Learning from Concept Drifting Data Streams with Unlabeled Data

Peipei Li(~), Xindong Wu(+,~), and Xuegang Hu(~)

(~) School of Computer Science and Information Engineering, Hefei University of Technology, 230009, Anhui, China
(+) Department of Computer Science, University of Vermont, Burlington VT 05405, USA
peipeili.hfut@gmail.com, xwu@cs.uvm.edu, jsjxhxg@hfut.edu.cn

Abstract

Contrary to the previous beliefs that all arrived streaming data are labeled and the class labels are immediately available, we propose a Semi-supervised classification algorithm for data streams with concept drifts and UNlabeled data, called SUN. SUN is based on an evolved decision tree. In terms of deviation between history concept clusters and new ones generated by a developed clustering algorithm of *k*-Modes, concept drifts are distinguished from noise at leaves. Extensive studies on both synthetic and real data demonstrate that SUN performs well compared to several known online algorithms on unlabeled data. A conclusion is hence drawn that a feasible reference framework is provided for tackling concept drifting data streams with unlabeled data.

1. Introduction

Most of the existing work relevant to classification on data streams always assumes that all arrived streaming data are completely labeled and these labels could be utilized at hand. Unfortunately, this assumption is violated in many practical applications, especially in the fields of intrusion detection, web user profiling and fraud identification. In such cases, if we only wait for the future labels passively, it is likely that much potentially useful information is lost. Thus, it is significant and necessary to learn actively and immediately.

Motivated by this, a Semi-supervised algorithm of SUN for concept drifting data streams with UNlabeled data is proposed in this paper. SUN provides several contributions. i) Unlike several existing semi-supervised algorithms for data streams (Wu et al. 2006; Ho and Wechsler 2007; Masud et al. 2008) with a clustering method, we develop a clustering algorithm based on *k*-Modes (Ng et al. 2007) to produce concept clusters at leaves in a decision tree built incrementally. Unlabeled data are predicted using these concept clusters and the labeled information is reused, as filling the gap in labeled data with relevant unlabeled data is conductive to reduce the drift rate, which is concluded in (Widyantoro 2007). ii) We utilize the deviations between history concepts and new ones in tandem of the bottom-up search to detect potential concept drifts from noise. To the best of our knowledge, this is a new method to detect concept drifts from data streams. iii) Evaluations on both synthetic and real-life data show that SUN is comparable to the state-of-the-art concept drifting algorithms of CVFDT (Hulten et al. 2001) and CDRDT (Li et al. 2009), even on unlabeled data streams. Meanwhile, SUN outperforms semi-supervised algorithms mentioned in (Wu et al. 2006).

2. The SUN Algorithm

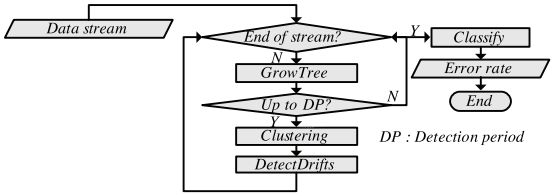

Figure 1. Processing flow of SUN

Our algorithm of SUN aims for utilizing the information of labeled data to label unlabeled data and handling scenarios where there are concept drifts in data streams. It mainly consists of four components as shown in Figure 1. i) *GrowTree*: With the arrival of streaming data, a decision tree is learned by recursively replacing leaves with decision nodes as in CVFDT. However, in contrast with CVFDT, instead of generating alternate sub-trees for each decision node, the statistical information of instances at leaves is maintained for future drifting detection. ii) *Clustering*: a clustering algorithm of k-Modes is developed to generate concept clusters at leaves, if the detection period is reached. And unlabeled data will be labeled with the majority-class in the cluster that they are assigned to by the minimum dissimilarity. iii) *DetectDrifts*: Based on concept clusters, a drifting detection is installed at each leaf in tandem of the bottom-up search. That is, the deviation between the history concept cluster and the new one is utilized to track concept drifts from noise. Several variables are defined to evaluate this deviation, such as the radius of the new/history concept cluster (e.g., r_{new}/r_{hist}) and the distance between clusters (e.g., dist). According to the relations between these variables, three cases in concept drifts are obtained as illustrated in Figure 2, including *potential drift*, *plausible drift* (impact from noise) and *abrupt drift*. iv) *Classify*: After training, the predictive ability is tested on the testing data by majority voting.

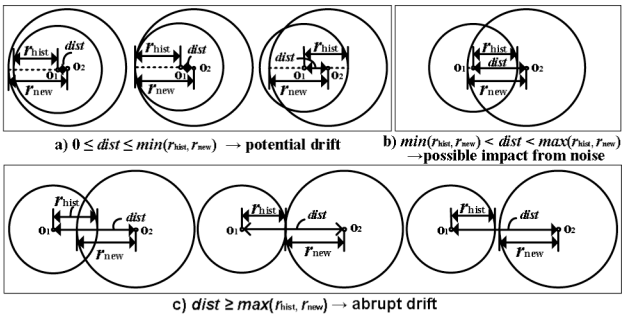

Figure 2. Cases of concept drifts

3. Experimental Evaluations

To validate the efficiency and effectiveness of SUN, we have compared SUN against concept drifting data stream algorithms of CVFDT and CDRDT and semi-supervised algorithms referred in (Wu et al. 2006) with a benchmark concept drift database (Kddcup99 1999) and a real database (Yahoo Shopping). All experiments are conducted on a P4, 3.00GHz PC with 2G main memory, running Windows XP Professional and all algorithms are implemented in C++.

Figure 3 shows the tracking curve over the sequential data chunks with 50% unlabeled data (i.e., *ulr* = 50%) (In this figure, the dotted lines reflect the distribution of class labels). And the statistical detection results on Figure 3 reveal that regarding the estimation metrics (Gama et al. 2009), namely i) probability of false alarms, ii) probability of true alarms, and iii) delay in detection, SUN could achieve an approximate performance compared to CDRDT (CVFDT has no function to report values of these three metrics. Hence, there are no comparisons between SUN and CVFDT). These cases infer that SUN could handle scenarios where there is a concept drift.

Meanwhile, Table 1 reports the predictive accuracies and the time overheads in SUN, CVFDT and CDRDT. First of all, a summary of experimental results presents that the predictive accuracy in SUN is improved by 3.85% and 18.16% on Kddcup99 respectively while it is improved by 12.55% and 21.27% respectively on Shopping compared to CDRDT and CVFDT (experiments in SUN are evaluated in the case of *ulr* = 50%). Secondly, in comparison to a similar semi-supervised algorithm (Wu et al. 2006) based on clustering for data streams (named as Clustering-training), SUN performs as well as Clustering-training and both of them outperform self-training, co-training and tri-training mentioned in (Wu et al. 2006) by a large margin (up to 30%).

Furthermore, regarding the consumptions of runtime, the total time overhead in SUN is only about 1/3~1/2 of those in CDRDT and CVFDT. It is obvious to conclude that SUN is the most efficient algorithm of the three.

4. Conclusions

This paper introduced a Semi-supervised classification

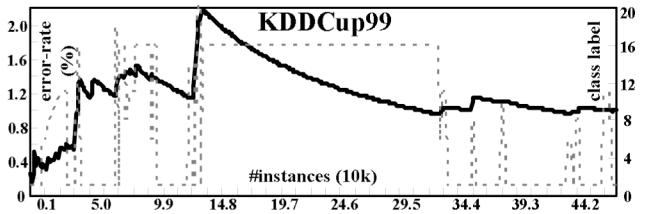

Figure 3. Drift tracking over sequential data chunks of Kddcup99

Table 1. Error rates of classification and overheads of time

Database	Error-rate (%)			Time (s)		
	SUN	CDRDT	CVFDT	SUN	CDRDT	CVFDT
Kddcup99	5.32	9.17	23.48	52	164	93
Shopping	2.08	14.63	23.35	6	11	13

algorithm for data streams with concept drifts and UNlabeled data (SUN). In this algorithm, we build a decision tree incrementally and generate concept clusters at leaves in a clustering algorithm developed from *k*-Modes. With the mechanism of bottom-up search and the deviation of classification using concept clusters, we detect concept drifts from noise. Experimental studies reveal the efficiency and effectiveness of SUN even in the cases with a large volume of unlabeled data. A preliminary application of SUN to a real database shows promising results. However, how to predict unknown concepts in advance and how to reduce the overheads of space are still challenging issues for our future work.

5. References

D. H. Widyantoro. 2007. *Exploiting unlabeled data in concept drift learning*. *Jurnal Informatika* 8(1): 54-62.

G. Hulten, L. Spencer, and P. Domingos. 2001. *Mining Time-changing Data Streams*. In *KDD'01*, 97-106.

J. Gama, R. Sebastião, and P. P. Rodrigues. 2009. *Issues in Evaluation of Stream Learning Algorithms*. In *KDD'09*, 329-338.

Kddcup99 data set. 1999. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

M. K. Ng, M. J. Li, Z. X. Huang, and Z. Y. He. 2007. *On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm*. IEEE Transactions on Pattern Analysis and Machine Intelligence 29: 503-507.

M. M. Masud, J. Gao, K. Latifur, and J. W. Han. 2008. *A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data.* In *ICDM'08*, 929-934.

P. P. Li, X. G. Hu, Q. H. Liang, and Y. J. Gao. 2009. *Concept Drifting Detection on Noisy Streaming Data in Random Ensemble Decision Trees*. In *MLDM'09*, 236-250.

S. S. Ho, and H. Wechsler. 2007. *Detecting Changes in Unlabeled Data Streams Using Martingale.* In *IJCAI'07*, 1912-1917.

S. Wu, C. Yang, and J. Zhou. 2006. *Clustering-training for Data Stream Mining.* In *ICDMW'06,* 653-656.

Yahoo Shopping Web Services. http://developer.yahoo.com.everything.html