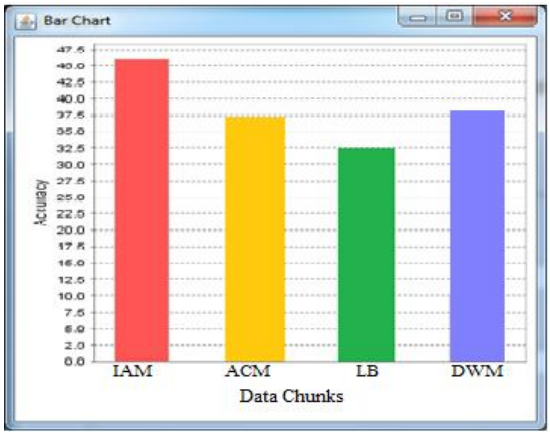

Figure-3. Performance evaluation.

7. CONCLUSIONS

In this research work, Incremental Aggregation Model is expounded not only to cope with concept drift, but also concept evolution and partial labeling. First, more specifically, Adaptive Probabilistic Neural Network is implemented to detect the novel classes in the data stream classification task. In addition, the ensemble model is cloned and undergone for training upon the arrival of novel class instances. The proposed system offers solution to reduce the classification error that occurs due to concept drifts in data stream classification. The proposed approach is experimentally tested on two datasets of UCI repository, and produced the satisfactory results in terms of accuracy than other three approaches, such as, ACM, LB, DWM. In our future work, it is planned to conduct the experiments on real time datasets.

REFERENCES

[1] Albert Hung-Ren Ko, and Robert Sabourin, **"**Single Classifier-based Multiple Classification Scheme for weak classifiers: An experimental comparison," ACM Journal Expert Systems with Applications: An International Journal, vol. 40(9), July, 2013, pp. 3606-3622.

[2] Albert Bifet, Geoff Holmes, Bernhard Pfahringer, and Ricard Gavalda, "Improving Adaptive Bagging Methods for Evolving Data Streams," ACM, Proceeding ACML „09, 2009, pp. 23-37.

[3] Charu C. Aggarwal, and Jianyong Wang, "A Framework for On-Demand Classification of Evolving Data Streams," IEEE Transactions on Knowledge and Data Engineering, vol.18 (5), 2006, pp. 577-589.

[4] Cheng-Jung Tsai, and Wei-Pang Yang, "An Efficient and Sensitive Decision Tree Approach to Mining Concept-Drifting Data Streams," Journal Informatica, Vol. 19(1), January 2008, pp. 135-156.

[5] Cheng-Jung Tsai, Chien-I. Lee, and Wei-Pang Yang, "Mining decision rules on data streams in the presence of concept drifts," Expert Systems with Applications, vol. 36: 2009, pp. 1164-1178.

[6] Chunquan Liang, Yang Zhang, Peng Shi, and Zhengguo Hu," Learning very fast decision tree from uncertain data streams with positive and unlabeled samples," ACM Journal Information Sciences: an International Journal, vol 213, December, 2012, pp. 50-67.

[7] Dariusz Brzezinski, and Jerzy Stefanowski, "Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm," IEEE Transactions on Neural Networks and Learning Systems, vol. 25(1), 2014, pp. 81-94.

[8] Dariusz Brzezinski, and Jerzy Stefanowski, "Combining block-based and online methods in learning ensembles from concept drifting data streams," Information Sciences, vol.265, 2014, pp. 50–67.

[9] Dayrelis Mena-Torres, Jesus S, Aguilar-Ruiz, "A similarity-based approach for data stream classification," Elsevier, Journal Expert Systems with Applications, vol. 41(9), July 2014, pp. 4224–4234.

[10] Dewan Md. Farid., Li Zhang., Alamgir Hossain., Chowdhury Mofizur Rahman., Rebecca Strachan, Graham Sexton, and Keshav Dahal, 2013. "An Adaptive Ensemble Classifier for Mining Concept-Drifting Data Streams," Elsevier, Expert Systems with Applications, vol. 40(15), pp. 5895–5906.

[11] Gama. J, Medas, P, Castillo. G, Rodrigues. P, "Learning with drift detection", Lecture Notes in Computer Science 3171, 2004, pp. 286-295.

[12] Haixun Wang, Wei Fan, Philip S. Yu, and Jiawei Han, "Mining concept-drifting data streams using ensemble classifiers," Proceedings of the 9th ACM SIGKDD, 2003, pp. 226–235.

[13] Hang Yang and Simon Fong, "Incremental Optimization Mechanism for Constructing a Decision Tree in Data Stream Mining", Hindawi Publishing Corporation, Mathematical Problems in Engineering, 2013, Article ID 580397.

[14] Hang Yang and Simon Fong, "Moderated VFDT in stream mining using adaptive tie threshold and incremental pruning," in Proceedings of the 13th International Conference on Data Warehousing and Knowledge Discovery, 2011, pp.471-483.

[15] S. Jayanthi, B. Karthikeyan, "A Recap on Data Stream Classification," Adv. in Nat. Appl. Sci.,8(17):76-82, 2014

[16] S. Jayanthi. B. Karthikeyan, "Aggregate Weighted Ensemble Model For Data Stream Classification," International Journal of Applied Engineering Research (IJAER), Volume 9, Number 21 (2014) Special Issues, pp.4945-4949

[17] Joao Gama, Raquel Sebastiao, and Pedro Pereira Rodrigues, "On evaluating stream learning algorithms,"Springer, Machine Learning, vol. 90(3), 2013, pp. 317-346.

[18] Jesse Read, Albert Bifet, Bernhard Pfahringer, and Geo Holmes, "Batch-incremental versus instance-incremental learning in dynamic and evolving data," Proceeding IDA"12, Springer-Verlag Berlin, 2012, pp.313-323.

[19] Jimenez Gonzalo Ramos, Jose del Campo-Ávila, Rafael Morales-Bueno, "Incremental Algorithm Driven by Error Margins", Proceedings 9th International Conference, DS 2006, Spain, 2006., pp 358-362

[20] Jing Liu, Xue Li, and Weicai Zhong, "Ambiguous decision trees for mining concept-drifting data streams," ACM Journal Pattern Recognition Letters, Vol.30(15), November, 2009, pp. 1347-1355

[21] Lior Rokach, "Ensemble-based classifiers," Springer, Artif Intell Rev. 33, 2010, pp.1–39.

[22] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani M. Thuraisingham, " A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data," IEEE ICDM, 2008, pp. 929-934.

[23] Mohammad M. Masud, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham, "Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints," IEEE Transactions on Knowledge And Data Engineering, vol.23(6), 2011, pp. 859-874.

[24] Mohammad M. Masud et al., "Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams,"IEEE Transactions On Knowledge And Data Engineering, vol. 25(7), 2013, pp. 1484-1497.

[25] Peng Zhang, Xingquan Zhu, Yong Shi, Li Guo, and Xindong Wu, "Robust Ensemble Learning for mining noisy data streams," Decision Support System, vol.50, 2011, pp. 469-479. [26]

[26] P. Domingos, and G. Hulten, "Mining high-speed data streams," Proceedings of 6th ACM SIGKDD, 2000, pp.71-80.

[27] W. Nick Street, and Yong Seog Kim, "A Streaming Ensemble Algorithm (SEA) for Large Scale Classification," Proceedings of the seventh ACM SIGKDD, 2001, pp. 377-382.

[28] Wenyu Zang, Peng Zhang, Chuan zhou and Li Guo, "Comparative study between incremental and ensemble learning on data streams: Case study", Journal of Big Data, 1:5, June 2014, pp.1-16.

[29] Xiaozhen Zhou, Shanping Li, Cheng Chang, Jianfeng Wu, and Kai Liu," Information-value-based feature selection algorithm for anomaly detection over data streams," Technical Gazette, vol.21, 2014, pp. 223-232.