Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 4617581, 15 pages
https://doi.org/10.1155/2020/4617581

Research Article

Research on Effect Evaluation of Online Advertisement Based on Resampling Method

Heyong Wang and Canxin Lin

Department of E-Business, South China University of Technology, Guangzhou, China

Correspondence should be addressed to Heyong Wang; wanghey@scut.edu.cn

Received 15 July 2020; Revised 14 October 2020; Accepted 5 December 2020; Published 21 December 2020

Academic Editor: Ioannis Kostavelis

With the rapid development of the Internet, the online advertising market has become larger and larger. Online advertisers often execute their advertising strategies based on the effect of online advertisements, so it is necessary to evaluate the advertising effect because it determines whether advertisers can display effective advertisements continually and remove ineffective advertisements timely. In practical scenarios, the quantity of ineffective online advertisements is always larger than that of effective online advertisements. The imbalanced distribution of them will bring serious bias to the evaluation models. We propose an improved undersampling method based on clustering (termed UBOC) to overcome the data imbalance. It can balance the advertising data into a more suitable data distribution. In addition, we adopt a new evaluation index for the effect evaluation of online advertisements based on C5.0 decision tree. Experimental results indicate the excellent performance of UBOC and the practical application of evaluation index for online advertisements. They can provide an effective evaluation of online advertisements and achieve the early removal of ineffective advertisements for advertisers, which will greatly increase the revenue brought by advertisements.

1. Introduction

Online advertising is an advertisement form which interacts with users through the Internet. With the rapid development of e-commerce, online advertising market has become larger and larger. Online advertising has become not only the main media for advertisers to promote products or services but also the way for many portals and long-tailed websites to make profits. It gradually reflects its distinct advantages of significant efficiency, instant interaction, and low cost.

In the study of online advertising, current literature focuses on two major research topics: how to effectively preprocess data and accurately evaluate effect of online advertisements. On the one hand, data preprocessing can eliminate the negative impacts on data model caused by the data problems of online advertisements. On the other hand, effect evaluation can further improve the commercial value brought by online advertisements. However, there are some crucial problems that need to be further studied in the actual business application. The problems can be summarized in two aspects as follows: data imbalance and optimization of effect evaluation.

1.1. Data Imbalance of Online Advertisements. Usually, most online advertising companies will divide each online advertisement into effective or ineffective one according to its economic profits. It always finds that the dataset made up of these advertisements is imbalanced because it contains a large proportion of ineffective advertisements and a small proportion of effective advertisements. It also means that the data distribution of online advertisements is imbalanced. The phenomenon of data imbalance also appears in many application scenarios [1], such as Internet personal credit evaluation [2], telecom customer churn prediction [3], and network bank fraud detection [4]. It will make the prediction model more partial to the data with a larger proportion, which directly causes poor performance. From the current study on data preprocessing of online advertisements, there are few literatures dealing with data imbalance to improve

the model performance. Therefore, it is necessary for advertising data to overcome the data imbalance during data preprocessing.

1.2. Optimization of Effect Evaluation on Online Advertisements. When optimizing the effect evaluation on online advertisements, there are two concepts that need to be considered: the cost difference that exists in different evaluation results and the evaluation index that reflects the effect of evaluation model. On the one hand, the cost difference in different advertisements will lead to the biased evaluation. The profit loss of removing the effective advertisements is much higher than displaying the ineffective advertisements. Accordingly, the cost of predicting effective advertisements as ineffective ones will be much higher. Advertisers expect to avoid the aforementioned situation. On the other hand, most previous evaluation indexes lack consideration of application scenario. An evaluation index available for online advertisements needs to be proposed in order to meet the actual requirements of e-commerce enterprises.

According to the above discussions, this paper focuses on resolving the data imbalance of online advertisements and then presents a practical evaluation index to reflect the effect of evaluation model. The main contributions can be described as follows. (1) For the data preprocessing, we propose an improved method of undersampling based on clustering (UBOC). It can balance the data distribution according to the data features. The comparison experiments on UBOC and other resampling methods demonstrate that UBOC can significantly improve the prediction performance of the evaluation model. (2) For the application scenario, we propose an evaluation index available for the evaluation model of online advertisements. Our study regards the evaluation of online advertisements as a binary classification based on the log data of online advertisements. The misclassification cost is combined into the evaluation model in order to fulfill the requirements of advertisers.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 provides the research methodology of our study. Section 4 describes the feature selection, resampling methods, and prediction models. Section 5 performs comparison experiments to demonstrate effectiveness of our methodology. Section 6 gives suggestions and conclusions.

2. Related Work

From the view of current research, most studies on online advertisements cover various applications, including the construction of effect prediction model of online advertisements [5, 6], the factor analysis affecting the online advertising effect [7, 8], and the establishment of complete evaluation system for online advertisements [9, 10]. Through the collection of these studies, it can be found that they mainly deal with the advertising data characteristics of high dimension and sparsity, instead of multidimension and imbalance. In other words, they ignore that advertising data also contain feature information from different dimensions and show an imbalanced distribution. At the same time, they adopt several evaluation models to evaluate the real effect of online advertisements. Although there are some literatures on the practical application at enterprise level [11–13], they rarely solve the problem of data imbalance of online advertisements and they rarely put forward a suitable evaluation index for the evaluation model. However, balancing the data distribution is the key step in data preprocessing and outputting the evaluation index is the key to evaluate the model performance. Therefore, we systemize the related works from the following aspects: data preprocessing method and effect evaluation model.

2.1. Data Preprocessing Method of Online Advertisements. In the procedure of data preprocessing, feature and data will be processed to obtain a valuable dataset. Usually, it is necessary to select the appropriate features at first and then filter out the useless data items. Our study mainly solves the imbalanced distribution of advertising data. For the feature selection, we consider the feature information from diverse sources and divide features into three types of online advertisements.

According to current literatures, many research studies have paid attention to the high dimension and sparsity of advertising data during data preprocessing [14–19]. Tu et al. [14] proposed the factor model (AdRec) based on joint probability matrix decomposition. They solved the problem of data sparsity by combining the information of users, advertisements, and web pages. Li et al. [15] constructed an advertising prediction model based on feature selection method of LASSO, which effectively overcame the obstacles of high dimension and sparse data processing. Shen et al. [16] proposed a sparse feature learning method for advertising data. It made full use of deep learning technology to study the nonlinear correlation and used tensor decomposition to achieve feature dimensionality reduction. The aforementioned methods have made great progress on solving the problems of high dimension and sparsity of online advertising data. However, the data problems of multidimension and imbalance are usually neglected.

Selection of multidimensional features plays an important role in feature selection. The key of feature selection is to find the most valuable features and delete other features with low correlation [20]. In fact, it neglects that feature selection needs not only high correlation but also diversification. At present, there are many literatures on the issue of multidimensional features in practical application [21–24]. Simultaneously, it is confirmed that the factors affecting advertising effect originate from diverse dimensions [25, 26], including the features of advertisements, users, and media platforms. Therefore, the features from different influence factors need to be selected when preprocessing the data of online advertisements.

In the real-world online advertising data, there is a phenomenon of imbalanced distribution over effective and ineffective advertisements [27]. The imbalance causes the problem that the general prediction models mainly emphasize information of majority class. Therefore, the

prediction performance of imbalanced data is usually biased. At present, the solutions for the imbalanced problem include resampling and cost-sensitive learning [28]. Resampling is more often to be adopted by academia in the practical applications [29–31]. It changes the sample distribution of the original dataset. Resampling is usually divided into the random method and heuristic method. Random methods copy or delete samples randomly while heuristic methods adopt the particular rules to generate new samples [32]. Cost-sensitive learning is used to distinguish the misclassification costs of samples of majority and minority classes. Although the cost-sensitive learning method seems simple and intuitive, it is difficult to define an appropriate cost-loss function in specific application [33].

Therefore, the effectiveness of the aforementioned methods needs to be verified in the field of online advertising. In this paper, we have done the following work in data preprocessing of online advertisements. For the processing of feature, in order to reduce the complexity of algorithms, we use the feature selection based on filtering to filter out features which are not related to the target classification. In addition, we also divide the selected features into three types according to data sources, thus making up for the defect that the traditional features of single type are not strong explanatory. For the strategy of resampling, we perform comparison analysis on three random resampling methods (random oversampling, random undersampling, and combined sampling) and two heuristic resampling methods (synthetic minority oversampling technique and random oversampling examples). In addition, we also propose an improved undersampling method based on the clustering method to output a sensible data distribution.

2.2. Effect Evaluation Method of Online Advertisements. In the current field of machine learning, there are many research studies on evaluating the advertising effect based on historical data. Richardson et al. [34] used the logistic regression model to perform the click-through rate prediction. Chapelle et al. [35] used dynamic Bayesian neural network to predict the click-through rate of advertisements, aiming at providing an unbiased estimation of the relevance of click logs for advertisements. Dave et al. [36] used the decision tree based on gradient descent algorithm to predict the click-through rate of advertisements, which significantly improves search engine ranking. It can be seen that many machine learning models have been used to evaluate the advertising effect and improve the performance of the prediction model. In addition, Chan et al. [37] showed that compared with other methods, decision tree has excellent learning efficiency and explanatory ability. According to the above discussions, we select decision tree as the evaluation model to evaluate the effect of online advertisements.

In addition, the aforementioned methods neglect the cost difference of different prediction situations. Some sensible enterprises will begin to emphasize the misclassification cost of effective and ineffective advertisements. On this foundation, based on the confusion matrix of decision tree, we can calculate a new evaluation index that reflects the cost of misclassification. It can be adjusted by setting different misclassification cost values to fulfill the needs of cost-sensitive advertising strategies from enterprises.

3. Research Methodology

This paper focuses on the data imbalance of online advertisements during data preprocessing and the effect evaluation of evaluation model suitable for online advertisements. The complete framework is shown in Figure 1.

3.1. Feature Selection. Feature selection is a process of selecting the most effective features of a group of features to realize the dimension reduction [20]. As a step of data preprocessing, it can improve the prediction accuracy, robustness, and interpretability of our learning algorithm. According to the combination ways of learners, feature selection can be divided into four methods including filter, wrapper, embedded, and ensemble methods [38]. As for the online advertising data, they contain numerous and complex features. Therefore, we adopt the feature selection based on the filter method in this paper because of its high efficiency. Furthermore, the selected features are interpreted as three types such as advertising information, user information, and media platform information, which can generate a set of features that comprehensively reflect the advertising effect.

3.2. Resampling Method. The resampling method is very critical because it solves the problem of data imbalance effectively. It also can improve the model performance by preventing the prediction model from emphasizing the features of majority class much more than minority class. Five common resampling methods (random oversampling, synthetic minority oversampling technique, random oversampling examples, random undersampling, and combined sampling) are considered for balancing advertising data in our study. After analyzing advantages and disadvantages of these five methods, we provide an improved undersampling method based on clustering (UBOC).

3.3. Decision Tree Model. The decision tree model is widely used in the task of binary classification. However, the cost of different classifications may exhibit difference depending on the actual situation. In online advertising, the cost of removing the effective advertisements will be higher than that of displaying the ineffective advertisements. Therefore, we set the misclassification cost in the decision tree model to build a cost-sensitive evaluation model, which can meet the needs of advertisers.

3.4. Effect Evaluation. Based on the confusion matrix, we propose an evaluation index for the evaluation model. It will be more applicable to the online advertising in terms of flexibility and interpretability. The evaluation model will provide a prediction label based on the effect of each advertisement. The prediction results are usually presented in the form of confusion matrix given by Table 1.

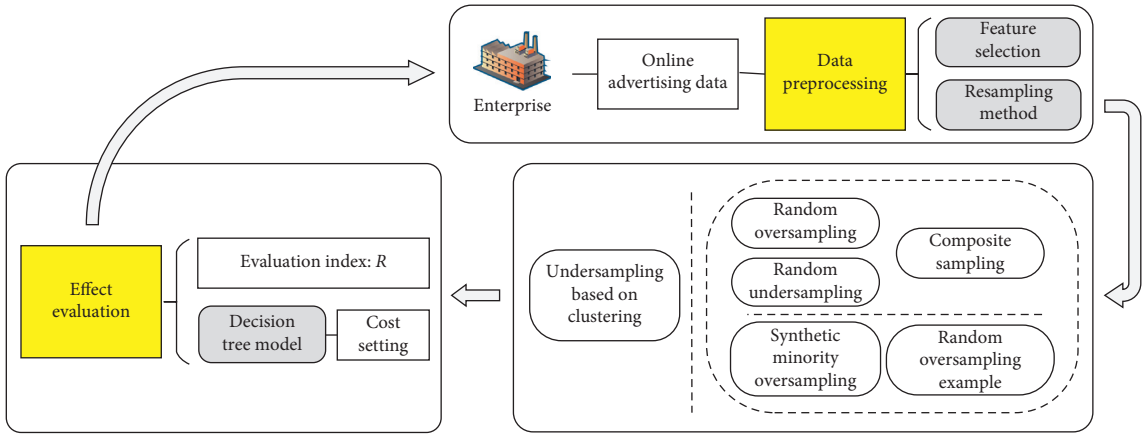

Figure 1: Framework for effect evaluation of online advertisements.

Table 1: The confusion matrix of effect evaluation of online advertisements.

Actual	Predicted	
	Ineffective advertisement	Effective advertisement
Ineffective advertisement	True positive (TP)	False negative (FN)
Effective advertisement	False positive (FP)	True negative (TN)

TP and TN are used to represent the numbers of ineffective advertisements which are predicted as ineffective ones and effective advertisements which are predicted as effective ones, respectively. FN and FP are used to represent the numbers of ineffective advertisements which are predicted as effective ones and effective advertisements which are predicted as ineffective ones, respectively. Since advertisers are more concerned about the cases of correct prediction of ineffective advertisements and incorrect prediction of effective advertisements, we mainly consider two evaluation indexes in our study: the true positive rate (TPR) and the false positive rate (FPR). TPR and FPR are, respectively, shown in equations (1) and (2).

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{1}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{2}$$

TPR represents the proportion of the correctly predicted ineffective advertisements in all actual ineffective advertisements. FPR represents the proportion of incorrectly predicted effective advertisements in all actual effective advertisements. Advertisers expect to reduce FPR as much as possible to avoid the loss of profits from the incorrectly predicted effective advertisements. Furthermore, in order to prevent the meaningless situation in which all advertisements are predicted as effective ones, TPR should be maintained at a certain high level at the same time. Therefore, in order to maintain high TPR and low FPR, we construct the ratio of true positive rate to false positive rate (**R**) as the evaluation index for online advertisements.

$$\mathbf{R} = \frac{\text{true positive rate}}{\text{false positive rate}} = \frac{\text{TP}/(\text{TP} + \text{FN})}{\text{FP}/(\text{FP} + \text{TN})} = \frac{\text{TP}(\text{FP} + \text{TN})}{\text{FP}(\text{TP} + \text{FN})}. \tag{3}$$

The index **R** is different from generalized traditional evaluation indexes such as accuracy rate, F-measure, AUC, and so on. The proposal of **R** is conducive to the cost-sensitive strategies of advertisers. The detailed reasons for adopting **R** are as follows:

(1) In the application scenario of online advertising, enterprises expect that more ineffective advertisements can be accurately identified and fewer effective advertisements are misclassified as ineffective ones at the same time. The expectation is reflected in the values of TPR and FPR. For efficiency and simplicity, we construct the evaluation index **R** by combining of TPR and FPR.

(2) The evaluation index for online advertisements is essentially to be constructed for all the relevant departments of enterprise to understand the evaluation results. Higher value of **R** indicates more ineffective advertisements are predicted accurately while fewer effective advertisements are misclassified. **R** is an explanatory evaluation index for enterprises because it directly reflects the effect of evaluation model and fulfills the demand of cost-sensitive advertising strategies.

(3) **R** is flexible when improving the model performance. Although the increase of TPR or decrease of FPR can both improve the value of **R**, we expect to avoid the misclassification of effective advertisements as much as possible without the loss of the correct classification of

ineffective advertisements. The adoption of **R** can directly achieve the relatively low values of FPR under the premise of a certain level of TPR.

4. Main Method

This section mainly introduces the principle of feature selection method based on filtering, five common resampling methods, and our modified resampling method. In addition, the C5.0 decision tree prediction model is also introduced in this section. For the evaluation index of feature selection, Pearson correlation coefficient is used to sort features. For resampling methods, they have two common parameters: the total number of sampling and the random seed. The total number of sampling determines the total number of samples in the final datasets after resampling. It changes the distribution of samples among majority and minority classes. Consider a simple example of an imbalanced dataset with 800 samples of majority class and 100 samples of minority class. In the oversampling strategy, the number of samples of minority class will be extended from 100 to 800, so the total number of samples in the dataset is extended to 1600. After resampling, the original imbalanced dataset has become balanced because the number of samples of majority and minority classes is equal. The random seed is used to determine random conditions during the resampling process. The resampling results of a dataset will not change after repeated resampling under the same value of random seed.

4.1. Feature Selection Based on Filtering. Compared with other feature selection methods, filtering has low complexity and high efficiency because it only evaluates the correlation between single feature and target classification. Under this method, every feature will be scored according to the specific evaluation index and be sorted in descending order. The first **k** features can be output as the feature subset.

In this paper, Pearson correlation coefficient is selected as the evaluation index of filtering. There are many continuous features of online advertising data, so Pearson correlation coefficient can be used to reflect the linear correlation between advertising features and target classification. The higher the value of coefficient, the more effective the feature is. If $\mathbf{f_i}$ and $\mathbf{f_j}$ represent two data features, Pearson correlation coefficient of them can be calculated as follows:

$$\mathbf{r}\left(\mathbf{f_i}, \mathbf{f_j}\right) = \frac{\mathbf{Cov}\left(\mathbf{f_i}, \mathbf{f_j}\right)}{\sqrt{\mathbf{Var}\left(\mathbf{f_i}\right)\mathbf{Var}\left(\mathbf{f_j}\right)}}, \tag{4}$$

$\mathbf{Cov}(\mathbf{f_i}, \mathbf{f_j})$ represents the covariance of $\mathbf{f_i}$ and $\mathbf{f_j}$ and $\mathbf{Var}(\mathbf{f_i})$ represents the variance of $\mathbf{f_i}$. On the basis of this index, we obtain the feature ranking of Pearson correlation coefficient in each group of training set. Taking the training data on September 1 as an example, Table 2 shows the coefficient value between the feature and target classification (effective or ineffective advertisements) of all the data features. We have filtered the ten training sets and eventually select 18 features from 29 features according to the average ranking. These 18 features can be classified as the following types: advertisements, users, and media platforms.

4.2. Random Oversampling. Random oversampling (OVER) is the simplest oversampling method and belongs to randomization resampling. It extends number of samples in minority class by randomly copying the samples in minority class for the purpose that the number of samples of minority class reaches a given proportion of samples of majority class. OVER possibly leads to overfitting of classification and increase of computational and storage cost because of the repeated samples in minority class. In summary, OVER is more suitable for datasets with small sample size and less noise in minority class.

According to the above discussion, OVER is not the suitable strategy for the task of this paper because the proportion of samples in minority class is not very low and there is a certain quantity of noise in the online advertisement dataset. Moreover, the strategy of OVER to increase the number of samples will occupy a large amount of memory and computational time.

The diagram of OVER is shown in Figure 2.

(1) Figure 2(a) shows an imbalanced dataset which contains the samples of minority class (solid triangles) and the samples of majority class (solid circles).

(2) Copy the samples in minority class randomly. The repeated samples are represented as hollow triangles Figure 2(b)

(3) Repeat step 2 until the final number of minority samples reaches a given proportion (e.g. 100%) of samples in majority class.

We can understand the changes of the imbalanced dataset after OVER. Before OVER, the dataset contains 6 samples of majority class (solid circles) and 3 samples of minority class (solid triangles). After OVER, the dataset becomes balanced because it contains 6 samples of majority class (solid circles) and 6 samples of minority class (solid triangles and hollow triangles). The samples of minority class have been extended by OVER.

4.3. Synthetic Minority Oversampling Technique. Synthetic minority oversampling technology (SMOTE) is an improved method of OVER and belongs to heuristic resampling. SMOTE is proposed to ease the overfitting problem of OVER. Firstly, SMOTE selects a random target sample of minority class and a random nearest neighbor of the target sample. Secondly, SMOTE selects a sample point on the connection of target sample and its selected neighbor as a synthetic sample for the imbalanced dataset, instead of simply replicating the original samples of minority class.

SMOTE may also lead to overfitting and high computational and storage cost. Moreover, SMOTE is unable to determine how many neighbor samples should be selected to achieve the optimal result. In addition, the synthetic samples generated by SMOTE will also be in the boundary position of sample space if the samples of minority class are mainly

Table 2: The Pearson correlation coefficient of training data features on September 1.

Feature	Pearson correlation coefficient
ROI on the day	0.197**
Amount of total payments	0.190**
Amount of total orders	0.189**
Average residence time of user	0.141**
Type of media platform	0.135**
Ad budget	0.127**
Ad type	0.127**
Number of total purchase volume	0.123**
Number of total search volume	0.121**
Ad position	0.119**
Number of total collection volume	0.112**
User amount	0.106**
Number of total recommendation volume	0.098**
Average views of home page	0.080**
Number of total register volume	0.079**
User gender	0.078**
Views of home page≥8	0.068**
Average views of business page	0.063**
Max of user age	0.061**
Min of user age	0.060**
Views of business page ≥ 8	0.043*
Number of order people	0.040*
Number of orders	0.035*
Number of payers	0.022
Number of payments	0.013
Residence time of user ≥ 8 min	0.001
Views of home page = 1	−0.001
Residence time of user ≥ 1 min	−0.002
Views of business page = 1	−0.022

**At the 0.01 level (double tails), the correlation is significant.
*At the 0.05 level (double tails), the correlation is significant.

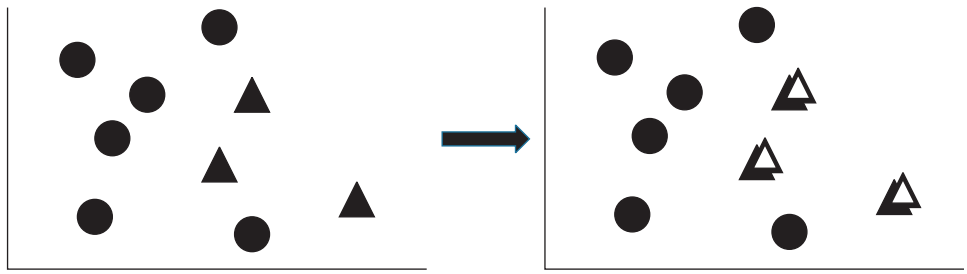

Figure 2: The diagram of OVER: (a) before sampling; (b) after sampling.

distributed on the boundary position, which causes distribution marginalization and increases the difficulty of classification.

The diagram of SMOTE is shown in Figure 3.

(1) Figure 3(a) shows an imbalanced dataset which contains the samples of minority class (solid triangles) and the samples of majority class (solid circles).

(2) For target sample of minority class (one of the solid triangles in Figure 3(b) select several similar nearest neighbors around target sample and randomly select one of these similar nearest neighbors (one of the other solid triangles in Figure 3(b))

(3) Connect target sample and its selected neighbor (represented as one of the dotted lines between solid triangles in Figure 3(b) and randomly select a location on the connection line to generate a new synthetic which is represented as one of the hollow triangles in Figure 3(b).

(4) Repeat steps 2 and 3 until the final number of samples of minority class reaches a given proportion (e.g., 100%) of samples of majority class.

The balanced dataset after SMOTE is of the same size as dataset after OVER (Figure 2(b)). Different from OVER, the generated samples of SMOTE are the synthetic samples

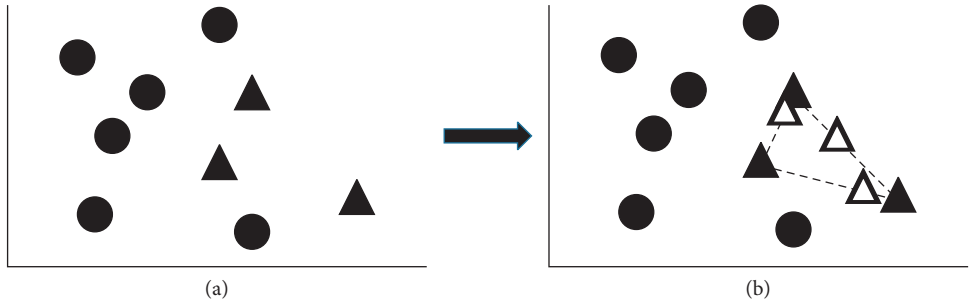

Figure 3: The diagram of SMOTE: (a) before sampling; (b) after sampling.

rather than the original samples. The synthetic samples are distributed around the original samples of minority class but do not coincide with any of the original samples.

4.4. Random Oversampling Examples. Random oversampling examples (ROSE) is a special method based on kernel function and belongs to heuristic resampling. ROSE uses kernel density estimation to generate new synthetic samples by expanding the feature space of samples of minority and majority classes. The kernel density estimation uses the method of kernel function to weight the average of relevant sample points in a certain range so as to obtain the probability density and distribution situation of the sample points and then determine the estimated value.

The diagram of ROSE is shown in Figure 4.

(1) Figure 4(a) shows an imbalanced dataset which contains the samples of minority class (solid triangles) and the samples of majority class (solid circles).

(2) For all the samples, the estimated value of kernel density $\widehat{\mathbf{p}_{\mathbf{n}(\mathbf{s})}}$ of the samples is calculated by the appropriate kernel function. Based on the kernel density estimated value, the new samples of majority and minority classes are generated. These new samples of majority and minority classes are, respectively, represented as hollow triangles and hollow circles in Figure 4(b).

(3) Repeat steps 2 and 3 until the final numbers of samples of majority and minority reach to given proportions. At this time, the number of total samples of the dataset after ROSE will be more or less than total samples of the original dataset.

According to Figure 4, the ratio of sample numbers of majority class (solid circles) to minority class (solid triangles) changes from 6 : 3 to 3 : 3 after ROSE. The total number of the dataset decreases from 9 to 6. ROSE automatically increases or deletes the original data according to the value of total sampling. The new sample points are generated based on kernel density estimation $\widehat{\mathbf{p}_{\mathbf{n}(\mathbf{s})}}$ which is shown in the following equation:

$$\widehat{\mathbf{p}_{\mathbf{n}(\mathbf{s})}} = \frac{1}{\mathbf{n} * \mathbf{h}} \sum_{\mathbf{j}=1}^{\mathbf{n}} \mathbf{K}\left(\frac{\mathbf{s} - \mathbf{s}_{\mathbf{j}}}{\mathbf{h}}\right), \tag{5}$$

where $\mathbf{s}$ represents a target sample, $\mathbf{n}$ represents the sample size of the dataset, and $\mathbf{h}$ represents the smooth parameters or bandwidth. The default value of $\mathbf{h}$ is set to 1. $\mathbf{K}(*)$ represents the kernel [39]. The common kernels include gauss kernels, uniform kernels, triangular kernels, and so on. From equation (5), we can see that the kernel density estimation is based on the central local function of each sample. In addition, on the basis of the weighted average effect of these local functions, the influence of the average effect on the sample density function is determined and the estimated value is obtained. Some studies have shown that the kernel function is actually a weight function, that is, the degree of influence in estimating the density of the remaining sample points depends on the distance between the points.

4.5. Random Undersampling. Random undersampling (UNDER) is the simplest undersampling method and belongs to randomization resampling. This method is contrary to OVER. UNDER randomly reduces the number of samples of majority class until the number of samples of minority class reaches a given proportion of samples of the majority class. UNDER is more suitable for datasets with large number of samples, and it can save computational and storage cost by reducing samples of datasets. However, UNDER is likely to delete the important samples of the majority class, especially in the case of small imbalanced datasets.

UNDER is more suitable for the case of large number of samples and much noise in the majority class. For the online advertising dataset in our study, there are a lot of noise samples in the majority class (ineffective advertisements). These noise samples are evaluated as ineffective advertisements even though they have good performance in a day, leading to classification difficulty. Moreover, UNDER highlights the characteristics of the minority class (effective advertisements) because part of samples of the majority class is deleted.

The diagram of UNDER is shown in Figure 5.

(1) Figure 5(a) shows an imbalanced dataset which contains the samples of minority class (solid triangles) and the samples of majority class (solid circles).

(2) Delete samples of the majority samples randomly.

(3) Repeat step 2 until the final number of minority samples reaches a given proportion (e.g., 100%) of the number of majority samples.

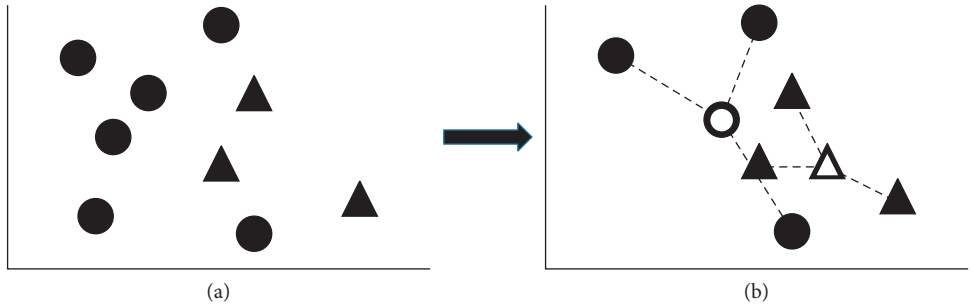

Figure 4: The diagram of ROSE: (a) before sampling; (b) after sampling.

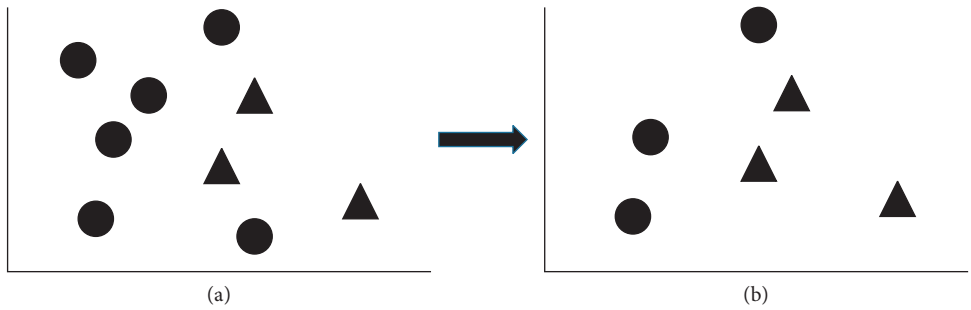

Figure 5: The diagram of UNDER: (a) before sampling; (b) after sampling.

It can be seen in Figure 5 that the ratio of sample numbers of majority class (solid circles) to minority class (solid triangles) changes from 6 : 3 to 3 : 3 after UNDER. For UNDER, the total number of sampling should not be set higher than the total number of samples of the original dataset because UNDER has to reduce part of the samples of majority class from the original dataset. The strategies of UNDER and OVER are opposite. UNDER balances a dataset by reducing samples of majority class while OVER balances a dataset by duplicating samples of minority class.

4.6. Composite Sampling. BOTH is the composition of OVER and UNDER. BOTH is designed to avoid the problems of overfitting caused by OVER and important information loss caused by UNDER. BOTH simultaneously duplicates samples of minority class by using OVER and reduces samples of majority class by using UNDER.

BOTH weakens the shortcomings of OVER and UNDER to a certain extent. However, the method cannot take good advantage of the two methods at the same time. In our comparison experiments, the performance of BOTH is between OVER and UNDER.

The diagram of BOTH is shown in Figure 6.

(1) Figure 6(a) shows an imbalanced dataset which contains the samples of minority class (solid triangles) and the samples of majority class (solid circles).

(2) Delete samples of majority class randomly.

(3) Reproduce the samples of minority class randomly and represent the newly duplicated sample points as hollow triangles.

(4) Repeat steps 2 and 3 until the numbers of majority and minority classes are equal. At this time, the number of total samples of the dataset after BOTH will be more or less than total samples of the original dataset.

Figure 6 shows that sample numbers of majority and minority classes are equal after BOTH. BOTH is a combination of OVER and UNDER, so the newly generated samples also retain the characteristics of the original dataset, that is, these samples coincide with the existing samples of the original dataset.

4.7. Undersampling Based on Clustering. As mentioned above, UNDER is more suitable to solve the imbalance problem of online advertising data. The log data are in days while the class labels of advertisements are based on observation during a period of time, which leads to the problem that effective advertisements are evaluated as ineffective advertisements even though they have good performance in a day. These effective advertisements are noise samples in majority class (ineffective advertisements). UNDER can exactly reduce the noise samples of majority class. Compared with the random resampling, undersampling after clustering can make the distribution of deleted samples more well distributed. Therefore, UNDER is likely to achieve better performance in the application research of our study.

In order to achieve better classification performance, we propose a modified method of UNDER based on clustering (UBOC). Compared with the UNDER, UBOC selects sample points regularly according to the data features instead of

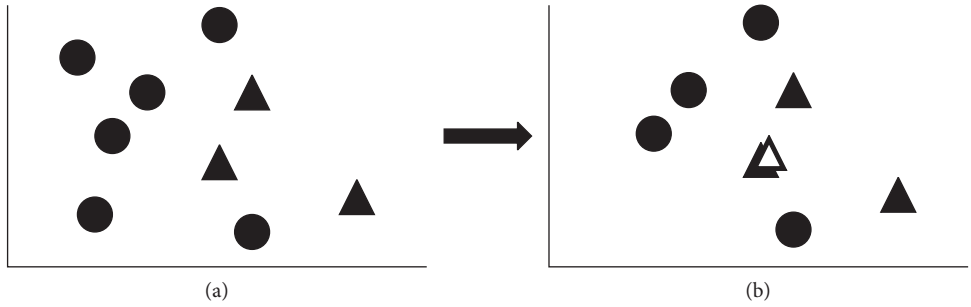

Figure 6: The diagram of BOTH: (a) before sampling; (b) after sampling.

randomly deleting them. We adopt the clustering strategy to group the samples with similar data features so that the sample data will become more balanced and representative.

The diagram of UBOC is shown in Figure 7.

(1) Figure 7(a) shows an imbalanced dataset which contains the samples of minority class (solid triangles) and the samples of majority class (solid circles).

(2) Divide the majority samples into different clusters by clustering methods, such as K-means.

(3) Delete samples in each cluster randomly.

(4) Repeat step 3 until the final number of minority samples reaches a given proportion (e.g., 100%) of the number of samples of majority class.

UBOC improves sampling performance on the basis of UNDER by considering the feature information of sample data. It divides the samples into many clusters before sampling, which makes the noise points more likely to be deleted and ensures that a certain number of points in each cluster are retained. Furthermore, UBOC overcomes the defect of randomness in randomization resampling. It also enables the process of undersampling more explicable.

4.8. C5.0 Decision Tree Model. Decision tree (DT) is a kind of supervised prediction model to represent the mapping relationship between object attributes and class labels. In the DT model, the internal nodes represent object attributes and the leaf nodes represents the class labels. The DT model is trained by the training set. The test set can be classified according to the classification rules of the trained DT model. When classifying a target sample by DT, the attributes of the target sample need to be distinguished by the classification rules of the internal nodes at each level. Finally, the target sample will be classified into a corresponding leaf node. The class label of the leaf node is the predicted class label for the target sample.

The C5.0 DT model applies boosting technology [40] to improve the accuracy of DT classification. More importantly, it constructs a smaller set of classification rules. The rules are still very robust and efficient in dealing with the case of large quantity of attributes [41]. Therefore, we use C5.0 DT model for the classification of online advertisements.

In addition, C5.0 DT has implemented cost-sensitive learning. We can use cost matrix to construct cost-sensitive DT to minimize the sum of misclassification cost when choosing split attributes for internal nodes of the DT model. Generally, the misclassification cost of effective advertisements and ineffective advertisements can be determined by changing the value of cost matrix. Instead of solely concerning the low error rates, we use the C5.0 DT model with cost-sensitive learning to minimize the misclassification cost of advertisements in order to help advertisers to achieve better profits.

5. Experiment

5.1. Experiment Data. In order to ensure the reliability of the experimental results, the experimental data used in our study are collected from the real enterprise environment of online advertising. The total number of samples (advertisements) is 45,683 after data preprocessing. According to the rate of return on investment (ROI), the online advertisements are divided into two classes: ineffective advertisements and effective advertisements. The proportion of effective advertisements is 21% and the proportion of ineffective advertisements is 79%. Obviously, the advertising data used in experiments are imbalanced. In addition, the data features are selected based on the filtering method in order to improve the operation efficiency of our algorithm. Table 3 shows the original features of online advertising data, and Table 4 shows the selected features of three types.

The experimental data in our study contain 11 days of online advertising data of a cross-border e-commerce enterprise from September 1 to 11 in 2018. The training set is constructed as follows. Firstly, the data of September 1 are used as training set and the data from September 2 are used as test set. Secondly, the dataset from September 1 and September 2 is merged as training set and the dataset from September 3 is used as test set. Thirdly, the dataset from September 1 to September 3 is also merged as training set and the dataset from September 4 is used as test set, and so on until dataset from September 1 to September 10 is used as training set and dataset from September 11 is used as test set. Ten training sets and ten corresponding test sets form ten experimental groups, which are labeled 1–10, respectively. The training sets and test sets are shown in Table 5.

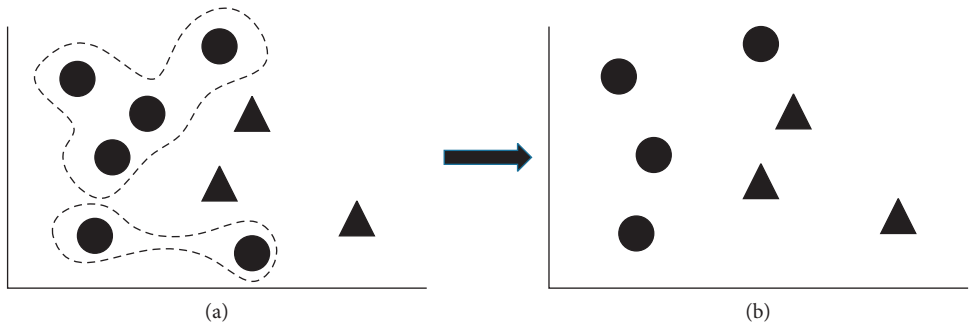

Figure 7: The diagram of UBOC: (a) before sampling; (b) after sampling.

Table 3: The original features of online advertising data.

Features of online advertising data		
Number of payers	Views of business page = 1	Number of total search volume
Number of payments	Views of business page ≥ 8	Number of total register volume
Amount of total payments	Average views of business page	Number of total purchase volume
Number of order people	Views of home page = 1	Number of total collection volume
Number of orders	Views of home page ≥ 8	Number of total recommendation volume
Amount of total orders	Average views of home page	Residence time of user ≥ 1 min
Ad type	User gender	Residence time of user ≥ 8 min
Ad budget	User amount	Average residence time of user
Ad position	Max of user age	Type of media platform
ROI on the day	Min of user age	

5.2. Experimental Process. We carry out an experimental analysis to compare the resampling effect of our proposed method UBOC and other common methods. We balance the distribution of advertisement data by using different resampling methods in order to compare their performance on balanced datasets. Detailed flowchart of our experimental process is shown in Figure 8.

From the experimental process, the balanced advertising data after resampling are used to construct the DT model. **R** can be further calculated by the results provided by the DT model. Performance of the resampling methods can be compared by comparing their corresponding values of **R**. The core pseudocode for the experimental process is given in Figure 9.

5.3. Experiment Result. In our study, we represent experiments of six resampling methods combined with DT as OVER-DT, SMOTE-DT, ROSE-DT, UNDER-DT, BOTH-DT, and UBOC-DT. As a control group, the experiment of DT without resampling is represented as NR-DT.

5.3.1. Comparison Experiment on Resampling Effect. Resampling methods will produce different results under different values of random seeds. In order to maintain robustness, the experiments are repeated five times under different values of random seeds to calculate **R**. The final results of **R** are the average values of the five times experiments. The **R** values of final results are shown in Table 6 and Figure 10.

According to Table 6 and Figure 10, the five common resampling methods are sorted according to their resampling effect (from the best to the worse): UNDER, BOTH, SMOTE, OVER, and ROSE. The resampling effect of UNDER is much better than other methods. The resampling effect of OVER, SMOTE, and BOTH does not differ from each other greatly. Table 7 shows values of TPR, FPR, and **R** in the five repeated experiments in group 10. It is obvious to find that UNDER performs the best while ROSE performs the worse.

From Table 6, the resampling methods can improve the prediction performance of the model except ROSE. Compared with the control group, ROSE has lower values of **R**. However, we cannot reach the conclusion that ROSE has no contribution on advertising effect evaluation. It merely means that ROSE cannot effectively reduce FPR to increase the value of **R**. From Table 7, values of TPR are up to 99% and values of FPR are up to 87% in ROSE. The strategy of ROSE is to improve the prediction accuracy of ineffective advertisements at the cost that most of the effective advertisements are predicted as ineffective ones. However, the resampling of ROSE does not meet the requirements of advertisers' cost-sensitive advertising strategies, so the value of **R** is only around 1.

All in all, compared with the experimental group without resampling, resampling can improve the effect evaluation of online advertisements by changing the distribution of advertising data. It deletes useless data from the ineffective advertisements or adds useful data into the effective advertisements, so as to make the features of effective or

Table 4: The selected features of three types.

Features of advertising information	Features of information on media platform	Features of user information
Advertising type	Advertising position	User gender
Amount of total payment		Max of user age
Amount of total order		User amount
Average views of business page	Type of media platform	Number of total search volume
Average residence time of user		Number of total collection volume
Average views of home page		Number of total recommendation volume
Advertising budget		Number of total purchase volume
ROI on the day		Number of total register volume

Table 5: The construction process of dataset.

Group	Training set	Test set
1	Sep.1	Sep.2
2	Sep.1-Sep.2	Sep.3
3	Sep.1–Sep.3	Sep.4
4	Sep.1–Sep.4	Sep.5
5	Sep.1–Sep.5	Sep.6
6	Sep.1–Sep.6	Sep.7
7	Sep.1–Sep.7	Sep.8
8	Sep.1–Sep.8	Sep.9
9	Sep.1–Sep.9	Sep.10
10	Sep.1–Sep.10	Sep.11

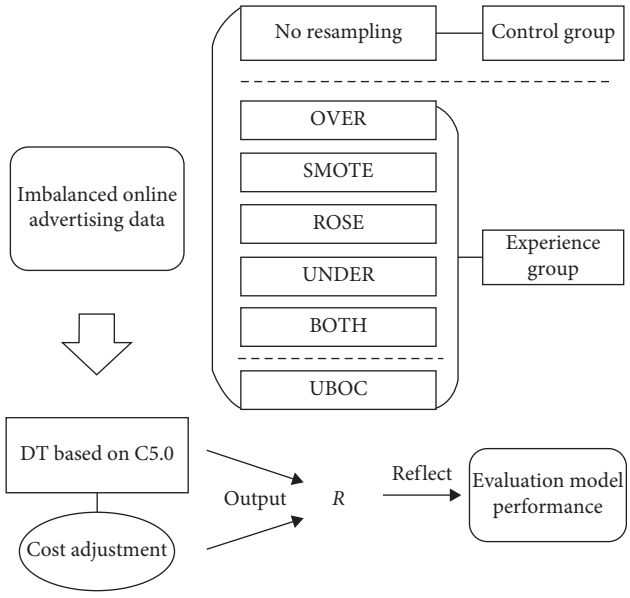

Figure 8: The schematic flowchart of resampling experiment.

ineffective advertisements much clearer. Besides, some of the ineffective advertisements perform well in a short time, which will affect the classification performance to a great extent. In theory, the strategy of undersampling is more suitable for advertising data than other sampling methods and our experimental results confirm this point.

5.3.2. Verification Experiment on Resampling Effect of UBOC. In previous experiments, UNDER outperforms other common methods. The disadvantage of UNDER is that it deletes a large number of samples of majority class to increase the importance of minority class. In actual online advertising data, there are too many ineffective advertisements. Some of the effective advertisements will also be labeled as ineffective advertisements because they do not show characteristics of effective advertisement in a short period of time. The classification results will be biased towards ineffective advertisements if the imbalanced advertising data are used for training the DT model. Thus, we can draw a conclusion that compared with other resampling methods, UNDER is more suitable for effect evaluation of online advertisements under the cost-sensitive advertising strategies. In order to further improve the performance of effect evaluation, we propose our method UBOC to improve UNDER.

UBOC uses K-means clustering to generate N clusters and then deletes the corresponding number of samples from each cluster. Samples in different clusters have different feature information. Resampling strategy according to clusters avoids the problem that the samples with important features are eliminated at all. The comparison experiments of resampling effect of UNDER and UBOC are shown in Table 8.

From Table 8, in all the experimental groups, the values **R** of UBOC are larger than those of UNDER, indicating better performance of UBOC. In fact, although UNDER deletes some advertising data whose features and classification label are inconsistent, the random deletion strategy is likely to accidentally remove the useful data. However, UBOC divides the advertising data into different clusters according to the data features in advance. In this way, we can remove the abnormal data and retain the useful data in the deletion step.

Meanwhile, the experimental results also demonstrate that the resampling based on the data features is more effective than the randomly resampling strategy for the effect evaluation of online advertisements. We can find that **R** of UBOC is more stable than that of UNDER in ten groups of experiments. It is because the resampling data will be different from the change of random seeds, which is difficult to ensure that the results definitely meet the expectations. However, UBOC deletes data based on the data features instead of the randomicity, so the experimental results can also verify that UBOC is more explanatory and practical than UNDER for the online advertising data.

5.3.3. Observation Experiment on Different Cost Settings. From the previous discussion, we can further increase the value of **R** by adjusting the misclassification cost parameter

```
Input: Online advertising dataset D
Procedure: Function Resampling, Decision Tree
Output: R
if D has not performed preprocessing then
   perform preprocessing for D; return
end if
if D is imbalanced then
   perform Resampling (D) get D*; return
end if
Set the value of cost in decision tree model
Generate confusion matrix by Decision Tree (D*)
Calculate the value of TPR and FPR
if FPR ≠ 0 then
   R = TPR/FPR; return
end if
```

Figure 9: The core pseudocode.

Table 6: The result of **R** from different resampling methods.

Group	1	2	3	4	5	6	7	8	9	10
NR-DT	2.09	2.01	1.71	2.18	2.37	2.41	2.79	2.81	2.20	2.00
OVER-DT	2.69	2.72	2.04	2.63	3.15	3.07	2.33	3.33	4.87	3.56
SMOTE-DT	2.89	2.79	2.42	2.26	3.00	2.81	3.99	3.65	4.38	2.72
ROSE-DT	1.32	1.22	1.12	1.19	1.20	1.12	1.25	1.28	1.15	1.14
UNDER-DT	3.38	3.59	2.63	3.45	3.85	3.94	3.06	4.32	6.11	3.77
BOTH-DT	3.07	3.02	2.16	2.77	3.07	3.43	2.40	3.41	4.85	3.79

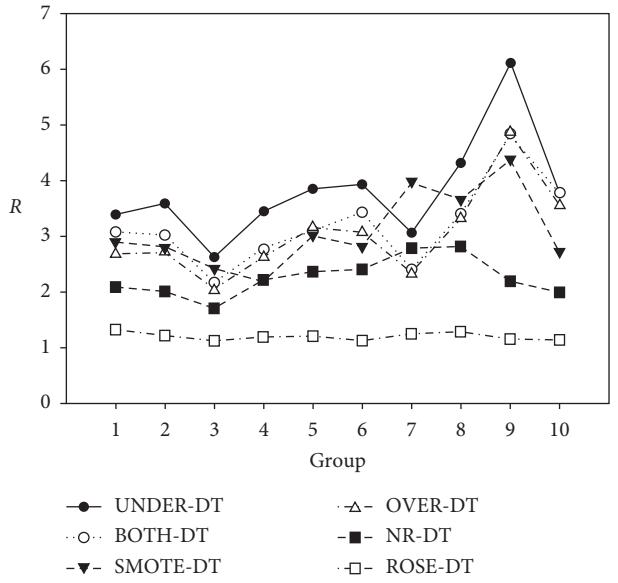

Figure 10: Comparison of impact on model of different resampling methods.

in the DT model according to cost-sensitive learning. Taking group 10 as an example, we observe the changes in the adjustment of different costs for UNDER. The cost of predicting effective advertisements as ineffective ones is regarded as a high level, so we conduct experiments of five different values of the cost (from 1 to 5). The experimental results are shown in Table 9.

Since the cost parameter defaults to 1, **R** of UNDER in the second column of Table 9 is the same as that in the fourth column of Table 7. It is obvious that value of **R** significantly increases as the cost increases. Although the FPR is reduced as much as possible, the TPR is also reduced at the same time. Therefore, we need to consider the acceptable level of TPR when setting the cost.

In summary, we can conclude that cost-sensitive learning is able to further increase values of **R** based on the effective resampling method. There are many ineffective advertisements performing well in a short term, so we

Table 7: Different indexes after five resampling methods in group 10.

Random seed	Index	BOTH-DT	UNDER-DT	OVER-DT	SMOTE-DT	ROSE-DT
1	TPR	77.48%	71.99%	78.49%	49.78%	99.91%
	FPR	17.42%	18.91%	23.17%	17.52%	87.58%
	R	4.4469	3.8076	3.3875	2.8422	1.1408
2	TPR	72.22%	72.45%	74.17%	49.81%	99.94%
	FPR	20.30%	20.11%	17.98%	17.79%	87.49%
	R	3.5581	3.6023	4.1254	2.7994	1.1424
3	TPR	72.22%	77.08%	75.12%	50.94%	99.88%
	FPR	19.83%	19.09%	23.54%	19.37%	85.73%
	R	3.6412	4.0372	3.1912	2.6296	1.1651
4	TPR	74.06%	75.99%	76.13%	62.35%	99.94%
	FPR	19.56%	19.37%	20.48%	20.39%	86.10%
	R	3.7871	3.9229	3.7169	3.0581	1.1608
5	TPR	79.15%	75.84%	77.05%	54.62%	99.94%
	FPR	22.61%	21.78%	22.80%	24.10%	92.86%
	R	3.5001	3.4822	3.3795	2.2666	1.0762
Average R		3.7867	3.7704	3.5601	2.7192	1.1371

Table 8: The results of **R** between UNDER and UBOC in 10 groups.

Group	1	2	3	4	5	6	7	8	9	10
UNDER-DT	3.38	3.59	2.63	3.45	3.85	3.94	3.06	4.32	6.11	3.77
UBOC-DT	3.56	3.84	2.70	3.82	4.13	4.16	3.15	6.39	6.83	5.19

Table 9: The different values after setting cost in decision tree.

	Cost				
Index	1	2	3	4	5
TPR	71.99%	70.23%	68.68%	64.97%	61.60%
FPR	18.91%	14.46%	12.33%	9.27%	10.19%
R	3.8076	4.8578	5.5719	7.0102	6.0429

increase the misclassification cost of ineffective advertisements to let the DT model pay more attention to the useful data information. The strategy of cost-sensitive learning significantly reduces FPR with the small loss of TPR, which can improve the evaluation index from the perspective of model parameters.

6. Suggestion and Conclusion

6.1. Suggestion. Enterprises often spend a lot of money on promoting online advertisements, so it is important to establish a perfect and convenient evaluation system. The following suggestions are put forward for future research studies.

On the one hand, more optimized heuristic resampling methods can be adopted for the strategies of resampling. The heuristic sampling method is robust and is able to achieve better performance. For example, the heuristic undersampling method also includes the simple integration method and balanced cascade method. The simple integration method extracts some independent samples of majority class to generate multiple subsets and then merges the minority class with these subsets. The method trains multiple classifiers based on the merged datasets. The balanced cascade method generates multiple classifiers and then systematically retains the samples of majority class based on certain rules. The practical effect of the above methods for advertising prediction needs further discussion.

On the other hand, some advanced methods can be introduced for the prediction of online advertising effect. Some literatures on online advertising have employed machine learning models, such as the neural network. Future studies should cover the topics of combining machine learning models with practical advertising strategies.

6.2. Conclusion. Based on the real-world log data, our study aims to improve the effect evaluation of online advertisements so as to provide practical guideline for e-commerce enterprises. Our contributions are mainly as follows. (1) For the imbalance of advertising data, we use the resampling method and propose UBOC to improve the prediction performance of online advertisements. (2) In order to meet the requirements of enterprises, we present the evaluation index (**R**) to better reflect the effect of evaluation model for online advertisements. Based on the above contributions, the following conclusions are drawn. (1) The data imbalance of online advertisements can be dealt with through the

resampling method, which improves the prediction performance. We compare several resampling methods and find that the strategy of undersampling is more suitable for the data preprocessing of online advertisements. On this basis, we propose UBOC to further eliminate the negative effects on the prediction model from data imbalance. (2) The optimized effect evaluation of online advertisements can fulfill the actual requirements of enterprises. We present a new evaluation index (**R**) available for online advertising, which has better flexibility and interpretability.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the Project of National Natural Science Foundation of China (grant no. 71731006), Fundamental Research Funds for Guangdong Natural Science Foundation (grant nos. 2018A030313795 and 2019A1515011386), Guangdong Soft Science Foundation (grant no. 2019B101001025), Guangdong Philosophy and Social Science (grant no. GD19CGL29), and Science and Technology Development Minister of Education (grant no. 2019J01001).

References

[1] F. Ye and F. Ding, "Research and application of unbalanced data classification," *Computer Applications and Software*, vol. 35, no. 1, pp. 132–136, 2018.

[2] Yi Li, T. Jiang, and Y. Liu, "Internet personal credit assessment research based on the perspective of unbalanced sample," *Statistics and Information Forum*, vol. 32, no. 2, pp. 84–90, 2017.

[3] J. Ding, G. Liu, and H. Li, "The application of improved random forest in the telecom customer churn prediction," *Pattern Recognition and Artificial Intelligence*, vol. 28, no. 11, pp. 1041–1049, 2015.

[4] W. Wei, J. Li, L. Cao, Y. Ou, and J. Chen, "Effective detection of sophisticated online banking fraud on extremely imbalanced data," *World Wide Web*, vol. 16, no. 4, pp. 449–475, 2013.

[5] S. Shahriar, O. Burkay, and D. Ali, "Online evaluation of bid prediction models in a large-scale computational advertising platform: decision making and insights," *Knowledge and Information Systems*, vol. 51, no. 1, pp. 37–60, 2017.

[6] L. Shan, L. Lin, C. Sun, and X. Wang, "Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization," *Electronic Commerce Research and Applications*, vol. 16, pp. 30–42, 2016.

[7] B. J. Jansen, K. Sobel, and M. Zhang, "The brand effect of key phrases and advertisements in sponsored search," *International Journal of Electronic Commerce*, vol. 16, no. 1, pp. 77–106, 2011.

[8] G. Zenetti, T. H. A. Bijmolt, P. S. H. Leeflang, and D. Klapper, "Search engine advertising effectiveness in a multimedia campaign," *International Journal of Electronic Commerce*, vol. 18, no. 3, pp. 7–38, 2014.

[9] G. R. Massey, P. Z. Wang, D. S. Waller, and E. V. Lanasier, "Best-worst scaling: a new method for advertisement evaluation," *Journal of Marketing Communications*, vol. 21, no. 6, pp. 425–449, 2015.

[10] N. Falciani-White and L. Tomcik, "Advertise with confidence: evaluating advertising assessment in light of business best practices," *College and Undergraduate Libraries*, vol. 22, no. 1, pp. 45–60, 2015.

[11] W. Zhou, S. Bandyopadhyay, H. K. Cheng, and P. Pathak, "A mechanism for on-line advertisement placement to deter click fraud," *International Journal of Electronic Commerce*, vol. 13, no. 2, pp. 9–28, 2008.

[12] K. Wang, E. T. G. Wang, and C.-K. Farn, "Influence of web advertising strategies, consumer goal-directedness, and consumer involvement on web advertising effectiveness," *International Journal of Electronic Commerce*, vol. 13, no. 4, pp. 67–96, 2009.

[13] Y.-M. Li, L. Lin, and S.-W. Chiu, "Enhancing targeted advertising with social context endorsement," *International Journal of Electronic Commerce*, vol. 19, no. 1, pp. 99–128, 2014.

[14] D. Tu, C. Shu, and H. Yu, "Context advertising recommendation based on joint probability matrix decomposition," *Journal of Software*, vol. 24, no. 3, pp. 454–464, 2013.

[15] C. Li, Y. Wu, and C. Qin, "Research on search engine advertisement click rate prediction model based on LASSO variable selection method," *Application of Statistics and Management*, vol. 35, no. 5, pp. 803–809, 2016.

[16] F. Shen, G. Dai, C. Dai et al., "CTR prediction for online advertising based on a features conjunction model," *Journal of Tsinghua University (Science and Technology)*, vol. 58, no. 4, pp. 374–379, 2018.

[17] Z. Zhang, Y. Zhou, X. Xie et al., "Research on advertising click-through rate estimation based on feature learning," *Chinese Journal of Computers*, vol. 39, no. 4, pp. 780–794, 2016.

[18] Y. Peng and P. Lai, "Evaluation of online advertising based on data mining," *Statistics and Decision*, vol. 15, no. 21, pp. 86–88, 2014.

[19] H. Guo, R. Shuai, X. Zhang et al., "Click fraud detection method based on ensemble feature selection," *Computer Engineering and Applications*, vol. 55, no. 17, pp. 246–251, 2019.

[20] X. Yao, X. Wang, Y. Zhang et al., "Summary of feature selection algorithms," *Control and Decision*, vol. 27, no. 2, pp. 161–166, 2012.

[21] Z. Hu, X. Yuan, J. Li et al., "Robust fragments-based tracking with multi-feature joint kernel sparse representation," *Journal of Computer Research and Development*, vol. 52, no. 7, pp. 1692–1704, 2015.

[22] F. Huang, F. Shi, D. Wang et al., "Mining topic sentiment in microblogging based on multi-feature fusion," *Chinese Journal of Computers*, vol. 40, no. 4, pp. 872–888, 2017.

[23] Z. Zeng, C. Wu, Q. Tang et al., "Classification of commodity image based on multi-feature fusion and depth learning," *Computer Engineering and Design*, vol. 38, no. 11, pp. 3093–3098, 2017.

[24] Y. Cui, A. F. Wise, and K. L. Allen, "Developing reflection analytics for health professions education: a multi-

dimensional framework to align critical concepts with data features," *Computers in Human Behavior*, vol. 100, pp. 305–324, 2019.

[25] T. J. Jason, M. Perry, and P. O. Kristensson, "Differentiation of online text-based advertising and the effect on users' click behavior," *Computers in Human Behavior*, vol. 50, pp. 535–543, 2015.

[26] Y.-C. Lee, "Comparing factors affecting attitudes toward LBA and SoLoMo advertising," *Information Systems and E-Business Management*, vol. 16, no. 2, pp. 357–381, 2018.

[27] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[28] C. Ji, "Solutions to unbalanced data classification," *Electronic Technology and Software Engineering*, vol. 14, no. 15, pp. 152-153, 2018.

[29] K. J. Wang, B. Makond, and K. M. Wang, "An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data," *BMC Medical Informatics and Decision Making*, vol. 13, no. 1, p. 124, 2013.

[30] S. Cateni, V. Colla, and M. Vannucci, "A method for resampling imbalanced datasets in binary classification tasks for real-world problems," *Neurocomputing*, vol. 135, no. 8, pp. 32–41, 2014.

[31] R. Blagus and L. Lusa, "Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models," *BMC Bioinformatics*, vol. 16, no. 1, p. 362, 2015.

[32] J. Ma, D. Hu, and X. Sun, "SMUP: synthetic minority using proximity of random forests," *Application of Statistics and Management*, vol. 34, no. 5, pp. 809–820, 2015.

[33] Q. Li, Y. Zhao, and Z. Gu, "Design of loss function for cost-sensitive learning," *Control Theory and Applications*, vol. 32, no. 5, pp. 689–694, 2015.

[34] M. Richardson, E. Dominowska, and R. Ragno, "Predicting clicks: estimating the click-through rate for new ads," in *Proceedings of the 16th International Conference on World Wide Web*, Alberta, Canada, May 2007.

[35] O. Chapelle and Y. Zhang, "A dynamic bayesian network click model for web search ranking," in *Proceedings of the 18th International Conference on World Wide Web*, Madrid, Spain, April 2009.

[36] K. S. Dave and V. Varma, "Learning the click-through rate for rare/new ads from similar ads," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, Geneva, Switzerland, July 2010.

[37] P. K. Chan, W. Fan, A. L. Prodromidis, and S. J. Stolfo, "Distributed data mining in credit card fraud detection," *IEEE Intelligent Systems*, vol. 14, no. 6, pp. 67–74, 1999.

[38] Z. Li, J. Du, B. Nie et al., "Summary of feature selection methods," *Computer Engineering and Applications*, vol. 55, no. 24, pp. 10–19, 2019.

[39] A. M. Charles and P. Massimiliano, "Learning the kernel function via regularization," *Journal of Machine Learning Research*, vol. 6, no. 6, pp. 1099–1125, 2005.

[40] T. Yu and W. Zhang, "Research on boosting theory and its applications," *Journal of University of Science and Technology of China*, vol. 46, no. 3, pp. 222–230, 2016.

[41] L. Luo, L. Duan, W. Duan et al., "An improvement and application of C5.0 algorithm," *Journal of Nanchang University (Engineering and Technology)*, vol. 39, no. 1, pp. 92–97, 2017.