Acta Technica *62*, No. 3B/2017, 263–276

An ensemble classification algorithm for imbalanced data streams with concept drifts[1]

Sun Gang[2,3], Zhao Jia[2,4], Wang Zhongxin[2]

Abstract. Concept drifts are often implied in the imbalanced data streams in practice. Most of the existing concept drift detection algorithms are based on the classification error rate of all instances and it can not be directly applied to the concept drift detection of imbalanced data stream. For this reason, this paper presents an ensemble classification algorithm for imbalanced data streams with concept drifts. First, an improved resampling method is used to establish a balanced training subset. Secondly, the support vector machine is used to create a base classifier on the training subset. Finally, an ensemble classifier is constructed using the WE integration model. The algorithm uses an improved resampling method to avoid merging the instances of different concept intervals into the same data block. The concept drift is detected by the double threshold determined by the Hoeffding Bounds inequality. The experimental results show that the proposed algorithm can detect the concept drifts in the imbalanced data streams, and not only has good classification performance for the positive instances, but also has good classification performance for all instances. It is an effective ensemble classification algorithm for imbalance data streams with concept drifts.

Key words. Data streams, classification, imbalanced, concept drifts, ensemble model.

1. Introduction

With the rapid development of information technology, more and more data needs to be processed in practical application. Thousands of daily shopping records

[1]The work was supported by the National Natural Science Foundation of China (51174257/F030504), and supported by the Fundamental Research Funds for the Central Universities of Hefei University of Technology (2013bhzx0040) and by the Natural Science Foundation of the Anhui Higher Education Institutions of China (KJ2016A549 and KJ2017A332) and Innovation Experiment Program for University Students (AH201610371053).

[2]School of Computer and Information Engineering, Fuyang Normal University, Fuyang, No. 100 Qinghexi Road, Yingzhou District, Fuyang Anhui, 236037, China

[3]School of Computer Science and Information, Hefei University of Technology, Hefei, 230009, China

[4]Corresponding author; e-mail: zhaojiafync@163.com

on the Taobao website and the fast-growing payment transaction records on the Alipay website and so on. Such continuous arrival, high-speed, dynamic and large-scale data is called data stream [1]. Data stream classification is widely used in the network monitoring, sensor networks, e-commerce and other practical applications. However, the class distribution of data streams in practical application is often imbalanced, that is, the number of instances of some classes is small, while the number of instances of other classes is relatively large, such as medical diagnosis, fraudulent credit card detection, anomaly detection and so on, such data streams are called imbalanced data streams [2].

The concept drift is often implied in the imbalanced data streams, which is the concept change of implied goals caused by the change in the context of the data stream, or even a fundamental change [3–4]. The concept drift of the imbalanced data stream is different from that of the traditional data stream. Most of the existing concept drift detection methods are based on the classification error rate of all instances. When the concept drift in the imbalanced data stream is occurred, the classification error rate of all instances does not change obviously but the classification error rate of the positive instances changes obviously, which leads to the concept drift can not be detected in time, and ultimately affects the classification performance. Therefore, the existing concept drift detection methods can not be directly applied to the concept drift detection of imbalanced data streams.

Therefore, how to construct an on-line classification algorithm with strong generalization ability under the environment of imbalanced data streams with concept drifts becomes the key and challenging task of data processing in practical applications. To solve this problem, this paper proposes an ensemble classification algorithm for imbalanced data streams with concept drifts, referred to as ECIDSCD. Firstly, an improved resampling method is used to establish balanced training subsets. Secondly, the support vector machine is used to create the base classifier on the training subsets. Finally, an ensemble classifier is constructed using the WE ensemble model. The algorithm uses an improved resampling method to avoid merging the instances in different concept intervals into the same data block, and the concept drift is detected by the double threshold determined by the Hoeffding Bounds inequality. The experimental results show that the proposed algorithm can detect the concept drifts in the imbalanced data streams, and not only has good classification performance for the positive instances, but also has good classification performance for all instances. It is an effective classification algorithm for imbalanced data streams with concept drifts.

2. Related works

As an important part of data stream classification, imbalanced data stream classification has aroused the concern of researchers. Gao et al. [5] proposed a method of integrating resampling and ensemble classifier, which divides the negative instances of the latest data block into several disjoint subsets and reassembles them with the positive instances, then uses the decision tree algorithm training the base classifiers and integrates them. Ouyang et al. [6] proposed an ensemble algorithm using weights

for imbalanced data stream, which divides the negative instances in each data block into n blocks and then reassembles them with the set of positive instances. Wang et al. [7] proposed CS algorithm, which uses k-means clustering method to select the negative instances that are several times as the positive instances, that is, down sampling the negative instances and adopting the ensemble method. Song et al. [8] proposed IDSS algorithm, which is also a resampling method based on clustering method. The clustering method can select representative negative instances and reduce the degree of imbalance, so that these algorithms can achieve better classification performance. Lichtenwalter et al. [9] proposed a simple resampling method, which randomly delete correctly predicted negative instances until the number of positive and negative instances satisfies a certain proportion. Aiming at the concept drift problem in data streams, Street et al. [10] proposed an ensemble algorithm based on multi-decision tree; Kolter et al. [11] proposed an incremental ensemble classification algorithm based on weights. Kuncheva et al. [12] proposed a new classification algorithm based on window mechanism. Hulten et al. [13] proposed a data stream classification algorithm based on Hoeffding decision tree. Liu et al. [14] proposed a data stream classification algorithm based on fuzzy decision tree on the basis of CVFDT. Zhang et al. [15] proposed an ensemble data stream classification algorithm based on feature drift. Liu et al. [16] proposed an incremental data stream classification algorithm based on uncertainty of the samples. Wang et al. [17] proposed an ensemble classification algorithm for data streams with noise and concept drifts.

From the above studies, it can be seen that the ensemble model is widely used in the classification of imbalanced data streams. Compared with the single model, the ensemble model has higher classification accuracy and can faster adapt to the concept drifts for data streams. Therefore, a good classification algorithm for imbalanced data streams with concept drifts should be able to detect different types of concept drifts from the imbalanced data streams, and not only have good classification performance for the positive instances, but also have a better classification performance for all instances.

3. An ensemble classification algorithm for imbalanced data steams with concept drifts

3.1. Resampling strategy

The data resampling is one of the key steps for imbalanced dada classification. The existing resampling methods can be divided into two categories: one is down sampling methods, which use clustering algorithm to select the negative instances. These methods are relatively-consuming, not suitable for data stream environment. Considering the real-time requirement of the algorithms under data stream environment, the other is the sampling methods, which divide the negative instances and combine the positive instances. However, the above two methods do not consider the problem of concept drifts.

Taking into account the existence of concept drifts, this paper adopts the method

of dividing negative instances to reduce the imbalance degree, which is inspired by references [18–19]. That is, this method divides the negative instances of the current data block and then combines them with the positive instances of the current data block. The method only divides negative instances, not down sampling, so do not loss the information of negative instances. In addition, this method does not merge the positive instances of different data blocks, and avoid merging the positive instances of different concept intervals into the same data block. Positive and negative instances merger in the same data block, and it is largest possible to maintain the original data distribution. Specific sampling steps are shown in Fig. 1.

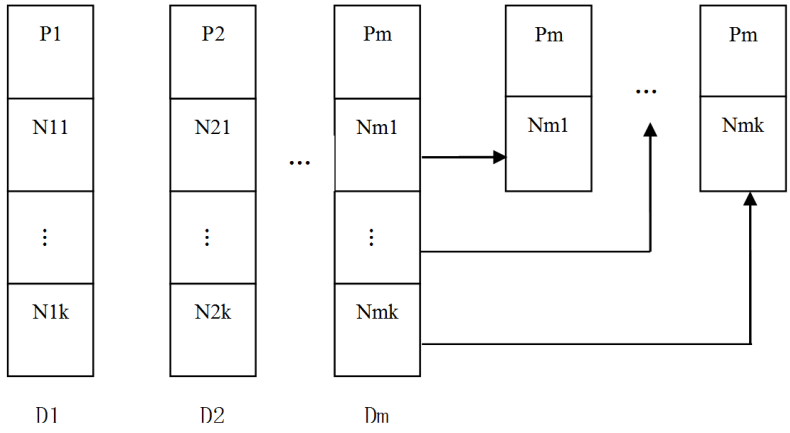

Fig. 1. Resampling method of ECIDSCD algorithm

Assume that the currently arriving data blocks are $D1, D2, \cdots, Dm$, each data block is equal in size, and contains the same number of positive stances. Firstly, the Nm of Dm is divided into k ($k = |Nm|/|Pm|$) disjoint subsets, in which the number of instances is equal to the number of instances of Pm, then k balanced training subsets are obtained by reassembling them with Pm, respectively. In the above operation, Nm is divided into k data blocks and the number of positive instances is the same. In addition, since the incorrectly classified positive instances often represent the direction of the concept drift, therefore, in the resampling process, these incorrectly classified positive instances are added to the training subsets to make the classifier more adaptable to the new concept.

3.2. Concept drift detection method

The concept drift detection mechanism of the ECIDSCD algorithm is based on the double threshold detection mechanism, and different concept drifts are detected from imbalanced data stream by calculating the classification error rate on the time window. Double threshold detection mechanism is an effective method for concept drift detection. In contrast to the other double threshold detection methods, this paper uses the double threshold determined by the Hoeffding Bounds inequality to detect concept drifts, which is inspired by references [20–21]. As the traditional

classification error rate in the imbalanced data streams is difficult to reflect the distribution change of the positive instances, therefore, there is no practical statistical significance for the imbalanced data streams, so the classification error rate of the positive instances is used instead of the traditional classification error rate.

In this paper, the method of detecting concept drift is implemented as follows: Firstly, the classification error rate of positive instances is calculated for the current data block and the previous data block, and then the difference Δe of the classification error rate of positive instances of 2 data blocks is calculated.

$$\Delta e = e_{\mathrm{c}} - e_{\mathrm{b}}\,, \tag{1}$$

where e_{c} is the classification error rate of the positive instances of the current data block and e_{b} is the classification error rate of the positive instances of the previous data block.

Assume that e_{b} and e_{c} are two independent variables, which are subject to the normal distribution. According to the nature of the normal distribution, Δe is also subject to the normal distribution. If there is no concept drift in the data stream, the probability distribution of the ensemble classifier EC on the current data block and the previous data block should be invariant. Therefore, according to the Hoeffding Bounds inequality:

$$P(e \geq \overline{e} - \varepsilon) = 1 - \delta, \quad \varepsilon = \sqrt{(R^2 \ln(1/\delta))/2n} \tag{2}$$

it can be obtained that

$$P(|e - \overline{e}| \leq \varepsilon) = 1 - \delta\,, \tag{3}$$

where $R = \log_2 C$, C being the number of categories.

The confidence level of the true value of Δe in the interval $e_{\mathrm{c}} - e_{\mathrm{b}} \pm k\varepsilon$ is $1 - \delta$. According to Hoeffding Bounds inequality, the relationship among the three variables k, Δe and δ can be obtained. If the value of k is large, the value of Δe is larger and the value of δ becomes smaller. The larger the value of Δe the larger the distribution change of the adjacent two data blocks, and the greater the probability of occurrence of concept drifts in the data stream. The ECIDSCD algorithm uses the double thresholds $k_1\varepsilon$ and $k_2\varepsilon$ determined by the Hoeffding Bounds inequality to detect the concept drifts, where $k_1 < k_2$. If $\Delta e \geq k_2\varepsilon$, the true concept drift occurs in the imbalanced data stream. If $\Delta e \leq k_1\varepsilon$, the potential concept drift occurs in the imbalanced data stream. If $k_1\varepsilon < \Delta e < k_2\varepsilon$, it is only affected by the noise data, and there is no concept drift in the imbalanced data stream.

3.3. Construction and updating of ensemble classifier

After obtaining the k balanced training subsets, the support vector machine algorithm is used to train the k base classifiers and construct an ensemble classifier EC. The weight of the base classifier is determined by the value of the F-value, the initial weight of which is $1/k$. The ensemble classifier EC using the weighted voting mechanism predicts the each instance in data block $Dm+1$, and the positive

instances which are incorrectly predicted are added to error set errInstP. After the prediction is completed, if the concept drift is detected, the classifier needs to be updated; otherwise the model do not needs to update which directly predicts the next data block, in which the weight of the each base classifier is adjusted by the F-value.

$$\mathrm{F-value} = \frac{(1+\beta^2) \times \mathrm{recall} \times \mathrm{precision}}{\beta^2 \times \mathrm{recall} + \mathrm{precision}}. \tag{4}$$

In the definition of F-value, the parameter β is adjustable, which is usually 1. It can be seen from formula (4) that only when the recall and the precision are large, F-value will increase, so the F-value can reflect the classification performance of the positive instances of the classifier.

The updating operation of the classifier is as follows:

1) The training set TS is updated to errInstP and $Dm+1$.

2) Divided the new training set TS to k training subsets TS1, TS2, ..., TSk;

3) Re-training k base classifiers, the weight of the base classifier is set to equal weight.

3.4. The framework of algorithm

Firstly, the symbols involved in the ECIDSCD algorithm are described. Data blocks $\{D1, D2, \cdots, Dm\}$ are obtained by continuous sampling from the imbalance data stream. Dm is the latest arriving data block, and the next data block $Dm+1$ is as test data block. Ej represents the jth instance of the data block, $Dm = Pm + Nm$, where Pm and Nm represent the positive instances and negative instances respectively in Dm. The number of instances in these data blocks are the same and $|Pm| \ll |Nm|$. Symbol k represents the number of base classifiers, EC represents the ensemble classifier, errInstP represents the set of positive instances that are incorrectly classified.

The framework of the ECIDSCD algorithm is then shown in Figure 2:

4. Experimental results and analysis

4.1. Concept drift detection analysis

In order to verify the effectiveness of the concept drift detection mechanism of ECIDSCD algorithm, the concept drift in the data stream with positive instances proportion of 5 % is experimented. In other cases, the change of concept drift is the same as that of the data stream with positive instances proportion of 5 %. The performance evaluation of the concept drift detection method in the data stream usually uses the probability that the concept drift is incorrectly detected and the number of the concept drift which is not detected during the detection of concept drift. Table 1 shows the statistics of concept drift detection used by concept drift detection method proposed in this paper in data sets SEA, HyperPlane and KDD-Cup. False alarms means the probability of the concept drift which is incorrectly detected during the detection of concept drift; and Missing means the number of the

ECIDSCD algorithm

Input: Inbalanced data stream, IDS; Ensemble classifier, EC = Null;

Output: trained ensemble classifier EC

Begin

While (new data block Dm arrives)

{If (EC = Null)

{The negative instance s and the positive instances in D m are reorganized to form k training subsets, k=|Nm|/|Pm|;

Support Vector Machine algorithm is used to train the base classifier s on training subsets to construct an ensemble classifier EC, and the weight of the base classifiers is set to 1/k;

else

{for (Ej ∈ Dm)

{If (Ej is incorrectly classified by EC && Ej is a positive instance)

Put Ej into the temporary misclassification buffer ErrInstP;}

if (concept drift occurred) // determine whether to update the classifier

{ Update the training set with the instance in ErrInstP + D_{m+1};

ErrInstP = Null;

Re-divide the new training set, and re -trains base classifiers. The weight of the base classifiers is set equal weight, and update the ensemble classifier EC;}

else

Adjust the weight of the base classifiers in the EC using the F-value of the prediction on D_{m+1};

}

}

End

Fig. 2. ECIDSCD algorithm

concept drift which is not detected during the detection of concept drift.

ECIDSCD algorithm uses the double threshold determined by the Hoeffding Bounds inequality to detect concept drifts. On the SEA data set, the probability of false prediction is 6.72 %, and the number of undetected concept drifts is 1. On the HyperPlane data set, the probability of false prediction is 8.16 %, and the number of undetected concept drifts is 8. On the KDDCup data set, the probability of false prediction is 7.24 %, and the number of undetected concept drifts is 6. The HyperPlane data set is a gradual concept drift data set. In the process of concept drift occurring gradually, the average error rate increases, and the number of false alarms is relatively large. False alarms usually occur at the beginning of training. Because at the beginning stage, the training data is insufficient, the classification error rate will produce relatively large fluctuation. Therefore, the false alarm of concept drifts is easy to happen. On the whole, the experimental results show that the concept drift detection method proposed in this paper has good performance and can detect most of the concept drifts in the data stream.

Table 1. Statistics of concept drift detection in data sets

Data set	False alarms (%)	Missing
SEA	6.72	1
HyperPlane	8.16	8
KDDCup	7.24	6

4.2. Classification performance

In order to verify the classification performance of ECIDSCD algorithm proposed in this paper for imbalanced data streams, ECIDSCD algorithm and SE algorithm, IDSS algorithm do some experiments on data sets SEA, Hyperplane and KDDCup. The evaluation indexes are F-value and G-mean. F-value can reflect the classification performance of the positive instances correctly, and the G-mean value reflects the classification performance of all instances. The following experiments are performed for imbalanced data streams with different proportion of positive instances.

The experiments are carried out in the case of positive instances proportion of 5 %, and the experimental results are shown in Figs. 3 and 4. It can be seen from Fig. 3 that the F-value of the ECIDSCD algorithm proposed in this paper is improved compared with other algorithms, that is, the positive classification performance of the algorithm proposed in this paper is better than the positive classification performance of other algorithms. On the experimental data sets, the ECIDSCD algorithm proposed in this paper is 3.52 % to 10.49 % higher than the F-value of the SE algorithm, which is 0.88 % to 8.17 % higher than the F-value of the IDSS algorithm. It can be seen from Fig. 4 that the G-mean of the ECIDSCD algorithm proposed in this paper is also proved compared with other algorithms, that is, the overall classification performance of the algorithm proposed in this paper is better than the overall classification performance of other algorithms. On the experimental data sets, the ECIDSCD algorithm proposed in this paper is 2.88 % to 9.21 % higher than the G-mean of the SE algorithm, which is 2.59 % to 8.46 % higher than the G-mean of the IDSS algorithm.

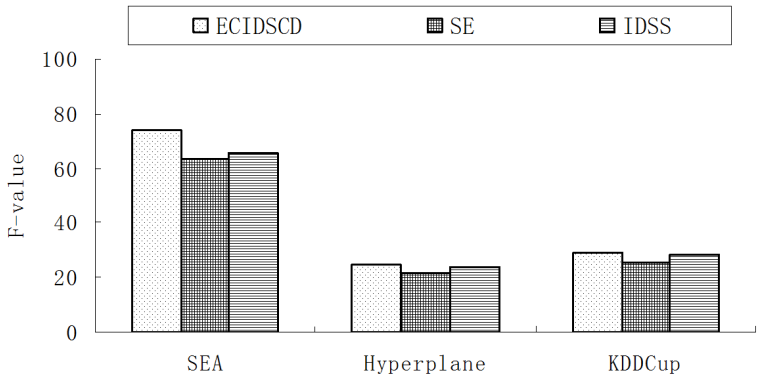

Fig. 3. F-value on data sets with positive instances proportion of 5 %

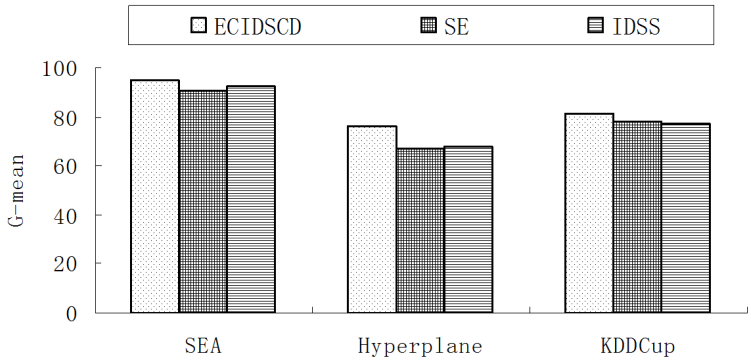

Fig. 4. G-mean on data sets with positive instances proportion of 5 %

The experiments are carried out in the case of positive instances proportion of 10 %, and the experimental results are shown in Figs. 5 and 6. It can be seen from Fig. 5 that the F-value of the ECIDSCD algorithm proposed in this paper is improved compared with other algorithms, that is, the positive classification performance of the algorithm proposed in this paper is better than the positive classification performance of other algorithms. On the experimental data sets, the ECIDSCD algorithm proposed in this paper is 3.36 % to 8.61 % higher than the F-value of the SE algorithm, which is 1.03 % to 7.38 % higher than the F-value of the IDSS algorithm. It can be seen from Fig. 6 that the G-mean of the ECIDSCD algorithm proposed in this paper is also proved compared with other algorithms, that is, the overall classification performance of the algorithm proposed in this paper is better than the overall classification performance of other algorithms. On the experimental data sets, the ECIDSCD algorithm proposed in this paper is 2.92 % to 4.05 % higher than the G-mean of the SE algorithm, which is 1.11 % to 4.67 % higher than the G-mean of the IDSS algorithm.

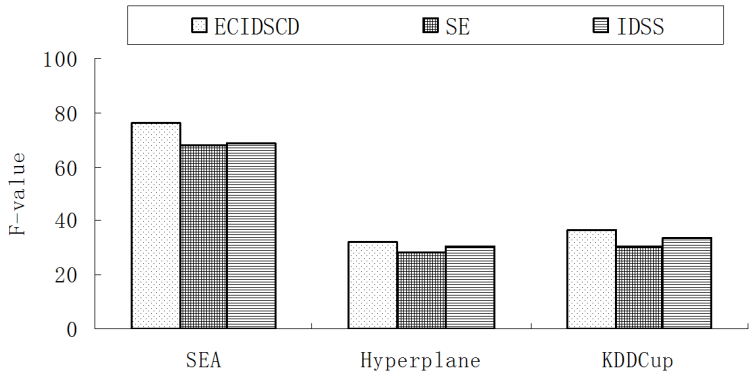

Fig. 5. F-value on data sets with positive instances proportion of 10 %

The experiments are carried out in the case of positive instances proportion of 15 %, and the experimental results are shown in Figs. 7 and 8. It can be seen from

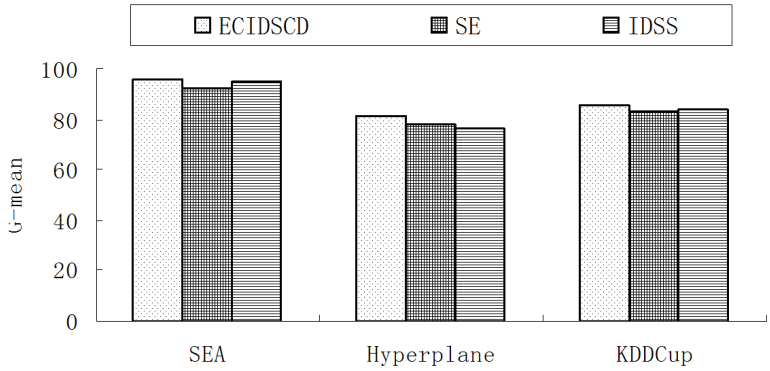

Fig. 6. G-mean on data sets with positive instances proportion of 10 %

Fig. 7 that the F-value of the ECIDSCD algorithm proposed in this paper is improved compared with other algorithms, that is, the positive classification performance of the algorithm proposed in this paper is better than the positive classification performance of other algorithms. On the experimental data sets, the ECIDSCD algorithm proposed in this paper is 1.91 % to 9.62 % higher than the F-value of the SE algorithm, which is 1.42 % to 7.01 % higher than the F-value of the IDSS algorithm. It can be seen from Fig. 8 that the G-mean of the ECIDSCD algorithm proposed in this paper is also proved compared with other algorithms, that is, the overall classification performance of the algorithm proposed in this paper is better than the overall classification performance of other algorithms. On the experimental data sets, the ECIDSCD algorithm proposed in this paper is 5.28 % to 11.52 % higher than the G-mean of the SE algorithm, which is 3.06 % to 6.44 % higher than the G-mean of the IDSS algorithm.

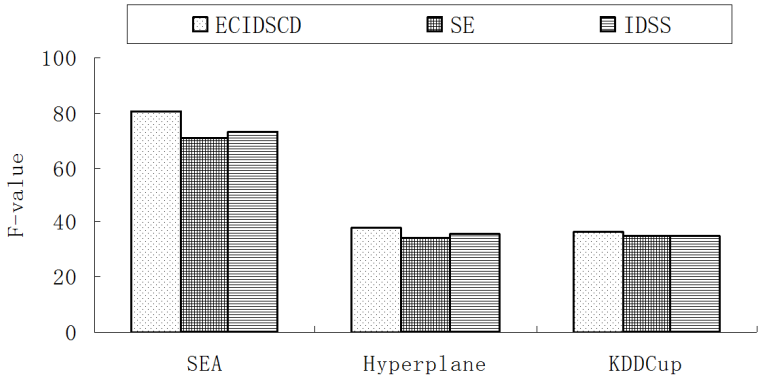

Fig. 7. F-value on data sets with positive instances proportion of 15 %

ECIDSCD algorithm proposed in this paper has good results on evaluation indexes F-value and G-mean for different datasets with different positive proportion. The algorithm not only has good classification performance for the positive instances,

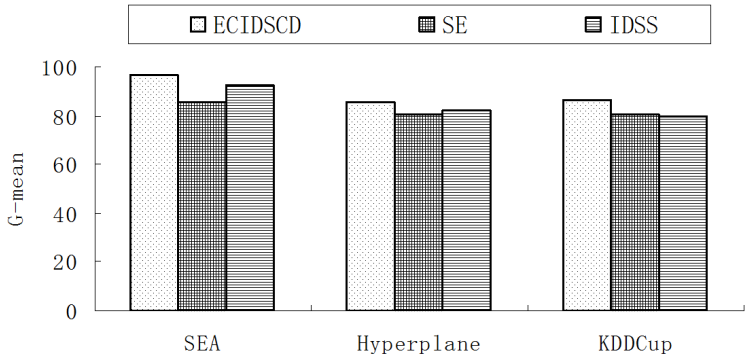

Fig. 8. G-mean on data sets with positive instances proportion of 15 %

but also has good classification performance for all instances.

In summary, ECIDSCD algorithm has better performance of concept drift detection, and can detect the implicit concept drifts in the data streams, and have better classification performance for positive instances. ECIDSCD algorithm meets the classification requirements of the imbalance data streams with concept drifts. Therefore, ECIDSCD algorithm is an effective classification algorithm for imbalanced data streams with concept drifts.

5. Conclusion

In the study of classification for imbalanced data streams, most of the algorithms do not consider the problem of concept drift in data stream. In the study of concept drift detection, most of the algorithms deal with the problem of balanced data streams, which can not directly be applied to concept drift detection of imbalanced data stream. For this reason, this paper presents an ensemble classification algorithm for imbalanced data streams with concept drifts. Firstly, an improved resampling method is used to establish a balanced training subset. Secondly, the support vector machine is used to create a base classifier on the training subset. Finally, an ensemble classifier is constructed using the WE integration model. The algorithm uses an improved resampling method to avoid merging the instances of different concept intervals into the same data block. The concept drift is detected by the double threshold determined by the Hoeffding Bounds inequality. The experimental results show that the proposed algorithm can detect the concept drift in the imbalanced data stream, and not only has good classification performance for the positive instances, but also has good classification performance for all instances. It is an effective ensemble classification algorithm for imbalance data streams with concept drifts. In general, there is a large number of unlabeled data in the real imbalanced data streams. These imbalanced data streams with unlabeled data contain only a small number of labeled instances and a large number of unlabeled instances. Therefore, how to design the concept drift detection method and classification algorithm for

imbalanced data stream with unlabeled data will be the main research contents in the future.

References

[1] M. M. Gaber, A. B. Zaslavsky, S. Krishnaswamy: *Mining data streams: A review.* ACM SIGMOD Record *34* (2005), No. 2, 18–26.

[2] Z. Z. Ouyang: *Research on classification technologies in mining unsteady data streams.* National University of Defense Technology, China, Ph.D. Thesis (2009).

[3] G. Widmer, M. Kubat: *Learning in the presence of concept drift and hidden contexts.* Machine Learning *23* (1996), No. 1, 69–101.

[4] J. Gama, P. Medas, G. Castillo, P. Rodrigues: *Learning with drift detection.* Proc. Brazilian Symposium on Artificial Intelligence, 29 Sept–1 Oct 2004, Sao Luis, Maranhao, Brazil, SBIA 2004 286–295.

[5] J. Gao, B. Ding, W. Fan, J. Han, P. S. Yu: *Classifying data stream with skewed class distribution and concept drift.* IEEE Internet Computing *12* (2008), No. 6, 37 to 49.

[6] Z. Z. Ouyang, J. Luo, D. Hu, Q. Wu: *An ensemble classifier framework for mining imbalanced data streams.* Chinese Journal of Electronics *38* (2010) No. 1, 184–189.

[7] Y. Wang, Y. Zhang, Y. Wang: *Mining data streams with skewed distribution by static classifier ensemble.* In book: Opportunities and Challenges for Next-Generation Applied Intelligence, Springer Berlin Heidelberg (2009), 65–71.

[8] Q. Song, J. Zhang, Z. Deng: *A better intrusion detection algorithm based on classification of skewed-data streams.* Journal of Northwestern Polytechnical University *27* (2009), No. 6, 859–862.

[9] A. Godase, V. Attar: *Classifier ensemble for imbalanced data stream classification.* Proc. CUBE International Information Technology Conference, 3–5 September 2012, Pune, India (2012), 284–289.

[10] W. N. Street, Y. S. Kim: *A streaming ensemble algorithm (SEA) for large-scale classification.* Proc. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 26–29 August 2001, San Francisco, California, USA (2001), 377–382.

[11] J. Z. Kolter, M. A. Maloof: *Dynamic weighted majority: a new ensemble method for tracking concept drift.* Proc. IEEE International Conference on Data Mining, 19-22 Nov. 2003, Melbourne, FL, USA, IEEE Conference Publications 123–130.

[12] L. I. Kuncheva, I. Žliobaité: *On the window size for classification in changing environments.* Journal Intelligent Data Analysis *13* (2009), No. 6, 861–872.

[13] G. Hulten, L. Spencer, P. Domingos: *Mining time-changing data streams.* Proc. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 26–29 August 2001, San Francisco, California, USA (2001), 97–106.

[14] J. Liu, X. Li, W. Zhong: *Ambiguous decision trees for mining concept-drifting data streams.* Pattern Recognition Letters *30* (2009), No. 15, 1347-1355.

[15] Y. P. Zhang, S. H. Liu: *Ensemble classification based on feature drifting in data streams.* Computer Engineering & Science *36* (2014), No. 5, 977–985.

[16] S. M. Liu, Z. X. Sun, T. Liu: *Research of incremental data stream classification based on sample uncertainty.* Journal of Chinese Computer Systems *36* (2015), No. 2, 193 to 196.

[17] Z. Wang, G. Sun, H. Wang: *Ensemble classification algorithm for data streams with noise and concept drifts.* Journal of Chinese Computer Systems *37* (2016), No. 7, 1445 to 1449.

[18] N. V. Chawla, N. Japkowicz, A. Kotcz: *Editorial: Special issue on learning from imbalanced data sets.* ACM SIGKDD Explorations Newsletter *6* (2004), No. 1, 1–6.

[19] J. Zhang: *Study of classification algorithms for skewed data streams based on ensemble framework.* Hefei University of Technology, Dissertation (2012).

[20] Y. Zhang: *A study on classification in data stream.* Hefei University of Technology, China, Ph.D. Thesis (2011).

[21] P. P. Li: *Concept drifting detection and classification on data streams.* Hefei University of Technology, China, Ph.D. Thesis (2012).

Received April 30, 2017