Classification and Adaptive Ensemble Models of Concept Drift

Nandhini. G
II[nd] Year – M.E. CSE
Srinivasan Engineering College
Peramabalur
Tamil Nadu, India

Jayanthi. S
Asst. Professor / CSE
Srinivasan Engineering College
Peramabalur
Tamil Nadu, India

Ayyappan. M
II[nd] Year - M.E. CSE
Srinivasan Engineering College
Peramabalur
Tamil Nadu, India

***Abstract*— Mining data streams with concept drifts using ensemble classifiers is a challenging task to cope with special properties of data streams in the field of data mining. The concept drift methods are used to identify classifiers. The accurate drift detection maintains the high performance. This project deals with to identify the concept drift problem. A new algorithm, Accuracy Updated Ensemble (AUE) is proposed which extends the different classifiers such as accuracy weighted ensemble by using classifiers and updating them according to the new arrival of data. In existing, Accuracy Weighted Ensemble (AWE) is used to train a new classifier on each incoming data block and use that block to evaluate all the existing classifiers in the ensemble. The component in the classifiers are weighted by their expected accuracy on the test data, the ensemble improves classification accuracy over a single classifier. Many Experiments have evolved with several data sets, lagging in processing time and memory aspects of mining AUE is more accurate than AWE which provides best average classification accuracy and less memory consuming than other ensemble approaches.**

Keywords*— *Concept drift, data stream mining, ensemble classifier.

I. INTRODUCTION

Ensembles have attracted many researchers as an approach for improving predictive accuracy. However, most of the research is devoted to static environments where the classification task is fixed and complete data is available for learning classifiers. On the other hand, a new type of problems is becoming more visible, one in which learning algorithms work in dynamic environments with data continuously generated in the form of a stream.

In the Ensemble learning algorithms have more popular then over the last several years because these algorithms, which generate multiple base models using traditional machine learning algorithms and combine them into an ensemble model, have often demonstrated significantly better performance than single models. Boosting and bagging are two of the most popular algorithms because of their good empirical results and theoretical support.

However, most ensemble algorithms operate in batch mode, i.e., they repeatedly read and process the entire training set. Typically, they require at least one pass through the training set for every base model to be included in the ensemble. The base model learning algorithms themselves may require several passes through the training set to create each base model. In situations where data is being generated continuously, storing data for batch learning is impractical, which makes using these ensemble learning algorithms impossible.

These algorithms are also impractical in situations where the training set is large enough that reading and processing it many times would be prohibitively expensive. Processing data streams implies new requirements concerning limited amount of memory, small processing time, and one scan of incoming data. Moreover, the data distributions and definitions of target classes change over time. These changes are categorized into sudden or gradual concept drift depending on appearance of novel classes in a stream and the rate of changing definitions of classes. Concept drifts directly influence algorithm classification abilities as classifiers generated prior to change have been trained on a different class distribution. As the reason of these changes is hidden and not known a priori, the task of learning classifiers becomes very difficult.

A classifier (individual or ensemble), if intended for such non-stationary environments, has to adapt to concept drifts. Several adaptation methods have been proposed including mainly: sliding window approaches, new online algorithms, special detection techniques, and adaptive ensembles. In the area of adaptive ensembles, component classifiers are generated from sequential blocks of training examples.

When a new block arrives, classifiers are evaluated and later updated, removed, or modified according to the result of the evaluation. Accuracy Weighted Ensemble (AWE) is the most popular method. However, defining an appropriate size of the data block can be problematic. Moreover, too many component classifiers can be excluded from the ensemble when they are not accurate enough and also noticed in preliminary experiments that AWE is not as accurate as other online classifiers.

Therefore, decided to propose a new algorithm, called Accuracy Updated Ensemble (AUE), which would improve over AWE on classification accuracy, while still keeping good computational efficiency

II. RELATED WORK

Mining streaming data is one of the recent challenges in data mining. Data streams are characterized by a large amount of data arriving at rapid rate and require efficient processing. Moreover, the data may come from non-stationary sources, where underlying data distribution changes over time. It causes modifications in the target concept definition, which is known as concept drift.

The main types of changes are usually divided into sudden or gradual concept drifts depending on the rate of changes. Classical static classifiers are incapable of adapting to concept drifts, because they were learned on the out-of-date examples. This is the reason why their predictions become less accurate with time. Some methods have already been proposed to deal with the concept drift problem.

They can be divided into two main groups: trigger based and evolving. Trigger-based methods use a change detector to identify the occurrence of a change. If the change is detected, then the online classifier, connected with the detector, is re-trained. One of the most popular detectors is DDM described. On the other hand, evolving methods attempt to update their knowledge without explicit information whether the change occurred. An example of such methods is an adaptive ensemble.

This paper focuses mainly on block-based ensembles, which component classifiers are constructed on blocks (chunks) of training data. In general, a block-based approach operates in a way that when a new block is available, it is used for evaluation of already existing component and for creation of a new classifier. The new component usually replaces the worst one in the ensemble.

III. OUR SYSTEM AND ASSUMPTIONS

Data stream classification is a major challenge to the data mining community. There are two key problems related to stream data classification. First, it is impractical to store and use all the historical data for training, since it would require infinite storage and running time. Second, there may be concept-drift in the data.

The solutions to these two problems are related. If there is a concept drift in the data, need to refine our hypothesis to accommodate the new concept. Thus, most of the old data must be discarded from the training set. Therefore, one of the main issues in mining concept-drifting data streams is to choose the appropriate training instances to learn the evolving concept.

One approach is to select and store the training data that are most consistent with the current concept. Some other approaches update the existing classification model when new data appear, such as the Very Fast Decision Tree (VFDT) approach. Another approach is to use an ensemble of classifiers and update the ensemble every time new data appear.

The ensemble classifier is often more robust at handling unexpected changes and concept drifts. In this project proposed a new stream classifier called Accuracy Updated Ensemble, inspired by an earlier proposed algorithm called Accuracy Weighted Ensemble.

AUE was more accurate than AWE on all data set and still requiring constant processing time and memory. Considering the updating technique of AUE can suspect that in periods of longer distribution stability, when no concept drift occurs, the component classifiers can be trained on more examples and should become more accurate.

However, updating many components with similar examples may reduce their diversity. Therefore, to refer this problem and possible modifications of AUE to ensure additional diversity of ensemble components. Then using mean square error on the most recent data chunk to weight component classifiers.

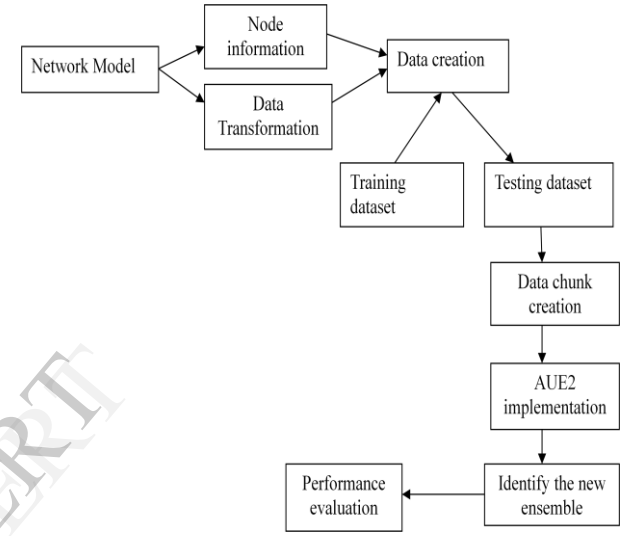

Fig 1 Architecture Diagram

A. *Network Model*

In this module, form the networks with various nodes. The nodes are used to transfer the data from one place to another. It can create the datasets based data transmission and compared with training datasets. The datasets contains the information such as location details, data details, attack details and so on.

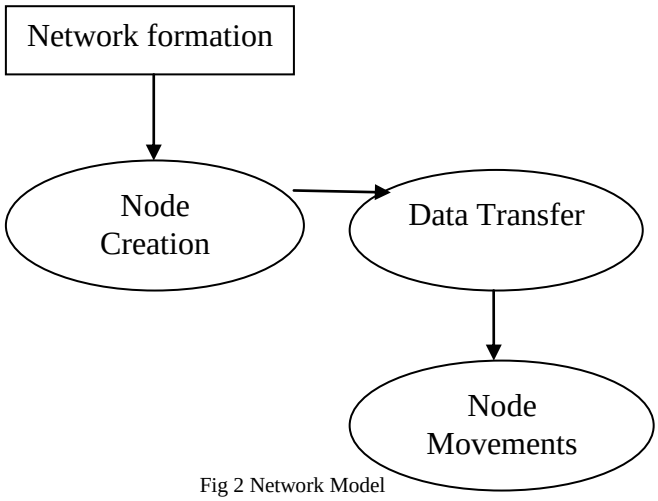

Fig 2 Network Model

B. Datasets Creation

To create two types of datasets such as training dataset and testing datasets. A training set is a set of data used in various areas of information science to discover network relationship. Training set has much the same role and is often used in conjunction with a test set. A test set is a set of data used in network to assess the strength and utility of a predictive relationship.

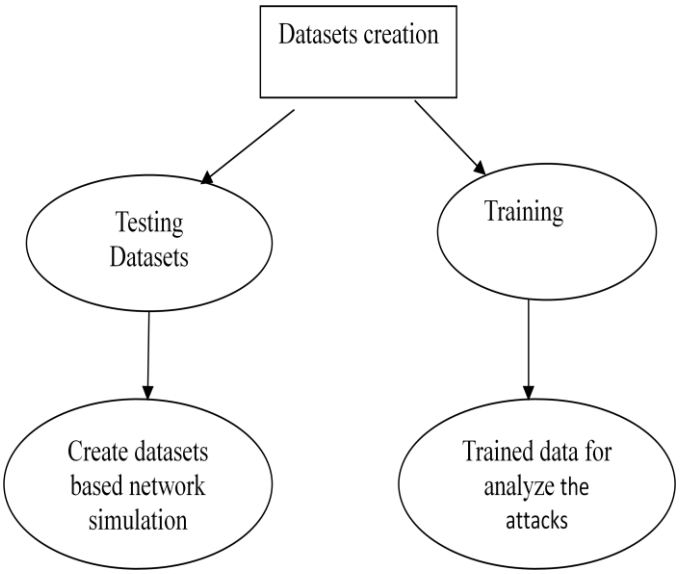

Fig 3 Datasets creation

C. Data Chunk Creation

Generally, data streams can be processed incrementally and they are divided into equally sized blocks (data chunks). The data streams are converted into datasets. These datasets are known as chunks. The chunks are inputted into AUE2 algorithm to classify the chunks which are used to analyze the various types of concept drifts

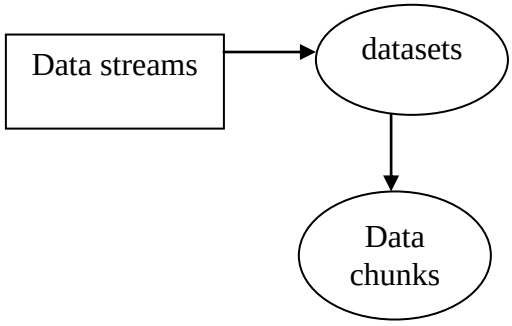

Fig 4 Data chunk creation

D. Accuracy Updated Ensemble (Aue2)

The Accuracy Updated Ensemble also differs from other data stream ensemble approaches. Ensemble members of AUE2 are weighted and can be removed. AUE2 can be considered as a hybrid approach it can react to sudden drifts and it can gradually evolve with slow changing concepts. The rapid adaptation after sudden drifts is achieved by weighting classifiers according to their prediction error and giving the highest possible weight to the newest classifier. AUE2 should protect the classifier from drastic accuracy losses in the presence of random blips, as a single "outlier" component can be over voted when the target concept stabilizes.

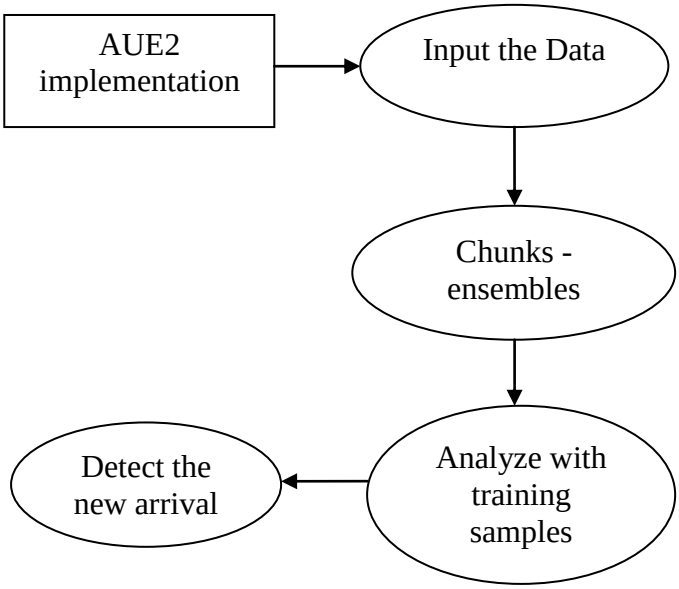

Fig 5 Accuracy updated ensemble

IV. SYSTEM PRELIMINARIES

Algorithm 1 Accuracy updated Ensemble

Input: S: data stream of examples

K: number of ensemble members

Output: *ε*: ensemble of k online classifiers with updated weights

1:C=Φ; //*C*: ← set of stored classifiers

2: for all data chunks *xi* ε S do

3: train classifiers *C'* on *x*i;

4: compute error MSE of *C'* via cross validation on *x*i;

5: derive weight *w*' for *C'* using(3);

6: for all classifiers *C*i ε *C do*

7: apply *C*i on *x*i to derive MSEi;

8: compute weight *w*i based on (3);

9: ε ← k of the top weighted classifiers in *C* U {*C'*};

10: *C* ← *C* U {*C'*};

11: for all classifiers *Ce* ε do

12: if *We> 1/MSEr and Ce* ⫞ *C'* then update classifiers *Ce with x*i

The algorithm is designed to perform well on cost-sensitive data, the MSEr threshold in Equation 2 cuts-of "risky" classifiers. In rapidly changing environments with sudden concept drifts this threshold can "mute" all ensemble members causing no class to be predicted. To avoid this, in AUE propose a simpler weighting function:

$$w_i = \frac{1}{MSE_i + \epsilon}$$

To update component classifiers according to the current distribution, while still keeping their diversity. To achieve this, update only selected classifiers. First of all, consider only current ensemble members - the k top weighted classifiers. Then use MSE_i as a threshold for allowing online updating of only "accurate enough" classifiers. Therefore, inaccurate classifiers can enter the ensemble, but will not be updated.

V. EXPERIMENTAL EVALUATION

To evaluate the performance with real time datasets. And analyze the AUE2 algorithm with existing approaches. The AUE not only selects classifiers, but also updates them according to the current distribution. AUE differs from AWE in the definition of the weight function, the use of online base classifiers, and updating components with incoming examples.

VI. CONCLUSION

Concept drift constitutes a challenging problem for the machine learning and data mining community .Concept drift describes changes in statistical properties of incoming data instances for online Classification over time. It decrease classifiers accuracy as time passes (non stationary data stream). By monitoring online classifier's performance measures over time, it would like to detect concept drift. Change detectors will monitor real-time measures and send alarms as soon as drifts are detected. A good test reduces detection delay (i.e., time needed to detect the change) and minimizes both false positive and negative detection numbers. Multiple or hierarchical detectors are sometimes applied to deal with different types of concept drifts and reduce false positive number. Monitoring the Measures are usually classifier's performance indicators or data properties in order to indicate the change point in time. Concluded to propose a new algorithm, called Accuracy Updated Ensemble (AUE), which would improve over AWE on classification accuracy, while still keeping good computational efficiency.

REFERENCES

[1]. J. Gama, Knowledge Discovery from Data Streams, 1st ed. Chapman & Hall/CRC, 2010.
[2]. L. I. Kuncheva, "Classifier ensembles for detecting concept change in streaming data: Overview and perspectives," in Proc. 2nd Workshop SUEMA 2008 (ECAI 2008), 2008, pp. 5–10.
[3]. W. N. Street and Y. Kim, "A streaming ensemble algorithm (SEA) for large-scale classification," in Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2001, pp. 377–382.
[4]. Y. Cao et al., "SOMKE: Kernel density estimation over data streams by sequences of self-organizing maps," IEEE Trans. Neural Netw. Learn. Syst., vol. 23, no. 8, pp. 1254–1268, Aug. 2012.
[5]. Zliobaite, "Adaptive training set formation," Ph.D. dissertation, Vilnius University, 2010.
[6]. P. Domingos and G. Hulten, "Mining high-speed data streams," in Proc. 6th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2000,pp. 71–80.
[7]. L. I. Kuncheva, "Classifier ensembles for changing environments," in Proc. 5th Int. Workshop Multiple Classifier Syst., 2004, pp. 1–15.
[8]. Combining Pattern Classifiers: Methods and Algorithms. Hoboken,NJ: Wiley-Interscience, Jul. 2004.
[9]. N. C. Oza, "Online ensemble learning," Ph.D. dissertation, The University of California, Berkeley, CA, Sep 2001.
[10]. H. Wang et al., "Mining concept-drifting data streams using ensemble classifiers," in Proc. 9th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2003, pp. 226–235.