IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. IV (Sep – Oct. 2014), PP 120-123
www.iosrjournals.org

Review paper on adapting data stream mining concept drift using ensemble classifier approach

Nilima Motghare[1], Arvind Mewada[2]
[1](Computer Science and Engineering, TIT Science, Bhopal / RGPV University, India)
[2](Computer Science and Engineering, TIT Science, Bhopal / RGPV University, India)

Abstract: *Data stream is massive, fast changing and infinite in nature. It is very natural that large amount of unlabeled data and small amount labeled are available in data stream environments. Storing and labeling all data is considered expensive and impractical. The objective is to label small portion of stream data and analyze data online without storing it. Concept drift, concept evolving, stream evolving is also the major challenging problem occurs while working with data stream. Online data stream active learning is needed to tackle these problems. Classification and clustering are two technical areas that are widely used to extract pattern from the large data stream, from that a classification model must endlessly adapt itself to the most recent concept. Hence, this paper gives the overview of various ensemble based classification algorithm techniques in the field of data stream mining.*
Keywords: *Concept drift, data stream mining, classifier ensemble.*

I. Introduction

Stream data mining faces an issue which is lack or limited amount of computational resources. Nowadays, the quantity of data that is created every two days is estimated to be 5 exabytes. Google receives over 2 million search queries, YouTube users upload 48 hours of new video, Facebook users share 684,000 bits of content, Twitter users send more than 100,000 tweets, Email users send more than 204 million messages, Mobile Web receives 217 new users, Apple receives around 47,000 application downloads, Brands receive more than 34,000 Facebook 'likes', Tumblr blog owners publish 27,000 new posts, WordPress users publish close to 350 new blog posts. Storing continuous data stream is a great challenge for storing devices. To generate knowledge from stream data, algorithm with different techniques are needed. Classification, clustering, web mining, graph mining are technical areas working on data stream. Classification and clustering are two technical areas that are widely used to extract pattern from the large data stream. Classification is supervised learning method as in this class labels are already defined. The classifier can be grouped into two main classes: single classifier approaches and ensemble classifier approach.

The paper is organized as, Ensemble classifier approach discusses in section II. Section III discusses the practical situation pose fundamental issue to be addressed by any continuous mining attempt with literature survey discussion in Section IV and section V conclusions.

II. Ensemble Classifier Approach

Single classifier approaches updates a single model with each new training instance. In contrast, ensemble classifier approaches build each model from a batch of training data using a traditional batch learning technique in such a way that this combination will improve the performance over a single classifier. Ensemble classifier has the advantage that model updates are typically far simpler than in single classifier. Classifiers in the ensemble can simply be removed or replaced. Studies show that to obtain high accuracy, it is important to diversify ensemble members from each other. Components can differ from each other by the data they have been trained on, the attributes they use, or the base learner they have been created. Fig.1 is showing the basic steps of ensemble based classification.

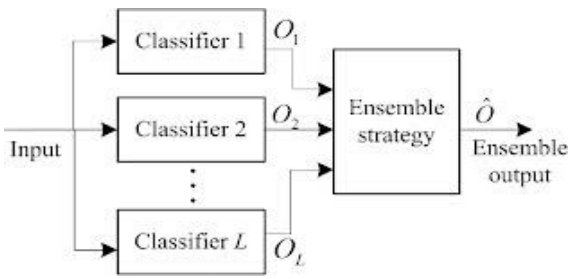

Figure1. Ensemble classifier approaches

III. Practical Situations Pose Fundamental Issues To Be Addressed By Any Continuous Mining Attempt

a) **Adaption Issue:**

In traditional learning tasks, data is stationary and the underlying concept that maps the attributes to class labels is unchanging. With data streams, however, the concept is not static but drifts with time due to changes in the environment. Learning algorithm often need to operate in dynamic environment which are changing unexpectedly. One desirable property of these algorithms is their ability of incorporating new data. Adaptive learning algorithms can be seen as advanced incremental learning algorithms that to adapt to evaluation of the data generating process over time.

b) **Robustness Issue**

Noise can severely impair the quality and speed of learning. It is difficult to distinguish noise from changes caused by concept drift. This problem is encountered in many applications where the source data can be unreliable, and also errors can be injected during data transmission. If an algorithm is too eager to adapt to concept changes, it may overfit noise by mistakenly interpreting it as data from a new concept. If the algorithm is too conservative and slow to adapt, it may overlook important changes.

c) **Performance Issue**

The requirement of on-line processing of data streams and by limited computation and memory resources, continuous data stream mining should conform to the following criteria: (a) Learning should be done very fast, preferably in one pass of the data. (b) Algorithms should use less memory resources, for the storage of either the intermediate results or the final decision models.

d) **Concept drift**

Concept drift primarily refer to an online supervise learning scenario when the relation between the input data and the target variable changes over time. It degrades the accuracy of classification system up to a point that the expected quality. The goal of predictions system is to predict class or the target value may change over time. Concept drifts can be grouped into two main families: abrupt and gradual.

e) **Concept drifts adaptation process**

Concept drift refers to the learning problem where the target concept to be predicted changes over time in some unforeseen behaviors. It is commonly found in many dynamic environments, such as data streams, P2P systems, etc. Real-world examples include network intrusion detection, spam detection, fraud detection, epidemiological, and climate or demographic data, etc.

III. Literature Survey

Kapil Wankhade et. al [1] proposed novel approach for data stream classification method. It uses weighted majority and adaptive sliding window approaches to handle noisy and concept drifting data streams for improving performance of classification in terms of accuracy. Weighted majority creates experts, remove them and updates their weights for accurate prediction and to recover from concept drift. Window technique is used to monitor the error rate of current model and helps to overcome the problem of storing and processing fast data.

Mayank Pal Singh [2] proposed a supervised learning model to quantify the concept drift in the network traffic. The ROC curves derived from classification of Naive Bayes classifiers identify the concept drift of the current class from the base class. The proposed model uses adaptive learning strategies with fixed training window to constantly evolve the model.

-Separating the data stream into chunk is popular solution for mining the data stream. Xingquan Zhu et. al [3] proposed algorithm that dynamically selects a single "best" classifier to classify each test instance at run time. Attribute-oriented Dynamic Classifier Statically partition the evaluation set into subsets by using the attribute values of the instances, Evaluate the classification accuracy of each base classifier on all subsets and for a test instance, use its attribute values to select the corresponding subsets and select the base classifier that has the highest classification accuracy from the selected subsets as the "best" classifier to classify the test instance. Attribute-oriented Dynamic Classifier outperforms most classifier combination (CC) technique or classifier selection (CS) technique in many situations and appears to be a good solution for mining real-world data.

-Gansen Zhao et. al [4] present SRASTREAM framework. It is an on-demand triggered clustering framework which will perform the clustering in due time when the concept drifting happens, It consists of four core components: Quick Computation Module (QCM), Evolving Detecting Module (EDM), Clustering Module (CM) and Resource Monitoring Module (RMM). CM is a core module in this framework will be triggered to

take the related clustering actions in time by identifying the concept drifting in an online and quick way. Hence the framework achieves improved clustering speed by losing the acceptable accuracy.

G. R. Marrs et. al [5] present two new algorithms that use a time of classification protocol for handling latency in data stream and improving classification in affected domains, that is CDTC versions 1 and 2.

Muhammad Shaheryar et. al [6] proposed an novel ensemble classifier Rot-SiLA which is developed by combining Rotation Forest algorithm and SiLA. Random Forest includes decision trees based on feature extraction where Principal Component Analysis (PCA) is used to rotate the feature subsets.

Zohre Karimi et. al [7] proposed a new classification algorithm for the classification of batch data called harmony-based classifier and then give its incremental version for classification of data streams called incremental harmony-based classifier. Finally, improve it to reduce its computational overhead in absence of drifts and increase its robustness in presence of noise. This improved version is called improved incremental harmony-based classifier.

Peng Zhang et. al [8] state that to classify each stream record in a timely manner the Ensemble-tree indexing for sublinear convert ensemble models into spatial databases and applies spatial indexing techniques to achieve fast prediction.

Amineh Amini et. al [9] present clustering base algorithm and more focus on Density-based method because it's ability to discover arbitrary shape clusters and noise detection. The algorithms were divided into two basic categorizations, micro-cluster and grid algorithms for easier investigation of the density-based clustering algorithms.

Chao-Wei Li et. al [10] proposed SA-Miner algorithm that discovers frequent itemsets through support approximation. SA-Miner learns a concept by constructing models for the support relationships that describe the concept. The proposed method not only performs efficiently in terms of time and memory but also preserves mining accuracy well on concept-drifting data streams.

Chien-I Lee et. al [11] proposed Concept Drift Rule Mining Tree called CDR-Tree to accurately mine rule of concept drift which was ignored in past. CDR-Tree initially integrating the new and old instances from different time point into pairs and build the CDR-Tree. During the building step, information gain is used as the criterion to select the best splitting attribute in each node. Higher the concept drift level would make the CDR-Tree more complex and less accurate so discretization algorithm used to reduce the complexity of CDR-Tree and able to accurately predict drifting instances.

Stephen H. Bach et. al [14] proposed a paired learner that pairs a stable online learner with a reactive one to cope with concept drift. A stable learner predicts based on all of its experience, whereas a reactive learner predicts based on its experience over a short, recent window of time.

In the case of unknown class labels of future data, it is expected that averaging probability ensemble (AP) ensemble classifier has higher classification accuracy although incapable of much prior knowledge of future data, but the application is limited to some extent because of neglecting the concept evolution caused by noise. Nevertheless, WE ensemble classifier is unable to solve the continuous concept evolution, but it has good capability for anti-noisy. In order to effectively solve these two problems arising from data stream mining, Zhenzheng Ouyang et. al. [12] proposed a novel ensemble classifier framework (WEAP-I) for mining concept-drifting data streams with noise based both on weighted ensemble (WE) and the averaging probability ensemble (AE) under the "Learnable Assumption". The method, called WEAP-I, which trains a weighted ensemble classifier on the most n data chunks and trains an averaging ensemble classifier on the most recent data chunk.

Yang Zhang et. al [13] firstly consider the problem of the one-class classification on text stream with respect to concept drift where a large volume of documents arrives at a high speed and with change of user interests and data distribution. An stacking style ensemble-based classifier designed to dealt with the problems of concept drift, small number of training examples, no negative examples, noisy data, and limited memory space on streaming data classification.

IV. Conclusion

In this paper, the state of the art on ensemble methodologies to deal with nonstationary data has been reviewed. Tracking concept drift is important for many applications. Ensemble classifier creates and removes base algorithm in response to change in performance, which makes it well suited for problem of concept drift. It achieved higher predictive accuracies and converged to those accuracies more quickly. We anticipate that these investigation lead to general, robust and scalable ensemble methods for tracking concept drift.

References

[1] Kapil Wankhade, Snehlata Dongre, Ravindra Thool, "New Evolving Ensemble Classifier for Handling Concept Drifting Data Streams", 2nd IEEE International Conference on Parallel, Distributed and Grid Computing, pp. 657-662, 2012

[2] Mayank Pal Singh, "Quantifying Concept Drifting in Network Traffic using ROC Curves from Naive Bayes Classifiers", Nirma University International Conference on Engineering (NUiCONE), 2013

[3] Xingquan Zhu, Xindong Wu, and Ying Yang, "Dynamic Classifier Selection for Effective Mining from Noisy Data Streams", Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04)

[4] Gansen Zhao, Ziliu Li, Fujiao Liu and Yong Tang, "A Concept Drifting based Clustering Framework for Data Streams", Fourth International Conference on Emerging Intelligent Data and Web Technologies, 2013.

[5] G. R. Marrs • M. M. Black • R. J. Hickey, "The use of time stamps in handling latency and concept drift in online learning", Springer-Verlag, 2012.

[6] Muhammad Shaheryar, Mehrosh Khalid and Ali Mustafa Qamar, "Rot-SiLA: A Novel Ensemble Classification approach based on Rotation Forest and Similarity Learning using Nearest Neighbor Algorithm", 12th International Conference on Machine Learning and Applications, pp. 46-51, 2013.

[7] Zohre Karimi · Hassan Abolhassani · Hamid Beigy, "A new method of mining data streams using harmony search", Springer Science+Business Media, LLC February 2012.

[8] Peng Zhang, Chuan Zhou, Peng Wang, Byron J. Gao, Xingquan Zhu, Li Guo, "E-Tree: An Efficient Indexing Structure for Ensemble Models on Data Streams", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, JUNE 2011 1.

[9] Amineh Amini, Member, IEEE, Teh Ying Wah, and Hadi Saboohi, Member, ACM, IEEE, "On Density-Based Data Streams Clustering Algorithms: A Survey", JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(1): 116–141 Jan. 2014.

[10] Chao-Wei Li , Kuen-Fang Jea, "An approach of support approximation to discover frequent patterns from concept-drifting data streams based on concept learning", Springer-Verlag London, 2013.

[11] Chien-I Lee, Cheng-Jung Tsai, Jhe-Hao Wu, Wei-Pang Yang "A Decision Tree-Based Approach to Mining the Rules of Concept Drift", Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007) IEEE.

[12] Zhenzheng Ouyang, Min Zhou, Tao Wang, Quanyuan Wu,"Mining Concept-Drifting and Noisy Data Streams using Ensemble Classifiers"International Conference on Artificial Intelligence and Computational Intelligence, 2009.

[13] Yang Zhang, Xue Li, "One-class Classification of Text Streams with Concept Drift" 2008 IEEE International Conference on Data Mining Workshops.

[14] Stephen H. Bach, Marcus A. Maloof, "Paired Learners for Concept Drift", Eighth IEEE International Conference on Data Mining, 2008.

[15] Zhou Tao, Lu Huiling, Liu Lihua,Yong Longquan,Tuo Shouheng, "A new Classification algorithm Based on Ensemble PSO SVM and Clustering analysis", IEEE International Conference on Granular Computing, 2012.