ATLANTIS PRESS
Advances in Computer Science Research, volume 74
2nd International Conference on Computer Engineering, Information Science & Application Technology (ICCIA 2017)

Weighted Ensemble with Dynamical Chunk Size for Imbalanced Data Streams in Nonstationary Environment

Nini Liu [a], Wen Zhu [b], Bo Liao [c] and Siqi Ren [d]

Hunan University, College of Computer Science and Electronic Engineering, Changsha 410082, China;

[a]1647013704@qq.com, [b]syzhuwen@163.com, [c]dragonbw@163.com, [d]siqirenzl@163.com

Keywords: Imbalanced data, concept drift, dynamic chunk, weighted ensemble.

Abstract. In recent years, learning from data stream has been more and more popular because of its extensive applications. However, most algorithms assume there are no concept drift in one chunk, as the performance of evaluation is sensitive to the chunk size. In this paper, we propose a new approach (WEDC) by introducing the concept drift detection mechanism to dynamically adjusting the chunk size. In addition, we add weighted mechanism to ensemble classifiers, which make WEDC could react to different types of concept drifts well. Experiments performed on the synthetic datasets show that our approach is competitive in the predication accuracy for data streams including different kinds of concept drifts.

1. Introduction

Generally, data streams can be described as an unlimited set of learning examples containing pairs{**x**,y}, where **x** is a vector of attribute values and y is the class label. Different from single classifier, ensemble classifiers usually learn from data stream based on data chunk, where data streams can be divided as $\{B_1, B_2, \dots, B_n\}$ in sequence. However, most of proposed algorithms is based on balanced data streams. What's more, proposed algorithms based on chunk is sensitive to the size of chunk, where they assumed that there is no concept drift in one chunk. However, when the data is about the detection of credit card frauds, financial services, credit card frauds, the detection of intrusion in computer networks, spam filtering, stock trading, the access of Web page as well as other scientific research fields, the class distribution is quite imbalanced and the cost of misclassifying the minority instances is more expensive comparing with the cost of misclassifying the majority instances. Especially when combined with concept drifts, it inspire bigger challenge.

Simply speaking, concept drifts means the information from current data become irrelevant from the past data, which would cause the decrease in prediction accuracy if the classification model cannot adapt to the new concept quickly. Generally speaking, the kinds of concept drift can be divided into sudden, gradual and recurring. In other words, Jing Gao et al.[1] summarized the causes of concept drift as three types by using the joint probability P(x) and P(y|x): feature change, conditional change and dual change. The feature change means the change happens to P(x) while the conditional probability P(y|x) remains the same. On the contrary, the conditional change means P(y|x) changes but P(x) remain the same. The last one is to say there are changes in both P(x) and P(y|x).

2. Algorithm Framework

2.1 Divide Dynamic Chunk.

In order to solve the drawbacks of the assumption that "there is no concept drift in one data chunk", we use a drift detector to find the concept drift stamp in current chunk, which is made to get the optimal size of instances. We use the very fast decision tree (VFDT or Hoeffding Tree) [2] as the drift detector because of its lower processing time and resource consumption. In order to find the concept drifts in both minority class and majority class, we use the G-mean as in formula (1) rather than using the overall accuracy to find drift. Before the max size of chunk we set at first is filled, there is a concept drift stampt_d, where the instances before t_d will be regard as current chunk.

$$\mathrm{G-mean} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \quad (1)$$

2.2 Weight Determination Strategy.

In WEDC, the weight determination strategy is based on a unified voting[3]. In one data stream $\{\ldots, d_{t-1}, d_t, d_{t+1}, \ldots\}$, assume the current training set is S_i, which contains m instances. MSE_{ij} represent the basic classifier C_j's mean square error of the attribute of instance x in the current training set S_i, which is calculated by formula(2). $f_y^j(x)$ means the probability that instances x is predicted by basic classifier C_j as y class.

$$MSE_{ij} = \frac{1}{|S_i|}\sum_{(x,y)\in S_i}(1 - f_y^j(x))^2 \quad (2)$$

Thus the weight W_{ij} of basic classifier C_j is as formula (3).

$$W_{ij} = \frac{1}{\log(MSE_{ij})} \quad (3)$$

The pseudo-code of the proposed WEDC algorithm for mining nonstationary imbalanced data streams is formulated as follows:

Input: Current data chunk S_i, max size of chunk C, number of ensembles k, the value of G-mean g_i, the concept drift time t_d

1. Initialization()
2. Using the drift detector to find the time stam t_d based on the value of G-mean g_i
   a)If there is a drift before filling the max chunk size C, storing the examples before t_d as current chunk S_i.
   b) else using the examples from the max chunk size C as the current chunk S_i
3. Split S_i into P_i and Q_i, where P_i denotes the minority example set and Q_i decribes the majority set.
4. Randomly under sampling Q_i to get $UQ_i$and train classifier C_i on $\{P_i, UQ_i\}$
5. Apply C_j on Si to derive w_{ij}
6. If(|E|<k)
   E=E∪Ci else
   substitute least accurate classifier inε with C_i; end if;
7. for all classifiers$C_j \in (E \backslash C_i)$ do
   incrementally train classifier C_j with S_i; end for;

3. Experiments Setup and Evaluation

In this paper, we use five synthetic dataset to testify our proposed algorithm. The detailed information can be seen in table 1.

Table 1. Characteristic of datasets

Dataset	No.Inst	No.Attrs	Skew_ratio	Noise	No.Drifts	Drift type
SEA_{RD}	1M	3	0.95	0%	4	recurring
SEA_{SD}	1M	3	0.95	0%	5	sudden
STA_{RD}	1M	3	0.95	0%	5	recurring
STA_{mix}	1M	3	0.95	0%	6	mixed
$RanTree_{SFD}$	100k	10	0.9	0%	15	sudden

In order to analyze the performance of algorithm, we use three representative algorithms to compare, including UB (Uncorrelated Bagging) [1], SERA (Selectively recursive approach towards nonstationary imbalanced stream data mining) [4], MuSeRA (Multiple Selectively Recursive Approach towards imbalanced stream data mining) [5]. Average values of the analyzed performance measures are given in Tables 2.

Table 2. Average calssification accuracies in percentage(%)

	UB	SERA	MuSeRA	WEDC
SEA_{SD}	75.9	96.34	96.5	96.02
SEA_{RD}	96.9	97.69	97.82	97.85
STA_{RD}	88.63	95.75	95.8	94.21
STA_{mix}	85.93	94.83	94.86	92.87
$RanTree_{SFD}$	63.85	82.47	82.91	83.98

Considering the length of paper, we choose three representative datasets to draw plots, as in Fig.1. The classification on STA_{RD} shows the capability to gradual and recurring concept drifts. After the example number 500k, recurring concept drifts could achieved by set the same Boolean function of STAGGER. We can see the plot of SERA and MuSeRA is nearly the same, as the two algorithms has almost the same mechanism. After 500k, both WEDC and MuSeRA rise because of their weighted mechanism, which keep the knowledge from old instances. From the classification accuracy on STA_{mix}, we can see obvious drops on 230K,690K and 790K, which means the sudden drifts occurring. It is worth noting that in Fig.2 on the $RanTree_{SFD}$ dataset, WEDC react sensitively to the fast sudden drifts, and most time keep higher than MuSeRA and SERA. The reason is that there is a drift detector in WEDC, which could find drift fast and make sure that the instances from divided dynamic chunk is of the same concept and train the classifiers well. The evaluation time on STA_{mix} shows that SERA is the lowest , followed by SERA, WEDC, the last is UB.

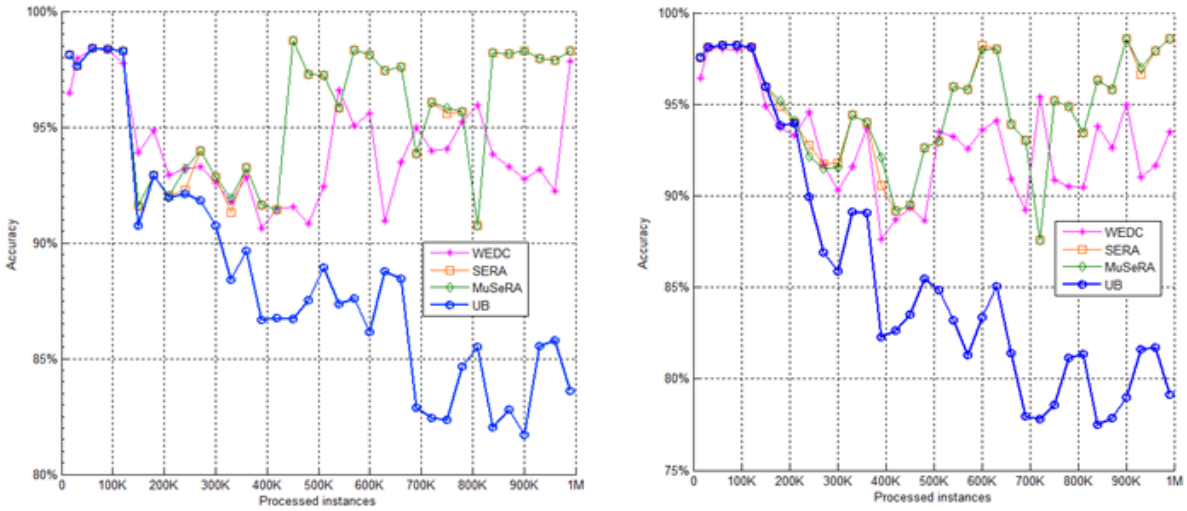

Fig.1 The classification accuracy on STA_{RD}, STA_{mix}

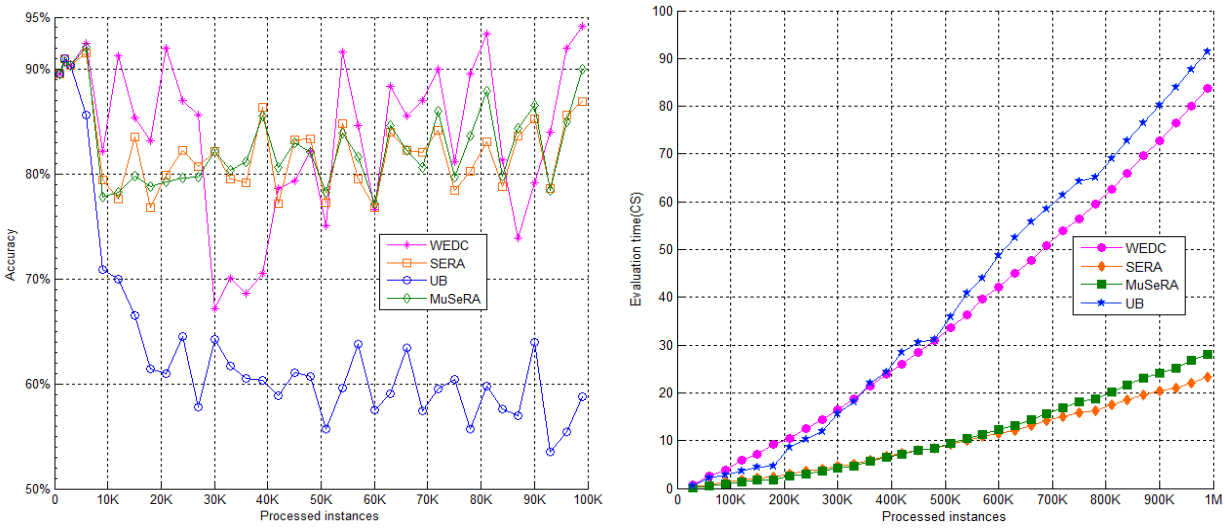

Fig.2 The classification accuracy on $RanTree_{SFD}$ and the evaluation time on STA_{mix}.

The drift detector and basic classifier of WEDC is VFDT, where assume v as the max number of attribute of instance, c is the kind of class, l is the number of leaves, L means the depth of the tree, and k is the number of basic classifiers. So the space complexity of WEDC is O(lkdvc) and the time complexity is O(Lkdvc).

4. Conclusion

In this paper, we proposed and evaluated a weighted ensemble with dynamical chunk size for imbalanced data streams in nonstationary environment, called WEDC, which was designed to react to different kinds of concept drifts. We also carried out an experimental study comparing WEDC with 3 additional state-of-the-art algorithms based on imbalanced data streams. The obtained results confirmed that WEDC could achieve comparable classification accuracy in imbalanced data streams and ahieve good robustness to fast sudden drifts. In our future work, we would apply the algorithm to multi-class in imbalanced data streams.

References

[1]. Gao, J., Fan, W., Han, J., et al. "A General Framework for Mining Concept-Drifting Data Streams with Skewed Distributions." Siam International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, Usa DBLP, 2007.
[2]. P. Domingos and G. Hulten, “Mining high-speed data streams,” in Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2000,pp. 71–80.
[3]. J. Gao, W. Fan, and J. Han, "On appropriate assumptions to mine data streams: Analysis and practice," in Proc. Int. Corif. Data Mining, Washington, DC, USA, 2007, pp. 143-152.
[4]. S. Chen and H. He, "Sera: Selectively recursive approach towards nonstationary imbalanced stream data mining," Neural Networks, IEEE-INNS-ENNS International Joint Conference on, vol. 0, pp. 522-529,2009.
[5]. Chen S, He H, Li K, et al. MuSeRA: Multiple Selectively Recursive Approach towards.
[6]. imbalanced stream data mining[C]// International Joint Conference on Neural Networks. IEEE, 2010:1-8.