HNUE JOURNAL OF SCIENCE DOI: 10.18173/2354-1059.2019-0045
Natural Sciences, 2019, Volume 64, Issue 6, pp. 173-182
This paper is available online at http://stdb.hnue.edu.vn

BAM: BORDER ADJUSTMENT METHOD IMPROVE THE EFFICIENCY OF IMBALANCED BIOLOGICAL DATA CLASSIFICATION

Nguyen Thi Hong and Dang Xuan Tho
Faculty of Information Technology, Hanoi National University of Education

Abstract. This paper presents a data classification problem and methods to improve imbalanced data classification. Especially, biomedical data has a very high imbalance rate and the sample identification of minority class is a very important. Many studies have shown that border elements are important in imbalanced data classification such as Borderline-SMOTE, Random Under Border Sampling. This paper provides a new method of adjusting data: generating synthetic elements on the borderline of the minority class, identify and eliminate noise elements of the majority class to achieve better classification efficiency. Experimental results of classification of SVM algorithm on six datasets of UCI international standard data warehouse: Blood, Haberman, Pima, Yeast, Ionosphere, and Glass showed that the adjustment of borderline has a positive effect on classification and the results are considered statistically significant.

Keywords**:** Classification, Imbalance Data, BAM.

1. Introduction

Today, data mining is an area of great interest to many scientists, especially in the age of technology 4.0, where problems with large data are popular. In particular, classification is one of the data mining problems with many practical applications such as email classification, financial fraud classification [1], classification of biomedical data [2-4], network intrusion detection [5]. There are many proposed classification algorithms and experiments show that the classification efficiency is good such as: Support Vector Machine (SVM) [6], K nearest neighbors (KNN), decision trees [7]. However, these standard classification algorithms are not highly effective with imbalance data that a class has more elements are called the majority and another class has fewer elements called minority classes. The difference in the number of elements of classes reduces the efficiency of class classification, especially identifying the minority class elements is not good.

Received May 30, 2019. Revised June 13, 2019. Accepted July 20, 2019.

Contact: Dang Xuan Tho, e-mail address: thodx@hnue.edu.vn

For example, for the financial fraud classification problem, the number of transactions in a time unit can be very large, here there is a small number of fraudulent transactions. Finding illegal transactions is very important in financial management. The accuracy of classification may be very high, but the efficiency of fraudulent transactions is very low due to the difference in the quantity between legitimate and illegal transactions. Misidentifying financial transactions, misclassifying emails can do damage on money or data, but in the biomedical field, misidentification can lead to great damage to human health. Therefore, improving classification efficiency in biomedical data is a great interest of many scientists. The problem of determining whether a person being sick or not is the classification problem based on the data collected from the patients' information and symptoms. Obviously, the number of people infected is very smaller than the number of healthy people. Specifically, the Blood dataset has about 748 samples, including 178 samples of minority classes (Positive) and 570 samples of majority class (Negative). If the efficiency of minority classes is 10%, only 18 samples of minority class are correctly identified and 160 samples are misclassified. The samples of minority class are misclassified means 160 disease peoples are diagnosed by the health. This has very serious consequences because patients are not timely detected and treated. Thus, due to the overwhelming number of majority class samples with minority class samples, the classification using traditional algorithms such as KNN, SVM, Naïve Bayes has the efficiency of identifying the minority class not high. Therefore, adjustment methods are needed to increase the efficiency of minority class identification.

There are many approaches to solving imbalanced data classification problems [8]. In which, there are two main approaches: algorithmic approach and data approach. Algorithm-based approach means adjusting standard classification algorithms such as SVM, Decision Trees, KNN to increase the ability to identify minority class elements. Data-based approach means adjusting data to reduce imbalance and increasing classification efficiency when applying standard classification algorithms. In addition, it is possible to combine data adjustment method with other methods such as reduce data dimension, feature selection [9] to increase classification efficiency. The data-based approach has been interested by many scientists and experiments were effective with many imbalance datasets. This approach aims to reduce the imbalance between the majority and minority classes by ways such as increasing the number of minority class elements or reducing the number of majority class elements. Increasing the number of minority class elements is done by duplicating or generating synthetic minority class elements, the method of generating synthetic elements has been proven to be effective as SMOTE [10], Borderline-SMOTE [11] Reducing the number of majority class elements is randomly remove or select the majority class elements to eliminate with some proposed algorithms such as: Condensed Nearest Neighbor Rule (CNN), the Neighborhood Cleaning Rule (NCL), the Tomek links [12].

In this paper, the authors select data-based approach that combines enhanced the minority class elements and reducing the majority class elements, and select data-based approach that combines enhanced the minority class elements and reducing the majority class elements. We realize that the border elements are very important in class identification and there are many noise elements of majority class that enterprise in the area of the minority class, reducing the efficiency of minority class identification. Therefore, we propose a new method of improvement from the Borderline-SMOTE algorithm by only generating synthetic elements of minority of the class on the line connecting two minority border elements, based on the original minority elements and the new minority elements to identify and eliminate noise majority class elements. The use of synthetic elements to erase the noise elements is the difference between our method and previous methods.

2. Content

Currently, SVM classification algorithms are evaluated as highly effective classification algorithms by selecting the classification line with the largest marginal distance from it to the border elements of the minority and the majority class. Thus, it can be seen that the boundary elements of classes are very important in class identification. Therefore, we choose approach base on data by defining border elements.

Borderline-SMOTE is a well-known and effective method to adjust the imbalance by adding synthetic elements around the boundary. However, we found that, in some cases, this method still has disadvantages because the synthetic elements of the minority class are added to be interspersed in the majority class area and there are still some noise elements of majority class in minority class area. That reduces the efficiency of minority class identification.

On that basis, we propose a new algorithm BAM (Borderline Adjustment Method) that improves from the Borderline-SMOTE algorithm by generating synthetic elements on the line between two border elements, use the original minority elements and synthetic elements to determine and remove majority class noise elements. Previous data preprocessing algorithms are often not interested in using newly synthetic elements. This is a new point in our proposed method. Thus, the efficiency of minority identification increases, but the majority class recognition efficiency is not reduced.

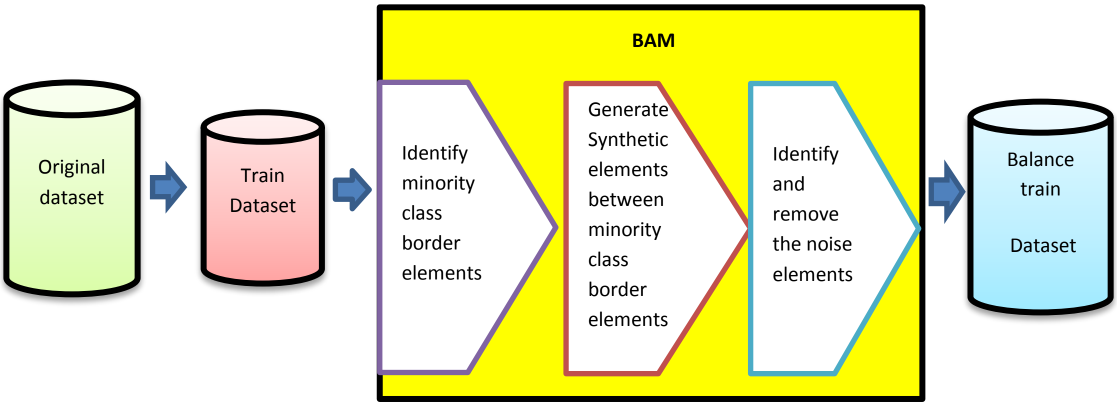
BAM
Original dataset
Train Dataset
Identify minority class border elements
Generate Synthetic elements between minority class border elements
Identify and remove the noise elements
Balance train Dataset

Figure 1. Diagram of preprocessing data by BAM method

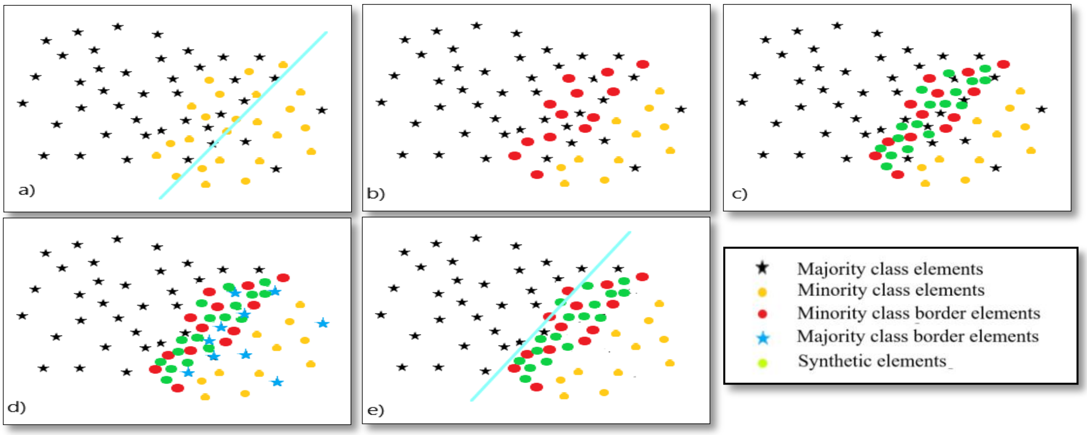

Figure 2. (a) The original dataset with the majority class elements are shaded in black, the minority class elements are colored in yellow. (b) The original dataset with minority class border elements is shown in red. (c) The dataset after generating the green minority class synthetic elements. (d) The data set defines the blue majority class border elements. (e) Dataset after BAM implementation

Figure 2 illustrates visually determining border elements of the two classes and generates the minority class synthetic elements and removes the border and noise elements of the majority class. Figure (a) is a description of the original dataset, in which yellow elements are minority class elements, black elements are majority class elements. Due to the overwhelming of majority class elements, the classification line is pushed towards the minority class, many minority class elements are mistakenly identified. Figure (b) marks the red minority class border elements. Figure (c) simulates the generation of synthetic elements (green) between minority class border elements. Figure (d) shows the determination of the majority class border elements on the dataset after the addition of artificial elements, the majority class elements within the minority are marked as border elements (blue). Figure (e) describes the data set after the addition of synthetic elements and the deletion of the majority class the border elements (noise) in the minority class areas. The adjustment reduces the overlap between the two classes and reduces the imbalance so the classification line is pushed towards the majority class, the efficiency of minority class identification increases and the effectiveness of the majority class recognition is not reduced.

**** BAM algorithm***

Input: Training data D consists of P minority class elements (Positive) and N majority class elements (Negative), M% (synthetic elements generated), K1 - the number of neighbors to determine the minority class border element, K2 - the number of neighbors to determine the majority class border element (select K2 greater than K1 to only remove noise elements and elements in the minority class area).

Output*:* The training data has been added the minority class synthetic elements and removed the majority class border elements (noise elements).

```
Step 1: Identify the border elements of the minority class
  - Calculate $K_1$ nearest neighbors of minority class elements in all training data D.
  - Count the nearest neighbors is the majority class element $K_1'$
  - If $K_1/2 \leq K_1' \leq K_1$, that element is the border element of the minority class.
Step 2: Generate synthetic elements of the minority class
  - For each border element $P_i$ of the minority class, randomly select $M\%$ in $K_1$ nearest neighbor elements of the minority border elements set: $P_j$
  - Calculate: $Dif = P_i - P_j$
  - Randomly select $Gap$ in [0, 1]
  - Generate synthetic element: $Synthetic = P_i + Gap*Dif$
  - $D' = D \cup Synthetic$
Step 3: Determine the noise and border elements of the majority class, then remove from the training data
  - Calculate $K_2$ nearest neighbors of majority class elements in all $D'$ training data.
  - Count the nearest neighbors that are the minority class element: $K_2'$
  - If $K'_2 \geq 3/4\, K_2$, that element is the majority of the border element;
  - Remove these elements from the training data set;
```

2.1. Experiments

In order to make statistics and evaluate the effectiveness of classification, we name the minority class label is Positive and the name of the majority class label is Negative. The confusion matrix [13] is used in determining the effectiveness measures for classification.

Table 1. Confusion matrix

	Predicted Positive	Predicted Negative
Reality Positive	TP	FN
Reality Positive	FP	TN

In Table 1, TP is the number of actual Positive class samples correctly predicted as Positive, FN is the number of actual Positive samples predicted as Negative, FP is the number of actual Negative samples predicted as Positive and TN is the number of actual Negative samples that are correctly predicted to be Negative.

Some measurements are determined based on the confusion matrix [13]:

$$TPrate = recall = TP/(TP + FN) \tag{1}$$

$$TNrate = TN/(TN + FP) \tag{2}$$

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{3}$$

With balance datasets, Accuracy is the measure used to evaluate classification performance. However, if the dataset is highly imbalanced, the value of Accuracy is high, but the rate of correctly identified minority class elements can be very low. Therefore, the G-mean measurement is used to evaluate the efficiency of classification with imbalanced datasets. G-mean is determined based on two values of TP rate and TN rate, according to the formula (4). When the rate of identifying majority and minority classes is high, the G-mean value will be high.

$$G - mean = \sqrt{TPrate.TNrate} \tag{4}$$

To evaluate the effectiveness of the algorithm, we experiment on R, Perl languages [14] with datasets UCI: Blood, Haberman, Pima, Yeast, Ionosphere, Glass (Table 2). After that, we compare the efficiency of the classification of datasets adjusted by proposing method with the original datasets and the datasets adjusted by the Borderline algorithm-SMOTE1, Borderline-SMOTE2, Random Border Undersampling (RBUS) [15] and Random Border Oversampling (RBOS) [16] because these methods adjust data based on border elements.

Table 2. Imbalanced datasets

Dataset	Number of elements	Number of attributes	Imbalanced ratio	Percentage minority class
Blood	748	4	1:3	23.8%
Haberman	306	3	1:3	26.47%
Pima	768	8	1:2	34.9%
Yeast	1484	8	1:28	3.4%
Ionosphere	351	34	1:2	35.9%
Glass	214	9	1:6	13.55%

In this paper, we use the SVM algorithm (Kernlab package in R [17]) to classify the original datasets and after adjusting. For objective evaluation and no over-fitting problem, the authors conducted cross validation method with 20 times 10-fold [18]. With a 10-fold implementation, the dataset is divided into 10 sections, each of which is selected as test data and the remaining 9 are training data. After classification, the G-mean value is calculated for each fold and the average G-mean of 10-fold. The final classification effectiveness is defined as the average value of the G-mean values when performed 20 times 10-fold. The method of calculating the T-test value (p-value < 0.05) is used to evaluate whether the G-mean values are statistically significant.

Figure 3 is a diagram showing the change of G-mean value calculated when performing classification on datasets with parameters N increased from 100% to 500% (the ratio of synthetic elements generated from the number of minority class border elements). With the Blood dataset and the Pima dataset, the G-mean values increase slowly when changing N from 100% to 500%. Haberman dataset has classification efficiency, increased quickly and peaked at parameter N = 300%. With Yeast dataset, the G-mean value skyrocket with N = 200%, after that, increases steadily and peaked when N = 500%. Ionosphere and Glass datasets, the classification efficiency is relatively stable when changing parameters.

Because the methods have different parameters, in Figure 4, we use a column chart to compare the calculated G-mean value when classify the original datasets with datasets adjusted by methods Borderline-SMOTE1 (BSM1), Borderline-SMOTE2 (BSM2), Random border Undersampling (RBUS), Random border Oversampling (RBOS) and BAM. Of the six datasets, there are five datasets Blood, Haberman, Pima, Yeast, Ionosphere have the best effective classification when adjusting by BAM. Thus, experiments show that the adjustment of data by the BAM method is better than the remaining methods.

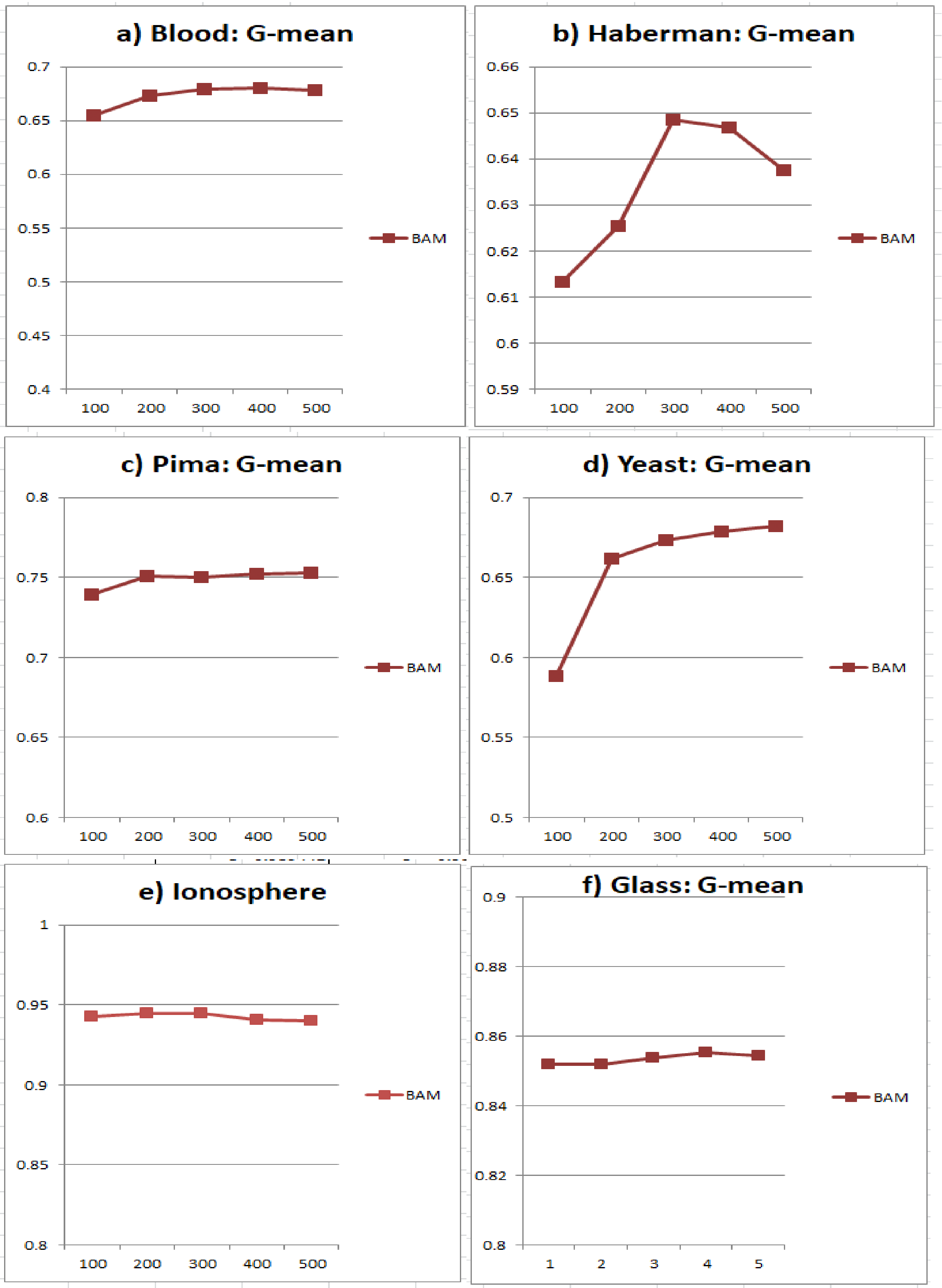

Figure 3. The diagram shows the change of G-mean value measured when changing parameter N% by BAM algorithm

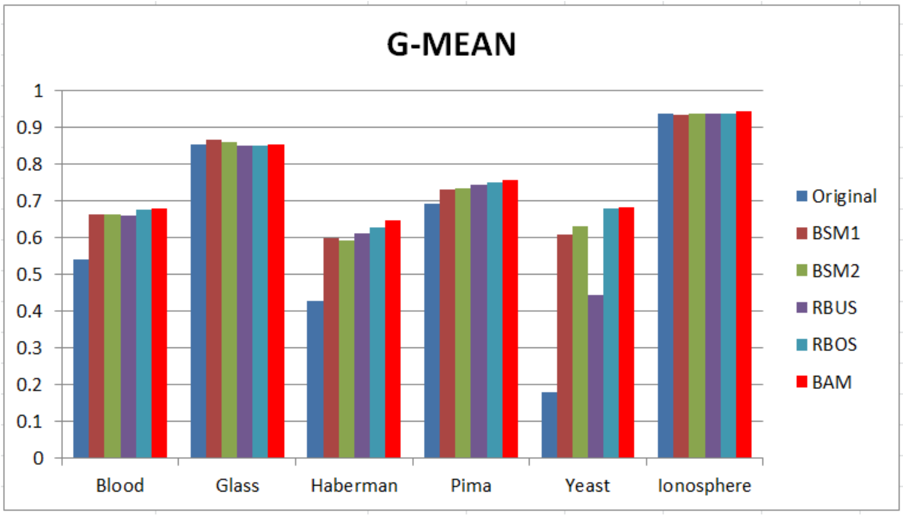

Figure 4. Chart show comparing the G-mean value calculated when classifying the original datasets with the datasets adjusted by Borderline-SMOTE1 (BSM1), Borderline-SMOTE2 (BSM2), RBUS, RBOS and BAM

Table 3 is a statistic on the number of minority class border elements, the number of synthetic elements generated, the number of majority class border elements deleted, the number of minority class elements and the number of majority class elements obtained after adjusting. Most datasets adjusted by BAM that have the number of minority class elements greater or approximating the number of majority class elements obtained better classification results. The Yeast dataset, after addition of synthetic elements five times the number of border elements, still has a large imbalance, but the classification efficiency is still significantly improved. With Glass dataset, only two border elements are identified, so the number of synthetic elements is very small. This has made the effectiveness of minority class identification not improved.

Table 3. Statistics of the number of border elements of the classes, the number of synthetic elements generated, the number of elements deleted and the total number of majority and minority class elements obtained of datasets

Dataset	**Number of minority class minority elements**	**Number of synthetic elements**	**Number of majority class deleted**	**Number of minority class obtained**	**Number of majority class obtained**
Blood	93	372	57	550	513
Haberman	53	159	54	240	201
Pima	100	500	67	768	433
Yeast	30	150	143	201	1290
Ionosphere	42	128	22	252	203
Glass	2	2	20	31	165

Table 4 shows the calculated P-value values when compare the average G-mean values of Blood, Haberman, Pima, Yeast, Ionosphere, and Glass datasets. Most of P-value values are less than 0.05, showing that these values are statistically significant and classification efficiency after adjusting the datasets by BAM better than the original datasets and datasets adjusted by Borderline-SMOTE1 and Borderline-SMOTE2, RBOS, RBUS algorithms.

Table 4. The p-values compare the G-mean value of the datasets after applying the BAM algorithm with the original data sets and the datasets after applying the Borderline-SMOTE, RBOS, RBUS

Dataset		Original	Borderline-SMOTE1	Borderline-SMOTE2	RBUS	RBOS
Blood	BAM	< 2.2e-16	0.00004873	0.001417	3.47e-07	0.1477
Haberman	BAM	< 2.2e-16	4.54e-09	6.1e-11	2.47e-11	1.40e-06
Pima	BAM	< 2.2e-16	1.01e-09	2.337e-08	3.26e-08	0.001268
Yeast	BAM	< 2.2e-16	4.97e-06	6.39e-05	< 2.2e-16	0.355
Ionosphere	BAM	0.000238	0.000747	0.001599	6.43e-06	0.000484
Glass	BAM	1.09e-10	2.49e-07	0.000939	0.1287	0.03546

3. Conclusions

In this paper, we presented about imbalanced data classification problem and proposed BAM data adjustment method based on the Borderline-SMOTE algorithm. The method increases the number of minority class border elements and deletes the majority class noise elements in the minority class area to reduce the rate of imbalance and increases the efficiency of minority class identification. The experiment showed that when using the new method, classifying effect on the Blood, Haberman, Pima, Yeast, Ionosphere, and Glass datasets were significantly improved and most value was statistically significant.

The reduction of imbalances in classification problem still needs to be further improved to increase the efficiency of data classification, so this will still be the main research direction of the authors in the future. In addition, to increase the efficiency of minority class identification, we will combine with some other methods such as clustering, reducing the dimensionality of attributes.

Acknowledgement. The research was supported by the Vietnam Ministry of Education and Training, project B2018-SPH-52.

REFERENCES

[1] Mahmoudi, Nader, and E. Duman, 2015. *Detecting credit card fraud by modified Fisher discriminant analysis*. Expert Syst. with Appl. *42.5*, pp. 2510-2516.

[2] C. Wang and *et al.*, 2015. *ImDC: an ensemble learning method for imbalanced classification with miRNA data*. Genet. Mol. Res. 14.1, pp. 123-133.

[3] Krawczyk, Bartosz, and *et al.*, 2014. *Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Appl. Soft Comput.*, Vol. 38, pp. 714-726.

[4] Lang, Philipp, and *et al*., 2016. *Methods and devices for evaluating and treating a bone condition on x-ray image analysis*. U.S. Pat., No. 9, pp. 275-496.

[5] Suthaharan and Shan, 2014. *Big data classification: Problems and challenges in network intrusion prediction with machine learning*. ACM Sigmetrics Perform. Eval. Rev., pp. 70-73,

[6] R. Batuwita and V. Palade, 2013. *Class Imbalance Learning Methods for Support Vector*. Imbalanced Learn. Found. Algorithms, Appl., pp. 83-100.

[7] Lior and Rokach, 2014. *Data mining with decision trees: theory and applications*. World Sci., Vol. 81.

[8] H. HE and E. a. *Garcia, 2010. Learning from Imbalanced Data Sets.*, IEEE Trans. Knowl. Data Eng., Vol. 21, No. 9, pp. 1263-1264,

[9] M. L. Bermingham *et al.*, 2015. *Application of high-dimensional feature selection: evaluation for genomic prediction in man.* Sci. Rep, Vol. 5: 10312.

[10] N. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 2002. *Synthetic minority over-sampling technique.* J. Artif. Intell. Res., Vol. 16, pp. 321-357.

[11] H. Han, W. Wang, and B. Mao, 2005. *Borderline-SMOTE: A New Over-Sampling Method in ICIC*, pp. 878-887.

[12] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, 2008. *On the Class Imbalance Problem,* Int. Conf. Nat. Comput, Vol. 4, pp. 192-201.

[13] Y. Sun, A. K. C. Wong, M. Kamel, and S., 2008. *Classification of Imbalanced Data: a Review*. Int. J. Pattern Recognit. Artif. Intell., Vol. 23, No. 4, pp. 687-719.

[14] B. d F. R. L.Schwartz, T. Phoenix, 2009. *Learning Perl*. O'reilly.

[15] N. M. Phuong, T. T. A.Tuyet, N. T. Hong, and D. X. Tho, 2015. *Random Border Under-sampling: a new method reduces random elements on the boundary in imbalanced data*. Fundamental And Applied IT Research (FAIR), DOI: 10.15625/vap.2015.000200, pp. 612-619 (in Vietnamese).

[16] B. D. Hung, V. V. Thoa, D. X. Tho, 2017. *Random Border-Over-Sampling: New border algorithm adds random samples on the boundary in imbalanced data.* Journal of Science and Technology on Information and Communications, No. 01 (CS.01) 2017, pp. 45-49 (in Vietnamse).

[17] A. Karatzoglou and *et al.*, 2015. *Package Kernlab Version 0.9-22. An S4 Package for Kernel Methods in R. Reference Manual.* J Stat Softw, pp. 1-20.

[18] Arlot, Sylvain, and A. Celisse, 2010. *A survey of cross-validation procedures for model selection.* Stat. Surv., Vol. 4, pp. 40-79.