e-ISSN: 2349-9745
p-ISSN: 2393-8161

Scientific Journal Impact Factor (SJIF): 1.711

International Journal of Modern Trends in Engineering and Research

www.ijmter.com

Mining Data Streams Using Accuracy Updated Ensemble Classifiers

Jones Merlin.E[1], Jayanthi.S[2], Ramya Devi.R[3], Sudha.C[4]

[1]*Department of ME-CSE, Srinivasan Engineering College*
[2]*Assistant Professor/CSE, Srinivasan Engineering College*
[3,4]*Department of ME-CSE, Srinivasan Engineering College,*

Abstract— Information stream mining garnered much attention owing to its manifestation in a extensive variety of assertions, such as sensor networks, banking, and telecommunication. One of the most vital tasks in knowledge from information streams is answering to idea implication, unexpected changes of the stream's core data distribution. Numerous classification procedures that manage with idea implication have been put forward, however, most of them concentrate in one type of change. Focus on the topic of adaptive ensembles that generate component classifiers sequentially from fixed-size blocks of training examples called data chunks. Compared to AUE1, forward a new weighting and updating mechanism as well as modify many other construction details to reduce computational costs and improve classification accuracy.

Keywords— *Concept Drift, Ensemble Classifiers, Adaptive Ensemble.*

I. INTRODUCTION

Data Mining has great potential for exploring the meaningful and hidden patterns in the data sets at the medical domain. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques is a remedy to this situation. Data mining functions include clustering, classification, prediction, and associations. One of the most important data mining applications is that of mining association rules. Association rules, introduced in 1993, are used to identify relationships among a set of items in databases. These relationships are not based on inherit properties of the data themselves, but rather based on co-occurrence of the data items. Emphasis in this research work is analysis of medical data. Medical profiles such as patient name, age, sex, disease name, address, time, date, etc., can be used to mining the frequent disease of patients in different geographical area at given time period.

Focus on the topic of adaptive ensembles that generate component classifiers sequentially from fixed-size blocks of training examples called data chunks. In such ensembles, when a new block arrives, existing component classifiers are evaluated and their combination weights are updated. 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. Moreover, standard, static learning algorithms, such as C4.5, are applied to generate classifiers from a given block. The SEA algorithm was the first of such adaptive ensembles and was soon followed by the Accuracy Weighted Ensemble which is currently the most representative method of this type. However, depending on the occurrence of concept drifts within the fixed-size data chunk, the mentioned block-based ensembles may not react sufficiently to changes. In particular, for sudden drifts, they may react too slowly as classifiers generated from outdated blocks still remain valid components even though they have inaccurate weights. This situation is connected with the problem of proper tuning of the data block size. Using small size chunks can partly help in reacting to sudden changes, but doing so will damage the performance of the ensemble in periods of stability and increase computational costs. New hybrid algorithm, called accuracy updated ensemble, which should react to different types of concept drift much better than related adaptive ensembles. our goal is to retain the simple schema of learning component classifiers and weighting their predictions, characteristic for block-based algorithms, while adding elements known from online methods. our main novel contribution is the introduction of incremental updating of component classifiers, which should improve the ensemble's reactions to different types of concept drift as well as reduce the impact of the chunk size.

II. RELATED WORK

First process is the input selection. Select the input dataset with different data streams (attributes). Then input dataset has been loaded into the database. After the dataset has been loaded into the database, based on the class labels, classify the data streams. In this classification based on class attribute in the dataset. After the data streams have been classified we have to apply the AUE 2 classifier algorithm. Based on this algorithm classify the chucks and cluster the data. Then, By using that algorithm predict the classification results and produce the accuracy results. This is done by both incremental value of data and decremented value of data.

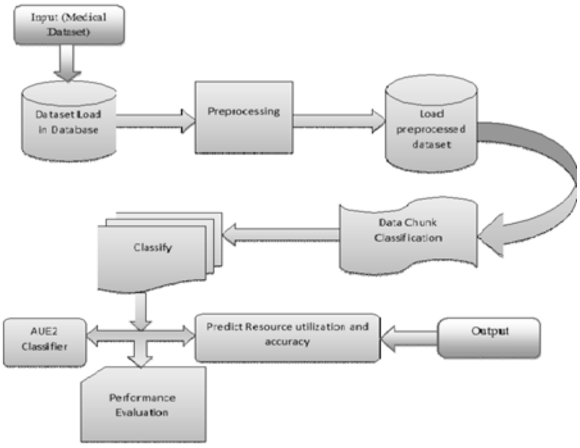

Figure 1 system architecture

The ideal classification scenario is to detect the changes when they come, and retrain the classifier automatically to suit the new distributions. Most methods for novelty detection rely on some form of modeling of the probability distributions of the data and monitoring the likelihood of the new-coming observation. Use direct detection of changes in data. After a drift is detecting older data are removed and a classifier is updated. However, the most widely used such direct approaches (called triggers) are based on observing decrease in the classification accuracy, which requires access to labeled stream of examples.

Generally, data streams can be processed either incrementally by single examples set or they are divided into equally sized blocks (data chunks) b1, b2,...,bn and the evaluation or updating of classifiers is performed after processing all examples from a block. That second perspective will be further discussed in our paper. each training example is generated by a source s j with a stationary distribution pj . If all the data in the stream are generated by the same distribution, we say that the concepts represented in incoming data are stable, otherwise, concept drift occurs. A sudden (abrupt) drift occurs when at a moment in time t the source distribution in st is suddenly replaced by a different distribution in st +1. Abrupt drifts directly deteriorate the classification abilities of a classifier, as a once generated classifier has been trained on a different class distribution. Gradual drifts are not so radical and they are connected with a slower rate of changes.

The mining concept-drifting data streams using weighted ensemble classifiers [1],[5]. An ensemble of classification models, such as C4.5, RIPPER, naive Bayesian, etc., from sequential chunks of the data stream. The classifiers in the ensemble are judiciously weighted based on their expected classification accuracy on the test data under the time-evolving environment. Thus, the ensemble approach improves both the efficiency in learning the model and the accuracy in performing classification. The empirical study shows that the proposed methods have substantial advantage over single-classifier approaches in prediction accuracy, and the ensemble framework is effective for a variety of classification models.

The main characteristics of data streams and discussed different types of changes that occur in streaming data. It focused on non-random class definition changes called concept drift[6]. The existing single classifier and ensemble approaches to mining data streams with concept drift. The analysis led to the development of a new algorithm called Accuracy Diversified Ensemble, which is based on our critique of the earlier developed Accuracy Weighted Ensemble. The MOA’s source code is not documented, it makes code re-usage more difficult. The data chunk size and the performance of bagging in the Accuracy not Diversified Ensemble. Furthermore, since bagging did not prove to provide additional accuracy in this method. It does not compare a larger number of algorithms to take a broader look at the performance of the stream mining techniques[9].

The problem of constructing accurate block-based ensemble classifiers from time evolving data streams.AWE is the best-known representative of these ensembles. A algorithm called Accuracy Updated Ensemble (AUE)[5], which extends AWE by using online component classifiers and updating them according to the current distribution.

Additional modifications of weighting functions solve problems with undesired classifier excluding seen in AWE[6].Experiments with several evolving data sets show that, while still requiring constant processing time and memory, AUE is more accurate than AWE. It still requiring constant processing time and memory. Updating many components may reduce their diversity. DDD maintains ensembles with different diversity levels and is able to attain better accuracy than other approaches[14]. Furthermore, it is very robust, outperforming other drift handling approaches in terms of accuracy when there are false positive drift detections. In all the experimental comparisons carried out, DDD always performed at least as well as other drift handling approaches under various conditions, with very few exceptions.The approaches are not yet prepared to take advantage of recurrent or predictable drifts.

III. SCOPE OF THE PAPER

Present and evaluate a block-based stream ensemble classifier, called AUE2, designed to react to different types of concept drift. The main contribution of the algorithm is the combination of an AWE inspired ensemble weighting mechanism with incremental training of component classifiers.

The algorithm was also optimized for memory usage by restricting ensemble size and incorporating a simple inner- component pruning mechanism. If recurrent changes are very frequent, a buffer can improve accuracy but in other cases it only increases memory requirements and algorithm processing time. The aim of the research is to put forward a data stream classifier that will react equally well to different types of drift.

Propose to combine accuracy-based weighting mechanisms known from block-based ensembles with the incremental nature of Hoeffeding Trees, in an algorithm called the Accuracy

Updated Ensemble (AUE2).The Accuracy Updated Ensemble maintains a weighted pool of component classifiers and predicts the class of incoming examples by aggregating the predictions of components using a weighted voting rule. After each data chunk of examples, a new classifier is created, which substitutes the weakest performing ensemble member.

The performance of each component classifier is evaluated by estimating its expected prediction error on the examples from the most recent data chunk. After substituting the poorest performing component, the remaining ensemble members are updated, i.e., incrementally trained, and their weights are adjusted according to their accuracy. We propose to use Hoeffding Trees as component classifiers, but the presented algorithm can be considered as a general method and in principle, one could use other online learning algorithms as base learners.

Most data stream classification algorithms tend to specialize in one type of drift. Some classifiers are more accurate on datasets with sudden drifts while others perform better in the presence of gradual changes. The aim of our research is to put forward a data stream classifier that will react equally well to different types of drift. To achieve this goal, we propose to combine accuracy-based weighting mechanisms known from block-based ensembles with the incremental nature of Hoeffding Trees, in an algorithm called the Accuracy Updated Ensemble (AUE2).

The Accuracy Updated Ensemble maintains a weighted pool of component classifiers and predicts the class of incoming examples by aggregating the predictions of components using a weighted voting rule. After each data chunk of examples, a new classifier is created, which substitutes the weakest performing ensemble member. The performance of each component classifier is evaluated by estimating its expected Prediction error on the examples from the most recent data chunk. After substituting the poorest performing component, the remaining ensemble members are updated, i.e., incrementally trained, and their weights are adjusted according to their Accuracy. We propose to use Hoeffding Trees as component classifiers, but the presented algorithm can be considered as a general method, and in principle, one could use other online learning algorithms as base learners.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. On the other hand, because components are updated after every chunk, they can react to gradual drifts. Additionally, the modular structure of 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.

Use sliding windows whose size, instead of being fixed a priori, is recomputed online according to the rate of change observed from the data in the window itself The window will grow automatically when the data is stationary, for greater accuracy, and will shrink automatically when change is taking place, to discard stale data. This delivers the user or programmer from having to guess a time-scale for change.It is a new approach for dealing with distribution change and concept drift when learning from data sequences that may vary with time. Proposed a time and memory efficient version of our algorithm, ADWIN2, that checks O (log W) cut points, use s O (log W) memory words , and whose processing time per example is O (log 2 W) (worst-case) and O (log W) (amortize d)[4].This approach using synthetic , time changing data streams and are dataset, showed the improvements of our estimator over fixed size windows and one of the most recently propose d variable length window strategies.

IV. CONCLSION

Main novel contribution is the introduction of incremental updating of component classifiers, which should improve the ensemble’s reactions to different types of concept drift as well as reduce the impact of the chunk size. Propose a new hybrid algorithm, called Accuracy Updated Ensemble, which should react to different types of concept drift much better than related adaptive ensembles.

Our goal is to retain the simple schema of learning component classifiers and weighting their predictions, characteristic for block-based algorithms.

Above all, the experimental study demonstrated that AUE2 can offer high classification accuracy in environments with different types of drift as well as in static environments. AUE2 provided best average classification accuracy out of all the tested algorithms, while proving less memory consuming than other ensemble approaches, such as Leveraging Bagging or Hoeffding Option Trees. As future work, we plan to investigate the possibility of adapting the proposed algorithm to work in a truly incremental fashion in partially labeled streams.

REFERENCES

[1] M. Baena-García, J. D. Campo-Ávila, R. Fidalgo, A. Bifet, R. Gavaldà, and R. Morales-Bueno,(2006), “Early drift detection method,” in Proc. 4th Int. WorkshopKnowl. Discovery Data Streams, pp. 1–10.

[2] A. Bifet, G. Holmes, and B. Pfahringer,(2010), “Leveraging bagging for evolving data streams,” in Proc. Eur. Conf. Mach. Learn./PKDD, I, pp. 135–150.

[3] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, (May 2010),“MOA: Massive online analysis,” J. Mach. Learn. Res., vol. 11, no. 5, pp. 1601–1604.

[4] A. Bifet and R. Gavaldà,(2007), “Learning from time-changing data with adaptive windowing,” in Proc. 7th SIAM Int. Conf. Data Mining, pp. 443–448.

[5] D. Brzezinski and J. Stefanowski, (2011),“Accuracy updated ensemble for data streams with concept drift,” in Proc. 6th HAIS Int. Conf. Hybrid Artif. Intell. Syst., II, pp. 155–163.

[6] D. Brzezinski,(2010),“Mining data streams with concept drift,” M.S. thesis, Inst. Comput. Sci., Poznan Univ. Technology, Poznan, Poland.

[7] Y. Cao, H. He, and H. Man,(Aug.2012), “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.

[8] E. Cohen and M. J. Strauss,(Apr. 2006), “Maintaining time-decaying stream aggre-gates,” J. Algorithms, vol. 59, no. 1, pp. 19–36.[9] P. Domingos and G. Hulten, (2000),“Mining high-speed data streams,” in Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 71–80.

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

[11] W. Fan,(2004), “Systematic data selection to mine concept-drifting data streams,” in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 128–137.

[12] W. Fan, Y. A. Huang, H. Wang, and P. S. Yu, (Apr. 2004), “Active mining of data streams,” in Proc. 4th SIAM Int. Conf. Data Mining, pp. 457–461.

[13] A. Frank and A. Asuncion. (2010), UCI Machine Learning Repository [Online]. Available: http://archive.ics.uci.edu/ml

[14] J. Gama,(2010) Knowledge Discovery from Data Streams, 1st ed. London, U.K.: Chapman & Hall,

[15] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, (2004), “Learning with drift detection,” in Proc. 17th SBIA Brazilian Symp. Artif.Intell., pp. 286–295.