Credit Card Fraud Detection Using Online Boosting with Extremely Fast Decision Tree

Aye Aye Khine
Associate Professor,ICTRC
Yangon,Myanmar
ayekhine71@gmail.com

Hint Wint Khin
Associate Professor,ICTRC
Yangon, Myanmar
snow.hwk@gmail.com

Abstract

Nowadays, data stream mining is a very hot and high attention research field due to the real-time industrial applications from different sources are generating amount of data continuously as the streaming style. To process these growing and large data streams, data stream mining, classification algorithms have been proposed. These algorithms have to deal with high processing time and memory costs, class imbalance, overfitting and concept drift and so on. It is sure that ensembles of classifiers are being effectively used to make improvement in the accuracy of single classifiers in either data mining or data stream mining. Thus, to get higher performance in prediction with largely no increasing memory and time costs, this paper proposes an Online Boosting(OLBoost) Approach, which is firstly use the Extremely Fast Decision Tree (EFDT) as base (weak) learner , in order to ensemble them into a single online strong learner. The experiments of the proposed method were carried out for credit card fraud detection domain with the sample benchmark datasets.

Keywords: *EFDT, Boosting, Credit Card Fraud, Data Stream Mining*

I. INTRODUCTON

Today, the volume of data generated from different sources is increasing exponentially and the analysis methods of such huge volumes give a competitive advantages for the today's business world. Many applications, information systems used in the modern business organizations are worked with the rapidly changes in environments, in which data are collected in the streaming form, i.e. "Data Streams" (DS). Some obvious such examples are network analysis, traffic control, GPS , mobile device tracking, user's click log mining, credit card fraud detection and so on. In contrast with static data mining, processing of streams need more computational requirements for algorithms to incrementally process incoming examples while using limited memory and time. In addition, by reason of the non-stationary characteristics, prediction models for data streams are regularly also vital to adapt to concept drifts [1].

Data stream mining uses two main analytical methods; classification and clustering. Among many data stream classification learning methods, decision tree learning method is ordinarily used, because of its fast and it can be easily understood. Normally, Decision Tree algorithms can be categorized based on learning style: batch and incremental (online). Likewise of batch data mining, there are data stream classification algorithms and data stream clustering algorithms. Typically, data stream classification is a different alternative of incremental learning classifiers that have to meets for massive streams of data with restrictive processing time, limited memory, and concept drift.

Based on our previous work [22], we have learned that the Very Fast Decision Tree (VFDT) algorithm [3] is one of the famous algorithm in classifying data streams , being an online decision tree with the advantage of a statistical property, the Hoeffding Bound (HB). In the last years, the authors [4, 5, and 6] have proposed a series of modifications to increase the predictive performance of the VFDT algorithm. Strict Very Fast Decision Tree (SVFDT) is proposed in the paper [3] so that to address the memory cost restrictions by keeping the predictive performance,.

Moreover, in the paper [7], the authors introduced the implementation of Hoeffding Anytime Tree- "Extremely Fast Decision Tree" to improve upon Hoeffding Tree (HT) by learning faster and ensuring merging to the asymptotic batch decision tree on a stationary distribution. Furthermore, the authors also concluded that EFDT is good to consider of ensemble by building as a forgetting, decay or subtree replacement approaches in order to deal with the concept drift.

In recent year, it is sure that researchers in data stream mining field are moving their focus to ensemble based learning for various application domains. An ensemble learning method; bagging, boosting and stacking, also called a multiple classifier, is a number of so-called base classifiers (weak learner), in which the prediction results are combined to forecast new received instances. Ensembles have been shown to be a proficient way of improving predictive accuracy or/and being a modeler for complex, difficult learning problem by decomposing the easier sub-problems [1].

In the paper [8], the authors have investigated the use of the SVFDT as base learner for ensemble solutions such as OzaBag, OzaBoost, Leveraging Bagging, Online Accuracy Updated Ensemble (OUAE) and Adaptive Radom Forests

(ARF), to reduce memory consumption without harming the predictive performance. The experimental and comparable results are shown using thirteen benchmark datasets.

Thus, inspired by these papers [7 and 8] , we propose the Online Boosting Approach by coupling with the Extremely Fast Decision Tree (EFDT) as base (weak) learner in order to ensemble them into a single online strong learner for credit card fraud detection.

The other remaining sections of this paper are organized as follows. Section 2 discuss how to retrieve information for our survey work of credit card fraud detection. Section 3 describes ensemble approaches for data stream and especially provides online boosting methods while Section 4 discusses the proposed methods in brief. Finally, Section 5 makes the conclusion and identifies the directions for the future work.

II. CREDIT CARD FRAUD DETECTION

Over the past few years, the usage of credit cards is widespread in modern day society. As a result, the fraud of credit card has been kept on growing. However, the major causes of great financial losses is credit card fraud although the careful and responsible usage of credit card provides enormous benefits. In addition, the improvement of new technologies offers additional ways of crime may commit fraud. Actually, Credit card fraud detection is considerably hard, but also prevalent problem to solve [20]. In the literature, many techniques have been proposed to conduct the credit card fraud detection problem. Some researchers use the advanced data mining methods while others use data streams mining in order to detect credit card fraud.

It is certain that credit card frauds can costs customers and banks billions of dollars totally. Thus, the society needs a great and perfect fraud detection system that not only can detect the fraud but also prevent it in advance. Therefore, credit card fraud detection methods need constant innovation and attention.

In addition, some researchers did the parametric comparisons of all the existing data mining technique while some researchers did the literature review of all the work done on fraud detection systems. Many AI and machine learning techniques, data mining, fuzzy logic, sequence alignment, neural network, logistic regression, naïve bayesian, genetic programming, decision tree etc., have developed in detecting for various credit card frauds from massive credit card transactions.

Normally, detection problems are solved with two different styles: batch and online style. In the batch learning style, a model detects occasionally relearnt from scratch (e.g. once a year or month) while the model is updated at once at the time of new data arrival in the online learning style, that credit card transaction in the form of data stream. Therefore, this also subsequently, the detection of fraudulent use by data stream mining is required.

In recent years, many researchers have paid a lot of attention in stream data mining. Thus, streaming data classification and clustering algorithms are discussed with their key features and comparative significance performances. Based upon the recent survey of fraud detection techniques for credit card, this paper mainly focus to present ensemble approach coupling with online(incremental) decision tree technique applied in credit card fraud detection mechanisms in detail.

There are a lot of different credit card datasets with different attributes such as type of fraud, number of fraudulent records and etc., in order to detect fraud. In this paper, we also follow the paper [21] to use dataset obtained from UCSD-FICO competition. It includes 100,000 records of credit card transactions. Each record consists of 20 data fields. The label data as legitimate and fraudulent, already defined by bank is also included. That dataset includes 97% legitimate transactions and the 3% of data records are fraudulent transactions.

III. ENSEMBLES APPROACHES FOR DATA STREAM

For the data stream mining, Ensemble algorithms are most widely used techniques because they can be combined with drift detection algorithms and include dynamic updates, such as selective removal or addition of classifiers. In paper [9], the authors proposed a taxonomy that is focused on ensemble learning for data streams.

Based on their taxonomy, ensemble techniques for data steams can be grouped as follow:

- Combination – Voting Method / Architecture
- Diversity–Inducer / Building Blocks for Ensemble Solution
- Base Learner – Any stable classifier : Batch / Incremental
- Update Dynamics – how learner takes place in the ensemble manner.

From the paper [9], we have learned the various base learners usage in ensemble approaches and also known that decision trees, especially Hoeffding Tree (HT) and its variations are the most common base learner for ensemble learning in a streaming setting. Likewise, in the paper [1] proposed the grouping of ensemble learning approaches for data streams with respect to the different points of view. The most common categorization are the following:

- stationary vs non-stationary stream classifiers
- active vs passive approaches
- chunk based vs on-line learning modes
- distinguishing different techniques for updating component classifiers and aggregating their predictions.

From these systematic survey, though there are serval not only bagging approaches but also boosting approaches for data streaming, this paper only focus on using the boosting ensemble methods as they offer several recognized guarantees and are mostly effective when the base models are simple.

A. Online Boosting Ensembles

In order to increase the predictive performance of Online Decision Tree (ODTs), a lot techniques have been proposed. Among them the three main groups such as the structural modification of the decision tree the additional prediction strategies with the same structure; and the ensembles approaches are famous [2]. From them, the last, ensembles methods have been many proposed not only bagging and boosting with the online style. Due to the focus of this paper is online boosting, the following paragraph is described some related works.

Oza and Russell introduced [10], a parallel boosting strategy by just following AdaBoost with the exception of weight calculation , because AdaBoost can do training with prior knowledge of the number of instances available and it cannot be use with data stream setting. It is called OzaBoost. Due to its simplicity and efficiency, OzaBoost has also got many great achievement in the real-world applications, especially for the computer vision [11],

Many other online boosting algorithms have been addressed for different application needs, such as semi-supervised learning [11], multi-instance learning [12], and feature selection [13].

Online Gradient Boost [11] is an online variant of Gradient Boost, which uses functional gradient descent to decide the optimal example weights and greedily minimizes the loss function of interest. It is denoted as OGBoost. Online Smooth Boosting (OSBoost) is proposed in paper [14].

In 2004, the paper [2], the authors proposed the Adaptive Boosting ensemble algorithm and did practical experiments with the real life data containing 100 k credit card transactions. In 2013, the paper [14] proposed OzaBoost Dynamic, in which the weight calculation is modified and the number of boosted “weak” learners are used in order to improve its performance in terms of memory consumption and also presented the pragmatic results showing the performance of all algorithms using data sets including fifty and sixty million instances.

In the paper [15], the author proposed an algorithm called Online Non-Stationary Boosting (ONSBoost) that are similar with Online Boosting, and also uses a static ensemble size without generating new members. At each time, new examples are presented and it adapts the changes in the data distribution. Using a parallelizing virtual machine, ONSBoost is evaluated with Online Boosting on the STAGGER dataset and the derived three other datasets.

In the paper [8], the author proposed SVFDT as base learner in not only OzaBoost but also bagging and shown the experimental results. Inspirations based on these literature, this paper propose Online Boosting (OLBoost) Approach by coupling with the Extremely Fast Decision Tree (EFDT) as base (weak) learner and explain detail in the next section.

IV. PROPOSED METHOD- ONLINE BOOSTING WITH EFDT

In 2000, the paper [18] presented Hoeffding Tree, namely Very Fast Decision Tree (VFDT) algorithm. For data stream classification, it is one of the most popular and first algorithms, being capable of inducing a decision tree in an online fashion with a statistical property, the Hoeffding Bound (HB). Hoeffding bounds is usually applied in the decision based on the number of instances to be trained to reach a certain level of confidence. After that, many variants or modifications of VFDT are emerged and that summary is shown in the table 1, as we described in our previous work [22].

TABLE I. SUMMARY OF ONLINE DECISION TREE

Algorithm	Description
VFDT (2000)	- incrementally learning from huge data streams in a single pass using constant memory per leaf
CVFDT(2001)	- extends VFDT to combine gradual changes in the basic data distribution for concept-drifting
EFDT (2018)	- improves the process of splitting data by letting revision on the split decisions and get great achievements in terms of performance on many datasets
One-Sided Minimum OSMDT (2018)	- Uses local node statistics to optimize the frequency of evaluation of split decisions.
VFDT-variants (2017, 2018)	-improve and enhance the performance of VFDT -to reduce memory cost

The paper [7] has introduced Hoeffding Anytime Tree (HATT) and denoted as “Extremely Fast Decision Tree (EFDT). Hoeffding Tree usually builds a tree increment manner, making the delay in selecting of a split at a node until not only it is confident but also it has identified the best split, and never revisiting that decision. Instead, HATT usually uses to select and split as soon as it reaches confident level and that split is useful, and then revisits that decision, make the replacement with the split if it subsequently becomes evident that split is better. Though EFDT is a learner not for concept drift, the authors [7] observed that it is highly effectively on benchmark dataset and has some inbuilt acceptance to concept drift.

TABLE II. EXPERIMENTAL PLAN FOR PROPOSED METHOD

Ensemble	Base Learner	Dataset
OzaBoost	EFDT(HATT)	Public Dataset for Credit Card Fraud Detection
OGBoost		
OSBoost		
OSBoost.OCP		
OSBoost.EXP		
ONSBoost		
OzaBoostDync		

V. CONCLUSION AND FUTUREWORK

With the aim to reduce costs of loss caused by credit card frauds in every year, this paper proposed an ensemble approach, online boosting using online decision tree, namely EFDT, as base learner. The experimental results of proposed method will be done as our plan and also plan to make comparative study of our proposed method with online boosting using others base learners by using not only UCSD-FICO credit card dataset but also another credit card dataset from Kaggle website.

ACKNOWLEDGMENT

This work is supported and encouraged me by my beloved supervisor, Dr. Hnit Wint Khin, Associate Professor, Information and Communication Technology Research Center (ICTRC) , Yangon Myanmar.

REFERENCES

[1] B. Krawczyk, L. Minku, J. Gama, and J. Stefanowski, “Ensemble learning for data stream analysis: A survey,” Information Fusion, vol. 37,pp. 1–86, 2017.

[2] F. Chu, C. Zaniolo, “Fast and light boosting for adaptive mining of data streams”, in: Proceedings of the Eight Pacific-Asia Knowledge Discovery and Data Mining Conference (PAKDD’04), 2004, pp. 282–292. Sydney

[3] V. G. T. Costa, A. C. P. L. Carvalho, and S. Barbon Junior, “Strict Very Fast Decision Tree: a memory conservative algorithm for data stream mining,” ArXiv e-prints, May 2018.

[4] Holmes, G., Richard, K., Pfahringer, B., 2005. “Tie-breaking in Hoeffding trees” , in: Proc. of the II Int. Workshop on Knowledge Discovery from Data Streams

[5] Yang, H., Fong, S., 2011b. “OVFDT with functional tree leaf - majority class, naive bayes and adaptive hybrid integrations”, in: Proc. of the III International Conference on Data Mining and Intelligent Information Technology Applications, pp. 65–70.

[6] Yang, H., Fong, S., 2013. “Incremental optimization mechanism for constructing a decision tree in data stream mining”, Mathematical Problems in Engineering 2013, 114–144.

[7] Manapragada, C., Webb, G. I., Salehi, M. (2018) “Extremely Fast Decision Tree”, Proceedings KDD 2018, ACM Press, New York, NY, USA, pp. 1953-1962.

[8] Victor G. Turrisi da Costa, Saulo Martiello Mastelini, Andre C. P. L. F. de Carvalho, and Sylvio Barbon. “Making data stream classification tree-based ensembles lighter” ,in 7th Brazilian Conference on Intelligent Systems, BRACIS 2018, Sao Paulo, Brazil, October 22-25, 2018, pages 480–485, 2018.

[9] H. M. Gomes, J. P. Barddal, F. Enembereck, A. Bieft, “ A Survey on Ensemble Learning for Data Stream Classification”, ACM Computing Surveys, Vol. 50, No. 2, Article 23, Publication date: March 2017

[10] N.C. Oza. , “Online bagging and boosting”, In Systems, Man and Cybernetics, 2005 IEEE International Conference on, Vol. 3. 2340–2345 Vol.3.DOI:http://dx.doi.org/10.1109/ICSMC.2005.157 1498

[11] Grabner, H., Leistner, C., and Bischof, H. “Semisupervised on-line boosting for robust tracking”, in Proceedings of ECCV, pp. 234–247, 2008.

[12] B. Babenko, M. Yang, and, S. J. Belongie , “Visual tracking with online multiple instance learning” , in Proceedings of CVPR, pp. 983–990, 2009b.

[13] Liu, X. and Yu, T., “Gradient feature selection for online boosting”, in Proceedings of ICCV, pp. 1–8, 2007.

[14] Erico N. de Souza and Stan Matwin , “Improvements to Boosting with Data Streams”,

[15] A. Pocock, P. Yiapanis, J. Singer, M. Luj´an, and G. Brown, “Online Non-Stationary Boosting”,

[16] A. Beygelzimer, S. Kale, and H. Luo. “Optimal and adaptive algorithms for online boosting”, in Proceedings of the 32nd International Conference on Machine Learning, 2015.

[17] S.Tse Chen, H. Tien Lin, and C.Jen Lu, “An Online Boosting Algorithm with Theoretical Justifications” , in Proceedings of the 29th International Conference on Machine Learning, 2012

[18] P. Domingos and G. Hulten. 2000. “Mining high-speed data streams”, in proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 71–80

[19] C. Marsh, “Boosting in the Online Setting”.

[20] D. Meenakshi. B, J. B. Gayathri. S , “Credit Card Fraud Detection Using Random Forest” , Volume:06, Issue:03, March 2019,p-ISSN:2395-0072

[21] M. Zareapoora, P. Shamsolmoalia, “Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier”, International Conference on

Intelligent Computing, Communication & Convergence (ICCC-2014)

[22] A. Aye Khine, H. Wint Khine,"A Survey of Decision Tree for Data Streams to Detect Credit Card Fraud", PROMAC-2019