Acta Technica *62*, No. 3B/2017, 317–326

A safety evaluation method of mine pressure based on concept drifting data stream classification[1]

Sun Gang[2,4], Wang Zhongxin[2,5,6], Zhao Jia[2],
Wang Hao[2], Ding Zhengqi[3]

Abstract. Mine pressure monitoring data is essentially a data stream. With the change of environment, mine pressure monitoring data stream implied concept drifts. The safety evaluation of mine pressure can be seen as concept drifting data stream classification, and classification labels are safety and unsafety. The safety evaluation method of mine pressure used in this paper is a concept drifting data stream classification algorithm based on double thresholds, which uses support vector machine as the basic classifier, and uses the Bayesian classifier to filter noise data, and uses double thresholds determined by Hoeffding bounds inequality to detect concept drifts. Experimental results show the method can better detect concept drifts in data stream, and it has better classification accuracy for data stream, and it can be applied to the safety evaluation of mine pressure.

Key words. Mine pressure; safety evaluation; concept drift; data stream classification..

[1] The work was jointly supported by the National Natural Science Foundation of China and Shenhua Group Co., Ltd. (51174257/E0422), and supported by the Fundamental Research Funds for the Central Universities of Hefei University of Technology (2013bhzx0040), and supported by the Natural Science Foundation of the Anhui Higher Education Institutions of China (KJ2016A549 and KJ2017A332), and supported by the Innovative Experiment Program for University Students (AH201610371053).

[2] School of Computer and Information Engineering, Fuyang Normal University, Fuyang, 236037, China

[3] Experimental and Training Management Center, Fuyang Normal University, Fuyang 236037, China

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

[5] Corresponding author; e-mail: `wzxfync@163.com`

[6] No. 100 Qinghexi Road, Yingzhou District, Fuyang Anhui 236037, China

1. Introduction

With the growing of the scale and the depth of coal mining, mine pressure damage is becoming serious, that has become an important problem need urgently to be solved. In order to prevent mine pressure damage, the coal mine invests a lot of manpower, material and financial resources to build all kinds of mine pressure monitoring systems, which monitor all kinds of mine pressure data and judges the safety of mine pressure through these monitoring data. Mine pressure monitoring data have characteristics of data stream as being real-time, continuous, orderly, time-varying and infinite [1], therefore, mine pressure safety judged through monitoring data can be seen as data stream classification, and classification labels are safety and unsafety.

In practical application, the most typical feature of data stream is concept drift, in addition to high-speed, continuous, varied, and infinite. The implied target concept change caused by change in the data stream context [2]–[4], and it is known as concept drift. With the ongoing mining work, internal and external environment in coal mine is constantly changing, and mine pressure safety is also changed. Mine pressure safety change caused by the internal and external environment can be seen as concept drift in data stream.

For internal and external environment in coal mine constantly changing, the safety evaluation method of mine pressure used in this paper is a concept drifting data stream classification algorithm based on double thresholds. The algorithm uses support vector machine as the basic classifier, and uses the Bayesian classifier to filter noise data, also uses double thresholds determined by Hoeffding bounds inequality to detect concept drifts. Mine pressure monitoring data stream is classified by this algorithm, and mine pressure safety can be judged by the results of data stream classification. Experimental results show the algorithm can better detect concept drifts in data stream, and it has better classification accuracy for data stream, and it can be applied to the safety evaluation of mine pressure.

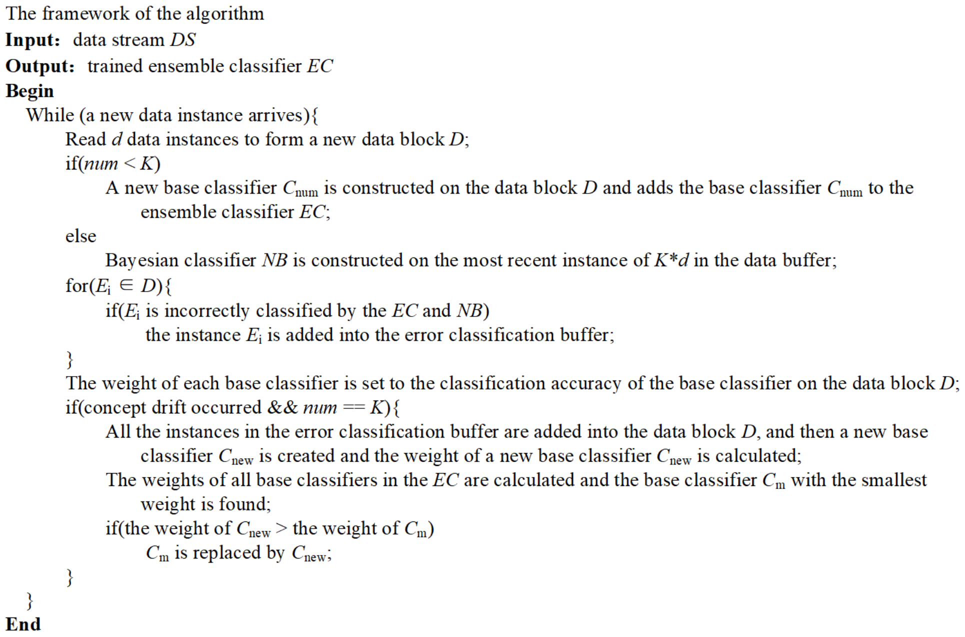
2. Concept drifting data stream classification algorithm

For the problem of concept drifts in data stream, the main problem is how to effectively detect concept drifts, how to quickly adapt to concept drifts and how to reduce the influence of noise data in data stream on classification model. Although there are some algorithms to solve these problems, few algorithms can achieve better balance between the two aspects, especially in the noise data stream classification environment [5]–[9]. The concept drifting data stream classification algorithm based on double thresholds firstly integrated horizontally K support vector machine base classifiers constructed in K data blocks in the data stream buffer, then a Bayesian classifier is constructed for K data blocks, which is used to filter the noise data. There are $K + 1$ classifiers in the model. The weight of each base classifier is dynamically weighted, the weight of which is the classification accuracy of the base classifier in the current data block. The double thresholds determined by the Hoeffding Bounds inequality are used to detect concept drifts. After the concept drift is detected, the base classifier with the smallest classification accuracy is discarded and the classification model is updated.

2.1. The framework of the algorithm

Firstly, the symbols involved in the algorithm are described. D denotes the current data block in the data stream, and E_i denotes the i-th instance in the data block D. EC denotes the set of all SVM base classifiers, C_i denotes the i-th base classifier in EC, K denotes the total number of base classifiers in EC, num denotes the number of base classifiers in EC at present, d represents the total number of instances in a data block.

The framework of the concept drifting data stream classification algorithm based on double thresholds is shown in Fig. 1



```
The framework of the algorithm
Input: data stream $DS$
Output: trained ensemble classifier $EC$
Begin
  While (a new data instance arrives){
      Read $d$ data instances to form a new data block $D$;
      if($num < K$)
          A new base classifier $C_{num}$ is constructed on the data block $D$ and adds the base classifier $C_{num}$ to the
              ensemble classifier $EC$;
      else
          Bayesian classifier $NB$ is constructed on the most recent instance of $K*d$ in the data buffer;
      for($E_i \in D$){
          if($E_i$ is incorrectly classified by the $EC$ and $NB$)
              the instance $E_i$ is added into the error classification buffer;
      }
      The weight of each base classifier is set to the classification accuracy of the base classifier on the data block $D$;
      if(concept drift occurred && $num == K$){
          All the instances in the error classification buffer are added into the data block $D$, and then a new base
              classifier $C_{new}$ is created and the weight of a new base classifier $C_{new}$ is calculated;
          The weights of all base classifiers in the $EC$ are calculated and the base classifier $C_m$ with the smallest
              weight is found;
          if(the weight of $C_{new}$ > the weight of $C_m$)
              $C_m$ is replaced by $C_{new}$;
      }
  }
End
```

Fig. 1. The framework of the algorithm

2.2. The concept drift detection mechanism of the algorithm

The concept drift detection mechanism of the algorithm is based on the double thresholds detection mechanism, which detects different concept drifts from the noise data stream by calculating the classification error rate on the time window. Double thresholds detection mechanism is an effective method for concept drift detection. In contrast to the other double threshold detection methods, this paper uses the double thresholds determined by the Hoeffding Bounds inequality to detect concept drifts, which is inspired by references [10], [11]. In this paper, the method of detecting concept drift is implemented as follows: Firstly, the classification error rate is calculated for the current data block and the previous data block, and then the difference Δe of the classification error rate 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 current data block and e_{b} is the classification error rate of the previous data block.

Assuming 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 drifts 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

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

It can be obtained

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

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 algorithm uses the double thresholds $k_1\varepsilon$ and $k_2\varepsilon$ determined by 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 data stream. If $\Delta e \leq k_1\varepsilon$, the potential concept drift occurs in the 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 data stream.

3. Safety evaluation method of mine pressure based on concept drifting data stream classification

The safety evaluation method of mine pressure based on concept drifting data stream classification is described as follows:

1. Interpolate the missing data of mine pressure monitoring data.

2. Chaotically analyze of mine pressure monitoring data, and reconstruct phase space.

3. Construct an ensemble classifier.

4. Detect concept drifts.

5. If concept drift occurs, return to 3 to reconstruct the classifier; if there is no concept drift, classify the data stream.

6. Classify the mine pressure monitoring data stream.

7. According to the classification results of mine pressure monitoring data stream, the safety evaluation of mine pressure is carried out.

The safety evaluation method of mine pressure based on concept drifting data stream classification is shown in Fig. 2

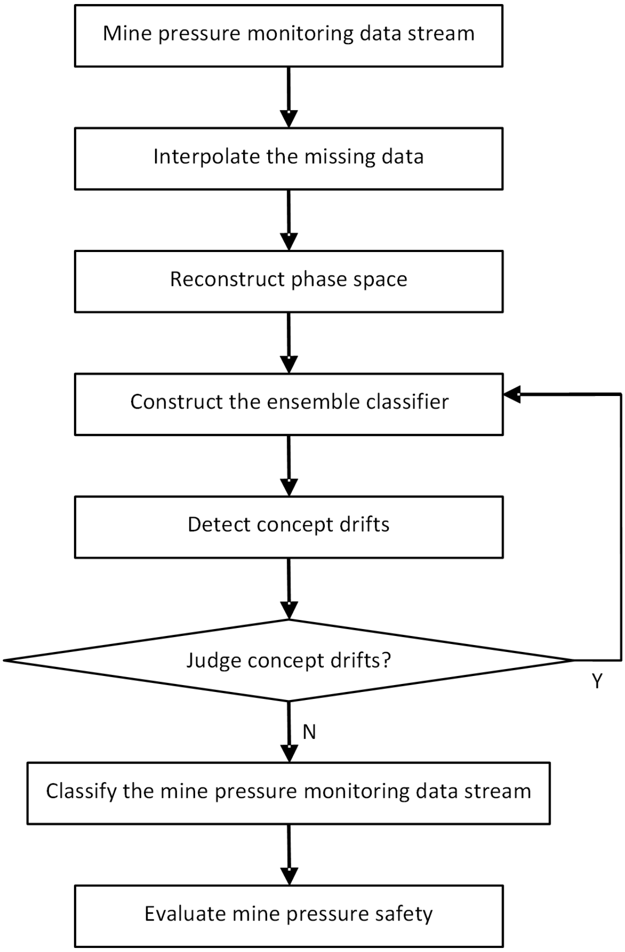

Mine pressure monitoring data stream
Interpolate the missing data
Reconstruct phase space
Construct the ensemble classifier
Detect concept drifts
Judge concept drifts?
Y
N
Classify the mine pressure monitoring data stream
Evaluate mine pressure safety

Fig. 2. The flow diagram of the safety evaluation of mine pressure

Table 1. Statistics of concept drift detection in data sets

Data set	False alarms (%)	Missing
SEA	5.29	0
HyperPlane	6.43	7
KDDCup	5.14	4

4. Experimental results and analysis

In order to verify the validity of mine pressure safety evaluation method based on concept drifting data stream classification, experiments are made in simulated data sets and real mine pressure data sets, and compared it with other data stream classification algorithms [12]-[14].

4.1. Concept drift detection analysis

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

The 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 5.29 %, and the number of undetected concept drifts is 0. On the HyperPlane data set, the probability of false prediction is 6.43 %, and the number of undetected concept drifts is 7. On the KDDCup data set, the probability of false prediction is 5.14 %, and the number of undetected concept drifts is 4. 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.

4.2. Classification accuracy analysis

In order to verify the classification accuracy of the algorithm (denoted as MPSE-CDDSC) proposed in this paper, the mine pressure monitoring data set Mine1 is classified by the MPSE-CDDSC algorithm. The average error rate of the 20 experimental results is 7.65 %. The mine pressure monitoring data set Mine1 are classified by other classical data stream classification algorithms CVFDT, HT-DDM, HT-EDDM and Bag-ASHT [15]–[18]. All experimental results are the average value of 20 times experiments, and the classification error rates are 15.84 %, 14.13 %, 14.58 % and 13.97 % respectively, as shown in Fig. 3. It can be seen from Fig. 3 that the MPSE-CDDSC algorithm used in this paper has better classification accuracy than the other data stream algorithms for mine pressure monitoring data set Mine1.

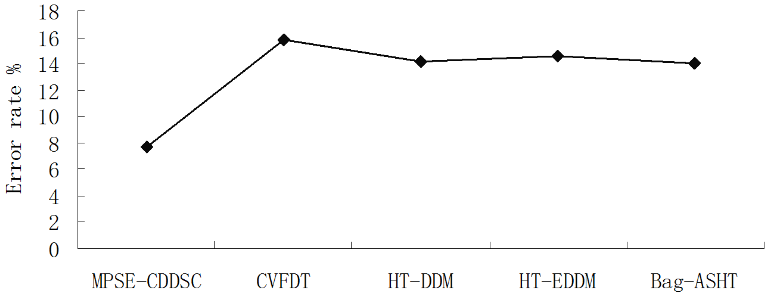


Fig. 3. Error rate of classification in data set Mine1

The mine pressure monitoring data set Mine2 is classified by the MPSE-CDDSC algorithm. The average error rate of the 20 experimental results is 8.46 %. The mine pressure monitoring data set Mine2 are classified by other data stream classification algorithms. All experimental results are the average value of 20 times experiments, and the classification error rates are 16.95 %, 15.87 %, 16.12 %, and 14.98 % respectively, as shown in Fig. 4. It can be seen from Fig. 4 that the MPSE-CDDSC algorithm has better classification accuracy than the other data stream algorithms for mine pressure monitoring data set Mine2.

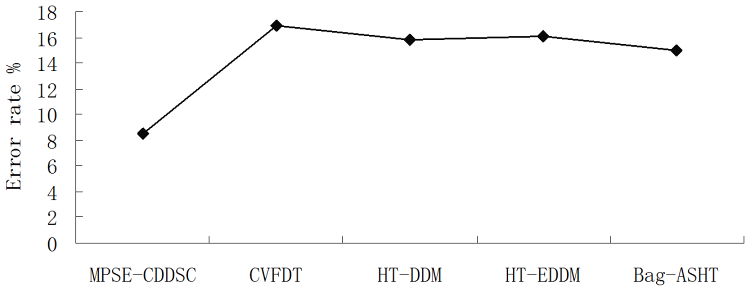


Fig. 4. Error rate of classification in data set Mine2

The mine pressure monitoring data set Mine3 is classified by the MPSE-CDDSC algorithm. The average error rate of the 20 experimental results is 4.38 %. The mine pressure monitoring data set Mine3 are classified by other data stream classification algorithms. All experimental results are the average value of 20 times experiments, and the classification error rates are 10.85 %, 14.68 %, 14.12 % and 13.46 % respectively, as shown in Fig. 5. It can be seen from Fig. 5 that the MPSE-CDDSC algorithm has better classification accuracy than the other data stream algorithms for mine pressure monitoring data set Mine3.

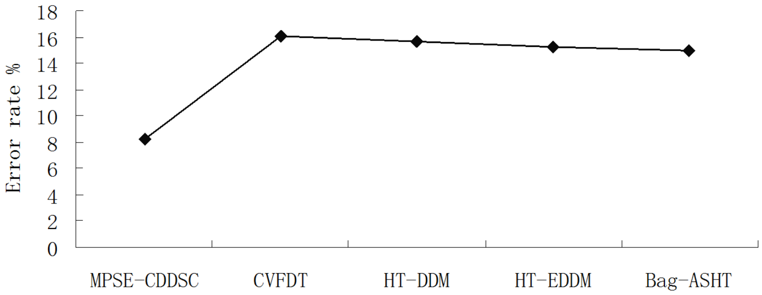


Fig. 5. Error rate of classification in data set Mine3

In order to verify the effectiveness of the MPSE-CDDSC algorithm, in addition to the mine pressure monitoring data sets Mine1, Mine2 and Mine3, the UCI data sets SEA, HyperPlane and KDDCup are classified by the MPSE-CDDSC algorithm and other algorithms CVFDT, HT-DDM, HT-EDDM and Bag-ASHT. Experimental results are shown in Fig. 6. It can be seen from Fig. 6 that the algorithm used in this paper has better classification accuracy in other data sets than other algorithms. Therefore, it shows that the algorithm is effective not only for real mine pressure monitoring data sets, but also for other simulated data sets, and it has better adaptability.

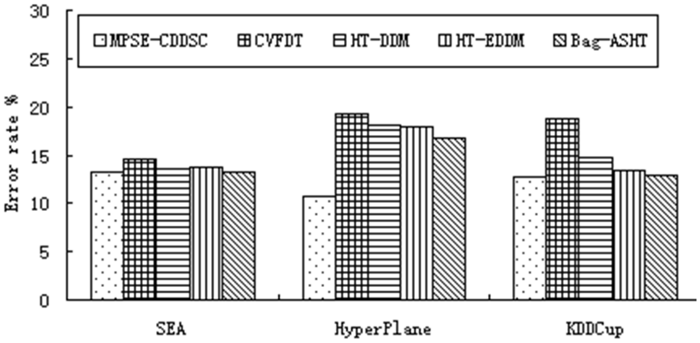


Fig. 6. Error rate of classification in other data sets

It can be seen from the above experimental results that the classification error rate of the MPSE-CDDSC algorithm used in this paper is less than the other classical algorithms in mine pressure monitoring data sets and other data sets. The higher classification accuracy of the algorithm is, and the more accurate safety evaluation of mine pressure is. Therefore, the algorithm can be used to classify mine pressure monitoring data stream to evaluate mine pressure safety.

5. Conclusions

Mine pressure monitoring data is essentially a data stream. With the change of coal mining environment, concept drift is implied in mine monitoring data stream. The problem of mine pressure safety evaluation can regarded as the problem of concept drifting data stream classification, and classification labels are safety and unsafety. For the constant change of the internal and external environment in coal mine, the safety evaluation model of mine pressure used in this paper is a concept drifting data stream classification algorithm based on double thresholds, which uses support vector machine as the base classifier, and uses Bayesian classifier to filter noise data, and uses the double thresholds determined by Hoeffding Bounds inequality to detect concept drift. Mine pressure monitoring data stream is classified by this algorithm, and mine pressure safety can be judged by the results of data stream classification. Experimental results show the algorithm can better detect concept drifts in data stream, and it has better classification accuracy for data stream, and it can be applied to mine pressure safety evaluation. Generally, because the class label is difficult to obtain in the mine pressure monitoring data, this paper will study the problem of incomplete label in mine pressure monitoring data in the future.

References

[1] L. Golab, M. T. Özsu: *Issues in data stream management.* ACM SIGMOD Record *32* (2003), No. 2, 5–14.

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

[3] J. Z. Kolter, M. A. Maloof: *Dynamic weighted majority: A new ensemble method for tracking concept drift.* IEEE International Conference on Data Mining, 19–22 Nov. 2003, Melbourne, FL, USA, IEEE Conference Publications (2008), 123–130.

[4] Q. Zhu, Y. H. Zhang, X. G. Hu, P. P. Li, X. Wu: *A double-window-based classification algorithm for concept drifting data streams.* IEEE International Conference on Granular Computing, 14–16 Aug. 2010, San Jose, CA, USA, IEEE Conference Publications (2010), 639–644.

[5] J. Gama, P. Medas, G. Castillo, P. Rodrigues: *Learning with Drift Detection.* Brazilian Symposium on Artificial Intelligence - SBIA'04, 29 September–1 October 2004, São Luís, Maranhão, Brazil, Advances in Artificial Intelligence (2004), 286–295.

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

[7] P. Zhang, X. Zhu, Y. Shi: *Categorizing and mining concept drifting data streams.* ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 24–27 August 2008, Las Vegas, Nevada, USA, ACM New York (2008), 812–820.

[8] Y. P. Zhang, S. H. Liu: *Ensemble classification based on feature drifting in data streams.* Computer Engineering and Science *36* (2014), No. 05, 977–985.

[9] 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, 193–196.

[10] Y. H. Zhang: *A Study on classification in data stream.* Hefei University of Technology, Dissertation (2011).

[11] P. P. Li: *Concept drifting detection and classification on data streams.* Hefei University of Technology, Dissertation (2012).

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

[13] J. C. Schlimmer, R. H. Granger: *Incremental learning from noisy data.* Machine Learning *1* (1986), No. 3, 317–354.

[14] W. Fan, H. Wang, P. S. Yu, S. Ma: *Is random model better? On its accuracy and efficiency.* IEEE International Conference on Data Mining, 19-22 Nov. 2003, Melbourne, FL, USA, IEEE Conference Publications 51–58.

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

[16] J. Gama, G. Castillo: *Learning with local drift detection.* International Conference on Advanced Data Mining and Applications, 14–16 August 2006, Xi'an, China, Lecture Notes in Computer Science *4093* (2006).

[17] M. Baena-García, J. del Campo-Ávila, R. Fidalgo-Merino, A. Bifet, R. Gavald, R. Morales-Bueno: *Early drift detection method.* Fourth international workshop on knowledge discovery from data streams (2006), 77–86.

[18] A. Bifet, G. Holmes, B. Pfahringer: *Leveraging bagging for evolving data streams.* Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 20–24 September 2010, Barcelona, Spain, Lecture Notes in Computer Science *6321* (2010), 135–150.

Received August 7, 2017