ENSEMBLE CLASSIFICATION OF DATA STREAMS

Yamini Kadwe

Student, Department of IT

M.I.T. College Of Engineering

Pune, India

E-mail: kadwe.yamini@gmail.com

Vaishali Suryawanshi

Assistant Professor, Department of IT

M.I.T. College Of Engineering

Pune, India

E-mail: vaishali.suryawanshi@mitcoe.edu.in

***Abstract*--Data stream classification is a challenging task. For real-time data concepts of instances keep varying with time, such as weather prediction or intrusion detection etc. The concept changes in continuously evolving data streams are termed as concept drifts. To address this issue, idea of an adaptive ensemble classifier is presented. The classification approach is ensemble-based rather than using a single classifier. The set of classifiers are updated with respect to accuracy achieved during classification. This work includes initial experiments and results carried out using MOA, a data stream mining framework.**

Keywords: concept drift, ensemble, experts, chunk, component classifiers

I. INTRODUCTION

The data streams are large volumes of high speed ordered, continuous data evolving from real-domain applications. Data Stream Mining is the process of extracting knowledge structures from continuous, rapid data records. The data stream can be read only once or a small number of times using limited computing and storage capabilities.

Classification is a supervised technique of [14] mining information from continuously generated data streams. In this, we provide a set of training examples in the form *(i,j)*, where *i* is the vector of *n* attributes with *j* being the discrete class label and aim in producing a model of the form *j=f(a)*.The function *f(a)* should accurately predict the class *j* for the future examples.

Clustering is an unsupervised technique of mining useful knowledge from data streams. It refers to the task of grouping a set of objects in such a way that objects in the same group, [14]referred to as a cluster, are more similar, to each other in some way than to those in other groups (clusters). The figure 1 shows various data stream mining tasks that can be carried out in continuously evolving data streams.

The real-domain applications[6] of data stream mining considering drifts are monitoring and control, personal assistance, decision support and artificial intelligence applications.

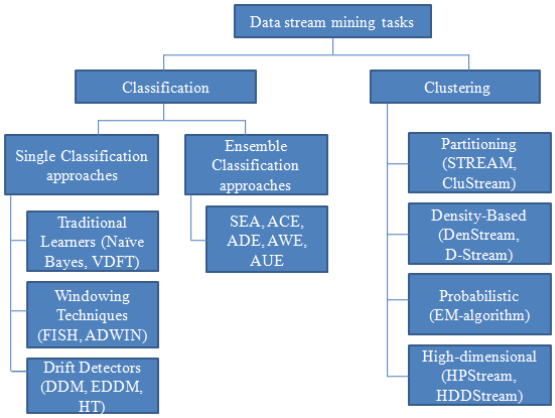

Figure 1: Data Stream Mining tasks

For instance, in an intrusion detection systems incoming network traffic is filtered in search of suspicious behavior. The source of concept drift in this application is mainly connected with the attacker. Adversary actions taken by the intruder evolve with time, to surpass the also evolving security systems.
Section I gives overview on data stream mining. The work in data stream classification algorithms is summarized in section II. The proposed algorithm and the objectives are presented in section III. Section IV and V gives the summary for initial experiments, results, followed by analysis.

II. RELATED WORK

The concepts in evolving data streams do not remain same. The change in concepts of the data is known as concept drift. [15]A drift can be *sudden* or *abrupt*, when concept switching is from one to another. The concept change can be *incremental*, consisting of many intermediate concepts in between. Drift may be *gradual;* change is not abrupt, but goes back to previous pattern for some time. Concept drift handling algorithms should not mix the true drift with an

outlier(blip) or noise, which refers to an anomaly. A recurring drifts is when new concepts that were not seen before, or previously seen concepts may *reoccur* after some time.

Several classification algorithms that cope with concept drift have been put forward, however, most of them specialize in one type of change.

A. *Single Classification approaches[6]*

Traditional Learners are the popular classifiers proposed for stationary data mining fulfill both of the stream mining requirements - have the qualities of an online learner and a forgetting mechanism. Some of the methods are neural networks, Naive Bayes, nearest neighbor methods, and decision rules.
Windowing technique is an approach to dealing with time changing data involves the use of sliding windows, that limits the amount of examples introduced to the learner. They include weighted windows, FISH, ADWIN and so on.
The drift detectors detect concept drift and alarm the base learner(using statistical test). DDM and EDDM are the drift detectors.
Hoeffding Tree or VFDT is the standard decision tree algorithm that uses the Hoeffding bound to decide the minimum number of arriving instances to achieve certain level of confidence in splitting the node.

B. *Ensemble Classification Approaches[10]*

In ensemble approaches, prediction of multiple classifiers are combined.

Streaming Ensemble Algorithm (SEA) [4] is a heuristic replacement strategy of the weakest base classifier based on accuracy and diversity with simple majority voting and base classifiers unpruned. *Accuracy Weighted Ensemble (AWE)* trains a new classifier C' on each incoming data chunk and use that chunk to evaluate all the existing ensemble members to select the best component classifiers.

Adaptive Classifier Ensemble (ACE) is a hybrid approach, in which a data chunk ensemble is aided by a drift detector. *Hoeffding Option Trees (HOT)* provide a compact structure that works like a set of weighted classifiers, and are built in an incremental fashion. It allows each training example to update a set of option nodes rather than just a single leaf. *Adaptive-Size Hoeffding Tree Bagging (ASHT Bagging)* diversifies ensemble members by using trees of different sizes and uses a forgetting mechanism.

Compared to AWE, the *Accuracy Updated Ensemble* (AUE1) conditionally updates component classifiers. It uses Hoeffding trees as component classifiers. Compared to AUE1, AUE2[11] introduces a new weighting function, does not require cross-validation of the candidate classifier, does not keep a classifier buffer, prunes its base learners, and always updates its components. It does not limit base classifier size and use any windows. *The Online Accuracy Updated Ensemble* (OAUE)[12], tries to combine block-based ensembles and online processing.

Generalizing, [1] there are 3 categories of ensemble classifiers.

Expert ensemble classifier method changes integration rules of experts with occurrence of concept drift. Ensemble classifier method updating the basic classifier set assigns “age” property to every basic classifier, then take the youngest basic classifier generated from the latest training sample to replace the oldest basic classifier. Ensemble classifier method with multi-algorithm use a variety of different classification algorithms. Such ensemble classifier methods are fit for dealing with data streams with continuous mutation.

III. PROPOSED WORK

The proposed ensemble approach uses Hoeffding trees as the basic classifier set. So, it is an expert ensemble classifier that uses a single algorithm and predicts class based on most recent chunk of data.

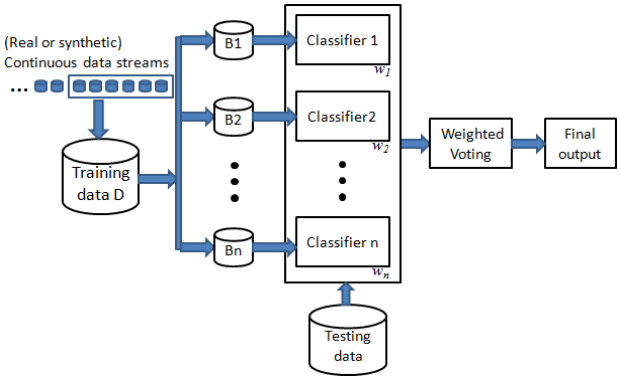

Fig 2 : Architecture of proposed system

The figure 2 shows working of the proposed adaptive ensemble classifier. It generates component classifiers sequentially from fixed size blocks of training examples called data chunks. A new classifier learned from the recent block is added to the ensemble and the weakest classifiers are removed according to the result of the evaluation.

From the fig. 2, the continuous data streams are given as input to the classifiers. This training data is divided into blocks $B_1...B_n$ and given to each component classifiers $C_1...C_n$. Based on the achieved classifier accuracy of each component classifier for each block of data, weights are calculated. The candidate classifier is then used to classify newly arriving data, i.e. testing data. In this way, the test data is classified based on component classification evaluation. The chunk size and number of classifiers will be 200 and 7 as proved in section IV and V.

The proposed algorithm will be able to handle different types of concept drifts.

The objectives of the proposed work are:

a) To develop an electricity pricing application
b) Comparative analysis of the proposed algorithm with state-of-the-art methods, evaluation based on parameter settings like classes, attributes, noise, drifts and evaluation based on performance parameters viz. accuracy, test time, train time.

An application on electricity pricing analysis will be developed. The electricity data that will be used for data stream classification originates from the New South Wales Electricity Market in Australia. In this market, prices are not fixed but set every five minutes, as determined by supply and demand. The data comprises 45,312 instances and 8 predictor attributes. The class label is determined based on the change of price relative to a moving average of the last 24 hours, yielding 2 class values.

IV. EXPERIMENTS & RESULTS

The proposed algorithm will be implemented in Java using Eclipse IDE. The algorithm, after successful implementation, can be added in the classifier list of the MOA framework.

This section summarizes experiments conducted on existing data stream classification algorithms, using various parameter settings and electricity dataset. The experiments are carried out on Massive Online Analysis(MOA) Tool, developed in Java. MOA is an open-source framework[8] for dealing with massive evolving data streams. It includes a collection of offline and online methods as well as tools for evaluation. The experiments were performed on a machine equipped with an Intel Core i7-2630QM @ 2.00 GHz processor and 8 GB of RAM.

Below are the results for the existing ensemble classification approach based on various parameter settings such as block size, number of component classifiers and performance measures viz. accuracy and time required. These results are carried out on electricity dataset.

In table 1, analysis on Accuracy Updated Ensemble(AUE2) is presented. This technique[11] has used chunk size as 500 and 10 component classifiers. The results are based on change in accuracy and time based on the change in chunk size and count of member classifiers. Member count is number of classifiers in an ensemble classifier.
In table 2, accuracy and time required for classification is checked for OAUE algorithm, the standard setting is same as AUE2 algorithm.

Table 1 : Analysis of AUE2 algorithm on Electricity dataset

	Member Count	**Chunk size**	**Accuracy (in %)**	**Time (in sec)**
1	10	500	77.44	3.28
2	10	300	77.89	3.70
3	10	200	78.29	3.78
4	7	500	76.82	2.48
5	5	500	76.13	1.86
6	5	300	76.86	2.06
7	7	200	78.07	2.79
8	13	300	78.10	4.52
9	13	200	78.75	4.65

Table 1 : Analysis of OAUE algorithm on Electricity dataset

	Member Count	**Window size**	**Accuracy (in %)**	**Time (in sec)**
1	10	500	87.74	4.35
2	10	300	88.61	3.76
3	10	200	88.76	3.84
4	7	500	87.80	2.59
5	5	500	87.60	2.01
6	5	300	88.08	2.14
7	7	200	88.37	2.85
8	13	300	88.50	4.51
9	13	200	88.98	4.70

V. ANALYSIS

From table 1, the AUE2 gives accuracy 77.44% in 3.28 seconds, for standard chunk size and number of classifiers. Also, from table 2, the OAUE on electricity dataset gives 87.74 accuracy in 4.35 seconds.
For smaller block sizes, accuracy has increased. With increase in set of classifier, there is a significant increase in accuracy. For instance, considering 7 classifiers and 200 chunk size, the accuracy has increased to 78.07% and minimized time of 2.79 seconds for AUE2; and accuracy has increased to 88.37% and minimized time of 2.85 seconds for OAUE. These varying of parameters has led to better performance than the standard parameter settings.

VI. CONCLUSION

In this paper, we have presented the idea of an adaptive ensemble classifier for classifying data streams. The data streams will be processed by the component classifiers in blocks, hence a block-based ensemble classifier. This component classifiers of the ensemble will be updated based on class prediction error. This classifier will be optimized for memory usage. From the results, inference can be drawn that the appropriate chunk size and number of component classifiers are 200 and 7 respectively, that will lead to better performance. The work mainly aimed analyzing ensemble

classification technique based on various parameter settings. Comparison is done based on chunk size and number of component classifiers; accuracy and time required for processing is compared.

REFERENCES:

[1] OUYANG Zhenzheng et al., "Study on the Classification of Data Streams with Concept Drift",Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD),2011.

[2] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, A. Bouchachia(2014), "A Survey on Concept Drift Adaptation ", ACM Computing Surveys, Vol. 46, No. 4, Article 44.

[3] D. Brzezinski and J. Stefanowski(2011), "Accuracy updated ensemble for data streams with concept drift," Proc. 6th HAIS Int. Conf. Hybrid Artificial Inteligent. Syst., II, pp. 155–163.

[4] W. N. Street and Y. Kim(2001), "A streaming ensemble algorithm (SEA) for large-scale classification," in Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 377–382.

[5] Sobhani P. and Beigy H.(2011), "New drift detection method for data streams", Adaptive and intelligent systems, Lecture notes in computer science, Vol. 6943, pp. 88–97.

[6] D Brzezinski, J Stefanowsk(2011), "Mining data streams with concept drift " Poznan University of Technology Faculty of Computing Science and Management Institute of Computing Science.

[7] I Žliobaite (2010), "Adaptive Training Set Formation", Doctoral dissertation Physical sciences, informatics (09P) Vilnius University.

[8] A Bifet(2009), "Adaptive Learning and Mining for Data Streams and Frequent Patterns", Doctoral Thesis.

[9] D Brzezinski, J Stefanowski(2012), "From Block-based Ensembles to Online Learners In Changing Data Streams: If- and How-To", ECML PKDD Workshop on Instant Interactive Data Mining, pp. 60–965.

[10] P. B. Dongre, L. G. Malik(2014), " A Review on Real Time Data Stream Classification and Adapting To Various Concept Drift Scenarios", IEEE International Advance Computing Conference (IACC), pp. 533-537.

[11] D Brzezinski, J Stefanowski (2014),"Reacting to Different Types of Concept Drift:The Accuracy Updated Ensemble Algorithm" IEEE Transactions On Neural Networks And Learning Systems, Vol. 25, pp. 81-94.

[12] D Brzezinski, J Stefanowski(2014), "Combining block-based and online methods in learning ensembles from concept drifting data streams", An International Journal: Information Sciences 265, 50–67.

[13] R. Elwell and R. Polikar(2011), "Incremental learning of concept drift in nonstationary environments," IEEE Trans. Neural Networks, vol. 22, no. 10, pp. 1517–1531.

[14] S S Khan et al., "Comparative Study of Streaming Data Mining Techniques", International Conference on Computing for Sustainable Global Development (INDIACom), 2014 , pp. 209-214.

[15] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, A. Bouchachia(2014), "A Survey on Concept Drift Adaptation ", ACM Computing Surveys, Vol. 46, No. 4, Article 44.