Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 15, no. 8, August 2019, pp. 2173-2181
DOI: 10.23940/ijpe.19.08.p17.21732181

Imbalanced Data Optimization Combining K-Means and SMOTE

Wenjie Li*

Hebei Vocational and Technical College of Building Materials, Qinhuangdao, 066000, China

Abstract

With the wide application of imbalanced data processing in various fields, such as credit card fraud identification, network intrusion detection, cancer detection, commodity recommendation, software defect prediction, and customer churn prediction, imbalanced data has become one of the current research hotspots. When classifying imbalanced data sets, aiming at the problems of low classification accuracy of negative class samples in the random forest algorithm and marginalization for selecting new samples in the SMOTE algorithm, a new algorithm, KMS_SMOTE, is proposed to deal with imbalanced data sets. In order to avoid the problem of marginalization of new samples, the K-Means algorithm is used to classify the negative class samples to obtain the centroid of the negative class samples, and then the new data set is obtained by selecting the samples near the centroid. Finally, in order to verify the effect of the KMS_SMOTE algorithm, it is compared with the SMOTE algorithm on the data sets of UCI machine learning. The experimental results show that the KMS_SMOTE algorithm effectively improves the classification performance of the random forest algorithm on the imbalanced data set.

Keywords: imbalanced data; random forest; SMOTE; K-Means; classification

(Submitted on June 11, 2019; Revised on July 7, 2019; Accepted on August 6, 2019)

1. Introduction

Imbalanced data sets [1-2] refer to data sets in which the number of samples in each class is different and the sample size differs greatly. That is, the number of samples in one certain class in the data set is particularly large, and the number of samples in another class is particularly small. Among them, the class with a large number of samples is called the majority class, and the class with a small number of samples is called the minority class. In the binary classification problem, the majority class is called the positive class, and the minority class is called the negative class.

There are many imbalance problems in real life. For example, in the prediction of whether to replace mobile phones, most users will continue to use their original mobile phone, while only a small number of users will choose to replace their mobile phone. In transmission line fault identification, most of the lines are in good operation, and only a small number of lines appear faulty. In natural fire prediction, fire will happen only in a few cases. In cancer identification and diabetes diagnosis, the majority of people are healthy, and only a few people will get sick. In the prediction of game user loss, the vast majority of players will continue to play the game, and only a very small number of players will stop playing. In the safety prediction of civil aviation, most flights are safe, while only a few cases are unsafe. In cyber security prediction, the vast majority of communication is secure, and only a few involve network attacks.

In dealing with the classification problem of imbalanced data sets, the traditional classifier has a good prediction effect on the positive class but a poor prediction effect on the negative class. As the degree of data imbalance increases, the classifier performance will drop significantly. However, in many cases, it is desirable for the classifier to have a better predictive effect on the negative class, such as when cancer patients are detected in cancer detection [3], when malicious requests can be accurately identified in network attacks [4], and when customer loss can be accurately predicted in customer churn prediction [5]. Therefore, how to improve the performance of classifiers on imbalanced data sets has attracted the attention of many scholars and become one of the current research hotspots.

* Corresponding author.
E-mail address: 1037700909@qq.com

To solve the problem that traditional classifiers cannot accurately predict negative class samples, scholars have proposed some improvement methods, which can be divided into two categories: (1) solving imbalanced problems by improving classifiers and (2) solving imbalanced problems by modifying data sets. In the first category, cost-sensitive learning, feature selection, single-class learning, and other methods are usually adopted to improve the classification algorithm, so that it can adapt to the imbalance of data sets and achieve the required performance. In imbalanced data sets, the cost of positive and negative prediction errors is different. According to the different costs, some improvements are made to the classifier to improve its prediction effect on imbalanced data sets. In the second category, random over-sampling and random under-sampling are usually used to modify data sets. In other words, sampling techniques are used to increase the number of negative class samples or reduce the number of positive class samples, so that the numbers of the two classes' samples are balanced.

The SMOTE algorithm is one of the classic techniques for processing imbalanced data sets using random over-sampling techniques. This algorithm selects new negative class samples from two adjacent original negative class samples. However, when these two original negative class samples are at the boundary of the negative class sample set, the new negative class samples obtained must also be at the edge of the negative class sample set. When the new negative class sample is closer to the boundary of the positive and negative class sample sets, the difficulty of dividing the boundary between positive and negative class samples will be greater. Therefore, it is necessary to propose more reasonable methods for imbalanced data sets.

This paper mainly studies the classification method of random forest on imbalanced data sets. Aiming at the marginalization problem of the SMOTE algorithm when selecting new negative class samples, the new data pre-processing algorithm KMS_SMOTE is proposed. The main idea of the algorithm is to obtain two samples close to the centroid in the original negative class by K-Means, select a new negative class sample between them and other samples, and then obtain the new data set from the new negative class samples set by SMOTE. Finally, the random forest model is trained on the new data set. In the process of data set balancing, the KMS_SMOTE algorithm preserves the sample distribution information of the original data set, which improves the performance of the random forest algorithm on the imbalanced data set.

The main contributions of this paper are summarized as follows. (1) A new sampling algorithm KMS_SMOTE is proposed, which combines the K-Means algorithm with the SMOTE algorithm. When the data set is balanced, the algorithm not only optimizes the sampling results of the new negative class samples, but also preserves the samples distribution information on the original data set, so that the performance of the random forest algorithm on the imbalanced data set can be improved. (2) Experiments are carried out on the data sets of UCI. The experimental results show that the improved method is accurate and effective.

The remainder of this paper is organized as follows. Section 2 discusses the related work. Section 3 introduces the SMOTE and K-Means algorithms. The data pre-processing algorithm KMS_SMOTE is proposed formally in Section 4. Section 5 checks the validity and effectiveness of the algorithm through experimentation. The last section concludes the paper and forecasts the future work.

2. Related Work

Since the problem of low accuracy of classification for negative class samples exists when the random forest model [6] is trained on an imbalanced data set, many scholars firstly process the data to improve the performance of the random forest algorithm. Relevant research is described in this section.

The simplest over-sampling is random over-sampling. To overcome the disadvantage of over-fitting, the SMOTE algorithm was proposed Chawla in 2002 [7]. The algorithm first finds the negative class samples that are close to all negative class samples clustered in the samples set and then generates a few new class samples by interpolating between them to reduce the imbalanced degree of the data set. Aiming at the problem of excessive boundary fitting of negative class samples in the SMOTE algorithm, Han proposed the data pre-processing algorithm BSMOTE (Borderline-SMOTE) [8]. This algorithm is suitable for the case where the samples of the negative class boundary region are more likely to be misclassified, and it over-samples in the appropriate region of the negative class boundary to make the newly obtained samples more efficient. This over-sampling area needs to be determined manually, so there are still some defects in the algorithm. To solve the problem of relatively centralized samples synthesized by the SMOTE algorithm, Dong proposed the Random-SMOTE algorithm, which applies the SMOTE algorithm in different directions for each negative class sample [9]. In order to improve the quality of negative class samples, Wang improved the SMOTE algorithm with support calculation and roulette technology [10]. Based on the Bootstrap resampling technique, Thanathamathee proposed a synthetic boundary negative class sampling algorithm [11]. In view of the different causes of data imbalance, Vorraboot proposed soft-hybrid to

solve the imbalance of classes [12]. Li proposed the PCBoost algorithm, which balances training data set by increasing the number of negative class samples in iterations. After the formation of sub-classifiers, it deletes the new negative class samples with classification errors and combines multiple base classifiers to solve the non-equilibrium problem. The algorithm has good generalization ability [13]. Yun proposed an over-sampling algorithm to automatically select the number of negative class samples according to the size of the neighborhood [14].

The commonly used under-sampling methods include the neighborhood cleaning rule, one-side selection, compressed nearest neighbor, and Tomeklinks. In view of the problems in the classification of imbalanced data, Wu [15] proposed a data pre-processing method based on NCL (neighborhood cleaning rule) technology. On this basis, the random forest algorithm was used to train the model. The experimental results show that the model learned on the processed training set has better classification performance. Devi proposed an improved Tomeklinks algorithm [16] to solve the problem of boundary and outliers in positive class samples. Xue [17] proposed an improved algorithm based on SMOTE that integrates under-sampling and over-sampling. It selects the nearest neighbor samples according to a certain method and uses a certain strategy to select new negative class samples. Random over-sampling and random under-sampling have some problems. The former increases the number of negative class samples, changes the balance of the data set by repeated sampling, and alters the distribution of samples in the original data set, thus affecting the performance of the training model. The latter selects only some samples from the original positive class and loses data, which results in incomplete information of the new data set and thereby affects the performance of the final model.

In cost-sensitive research, Fan proposed the AdaCost algorithm to solve the problems in the Boosting method [18]. The algorithm is superior to other methods in accuracy and recall rate evaluation. Wang proposed a method of updating sample weights based on gradient weighting and direct weighting [19]. In order to improve the detection rate of a few classes, Cheng proposed a cost-sensitive LDM method [20]. Datta solved the non-equilibrium problem by combining marginal transfer and cost sensitivity in support vector machines [21]. Du [22] proposed a cost-sensitive random forest algorithm, which combines the cost-sensitive decision tree algorithm with the random forest algorithm to deal with the classification problem of imbalanced data. It first randomly samples the training set to obtain multiple bags, establishes a cost-sensitive decision tree in all the bags, and then integrates them with certain strategies to obtain the final integrated learning algorithm. In integrated learning, Seiffert proposed an USBoost algorithm based on random under-sampling and the AdaBoost algorithm [23]. Galar improved the classification efficiency by selecting the base classifier [24]. Kim changed the evaluation criteria of the weak classifier from the error rate to the geometric mean of the positive and negative class samples error rate and, combined with the sampling algorithm, proposed the GMBoost algorithm to solve the non-equilibrium classification problem [25]. Hu combined the random under-sampling and SMOTE algorithms to solve the problem that integrated learning is not efficient for imbalanced data [26].

In one class learning [27], the negative class samples are usually selected according to different classification objectives, and then the data set composed of this negative class is used to train the classifier, which is only used to identify the negative class samples in the classification. Wang [28] proposed a new data pre-processing algorithm based on the SVM classifier. The algorithm marks the misclassified samples in boosting iteration and randomly selects some new samples of the same class as the classified error samples between the classified samples and their neighbors. Finally, the newly obtained samples and the original training set are combined to obtain a new data set, and the model is trained on the new data set to improve the recognition rate of negative class samples.

For processing imbalanced data, the above methods still have certain deficiencies. In this paper, the marginalization problem existing in the SMOTE algorithm when selecting new negative class samples is improved to enhance the classification performance.

3. Preliminary

3.1. Smote Algorithm

The SMOTE algorithm is an improved algorithm based on random over-sampling technology. Its basic idea is to change the sample distribution by synthesizing new negative class samples artificially. The basic principle of this method is to interpolate the negative class samples that are close to each other in order to generate new negative class samples. Here is a brief introduction of the key steps of the SMOTE algorithm.

Given a data set D, assuming that the original negative class samples number in the training set is n, the $m \times n$ new negative class samples are selected by the SMOTE algorithm from two adjacent original negative class samples. m must be a positive integer; if $m < 1$, then force $m = 1$. In the original negative class, a sample x_i is taken, and its eigenvector is $f_j(x_i)$, j

$\in \{1, 2, \cdots, J\}$. The specific process of the SMOTE algorithm is as follows.

(1) Take any sample x_i and find its K neighbours by some measure (e.g. Euclidean distance), which are marked as x_{ik}, $k \in \{1, 2, \cdots, K\}$.

(2) A sample is randomly selected among the K neighbours, and a random number ε between 0 and 1 is generated. Thus, a new negative class sample p_{ik} is obtained, as shown in Equation (1).

$$p_{ik} = x_i + \varepsilon \times (x_{ik} - x_i) \tag{1}$$

(3) Step (2) is repeated m times to get m new negative class samples.

(4) Repeat steps (1)-(3) above for the other n-1 samples in the original negative class. The $m \times n$ new negative class samples can be obtained.

It can be seen that the new data set does not have duplicate samples, and the new negative class samples contains the characteristics of multiple original negative class samples. Therefore, the new data set obtained by SMOTE can better reflect the information of the original data set, thereby improving the performance of the classifier.

3.2. K-Means Algorithm

K-Means is one of the classical clustering algorithms. For the given data set and clusters number K, all samples are divided into K classes. The specific process of the algorithm is as follows.

(1) Initialize the constant K and randomly select K initial samples in the data set as the centroid.

(2) Calculate the distance between the sample and each centroid, and classify the sample and the centroid with its minimum distance into one class.

(3) Recalculate the centroid of each class.

(4) Repeat steps (2) and (3) until the centroid no longer changes.

(5) Output the centroids of the K classes and the corresponding class members.

Therefore, the K-Means algorithm must calculate the distance between the sample and each centroid repeatedly. As the amount of data increases, the efficiency of the algorithm increases exponentially.

4. KMS_SMOTE Algorithm

In view of the shortcomings of random forest in dealing with imbalanced data and the marginalization problem of the SMOTE algorithm in selecting new negative class samples, this paper proposes the KMS_SMOTE algorithm. This algorithm can greatly improve the imbalance of data sets, and further improve the performance of the random forest algorithm for the classification of imbalanced data sets.

The core idea of the algorithm is as follows. Firstly, the K-Means algorithm is used to classify the original negative classes and calculate the centroid of each class. Secondly, new negative class samples are selected between the two centroids and other samples to make the selected new samples close to the centroid of the original negative class. Finally, the improved method is used to obtain a new data set on the new negative class samples set. To facilitate the description, the calculation methods of class centroid and new negative class samples are given.

Defintion 1 (Class Centroid) Given training set $T = \{x_1, x_2, \cdots, x_n\}$, the sample size is n. If each sample x_i has r attributes, the eigenvector of the sample is denoted as $f(x_i) = \{f_{i1}, f_{i2}, \cdots, f_{jr}\}$; if the sample is divided into L classes, take a class $S_l = \{x_{l1}, x_{l2}, \cdots, x_{ly}\}$, ($l = 1, 2, \cdots, L$, $y < n$), and the centroid of this class is $X_i^{centroid}$, as shown in Equation (2).

$$X_i^{centroid} = (\frac{1}{y}\sum_{i=1}^{y} f_{i1}, \frac{1}{y}\sum_{i=1}^{y} f_{i2}, \ldots, \frac{1}{y}\sum_{i=1}^{y} f_{ir}) \tag{2}$$

In Section 3.1, the method of acquiring new negative class samples using the SMOTE algorithm was introduced. In order to make the new negative class samples approach the centroid of the original negative class purposefully and avoid the marginalization problem, an improved calculation method of the new negative class samples is given as follows.

Defintion 2 (New Negative Class Samples) Given the original negative class sample training set and taking any training samples set $S_l = \{x_{l1}, x_{l2}, \cdots, x_{ly}\}$ in a certain class, the new negative class sample is recorded as p_{ik}, ($k = 1, 2, \cdots, m$), and the calculation method is shown in Equation (3).

$$p_{ik} = x_{li} + \varepsilon \times (X_l^{centroid} - x_{li}) \tag{3}$$

Where $X_l^{centroid}$ is the centroid of the samples and ε represents a random number between 0 and 1.

The KMS_SMOTE algorithm is shown Algorithm 1. Based on the above two definitions, the algorithm effectively solves the marginalization problem of the SMOTE algorithm, which not only ensures the original content in the data set but also improves the imbalanced problem of the data set, thereby improving the performance of the random forest algorithm.

Algorithm 1 KMS_SMOTE

Input: The data set D

Output: The new data set D'

(1) The data set is grouped according to classes, and the original negative class samples are obtained.

(2) The original negative class sample set is classified into two classes by K-Means, and they are clustered into two classes, S_1 and S_2, and their corresponding centroids are $X_1^{centroid}$ and $X_2^{centroid}$, respectively.

(3) Add the new negative class sample p_{ik} according to Equation (3). Random sampling is performed on the line between the sample in S_1 and $X_1^{centroid}$ and on the line between the sample in S_2 and $X_2^{centroid}$. Thus, a new negative sample near the centroid point $X_l^{centroid}$ is obtained, and the specific position of the sample is determined by the random number ε.

(4) Based on the new negative class sample set obtained in step (3), a new data set is obtained by SMOTE.

The specific process of the KMS_SMOTE algorithm is shown in Figure 1. The algorithm processing ends, and the random forest model can be obtained by training the new data set. It can be seen that the core of the KMS_SMOTE algorithm is the central point of the two classes, so that new samples can be selected in the central region of the negative class samples set. Therefore, the algorithm avoids the problem of new sample marginalization and can greatly improve the imbalanced data set, so that the performance of the random forest algorithm in imbalanced data set classification is further improved.

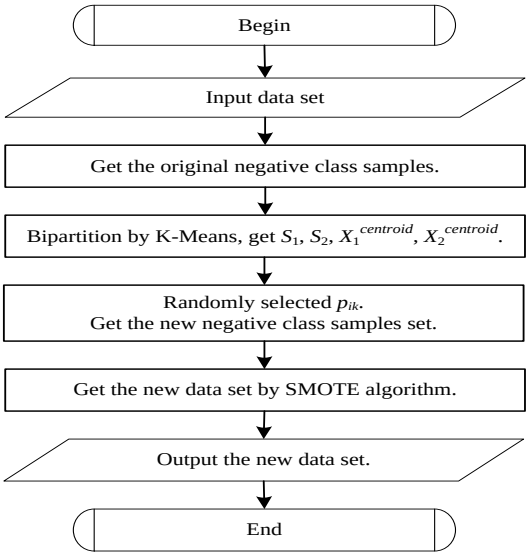

Figure 1. The process of algorithm KMS_SMOTE

5. Experiments

In order to verify the validity of the KMS_SMOTE algorithm, the original data set of UCI, the UCI data set processed by the SMOTE algorithm, and the UCI data set processed by the KMS_SMOTE algorithm are used to perform the binary classification experiment based on random forest. The analysis is as follows.

5.1. Experiment Data Set

The experimental data is four imbalanced data sets in the UCI machine learning database, namely pima, parkinsons, vertebral, and ionosphere. The basic information of these four data sets, as shown in Table 1, describes the characteristics of the data set name, size of samples, size of features, size of positive and negative class samples, imbalanced rate, and so on, in which the imbalanced rate refers to the ratio of negative class samples to total samples. They are all classified into two classes.

Table 1. The basic information of imbalanced data sets

Data set	Size of samples	Size of features	Size of negative class samples	Size of positive class samples	Imbalanced rate
Pima	768	8	268	500	34.9%
Parkinsons	195	22	48	147	24.6%
Vertebral	310	6	100	210	32.2%
Ionosphere	351	34	126	225	35.9%

5.2. Evaluation Indicators

At present, the classification algorithms have many evaluation indicators, such as classification accuracy, geometric mean (G-mean), negative class recall rate (Recall), and negative class test value (F-value). In this experiment, G-mean, F-value, and OOB (out of bag) were selected as evaluation indicators. The specific calculation methods of the three evaluation indicators are as follows.

(1) G-Mean

Geometric mean is the square root of the product of the positive class correct rate and the negative class correct rate, as shown in Equation (4).

$$G\text{-}mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{4}$$

TP and *TN* respectively represent the number of samples of the positGive class and negative class that are correctly classified, and *FP* and *FN* respectively represent the number of samples of the negative class and positive class that are misclassified, as shown in Table 2.

Table 2. The confusion matrix of binary classification

	Positive of prediction	Negation of prediction
Actually positive	*TP* (true positives)	*FN* (false negatives)
Actually negative	*FP* (false positives)	*TN* (true negatives)

In Table 2, the row represents the real class of the sample, the column represents the predicted class, and there are two classes of positive and negative.

$TP/(TP+FN)$ indicates the accuracy of the classification algorithm for the positive class, which is also called sensitivity.

$TN/(TN+FP)$ indicates the accuracy of the classification algorithm for the negative class, also known as specificity.

G-means is a geometric mean of sensitivity and specificity, which considers both sensitivity and specificity. Therefore, when the classification accuracy of positive class samples is higher but the classification accuracy of negative class samples

is lower, the geometric mean value will not be high; only when the classification accuracy of both positive class samples and negative class samples is higher can the geometric mean value obtain a higher value.

(2) F-Value

Negative class test value is an evaluation indicator that comprehensively evaluates the performance of random forests from the perspective of negative categories. It is a combination of the recall and precision of the negative class. The calculation method is shown in Equation (5).

$$F\text{-}Value = \frac{(1+\beta^2) \times Recall \times Precision}{\beta^2 \times Recall + Precision} \tag{5}$$

$Recall = TN/(TN + FP)$ is called the negative recall rate, which indicates the proportion of the samples of the negative class classified correctly to the real negative class.

$Precision = TP/(TP + FN)$ is called the precision rate of the negative class, which indicates the proportion of the samples of the negative class classified correctly to the samples of the negative class.

The value range of β is (0, 1]. The selection of its value depends on the actual situation, but it is usually chosen as 1.

(3) OOB

Out of bag estimate is a classic unbiased estimate of the random forest generalization error. The calculation method is described as follows: for the training samples set, the classification results of this sample are first determined by the statistical decision tree, and the classification structure is usually determined by a simple majority voting method. Then, the OOB error rate of the random forest is recorded by counting the proportion of the wrong samples to the total samples. Therefore, OOB does not need cross validation or separate test sets to obtain unbiased estimates of test set errors.

5.3. Experimental Results and Analysis

In the experiment, the SMOTE algorithm and KMS_SMOTE algorithm are used to balance the four data sets. Then, the random forest classification model is established on the original data set, and the data set is processed by the SMOTE algorithm and KMS_SMOTE algorithm. Finally, OOB, G-means, and F-value are used to evaluate the classification performance of each model. In this way, the effects of the SMOTE algorithm and KMS_SMOTE algorithm are evaluated. In order to ensure the stability of the algorithm results, 100 experiments were performed on each data set, and then the average of 100 experimental results was taken for each evaluation indicator.

(1) OOB

On the four data sets, after adding the new negative class samples, the random forest algorithm was used for the classification experiment. The experimental results based on the OOB indicator are shown in Table 3. It can be seen from Table 3 that the data set processed by the SMOTE algorithm or the KMS_SMOTE algorithm has a higher OOB accuracy than the original data set, and the data set processed by the KMS_SMOTE algorithm has the highest OOB value. In particular, the improvement on the pima data set is the largest, with the SMOTE algorithm and the KMS_SMOTE algorithm increasing by 3.4% and 6%, respectively, and the KMS_SMOTE algorithm is improved by 2.6% compared with SMOTE.

Table 3. OOB accuracy on individual data sets

Data sets	Pima	Ionosphere	Parkinsons	Vertebral
The original data set	0.7556	0.9252	0.9126	0.7680
The data set processed by SMOTE	0.7892	0.9313	0.9197	0.7796
The data set processed by KMS_SMOTE	0.8156	0.9426	0.9391	0.7854

(2) G-Mean

On the four data sets, after adding the new negative class samples, the random forest algorithm was used for the

classification experiment. The experimental results based on the G-mean indicator are shown in Table 4. It can be seen from Table 4 that the data set processed by the SMOTE algorithm or the KMS_SMOTE algorithm has a higher G-mean than the original data set, and the data set processed by the KMS_SMOTE algorithm has the highest G-mean value. In particular, the improvement on the vertebral data set is the largest, with the SMOTE algorithm and the KMS_SMOTE algorithm increasing by 6.89% and 8%, respectively, and the KMS_SMOTE algorithm is improved by 2.6% on the parkinsons data set compared with SMOTE.

Table 4. G-mean value on individual data sets

Data sets	Pima	Ionosphere	Parkinsons	Vertebral
The original data set	0.7109	0.9104	0.8431	0.6931
The data set processed by SMOTE	0.7570	0.9292	0.8664	0.7620
The data set processed by KMS_SMOTE	0.7779	0.9328	0.9037	0.7734

(3) F-Value

On the four data sets, after adding the new negative class samples, the random forest algorithm was used for the classification experiment. The experimental results based on the F-value indicator are shown in Table 5. It can be seen from Table 5 that the data set processed by the SMOTE algorithm or the KMS_SMOTE algorithm has a higher F-value than the original data set, and the data set processed by the KMS_SMOTE algorithm has the highest F-value. In particular, the improvement on the vertebral data set is the largest, with the SMOTE algorithm and the KMS_SMOTE algorithm increasing by 7.22% and 8.52%, respectively, and the KMS_SMOTE algorithm is improved by 3.62% on the parkinsons data set compared with SMOTE.

Table 5. F-value on individual data sets

Data sets	Pima	Ionosphere	Parkinsons	Vertebral
The original data set	0.6329	0.8924	0.8047	0.6054
The data set processed by SMOTE	0.6849	0.9061	0.8211	0.6776
The data set processed by KMS_SMOTE	0.7049	0.9127	0.8573	0.6909

In summary, (1) in the case of adding new negative class samples to each data set, regardless of whether the SMOTE algorithm or the KMS_SMOTE algorithm is used, the performance of the random forest algorithm is improved on each evaluation indicator because they reduce the imbalance of the data set; (2) the KMS_SMOTE algorithm has a better effect than the SMOTE algorithm when dealing with imbalanced data. When the new negative class sample is selected, the KMS_SMOTE algorithm is purposefully approached to the centroid of the original negative class, thus avoiding the marginalization problem of the SMOTE algorithm when selecting the new negative class samples. Thus, the KMS_SMOTE algorithm has good performance on all indicators.

6. Conclusions

In this paper, the K-Means algorithm and SMOTE algorithm are combined to propose a data preprocessing algorithm, KMS_SMOTE. New negative class samples are purposefully close to the centroid of the original negative class, effectively avoiding the marginalization problem of the SMOTE algorithm in selecting new negative class samples. The experimental results show that the KMS_SMOTE algorithm has better effects in the four data sets of UCI than the data processed by the SMOTE algorithm. Therefore, the algorithm improves the classification performance of the random forest algorithm on the imbalanced data set. Next, how to consider the discretization technology of continuous variables and improve the operating efficiency of the algorithm will be the focus of future work.

Acknowledgements

This work is supported by the National Youth Science Foundation of Hebei (No. F2017209070).

References

1. Q. Jing, X. Z. Qian, and W. T. Wang, "A Parallel Random Forest Algorithm for Imbalanced Big Data," *Microelectronics and Computer*, Vol. 34, No. 4, pp. 22-27, April 2017
2. L. Xue and S. W. Zhang, "Imbalanced Data Classification Algorithm based on Quadratic Random Forest," *Software*, Vol. 37, No. 7, pp. 75-79, July 2016
3. R. F. Chang, W. J. Wu, and W. K. Moon, "Support Vector Machines for Diagnosis of Breast Tumors on US Images," *Academic*

Radiology, Vol. 10, No. 2, pp. 189-197, February 2003
4. Y. Shi, X. M. Li, and X. H. Qi, "Classification Research of SVM with Imbalanced Data based on a New Type of under Sampling Samples," *Computer Measurement and Control*, Vol. 20, No. 5, pp. 1203-1235, May 2012
5. P. K. Chan and S. J. Stolfo, "Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection," in *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pp. 164-168, New York, American, August 1998
6. G. L. Sun, S. Li, Y. Cao, and F. Lang, "Cervical Cancer Diagnosis based on Random Forest," *International Journal of Performability Engineering*, Vol. 13, No. 1, pp. 446-457, July 2017
7. N. V. Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE: Synthetic Minority over-Sampling Technique," *Journal of Artificial Intelligence Research*, Vol. 16, No. 1, pp. 321-357, January 2011
8. H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A New over-Sampling Method in Imbalanced Data Sets Learning," in *Proceedings of the 1th International Conference on Intelligent Computing*, pp. 878-887, Heidelberg, Germany, July 2005
9. Y. J. Dong and X. H. Wang, "A New Over-Sampling Approach: Random-SMOTE for Learning from Imbalanced Data Sets," *Knowledge Science, Engineering and Management*, Vol. 7091, pp. 343-352, December 2011
10. X. C. Wang, Z. M. Pan, and L. L. Dong, "Research on Classification for Imbalanced Dataset based on Improved SMOTE," *Computer Engineering and Applications*, Vol. 49, No. 2, pp. 184-187, February 2013
11. P. Thanathamathee and C. Lursinsap, "Handling Imbalanced Data Sets with Synthetic Boundary Data Generation using Bootstrap Re-Sampling and Adaboost Techniques," *Pattern Recognition Letters*, Vol. 34, No. 12, pp. 1339-1347, December 2013
12. P. Vorraboot, S. Rasmequan, and K. Chinnasarn, "Improving Classification Rate Constrained to Imbalanced Data Between Overlapped and Non-Overlapped Regions by Hybrid Algorithms," *Neurocomputing*, Vol. 152, pp. 429-443, March 2015
13. X. F. Li, J. Li, Y. F. Dong, and C. W. Qu, "A New Learning Algorithm for Imbalanced Data-Pcboost," *Chinese Journal of Computers*, Vol. 35, No. 2, pp. 202-209, February 2012
14. J. Yun, J. Ha, and J. S. Lee, "Automatic Determination of Neighborhood Size in Smote," in *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, *Association for Computing Machinery*, Vol. 100, pp. 1-8, New York, NY, USA, January 2016
15. W. Qiong, Y. T. Li, and X. W. Zheng, "Optimization of Random Forest Algorithm for Classification of Imbalanced Training Sets," *Industrial Control Computer*, Vol. 26, No. 7, pp. 89-90, July 2013
16. D. Devi, S. K. Biswas, and B. Purkayastha, "Redundancy-Driven Modified Tomek-Link based undersampling: A Solution to Class Imbalance," *Pattern Recognition Letters*, Vol. 93, pp. 3-12, July 2017
17. W. Xue, "Improvement SMOTE Resampling Algorithm of Imbalanced Data Sets," *Statistical Research*, Vol. 29, No. 6, pp. 95-98, June 2012
18. W. Fan, S. J. Stolfo, and J. X. Zhang, "Adacost: Misclassification Cost-Sensitive Boosting," in *Proceedings of the 6th International Conference on Machine Learning*, pp. 97-105, San Francisco, CA, USA, June 1999
19. X. L. Wang and J. L. Wang, "Improving Adaboost Algorithm based on Cost-Sensitive," *Computer Applications and Software*, Vol. 30, No. 10, pp. 123-125, October 2013
20. F. Y. Cheng, J. Zhang, and C. H. Wen, "Cost-Sensitive Large Margin Distribution Machine for Classification of Imbalanced Data," *Pattern Recognition Letters*, Vol. 80, pp. 107-112, February 2016
21. S. Datta and S. Das, "Near-Bayesian Support Vector Machines for Imbalanced Data Classification with Equal or Unequal Misclassification Costs," *Neural Networks the Official Journal of the International Neural Network Society*, Vol. 70, pp. 39-52, October 2015
22. J. Du, "Cost-Sensitive Learning and Its Application," China University of Geosciences Doctoral Dissertation, Wuhan, China, December 2009
23. C. Seiffert, T. M. Khoshgoftaar, and J. VanHulse, "Rusboost: A Hybrid Approach to Alleviating Class Imbalance," in *Proceedings of the IEEE Transactions on Systems*, Vol. 40, No. 1, pp. 185-197, Piscataway, NJ, USA, January 2010
24. M. Galar, A. Fernandez, and E. Barrenechea, "Ordering-based Pruning for Improving the Performance of Ensembles of Classifiers in the Framework of Imbalanced Data Sets," *Information Sciences*, Vol. 354, No. C, pp. 178-196, March 2016
25. M. J. Kim, D. K. Kang, and B. K. Hong, "Geometric Mean based Boosting Algorithm with over-Sampling to Resolve Data Imbalance Problem for Bankruptcy Prediction," *Expert Systems With Applications*, Vol. 42, No. 3, pp. 1074-1082, March 2015
26. X. S. Hu, J. P. Wen, and Y. Zhong, "Imbalanced Data Ensemble Classification using Dynamic Balance Sampling," *CAAI Transactions on Intelligent Systems*, Vol. 11, No. 2, pp. 257-263, February 2016
27. B. Scholkopf, J. C. Platt, and J. Shawetaylor, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation*, Vol. 13, No. 7, pp. 1443-1471, July 2001
28. C. Y. Wang, "Research on Classification Method of Imbalanced Data Sets and Its Application in Telecom Industry," Zhejiang University Master Thesis, Hang Zhou, China, June 2011

Wenjie Li graduated from Northeast University and Yanshan University, China with a bachelor's degree and master's degree in engineering, respectively. She is currently a lecturer at Hebei Vocational & Technical College of Building Materials. Her research interests include graphic image processing and social network analysis.