International Journal for Research in Applied Science & Engineering Technology (IJRASET)

An Assembly Learning Approaches For Assorted Types of Concept Drift

G.Bakkiyaraj[1] , P.Ayesha Barvin[2]
[1]P.G Student, Department of CSE, Roever Engineering College, Perambalur, India
[2]Asst.Prof, Department of CSE, Dhanalakshmi Srinivasan College of Engineering, Perambalur, India

Abstract— To Detecting and monitoring changes during the learning process are important areas of research in many industrial applications. The challenging issue is how to diagnose and analyze these changes so that the accuracy of the learning model can be preserved. Recently, ensemble classifiers have achieved good results when dealing with concept drifts. The Information flow mining garners much attention owing to its manifestation in an 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. Recently, concept drift has become an important issue while analyzing non-stationary distribution data in data mining. For example, data streams carry a characteristic that data vary by time, and there is probably concept drift in this type of data.
Keywords: Concept Drift, Ensemble Approaches, 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.

Changes of target concepts are categorized into sudden, gradual, or recurring drifts. A good classifier should be able to learn incrementally and adapt to such changes. Standard static classifiers are not capable of fulfilling these conditions, although the issue of incremental learning has already been studied. For example, neural networks or Bayesian classifiers can naturally incorporate incoming examples, while other approaches, such as decision trees, have been adapted to work online (see VFDT [6]). However, simple incremental learning is not sufficient for dealing with concept drifts as forgetting outdated data and quick adaptation to most recent states are a necessity in no stationary environments [1].

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.

The 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
Results and produce the accuracy results. This is done by both incremental value of data and decremented value of data.

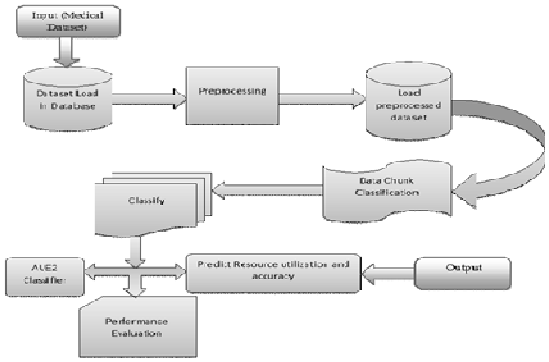

Figure 1 system architecture

A. *Basic Concepts and Notation*

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.

B. *Classifiers for Data Streams With Concept Drift*

In static classification problems, a set of learning examples contains pairs {x, y}, where x is a vector of attribute values and y is a class label (y ∈ {k1, kl}). Classes' k j is known a priori. The learning algorithm constructs a classifier, which outputs a class prediction for a given example. In incremental learning, examples arrive continuously inthe form of a data stream s. A learning algorithm is presented with a sequence of labeled examples st= {xt, yt} fort = 1, 2, t. At each time step t, a learner can analyze historical labeled training examples (s1, s2, st) and an incoming instance st +1, which is treated as a testing example. The classifier predicts its class label ˆyt +1. It is assumed that after some time, the true class label yt +1 is provided.1 having both yt +1 and ˆyt +1, the learning algorithm can update its hypothesis about a classifier if necessary. Then, example st +1with its class yt +1 becomes a part of the training data and the process is repeated when the next instance is observed. Generally, data streams can be processed either incrementally by single examples st (as described above) 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.

III. ACCURACY UPDATED ENSEMBLE

In 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 exam-ples, a new classifier is created, which substitutes the weak-est 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., incremen-tally 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.

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.

IV. EXPERIMENTAL EVALUATION

To 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.

A. Datasets

Most of the common benchmarks for machine learning algorithms, e.g., gathered in the UCI repository, contain too few examples to be concerned suitable for evaluating data stream classification methods, especially in terms of algorithm efficiency. Furthermore, datasets used to test algorithms designed for static environments usually do not contain any type of concept drift. In terms of real-world data, there is still a shortage of suitable and publicly available benchmark dataset. 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.

B. Experimental Setup

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 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 general method and in principle, one could use other online learning algorithms as base learners.

C. Analysis of the Components of the Proposed Algorithm

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.

D. Comparative Study of Classifiers

After establishing the properties of AUE2, a set of experiments was conducted to compare the newly proposed algorithm against 11 classifiers: the Hoeffding Option Tree (HOT), ACE, the previous version of the Accuracy Updated Ensemble (AUE1), the Accuracy Weighted Ensemble (AWE), Leveraging Bagging (Lev), Online Bagging (Oza), Dynamic Weighted Majority (DWM), Learn++.NSE (NSE), Drift Detection Method with a Hoeffding Tree (DDM), a single Hoeffding Tree with a static window (Win), and the Naive Bayes algorithm (NB). We chose AWE and AUE1 as those are the classifiers we tried to improve upon. HOT and ACE were selected as they can be considered as hybrid ensemble algorithms combining elements of incremental learning. Oza, Lev, NSE, and DWM were chosen as strong representatives of online ensembles. The DDM algorithm and the windowed Hoeffding Tree were chosen as representatives of single classifiers. Additionally, the Naive Bayes algorithm is added to the comparison as a reference for using an algorithm without any drift reaction mechanism. All the studied algorithms were evaluated in terms of classification accuracy, memory usage, chunk training time, and testing time.

V. CONCLUSION

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

VI. FUTURE ENCHANCEMENTS

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 no stationary 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.
[16] G. Hulten, L. Spencer, and P. Domingos, (2001), "Mining time-changing data streams," in Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 97–106.
[17] A. Kehagias and V. Petridis, (Jan. 1997) "Predictive modular neural networks for time series classification," Neural Netw., vol. 10, no. 1, pp. 31–49.
[18] R. Kirkby, "Improving Hoeffding trees, (2007), Ph.D. dissertation, Dept. Com-put. Sci., Univ. Waikato, Hamilton, New Zealand.
[19] M. Kmieciak and J. Stefanowski,(Mar.2011) "Handling sudden concept drift in Enron message data streams," Control Cybern., vol. 40, no. 3, pp. 667–695.
[20] J. Z. Kolter and M. A. Maloof, (Dec.2007) "Dynamic weighted majority: An ensemble method for drifting concepts," J. Mach. Learn. Res., vol.8, pp. 2755–2790.
[21] L. I. Kuncheva, (2008), "Classifier ensembles for detecting concept change in streaming data: Overview and perspectives," in Proc. 2nd Workshop SUEMA, pp. 5–10.
[22] L. I. Kuncheva, (2004), Combining Pattern Classifiers: Methods and Algorithms. New York, USA: Wiley.
[23] M. M. Masud, J. Gao, L. Khan, J. Han, and B. Thuraisingham,(Dec.2008), "A practical approach to classify evolving data streams: Training with limited amount of labeled data," in Proc. 8th IEEE Int. Conf. Data Mining, pp. 929–934.
[24] L. L. Minku, A. P. White, and X. Yao, (May 2010) "The impact of diversity on online ensemble learning in the presence of concept drifts," IEEE Trans. Knowl. Data Eng., vol. 22, no. 5, pp. 730–742.