mathematics

MDPI

Article

An AdaBoost Method with K′K-Means Bayes Classifier for Imbalanced Data

Yanfeng Zhang * and Lichun Wang

Department of Statistics, Beijing Jiaotong University, Beijing 100044, China
* Correspondence: 17118445@bjtu.edu.cn

Abstract: This article proposes a new AdaBoost method with k′k-means Bayes classifier for imbalanced data. It reduces the imbalance degree of training data through the k′k-means Bayes method and then deals with the imbalanced classification problem using multiple iterations with weight control, achieving a good effect without losing any raw data information or needing to generate more relevant data manually. The effectiveness of the proposed method is verified by comparing it with other traditional methods based on numerical experiments. In the NSL-KDD data experiment, the F-score values of each minority class are also greater than the other methods.

Keywords: imbalanced data; naive Bayes; imbalanced classifiers; AdaBoost method

MSC: 68T09; 94A16

1. Introduction

Classification technology is widely used in finance [1], medicine [2], meteorology [3], physics [4], road safety [5–7], and many other fields. The problem of imbalanced classification has attracted much attention for a long time in machine learning and artificial intelligence. In fact, imbalanced data occurs more often than balanced data, and the minority category is more of a concern. For example, the number of fraudulent transactions in a bank is far lower than that of normal, but abnormal transactions bring losses to the bank and have to be detected. The current number of COVID-19 patients is far lower than healthy people, but it is particularly important to detect the COVID-19 patients accurately as much as possible. In the field of road safety, fatal crashes represent roughly 2–3% of the total crashes, however, the correct classification of the minority class makes more sense. In the face of imbalanced data, the traditional algorithms cannot detect the minority classes accurately; they can only get high accuracy and recall on the majority classes. Hence, we need to improve the traditional algorithms and pay more attention to the detection of the minority classes rather than the majority classes.

A lot of improvement methods based on data and algorithms have been proposed. Resampling methods based on data mainly include over-sampling [8], under-sampling [9], and data synthesis [10]. The over-sampling method reduces the imbalanced degree by copying the minority data, but it is easy to cause over-fitting. The under-sampling method deletes most of the majority of data to achieve the balance between different categories, which makes it easy to lose important data information. The data synthesis methods like SMOTE(Synthetic Minority Oversampling Technique) [11] and ADASYN(Adaptive Synthetic Sampling) [12] methods synthesize new data based on the raw data, which avoid duplication and loss of data compared to the over-sampling method and the under-sampling method, and get better results. Based on an algorithm, there are mainly cost-sensitive learning approaches [13], boosting algorithms [14], bagging algorithms [15], and random forest algorithms [16]. The cost-sensitive learning approach obtains more recall and higher accuracy by giving wrong classification costs. The boosting method increases the weight of the wrong sample to improve the prediction results. The random forest

check for updates

Citation: Zhang, Y.; Wang, L. An AdaBoost Method with K′K-Means Bayes Classifier for Imbalanced Data. *Mathematics* **2023**, *11*, 1878. https://doi.org/10.3390/math11081878

Academic Editor: Palle E.T. Jorgensen

Received: 24 March 2023
Revised: 14 April 2023
Accepted: 14 April 2023
Published: 15 April 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

improves the prediction result by integrating the decision results of multiple decision trees. However, most of the existing methods have good performance in binary problems but work badly in multi-class problems.

A new method that makes improvements to both the algorithm and data is proposed in this paper. First, we reduce the imbalance degree between categories by clustering method without changing the data information; second, we increase the weight training for the samples that are difficult to predict to make them more likely to be accurately predicted using the AdaBoost algorithm. Additionally, based on the advantage of naive Bayes [17] in handling multiple classifications, this algorithm can solve both binary and multi-classification problems.

Given an imbalanced data set, we follow the k′k-means method proposed in 2021 in [18] to reduce the imbalance degree between categories to a reasonable level, then use the AdaBoost [19] method based on the naive Bayes method as the based classifier to improve the prediction effect of the minority. The key to this method is to restore the new categories generated by clustering to the raw labels at each iteration. Because we do not care whether the predictions of the newly generated categories are correct or not, what we really care about is whether the raw minority categories can be accurately identified.

In the following sections, we introduce three methods in Section 2, they are the naive Bayes method, the k′k-means method, and the proposed AdaBoost Bayes method with a k′k-means Bayes classifier. Then, In Section 3, we prove the advanced nature of the proposed method by numerical experiments. Finally, the last section summarizes the conclusions.

2. The AdaBoost k′k-Means Bayes Method

2.1. The Naive Bayes Classification

There is a training set $T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)\}$, where the number of samples is N, $\mathbf{x}_i \in R^m$ represents a dimensional array, y_j represents the class label with $y_j \in \{0, 1, 2, \cdots, t\}$. From the Bayes theorem, a test sample $\mathbf{x}(\mathbf{x} \in R^m)$ will be predicted to belong to the class y_j when it has the highest posterior probability conditional on $\mathbf{x}$. The predicted posterior probability is denoted as follows,

$$\psi(\mathbf{x}) = \arg\max_j P(y_j|\mathbf{x}) = \arg\max_j \frac{P(y_j)P(\mathbf{x}|y_j)}{P(\mathbf{x})}, \tag{1}$$

where $P(y_j)$ refers to the prior probability of class y_j, $P(\mathbf{x}|y_j)$ refers to the priori probability of $\mathbf{x}$ conditional on y_j, and $P(\mathbf{x})$ refers to the prior probability of $\mathbf{x}$ also named the normalization factor.

Making the naive assumption of class condition independence can reduce the computation in evaluating $P(y_j)P(\mathbf{x}|y_j)$. Mathematically, it means that

$$P(\mathbf{x}|y_j) = \prod_{i=1}^{n} P(\mathbf{x}_i|y_j), \tag{2}$$

where $\mathbf{x}_i$ denotes the value of the i-th attribute of test sample $\mathbf{x}$ and $P(\mathbf{x}_i|y_j)$ denotes the conditional probability of the i-th attribute on y_j.

To calculate each $P(\mathbf{x}_i|y_j)$, the numerical variable is usually typically supposed to follow a normal distribution with the following probability density function:

$$P(\mathbf{x}_i|y_j) = \frac{1}{\sqrt{2\pi}\sigma_{(i|y_j)}} \exp\{-\frac{[\mathbf{x}_i - \mu_{(i|y_j)}]^2}{2\sigma^2_{(i|y_j)}}\}, \tag{3}$$

where $\mu_{(i|y_j)}$ is the mean and $\sigma^2_{(i|y_j)}$ is the variance of the i-th variable for class y_j.

Hence, the classification result of the naive Bayes method can be achieved as follows:

$$\psi(\mathbf{x}) = \arg\max_j P(y_j) \prod_{i=1}^{n} P(\mathbf{x}_i|y_j). \tag{4}$$

2.2. The k′ k-Means Bayes Method

For any classification problems, the number of normal categories is usually very large, where the largest category is named the majority; the number of abnormal categories that we really care about is very small, in which the smallest category is called the minority. Let $N(y_j)$ represent the number of samples in the n-th class, then the number of the majority category is $\max\{N(y_j)\}$, the number of the minority class is $\min\{N(y_j)\}$. The degree of imbalance ζ is defined as follows:

$$\zeta = \max\{N(y_j)\}/\min\{N(y_j)\}. \tag{5}$$

Obviously, the large value of ζ means that the number of each category varies significantly. First, after the majority is updated into k subsets using the k-means method, the imbalance degree ζ is calculated. When ζ falls within a reasonable range, the k′k-means method stops. Generally speaking, an imbalance degree greater than 99 is considered an extreme imbalance classification problem. If $\zeta \geq 99$, the threshold is set to 50; otherwise it is set to $\frac{\zeta}{2}$. Specifically, $\max\{N(y_j)\}$ needs to be recalculated after each update. However, $\min\{N(y_j)\}$ is fixed. Due to the k′k-means method can eliminate the influence caused by imbalanced data, smaller categories do not have to be divided from the raw category into smaller categories which will effectively prevent invalid loops in the program. Finally, after the k′k-means method, the raw data set is updated into $T = \{(\mathbf{x}_1, y_1'), (\mathbf{x}_2, y_2'), \cdots, (\mathbf{x}_N, y_N')\}$ where $y_j' \in \{1, 2, \cdots, (k'k - k' + t)\}$ contains $(k'k - k' + t)$ categories. A definite one-to-many mapping relationship f between y_j' and y_j is denoted as $y_j = f(y_j')$. In the process of data classification being updated, the Bayes classifier can calculate the probability value of the test sample belonging to each category, and the sample will be predicted to belong to the corresponding category with the maximum probability value. For a test sample $\mathbf{x}_0$, the final predicted label $\hat{y}$ can be achieved as

$$\hat{y}(\mathbf{x}_0) = f(\psi(\mathbf{x}_0)). \tag{6}$$

The procedure of the k′k-means Bayes method is presented as follows.

The proposed k′ k-means Bayes method keeps the complete information of raw data without changing the samples, such as deleting or adding samples. Using the k′ times k-means method eliminates the impact of data imbalance on the model. The classification boundary is obtained by training in a relatively balanced sample environment. It deals with imbalanced data effectively from the perspective of data.

2.3. The AdaBoost k′k-Means Bayes Classification

AdaBoost (Adaptive Boosting) is an iterative algorithm. It assigns higher weights to hard-to-train samples by updating again to obtain a series of base classifiers. Let $\psi(\mathbf{x})$ obtained using k′k-means Bayes in Section 2.2 act as the basic classifier in AdaBoost. First, initialize the weight distribution of the training data T as follows:

$$D_1 = (\theta_{11}, \theta_{12}, \cdots, \theta_{1i}, \cdots, \theta_{1N}) = (\frac{1}{N}, \frac{1}{N}, \cdots, \frac{1}{N}). \tag{7}$$

For the m-th iterative, the weight of training data is denoted as D_m. The k′k-means Bayes classifier is represented as $\psi_m(\mathbf{x})$ whose classification error on the training data is given as

$$\alpha_m = \frac{1}{2} \log \frac{1 - e_m}{e_m}, e_m = P(\psi_m(x_i) \neq y_i). \tag{8}$$

Then, the weight allocation of the data is further updated as follows

$$D_{m+1} = (\theta_{(m+1),1}, \theta_{(m+1),2}, \cdots, \theta_{(m+1),i}, \cdots, \theta_{(m+1),N}), \theta_{(m+1),i} = \frac{\theta_{mi}}{Z_m} \exp(-\alpha_m y_i \psi_m(x_i)), \tag{9}$$

where $Z_m = \sum_{i=1}^{N} \theta_{mi} \exp(-\alpha_m y_i \psi_m(x_i))$ is the regularization factor.

When the classification result is wrong, it is clear that $y_i \psi_m(\mathbf{x}_i) < 0$, which can make $\theta_{(m+1),i}$ bigger. This is in line with our expectation of increasing the weight of the misclassified samples in the next classifier.

Finally, the AdaBoost classifier based on a k'k-means Bayes is represented as

$$H(\mathbf{x}) = \text{sign}(\sum_{m=1}^{M} \alpha_m \psi_m(\mathbf{x})). \tag{10}$$

In this section, the prediction results of difficult training samples can be improved by increasing the weight of wrong score data from the perspective of the algorithm. Let the naive Bayes method be used as the based classifier when the imbalance degree is effectively reduced using the k′k-means method; the method combined with AdaBoost can significantly improve the classification of imbalanced sample data.

3. Numerical Experiments

In this section, the effectiveness and superiority of the proposed new method will be demonstrated in contrast to the traditional naive Bayes method, the under-sampling method, the over-sampling method, the k-means Bayes method [20], and the AdaBoost Bayes method using simulated data and real data respectively. It is obvious that the prediction results depend on feature engineering largely, while in this paper, without considering the optimization scheme caused by variable selection, we discuss the progressiveness and effectiveness of the proposed method when compared with the traditional method. First, in Section 3.1, a multi-class simulation imbalanced data set is considered, and the prediction results of each method are compared. Then, the prediction results of the proposed method and other traditional methods are compared and analyzed based on network intrusion NSL-KDD data.

3.1. Performance Metrics

In this section, we introduce some common metrics to evaluate the performance of imbalanced classification methods. A confusion matrix is a visual matrix that can represent the relationship between predicted and true values, as shown in the figure. The elements on the main diagonal represent the number of samples correctly predicted, and the remaining elements represent cases where the predicted values and true values do not match. According to the elements in the matrix, several common evaluation metrics of the imbalanced classification algorithm can be calculated according to Table 1.

Table 1. The results of different methods.

True Value	Predict c_0	Predict c_1	Predict c_2
c_0	a_{11}	a_{12}	a_{13}
c_1	a_{21}	a_{22}	a_{23}
c_2	a_{31}	a_{32}	a_{33}

Using class c_0 as an example, we can calculate the following metrics where FPR refers to the false positive rate.

$$\begin{cases} Recall = \dfrac{a_{11}}{a_{11}+a_{12}+a_{13}} \\ Precision = \dfrac{a_{11}}{a_{11}+a_{21}+a_{31}} \\ FPR = \dfrac{a_{21}+a_{31}}{a_{11}+a_{21}+a_{31}} \\ F-score = 2\dfrac{Precision \cdot Recall}{Precision + Recall} \end{cases} \tag{11}$$

3.2. On Simulation Study

Without loss of generality, a data set containing 20 characteristic variables from the Gaussian distribution is constructed, the number of which is 30,000, including three categories. The sample sizes of the three categories are presented as $N(c_0) = 25{,}000$, $N(c_1) = 4950$, and $N(c_2) = 50$, respectively. It is obvious that it is an imbalanced multi-classification problem, and the imbalance degree is as high as $\zeta = 500$. The samples are divided into a training set and test set according to 5:1, and the prediction results obtained using the six methods are shown in Table 2.

Table 2. The results of different methods.

Method	Classification	FPR	Recall	Precision	F-Score
Naive Bayes	c_0	0.169	1	0.831	0.908
	c_1	-	0	0	0
	c_2	-	0	0	0
Under-sampling Bayes	c_0	0.166	0.371	0.834	0.513
	c_1	0.839	0.276	0.161	0.203
	c_2	0.998	0.5	0.002	0.005
Over-sampling Bayes	c_0	0.169	0.544	0.831	0.657
	c_1	0.835	0.261	0.165	0.202
	c_2	0.998	0.286	0.002	0.005
K-means Bayes	c_0	0.139	0.925	0.861	0.891
	c_1	0.584	0.267	0.416	0.325
	c_2	-	0	0	0
AdaBoost Bayes	c_0	0.169	0.564	0.831	0.672
	c_1	0.841	0.432	0.167	0.24
	c_2	1.000	0	0	0
AdaBoost K′K-means Bayes	c_0	0.168	0.713	0.832	0.768
	c_1	0.829	0.274	0.171	0.211
	c_2	0.994	0.071	0.006	0.011

From Table 2, it can be seen that when faced with imbalanced data, the naive Bayes method predicts all samples to be the majority class. Since the number of the minority takes up a small proportion, the method can still achieve a higher index result. However, this is not the method that we want. Both the k-means Bayes method and AdaBoost Bayes method fail to detect the "c_2" class. As can be seen from the confusion matrix in Figure 1, the under-sampling Bayes method has a high recall rate, but also has a very high false positive rate. The over-sampling Bayes method has not only a stable false positive rate but also a moderate number of minority classes. The proposed AdaBoost k′k-means Bayes method has the largest recall and the lowest false positive rate. Among the performances of the six methods, the AdaBoost k′k-means Bayes method is the best.

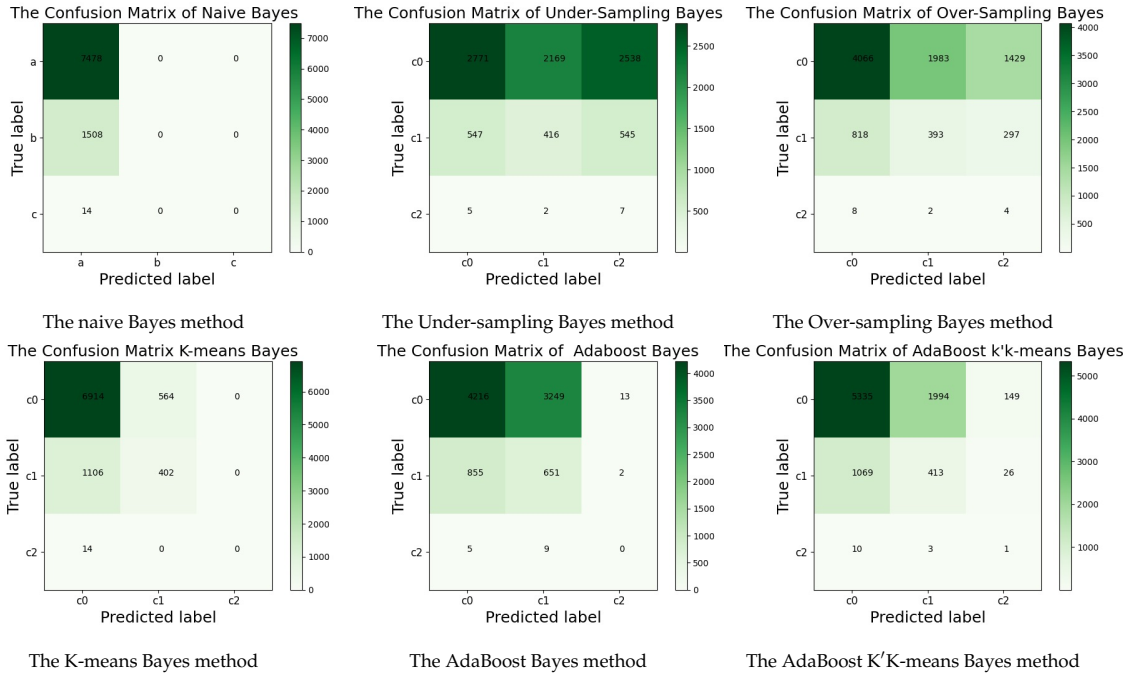


Figure 1. The confusion matrices of the six methods.

3.3. On Real Data Study

In this subsection, real imbalanced multi-class data will be used to demonstrate the effectiveness of the proposed method. The data is the famous NSL-KDD data set which comes from Canadian Institute for Cybersecurity (https://www.unb.ca/cic/datasets/nsl.html, accessed on 27 March 2023).

We will determine whether a piece of network data is normal behavior or a cyber-attack, and predict what type of attack it is. There are 42 variables in the data test where the target variable attack type contains "normal," "DOS" (Denial of Service), "U2R" (User-To-Root), "R2L" (Remote-to-Local), and "probe" 5 values shown in Table 2. The total sample size of the training data is 125,973, where the number of normal samples is 67,343, accounting for 53.46%, and the number of class "U2R" is only 52, accounting for 0.041%. The imbalance degree ζ is 1295 and can be calculated clearly so that it can be analyzed as a typical imbalanced multi-classification problem. Then, we will simultaneously compare the proposed method with the traditional naive Bayes, under-sampling, over-sampling, k-means Bayes, and AdaBoost Bayes methods based on the NSL-KDD data simultaneously to discuss which one has superiority (Table 3). The results are provided in Table 4.

Table 3. The distribution of attack type.

Data set	Classification	Training Data	Percentage	Testing Data	Percentage
NSL-KDD	normal	67,343	53.46%	9711	43.08%
	DOS	45,926	36.46%	7635	33.87%
	U2R	52	0.04%	200	0.89%
	R2L	995	0.79%	2574	11.42%
	probe	11,656	9.25%	2423	10.75%

Table 4. The results of different methods.

Method	Classification	FPR	Recall	Precision	F-Score
Naive Bayes	normal	0.780	0.032	0.22	0.056
	DOS	0.670	0.828	0.33	0.472
	U2R	0.992	0.06	0.008	0.014
	R2L	0.997	0	0.003	0.001
	probe	1.000	0	0	0
Under-sampling Bayes	normal	0.116	0.498	0.814	0.618
	DOS	0.780	0.163	0.22	0.187
	U2R	0.963	0.3	0.037	0.067
	R2L	0.941	0.05	0.059	0.009
	probe	0.746	0.96	0.254	0.402
Over-sampling Bayes	normal	0.861	0.017	0.139	0.031
	DOS	0.510	0.816	0.49	0.612
	U2R	0.990	0.41	0.01	0.02
	R2L	0.996	0.001	0.004	0.001
	probe	1.000	0	0	0
K-means Bayes	normal	0.373	0.811	0.647	0.72
	DOS	0.348	0.009	0.184	0.016
	U2R	0.797	0.545	0.014	0.028
	R2L	0.989	0.029	0.395	0.054
	probe	1.000	0.136	0.155	0.145
AdaBoost Bayes	normal	0.353	0.874	0.63	0.732
	DOS	0.816	0.728	0.652	0.688
	U2R	0.986	0.175	0.203	0.188
	R2L	0.605	0.002	0.011	0.003
	probe	0.845	0	0	0
AdaBoost K'K-means Bayes	normal	0.188	0.919	0.817	0.865
	DOS	0.093	0.708	0.84	0.768
	U2R	0.273	0.055	0.367	0.096
	R2L	0.533	0.379	0.463	0.417
	probe	0.466	0.642	0.512	0.57

For this data case, we are more concerned about how much more the attack data is detected than the normal type. As shown in Table 4, the type "normal" always performs well, and the type "DOS" with a large amount of data performs well in all of the six methods; however, what we are more concerned about is the performance of the other three types with a smaller amount of data. The AdaBoost Bayes method has the highest recall on the "U2R" type, but the precision is very low. The method with the best performance on the "probe" is the under-sampling; it is disappointing that it performs very poorly on the other types of indicators and loses a lot of raw data information in order to achieve the balance with "probe." The new method proposed in this paper is slightly worse than the AdaBoost Bayes method in the recall of "U2R," which has the best performance among the six methods for the other categories. In terms of the results of the F-score for each category, the proposed method is the best.

Next, we specifically analyze the results obtained from each method using the confusion matrices presented in Figure 2. The lager values on the diagonal suggest better prediction. It is obvious that (a), (b), and (c) do not satisfy this rule. The greater the value beyond the diagonal, the higher the false positive rate. The results show that the proposed method has the highest precision and the best recall in each index.

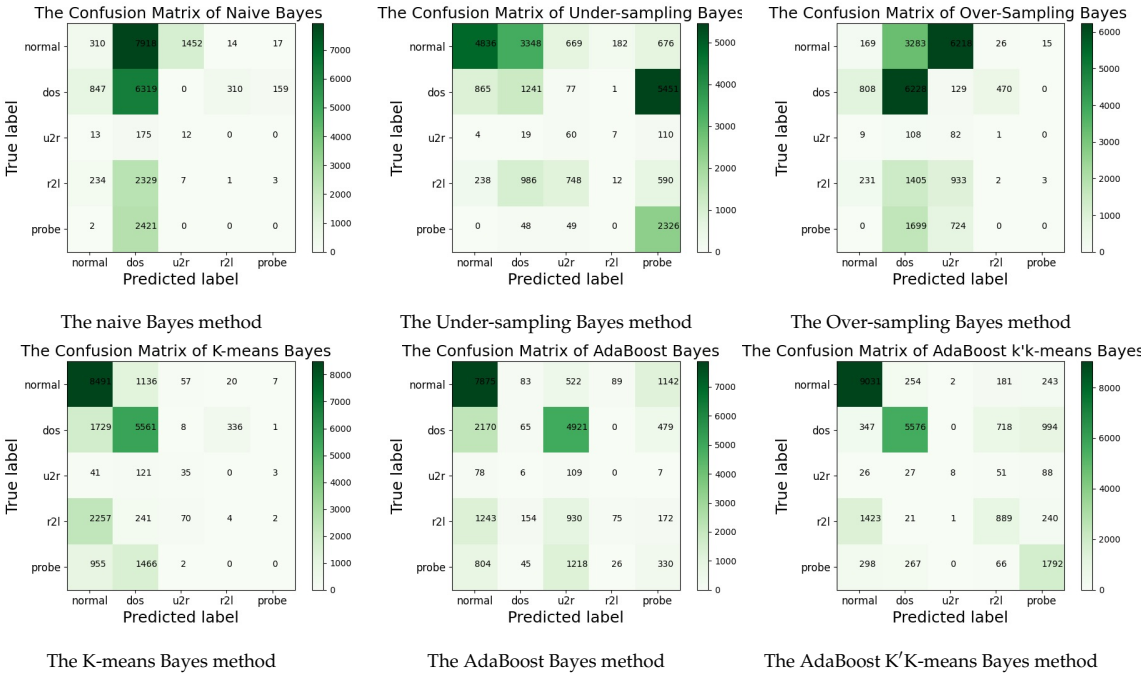


Figure 2. The confusion matrices of the six methods.

There are four parameters that need to be determined in two steps for the proposed method. First, the base classifier can be determined by determining the appropriate k and k′. It is well known that one of the difficulties in the k-means clustering method is determining the value of k. However, in the application of this paper, we only divide the majority class into many smaller subsets using the k-means method to reduce the imbalance degree, and do not care about the rationality and accuracy of the classification result, which will not affect our prediction results. So the determination is not as difficult and complicated as in the k-means clustering method. The values of k′ and ζ are interdependent. We cannot determine the value of k′ subjectively, but we can obtain a reasonable value k′ by adjusting the imbalance parameter of the data. Hence, we can give a reasonable ζ as 50 empirically, an initial k as $\frac{\zeta}{50}$, and then adjust it.

Second, the two parameters n-estimators and learning-rate need to be determined in AdaBoost method. The parameter n-estimators in the method is the maximum number of iterations of the base classifier, and the larger the value, the more likely it is to be overfitting. The parameter learning-rate represents the weight reduction coefficient of the base classifier, whose meaning is equivalent to that of the regularization term, and the value range is in [0, 1]. These two parameters require joint debugging. In the NSL-KDD data, the values of the four parameters are respectively k' = 18, k = 50, n-estimators = 10, and learning-rate = 0.05.

Then, after combining the test data and training data of KDD-NSL into a new data set, the generalized method of the proposed method is discussed through multiple cross-validation experiments. The data sets for cross-validation are obtained by changing the percentage of test samples and random seeds. Table 5 shows the results of 6 cross-validation experiments. The test data percentages and random seeds are (30%, 100), (35%, 180), (20%, 260), (40%, 40), (20%, 47), and (30%, 106). It can be found that the result fluctuation of each experiment is very small, and the prediction ability is stable. It is shown that the proposed method is generalized.

Table 5. The results of 6 cross-validation experiments.

Classification	Cross Validation Number	Recall	Precision	F-Score	Testing Data
normal	1	0.734	0.945	0.826	23,134
	2	0.769	0.936	0.844	26,939
	3	0.777	0.948	0.854	15,488
	4	0.776	0.938	0.85	30,607
	5	0.782	0.942	0.855	15,555
	6	0.765	0.936	0.842	23,221
DOS	1	0.887	0.762	0.82	16,022
	2	0.911	0.748	0.821	18,749
	3	0.873	0.774	0.82	10,692
	4	0.843	0.742	0.789	21,604
	5	0.874	0.726	0.793	10,595
	6	0.889	0.755	0.816	16,017
U2R	1	0.024	0.028	0.026	82
	2	0.046	0.062	0.053	87
	3	0.073	0.1	0.084	55
	4	0.031	0.055	0.039	98
	5	0.043	0.045	0.044	46
	6	0.041	0.073	0.052	74
R2L	1	0.68	0.258	0.374	1128
	2	0.656	0.29	0.402	1225
	3	0.66	0.313	0.424	688
	4	0.667	0.314	0.427	1418
	5	0.646	0.318	0.426	673
	6	0.62	0.285	0.39	1067
probe	1	0.491	0.419	0.452	4189
	2	0.396	0.471	0.43	4981
	3	0.538	0.433	0.48	2780
	4	0.392	0.344	0.367	5679
	5	0.339	0.366	0.352	2834
	6	0.411	0.395	0.403	4176

3.4. Experiments Result

In the simulation data experiment, the imbalance degree of the generated data is 500. There are three methods that can not recognize any "c_2" class, while the AdaBoost k′k-means Bayes method has the largest F-score of each class of the remaining three methods.

In the real data experiment, the imbalance degree of the NSL-KDD data is 1295. Three methods can not identify the "probe" class at all; however, the AdaBoost k′k-means Bayes method has the largest F values of each of the six classes.

Based on the classification results of different methods and cross-experiments, it can be concluded that the proposed method is more effective than other methods and generalized as well.

4. Conclusions

For the imbalanced multi-classification problem, a new AdaBoost method with a k′k-means Bayse classifier is proposed, which makes improvements on both data and algorithm to achieve better results compared to the traditional classification methods. It reduces the imbalance degree of training data using the k′k-means Bayes method, which avoids the change of the raw information caused by adding or removing samples in the resampling method. It increases the weights of samples that are difficult to predict using a multi-iterative method by improving its prediction accuracy. It is worth mentioning that in each iteration, the new categories generated with the k′k-means method need to be mapped to the raw categories to calculate the prediction results. The performances of the

proposed method are proven to be the best using some detailed comparisons with some other traditional methods based on simulation data and the NSL-KDD data.

We have proposed a more effective solution for multi-class imbalanced classification for both continuous and discrete variables, which can be used for both binary classification and multi-classification. In the future, it may be possible to generalize this idea to combine it with other classical algorithms to solve more complex problems.

Author Contributions: Methodology, Y.Z. and L.W.; software, Y.Z. and L.W.; writing—original draft preparation, Y.Z. and L.W.; writing—review and editing, Y.Z. and L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No.11371051).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, J.; Chai, J.; Cho, S. Deep learning in finance and banking: A literature review and classification. *Front. Bus. Res. China* **2020**, *14*, 1–24. [CrossRef]
2. Singh, S.; Jangir, S.; Kumar, M.; Verma, M.; Kumar, S.; Walia, T.; Kamal, S. Feature Importance Score-Based Functional Link Artificial Neural Networks for Breast Cancer Classification. *BioMed Res. Int.* **2022**, *2022*, 1–24. [CrossRef]
3. Kumar, S.; Sastry, H.; Marriboyina, V. Information extraction from the agricultural and weather domains using deep learning approaches. *Int. J. Softw. Innov.* **2022**, *10*, 1–112. [CrossRef]
4. Lombacher, J.; Hahn, M.; Dickmann, J.; Wöhler, C. Object classification in radar using ensemble methods. *Int. J. Softw. Innov.* **2017**, 87–90. [CrossRef]
5. Rella, R.; Mauriello, F.; Sarkar, S.; Galante, F.; Scarano, A.; Montella, A. Parametric and non-parametric analyses for pedestrian crash severity prediction in Great Britain. *Sustainability* **2022**, *14*, 3188. [CrossRef]
6. Gao, L.; Lu, P.; Ren, Y. A deep learning approach for imbalanced crash data in predicting highway-rail grade crossings accidents. *Reliab. Eng. Syst. Saf.* **2021**, *216*, 108019. [CrossRef]
7. Yahaya, M.; Jiang, X.; Fu, C.; Bashir, K.; Fan, W. Enhancing crash injury severity prediction on imbalanced crash data by sampling technique with variable selection. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference, Auckland, New Zealan, 27–30 October 2019; pp. 363–368.
8. Junsomboon, N.; Phienthrakul, T. Combining over-sampling and under-sampling techniques for imbalance dataset. In Proceedings of the 9th International Conference on Machine Learning and Computing, Singapore, 24–26 February 2017; pp. 243–247.
9. Tsai, C.; Lin, W.; Hu, Y.; Yao, G. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. *Inf. Sci.* **2019**, *477*, 47–54. [CrossRef]
10. Rees, E.; Nightingale, E.; Jafari, Y. COVID-19 length of hospital stay: A systematic review and data synthesis. *BMC Med.* **2020**, *18*, 270. [CrossRef] [PubMed]
11. Dablain, D.; Krawczyk, B.; Chawla, N. DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–15. [CrossRef] [PubMed]
12. Lu, C.; Lin, S.; Liu, X.; Shi, H. Telecom fraud identification based on ADASYN and random forest. In Proceedings of the International Conference on Computer and Communication Systems, Shanghai, China, 15–18 May 2020; pp. 447–452.
13. Mienye, I.; Sun, Y. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Inform. Med. Unlocked* **2021**, *25*, 100690. [CrossRef]
14. Tyralis, H.; Papacharalampous, G. Boosting algorithms in energy research: A systematic review. *Neural Comput. Appl.* **2021**, *33*, 14101–14117. [CrossRef]
15. Andiojaya, A.; Demirhan, H. A bagging algorithm for the imputation of missing values in time series. *Expert Syst. Appl.* **2019**, *129*, 10–26. [CrossRef]
16. Tyralis, H.; Papacharalampous, G.; Langousis, A. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water* **2019**, *11*, 910. [CrossRef]
17. Salmi, N.; Rustam, Z. Naive Bayes classifier models for predicting the colon cancer. *Mater. Sci. Eng.* **2019**, *546*, 052068. [CrossRef]
18. Zhang, Y.; Wang, L. K′ times k-means logistic regression algorithm for imbalanced classification. *Commun.-Stat.-Simul. Comput.* **2021**, 1–8. [CrossRef]

19. Wang, W.; Sun, D. The improved AdaBoost algorithms for imbalanced data classification. *Inf. Sci.* **2021**, *563*, 358–374. [CrossRef]
20. Chen, G.; Liu, Y.; Ge, Z. K-means Bayes algorithm for imbalanced fault classification and big data application. *J. Process Control* **2019**, *81*, 54–64. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.