Hindawi
Security and Communication Networks
Volume 2023, Article ID 9837989, 1 page
https://doi.org/10.1155/2023/9837989

Retraction

Retracted: Construction of Digital Marketing Recommendation Model Based on Random Forest Algorithm

Security and Communication Networks

Received 26 December 2023; Accepted 26 December 2023; Published 29 December 2023

This article has been retracted by Hindawi, as publisher, following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of systematic manipulation of the publication and peer-review process. We cannot, therefore, vouch for the reliability or integrity of this article.

Please note that this notice is intended solely to alert readers that the peer-review process of this article has been compromised.

Wiley and Hindawi regret that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] W. Gao and Z. Ding, "Construction of Digital Marketing Recommendation Model Based on Random Forest Algorithm," *Security and Communication Networks*, vol. 2022, Article ID 1871060, 9 pages, 2022.

Hindawi
Security and Communication Networks
Volume 2022, Article ID 1871060, 9 pages
https://doi.org/10.1155/2022/1871060

Research Article

Construction of Digital Marketing Recommendation Model Based on Random Forest Algorithm

Weiji Gao [1,2] **and Zhihua Ding** [1]

[1]*School of Economics and Management, China University of Mining and Technology, Xuzhou 221116, China*
[2]*School of Business, Jiangsu Vocational College of Electronic and Information, Huai'an 223001, China*

Correspondence should be addressed to Weiji Gao; weijigao1913@sina.com

Received 18 May 2022; Revised 29 June 2022; Accepted 6 July 2022; Published 12 August 2022

Academic Editor: Tao Cui

The traditional marketing model can no longer meet the needs of users and can not add more benefits to the enterprise, and digital marketing came into being. At present, most of the marketing focus of various enterprises is still mainly on products, and the reflection arc to market changes is long. Therefore, the formulation of marketing activities should always pay attention to changes in user needs and combine corresponding activity planning, product planning, brand building, etc., according to Changes in the target market adjust the content of marketing activities and products in real-time and, at the same time, pay attention to user feedback on products in order to iteratively update products in time, improve product market competitiveness, and optimize the user experience. In this paper, through the study and research of the traditional random forest method and some data processing algorithms, the feature selection and class imbalance problems of random forest are improved, respectively. Through the study of feature selection methods, we can maintain a balance between feature strength and relevance during feature selection and improve the final model classification effect. And through the research and experiment of the imbalanced data classification problem and the random forest algorithm, the method of the random forest model to deal with the imbalanced problem has been improved. After experimental calculation and analysis, it is found that for the effect of the minimum number of samples required for node splitting with different numbers, the best results are obtained when 2 samples are taken as the minimum number of samples required for node splitting, and the average value of the $F1$ evaluation is 0.1038; for different specifications, the effect of the random forest is the best using the Gini index, and the average value of its $F1$ evaluation is 0.1033; for the effect analysis of random forests with different numbers of trees, 7 to 10 decision trees are the best, and the $F1$ evaluation is the best. The average is 0.10175.

1. Introduction

Today is an era of information explosion. Compared with the previous situation of lack of information, massive data resources present a colorful Internet world for users, but a large amount of data presents a severe test to users' screening ability. Faced with the massive amount of information data at this stage, the selection and filtering of information have become a priority target to weigh the pros and cons of a system. A platform with excellent user experience will optimize a large amount of information in various ways and present the information that users need and are most interested in their search results [1]. Recommender systems and search engines can play complementary roles in terms of users' information retrieval needs. When the user's information needs are clear, the search engine can meet the user's goal, and when the user does not have a clear goal, the recommender system can provide some information for the user to use. A good recommender system can not only provide users with personalized information service needs, but also allow users to have a high degree of trust in it and improve user stickiness. Whenever users cannot clarify their needs, they think of the recommender system [2]. Recommender systems have been widely used in many fields, among which the most typical and promising application field is digital marketing. At the same time, the research enthusiasm for recommender systems has always been high in academia and has gradually formed an independent discipline [3].

The digital marketing recommendation system solves the problem of information overload by recommending new items that the user has little contact with for digital marketing, and these new items are related to the current needs of the user. The digital marketing recommendation system utilizes the data about users, available items, existing transactions, and various types of other data stored in the existing database environment for digital marketing recommendations. Users may or may not accept content recommended by digital marketing, or may provide explicit or implicit feedback over a period of time, which is very valuable to digital marketing recommendation systems [4]. Therefore, all the feedback information of users can be stored in the corresponding database of digital marketing recommendations, so that by using these feedbacks information, the digital marketing recommendation system can generate new digital marketing recommendation when the user has relevant behavior next time. The random forest used at this stage is purely random in feature selection, using a combination of multiple classifiers, each individual classifier votes, and the combined classifier makes predictions based on the returned class label of the vote. The benefit of the combined classifier is that it is more accurate than the individual classifiers in it [5]. Here, decision trees are used as individual classifiers, which are assembled into a forest. Individual decision trees use random selection features to decide the partition at each node. Each tree relies on independent sampling and has the same distribution of random vector values as all trees in the forest.

The problem of classification is a common problem in digital marketing. The so-called classification is to find a set of models that can describe the common characteristics of all the data from the known data, so as to be able to identify the category of the unknown data. When solving a classification problem, a classification algorithm is usually used to build a classification model, which represents a collection of classification knowledge. To solve the classification problem, researchers have proposed hundreds of classification algorithms, which come from different fields and have different functions [6]. Various classification problems in digital marketing can be solved by these classification algorithms. This article will introduce several commonly used classification algorithms in detail and explain the use and feasibility of classification algorithms in digital marketing in combination with marketing cases in actual work. Digital marketing theory is the basic theoretical framework, and the focus of marketing activities is to effectively analyze marketing data and how to use random forests to improve data analysis capabilities. Taking the common classification problem in digital marketing as an example, this paper introduces the principle and definition of the classification problem in detail and the classification problem in digital marketing and uses a random forest algorithm to solve the classification problem.

Chapter arrangement of this paper: Chapter 1 introduces the related scholars' research on random forest algorithm and recommendation system; Chapter 2 introduces random forest algorithm technology; Chapter 3 classifies users based on the actual situation of digital marketing design; Chapter 4 conducts experiments on the optimal value range of each specification of the proposed random forest algorithm for recommendation model and draws the results; and Chapter 5 summarizes the full text.

The innovation of this paper: Through the study and research of the traditional random forest method and some data processing algorithms, the feature selection and class imbalance problems of random forest are improved, respectively. Through the study of feature selection methods, we can maintain a balance between feature strength and relevance during feature selection and improve the final model classification effect. And through the research and experiment of the imbalanced data classification problem and the random forest algorithm, the method of the random forest model to deal with the imbalanced problem has been improved.

2. Related Work

Whether a recommendation system can be accepted by users depends on whether the recommendation algorithm it adopts is accurate and efficient. At present, in the field of mainstream recommendation systems, the random forest algorithm is mainly used. Due to the excellent characteristics of the random forest algorithm, it has been widely used in many fields.

Rumín introduced the Hough transform into the voting mechanism of random forest and obtained the Hough forest algorithm, which is widely used in various target tracking and behavior recognition. Hough forest optimizes and improves the voting process of random forest and uses Hough voting to replace the majority voting mechanism. In addition to the category test, the leaf nodes of the Hough forest also use the generalized Hough transform to estimate the maximum posterior probability in the Hough space to detect the impurity of the offset [7]. Liu et al introduced the concept of the survival tree, improved the random forest construction process, and obtained the random survival forest algorithm. The algorithm also uses the bootstrap resampling method to select N training sets during sampling. The difference is that a generative analysis tree is built for each training set. When voting, the survival function of each tree is calculated. The combined value is the mean survival function, which combines the content of the resulting parse tree and the predictions. When calculating the survival function, the algorithm adopts the method of KM estimation [8]. Hasanat et al. also theoretically proved the consistency of random survival forest and believed that random survival forest (RSF) is significantly better than other survival analysis methods when dealing with high-dimensional data [9]. Ohtake, Seki, and Kodaka proposed to introduce NCL technology into the random forest algorithm, mainly for the unbalanced training set, perform NCL technology processing on the data, and then classify the processed data with the random forest algorithm. The test results show that the improved random forest algorithm. The forest algorithm classification effect is better [10]. Prabhu developed a new cost-sensitive random forest algorithm from the perspective of applying cost-sensitive learning algorithm to solve the

problem of unbalanced dataset classification. The cost-sensitive random forest algorithm first forms multiple bags by performing an alternative random sampling method on the original training set; then randomly selects some attributes in each bag to prune, thereby generating a cost-sensitive decision tree. The bags are combined to form an ensemble algorithm [11]. Mathai and Jeswani proposed a new random forest feature selection algorithm by analyzing the relationship between the strength and correlation coefficient of each tree in random forest. The main idea of the algorithm is to find out through the analysis of the upper bound of the generalization error of random forest: when building a random forest, increasing the strength of the decision tree in the forest can achieve the goal of reducing the generalization error of random forest. But they also found that while increasing the strength, the correlation between decision trees should be minimized [12]. Xiang et al. proposed a method of generating multiple attributes at each node of the decision tree in view of the fact that each node of the traditional decision tree algorithm has only one attribute, and the process of generation and decision-making is complicated. Random forest algorithm uses fewer and targeted attributes to randomly select each data clustering heap to generate multiple multivariate decision trees. Finally, a random forest is formed according to the multiple multivariate decision trees constructed by each clustering heap to carry out the weighted integration of the classifiers, thus trying to cover all the concepts in the dataset and further ensuring the classification accuracy of the random forest algorithm [13]. Feng et al. focused on the relationship between the strength and correlation of each sub-classifier in the RF algorithm, but cannot effectively predict the strength and correlation of the forest before constructing a subtree [14]. Rewade adopts the backward selection method of sequence and generalized sequence for feature selection and proposes an encapsulated feature selection algorithm based on random forest, but this algorithm cannot determine the generalized backward search method in high-dimensional data sets L value [15]. In order to reduce the generalization error of the balanced sample set, the error is reduced by about 15% by changing the sample size of the training set and the sample sampling method, but this method fails to obtain the quantitative relationship between the degree of balance and the sampling interval[16].

From the above aspects of optimization, the improvement of the classification performance of the random forest algorithm mainly focuses on the combination research of various algorithms. The research of these combinations is generally applied in a specific field; the second aspect focuses on the special data set in application optimization in random forest; and the optimization method of the third aspect is mainly through the improvement of the algorithm itself, so that it has certain versatility and can be applied in different fields. This paper intends to start from the second and third aspects to optimize the random forest algorithm to build a digital marketing recommendation model.

3. Introduction to Random Forest Algorithm

3.1. Decision Tree. Before understanding random forest, we must first understand decision tree. Decision tree is a tree structure classification from top to bottom. Its core idea is to find purity classification. Purity means that the target variables can be completely separated, that is, y can only be equal to 0 during classification. or 1. Decision tree is one of the most classic data mining algorithms. It shows the decision-making and classification process in a tree structure, which is simple, intuitive, and highly interpretable. In this article, classification trees are used. In the decision-making process, when the root node is split into leaf nodes according to the characteristic attributes, how to judge the quality of the leaf nodes obtained by splitting? Usually, the Gini index (Gini impurity) is used to judge. Gini index (Gini impurity): indicates that in the sample The probability that a randomly selected sample in the set is wrongly classified, so the smaller the Gini index, the smaller the probability that the selected sample in the set is wrongly classified, that is to say, the higher the purity of the set, the more impure the set is, the more That is, the more impure the split leaf node is [17, 18]. Its mathematical formula is shown as follows:

$$\mathrm{Gini}(p) = \sum_{p=1}^{K} p_k(p_k - 1) + p_k^2. \tag{1}$$

Among them, p represents the probability that the randomly selected sample belongs to the k category. After the root node is split into two leaf nodes, the purity of the Gini index is calculated. If the purity is small enough and the Gini value is small enough, the two leaf nodes can be regarded as leaves. The nodes are no longer classified. If the purity is high, the above two leaf nodes will be used as a new set and continue to be split until the purity is small enough to meet the standard [19].

Random forest is to create a forest by random method. The forest consists of many decision trees and each decision tree is not correlated. The result is shown in Figure 1.

After the random forest is established, when a new sample enters the forest, each decision tree in the forest performs decision classification to see which class is selected the most to predict which class the new sample belongs to. Since the random forest will randomly sample the input data set with the replacement of rows and columns, the phenomenon of repeated sampling may occur [20]. Assuming that there are m decision trees, m sample sets are required to train each tree. It is not advisable to use full samples to train m trees, because the full samples will ignore the local sample rules, resulting in a weaker generalization ability of the model. With replacement, n samples are extracted, and m decision trees are trained with these n samples, and finally, the prediction classification is obtained. Multiple random variables are usually calculated using the chain rule, as shown in equation (2).

$$p(x_1, x_2 \ldots x_n) = p(x_1) \cup px(x_i | x_1 \ldots x_{i-1}). \tag{2}$$

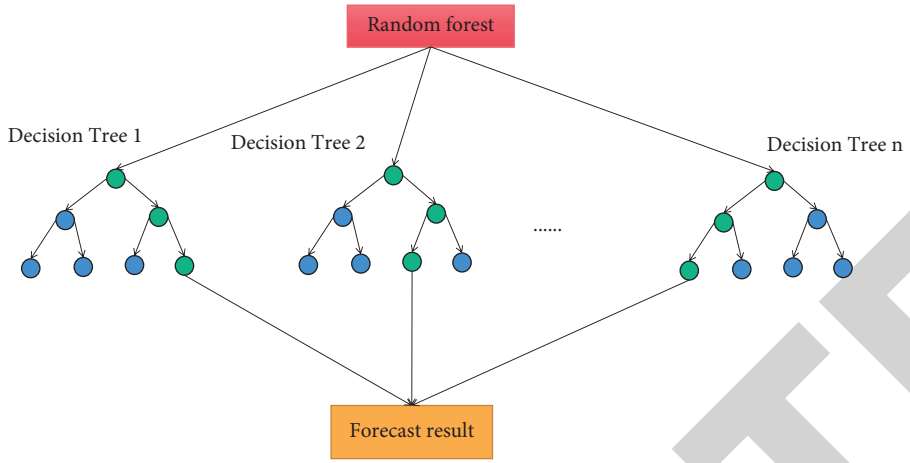

Figure 1: Schematic diagram of the principle structure of decision tree.

Random forest has high operating efficiency, simple implementation, and easy to understand, and users do not need too much mathematical or statistical knowledge to operate. In addition to the advantages mentioned above, random forest has a strong anti-overfitting ability and can be applied at the same time to the advantages of classification and regression problems. However, the application effect of random forest in the regression problem is not as good as the classification problem. Because it cannot give a continuous output, it cannot give predictions beyond the data range of the training sample set when solving the regression problem. Data are prone to overfitting. For low-dimensional data sets, the effect will be reduced when predicting, because the advantage of random forest is to deal with high-dimensional and imbalanced data sets.

3.2. Random Forest Generalization Error Estimation. Generalization ability refers to the ability of the trained model to predict unknown data, that is, the ability of the model to correctly reflect fresh samples. The purpose of the machine learning training model is to obtain the laws behind the data that are not easy to find. For new data other than the learning samples with the same laws (same distribution), the model can also give better classification results [21]. If the generalization error is large, the learning performance of the model will be worse, and vice versa, the performance will be better. The generalization error is defined in the following equation:

$$\mathrm{Re}xp(f) = Ep[L(Y, f(X)p(x, y)\mathrm{d}x\mathrm{d}y]. \tag{3}$$

From the formula, it can be concluded that the generalization error is the loss function expectation of the model. In theory, the generalization error of random forest can be calculated by the above formula, but in practical scenarios, some indicators cannot be calculated directly by calculating the generalization error. Assess the generalization ability of the model, such as the distribution of samples and the expected output are generally unknown.

From the inside, it can be considered from the number of features, the maximum depth of each decision tree, the minimum number of samples required for internal node subdivision, the classification strength of the tree, the correlation between trees, and the minimum impurity of node division. From the perspective of external factors, it is mainly affected by the imbalance of the data of the sample itself, the size of the training sample, the number of features, and the type of features. Generally speaking, the higher the separation accuracy, the better the classification of the algorithm, and its verification index formula is shown in thefollowing equation:

$$J(y_i, y_i) = \frac{(y_i \cap y_i)}{|y_i \cup y_i|}. \tag{4}$$

The overall performance of the algorithm on the test set is obtained through the sample prediction situation. If the actual situation is completely consistent with the predicted result, the coefficient is 1. On the contrary, if it does not match, the coefficient is 0. The data binary classification matrix is shown in Table 1.

Suppose there are two categories in the data set, in which the row represents the required predicted value for prediction, and the column represents the true classification label value, which is called positive and negative classes, respectively. The classification evaluation index is shown in the following formula:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{5}$$

This indicator is used to evaluate the correct ratio of random forests to the overall classification of the test set. The closer the prediction results are to the real situation, the closer to 1. Generally speaking, the higher the separation accuracy, the better the classification of the algorithm.

4. Construction of User Classification Model Based on Digital Marketing

4.1. Classification of User Behavior. User behaviors are generally divided into two types in personalized recommendation systems, namely, explicit feedback behaviors and

Table 1: Confusion matrix for binary data.

C1assification of data sets	Classified positive	C1assified negative
Positive	TP	FN
Negative	FP	TN

implicit feedback behaviors. Explicit feedback behaviors are those behaviors that clearly show how much the user likes or dislikes the item. The most commonly used explicit feedback method is the user's rating or intuitive feedback on a product. The user's rating results are often at the extremes of the highest score and the lowest score, and other options are rarely clicked by users.

Corresponding to explicit feedback is implicit feedback. Implicit feedback also reflects the user's behavior habits, but it refers to those behaviors that cannot clearly reflect the user's preferences. The most representative invisible feedback behavior is the website visit data. When a user is visiting, browsing a page of an item does not mean that the user must like the item displayed on this page. Compared with explicit feedback, although implicit feedback cannot clearly show the user's preferences and interests, the amount of implicit feedback data is very large, and at the same time, it has a great impact on the smoothness of analysis and modeling. And in many platforms, many users even only have implicit feedback data, but no explicit feedback data. For example, it may be because the product link is displayed on the home page, and the user is more likely to click it.himself. The difference between explicit feedback and invisible feedback behavior is that explicit behavior can more intuitively express the user's liking or dislike for a product, thereby generating interactive behavior, while invisible behavior relies on a larger amount of data to count users' in-depth behaviors, habits, and the differences between them are shown in Table 2.

User interests are constantly changing over time, and the changes in user interests mentioned here are active factors generated by the user's own reasons. The most common example is that as users grow older, they like to watch cartoons when they are young and like to watch action movies when they grow up. Another R & D staff gradually transitioned from reading entry books to reading professional books with the increase in working years. Another person has graduated and started working, and his interests after work have changed compared with those of his student days.

The other is the item life cycle. Like movie predictions, a movie's popularity is affected by its life cycle and will soon: out of people's sight. Affected by this, when the recommendation system decides to recommend an item to a user at a certain moment, it needs to consider whether the item is out of date at that moment. For example, the recommendation result for a football fan in the news recommendation includes the news of a certain team ten years ago, which obviously lacks consideration of the time factor. Items in different industries have different life cycles. For example, the life cycle of news is very short, while the life cycle of movies is relatively long and other factors must be considered comprehensively.

4.2. Analysis of Cold Start and Long Tail Effect. The recommender system builds a model based on the user's historical behavior data and then obtains the possible behavior or interest points of the user in the future through the model. Therefore, the historical behavior data become an important part and prerequisite of the recommender system. However, many websites or scholars from research institutions research the performance of recommender systems. For those platforms that hope to have personalized recommendation applications in the operation stage, how to design a personalized recommendation system without a large amount of behavioral data are the content that needs to be researched in the problem of cold and late action. In general cold start, problems fall into the following categories. The first is the user-level cold start. The user's cold start problem is mainly aimed at new users. This problem exists on both new and old platforms. When a user enters the system, the user's behavior data do not exist in the data set, and there is no way to model the user. The second is the cold start of items. The cold start of items is the same as the cold start of users, and it is also for recent items. The last is the cold start of the system. The cold start of the system mainly occurs when the new platform starts to provide personalized recommendation services.

When conducting multi-attribute decision-making analysis, first, the specific problem to be solved should be put forward by the decision-maker, and it should be analyzed whether the problem belongs to this range; second, the possible solutions to the problem should be listed; then, the attribute values of each solution should be listed information and the relationship information between attributes and attributes; then all schemes are sorted and the best scheme is selected; finally, the decision maker selects one or more schemes according to objective conditions, as shown in Figure 2.

Generally speaking, non-personalized recommendations can be provided for cold start. The simplest example of non-personalized recommendations is the popular rankings. The recommendation system can recommend popular rankings to users, and then wait until a certain amount of user data is collected before switching for personalized recommendations. For the cold start of users, the age, gender, and other data provided by the user during registration can be used to personalize the user portrait. In addition, with the authorization of the social network, the user's friend information and related knowledge on the social networking site can be introduced, and then the recommendation can be made according to his friend's hobbies. The other is the user's point of interest feedback, which allows users to select their own points of interest when entering the system, and then make recommendations based on the points of interest. For newly added items, we can make relevant recommendations based on the attribute information of the items. In the face of the system cold start problem, the best method is expert recommendation, and expert knowledge can be used to display the results. The above schemes also have good applications in the later data to meet the personalized recommendation.

Table 2: The difference between explicit feedback and implicit feedback.

	Explicit feedback data	Implicit feedback data
User interest response	Clear	Unclear
Amount of data	Smaller	Huge
Real-time	Real time	With delay
Positive and negative feedback	Both include	Only positive feedback

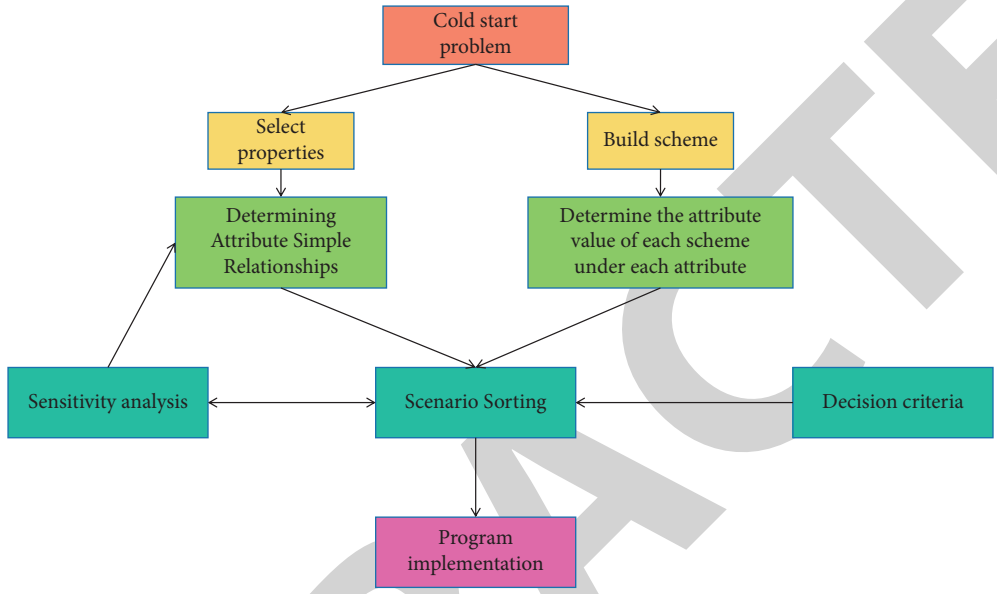

Figure 2: Cold start decision analysis.

4.3. Iterative Fusion of Classified Data and Data Feature Selection. In feature selection, select intuitive features such as the number of clicks, the number of favorites, the number of shopping carts, and the number of purchases. Calculation of interactive behavior between users and brands, such as click-to-purchase ratio, purchase days, and distance variance of purchase days. Based on in-depth mining of user behavior, combined with collaborative filtering and feature selection of solid theory algorithm results, such as brand popularity, user behavior similarity, and other related features. Finally, the selection of features is based on forgetting time. Through data aggregation, this paper found that some users have millions of clicks but they have any purchase behavior. Therefore, this paper thinks that these users should be dealt with separately. Later, after analysis, it is found that these users may use crawler programs to automatically visit the target website. Recommendation results have a big impact. In this paper, noise removal is carried out. In this paper, the user's behavior is regarded as noise data, and the user's behavior data are removed from the data set.

This paper uses OPSD cluster and conducts three experiments. Compared with a single algorithm, the bosting-based algorithm has obtained good results in the experiment, while the random forest has poor robustness at some times, as shown in Figure 3.

In this paper, the fusion algorithm model based on random forest and boosting ideas has been tested many times on the big data platform, and the algorithm has also achieved stable results. The experimental results are shown in Figure 4.

Figure 4 shows the distribution diagram of the results of the 9 tests. It is found that the $F1$ score results are all stable at about 5%. A very important algorithm validity test under big data is the random sampling result. During the tuning process, in the sixth result, because the median TOPN used is smaller, the recall rate is reduced and the accuracy rate is higher, but the overall result remains at 5%.

Overfitting is the process of model parameter fitting. Since the training data contain sampling error, the complex model during training also takes the sampling error into account, and the sampling error is also well fitted, resulting in the model performing well on the training set. But the effect is poor when applied in practice. The generalization ability of the model is weak. Generalization ability refers to the predictive ability of the model learned by the algorithm to unknown data. The number of sample sets is too small, and the main reason for overfitting occurs when there is a large amount of noisy data in the sample training set. There are usually three solutions to overfitting. Because the sample data dimension is too small, it is easy to cause overfitting. Therefore, according to the independent and identical distribution assumption, more data tend to be more accurate in estimating the overall distribution of the sample space. However, in practical applications, due to various reasons, it is not always possible to obtain enough data. The popular

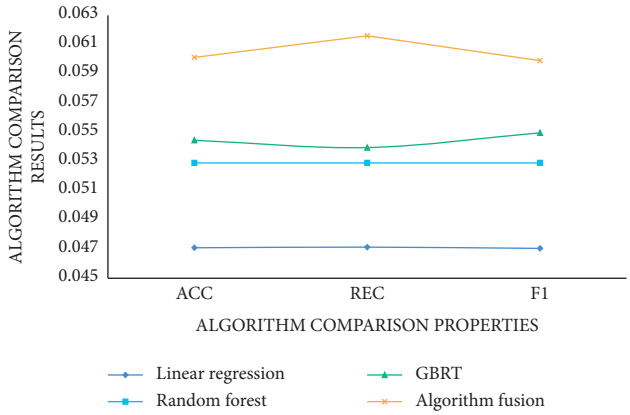

Figure 3: Comparison of algorithm effects.

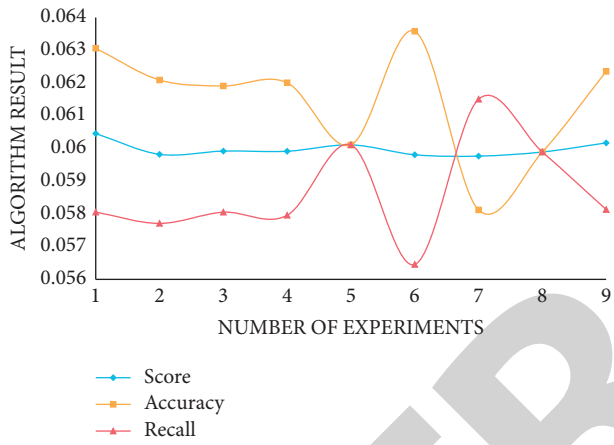

Figure 4: Results of multiple algorithms.

understanding of data expansion is to add data with the same characteristics or similar characteristics according to the characteristics of the existing sample training set on the basis of the existing sample training set. You can collect more data from the data source or copy the original sample data set and randomly add a small amount of noise data, etc.

5. Recommended Model Experimental Analysis

5.1. Classification Indicators Based on Digital Marketing Commodities. For random forest models, the decision trees in each ensemble model are sampled from the training set using Bootstrap's method. When a node is split, it does not need to consider all the features, but extracts a part of all the features as a subset, and finds the best splitting feature and the best splitting point from the feature subset. Due to the randomness of random forests, the bias of random forests tends to increase slightly than that of a single decision tree, and due to the combined averaging of multiple trees, the variance of random forests tends to decrease and is sufficient to compensate for the decrease in bias, so we can get a better model.

The classic random forest method uses the combined voting method of decision trees to predict the results. Here, multiple decision trees are used to predict the average of the probability of the results. As the final result, the default probability of 50% will be used as the two categories demarcation line. Usually, the implementation of random forest is based on decision tree, using chaotic combination technology. When the model is built, a series of classifiers are composed of randomly extracted data and features, and the results of multiple classifiers are averaged as the final result output. The input to the random forest classifier consists of two arrays, which together form the training set. In addition, for a random forest classifier, it also saves some parameters required for the establishment of the entire model, including the number of trees used, the feature dimension randomly extracted from each node, and the child nodes after the decision tree are split.

In the first stage, predict which sessions have bribe behaviors, use the extracted session-related features, use the default parameters of the random forest model, and set the number of trees to 5. Since in the problem of the recommender system, the purchase proportion of behaviors is generally very small, so it is an imbalance problem. Here, other indicators need to be used to judge the quality of the model. Parameter optimization is carried out for the model that predicts whether there is a purchase behavior in a session, and the features used include session features and date features, and the number of features is 10. First, adjust the number of the most important parameter trees, as shown in Figure 5.

It can be found from Figure 5 that for this problem, the best effect is obtained when the number of trees is controlled at about 7–10 of the number of features. There are many different implementation algorithms for decision trees in random forests, and CART uses the Gini index to complete the generation of decision trees. In random forests, you can also compare and choose a better decision tree to use to complete model training. The results are shown in Figure 6.

It can be seen from Figure 6 that the effect of using the Gini index is the best. Through the research and improvement of the random forest class imbalance problem, the accuracy of the model can be improved to a certain extent. By combining the balanced random forest and the weighted random forest, the loss and redundancy of data information caused by resampling can be solved, the data of all rare classes can be selected, and the accuracy of the model can be guaranteed by weighting. Furthermore, the difference between decision trees is guaranteed.

5.2. Recommendation Improvement Experiment Based on Feature Index. Random forests also come in different forms, including normal random forests and random forests that include additional trees. In the latter, the tree does one more random step when it splits. When selecting split points, a series of random feature subsets are used. However, instead of finding the most discriminative threshold, a threshold is randomly selected for each candidate feature, and then selected from these randomly generated thresholds. The optimal threshold is used as the split judgment condition.

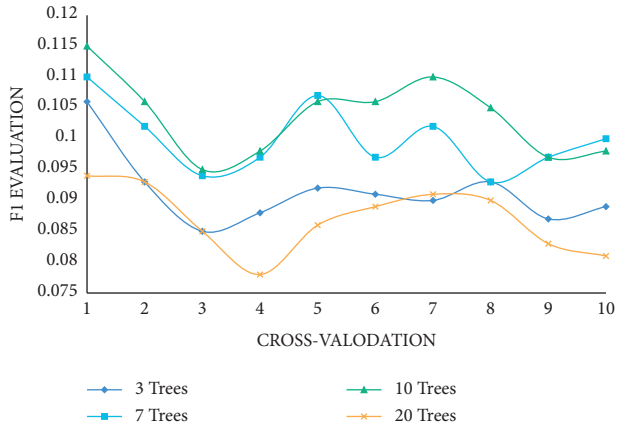

Figure 5: Effect analysis of random forests with different numbers of trees.

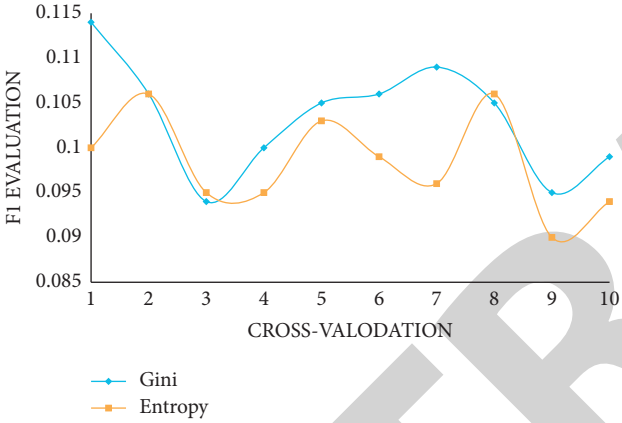

Figure 6: The effect of random forests with different norms.

Doing so usually reduces the variance of the model while slightly increasing the bias of the model.

In a decision tree, when finding the best split point, it is necessary to use the relevant indicators of the calculated features to select appropriate features, and the quantification of the selected features will affect the final result. Commonly used selection methods include the square root of the total number of features, the logarithm value, and the number of samples required for splitting all feature extraction nodes, which determine the degree of splitting of a tree and also affect the height of the tree. The selection of different nodes is experimentally analyzed. The experiments are shown in Figure 7.

It can be seen from the results in Figure 7 that the minimum number of samples that needs to be included when a new leaf node is created affects the purity of each node of the tree. If the node obtained after splitting contains samples smaller than this value, it will be discarded. If the value is too large, it will result in a lack of discrimination between nodes. The best results are obtained when 2 samples are taken as the minimum number of samples required for node splitting, which ensures that the decision tree is completely split.

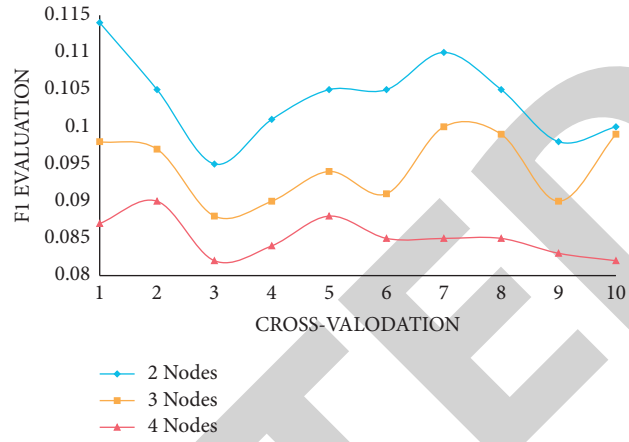

Figure 7: The effect of the minimum number of samples required for different number of node splits.

6. Conclusions

With the development of modern e-commerce websites, recommender systems have become the core business of many companies in the modern Internet industry, and combinatorial models in machine random forest algorithms are often used in recommender systems. As an ensemble learning method, the random forest algorithm votes through the combined prediction results of multiple decision trees to improve the prediction effect. The random forest algorithm has many advantages over other linear classifiers, not only the accuracy of the result prediction is improved, but the generalization error is also smaller, and the processing of high-dimensional data is also efficient, the training process is fast, and can be parallelized accomplish. Therefore, it is necessary to study the optimization and application of random forests in recommender systems, in which feature selection and imbalanced classification are both problems that data mining often encounters in real big data. Through experimental analysis, this paper finds the effect of the digital marketing recommendation model on the minimum number of samples required for node splitting with different numbers of nodes. The best results are obtained when 2 samples are taken as the minimum number of samples required for node splitting, and the average value of the $F1$ evaluation is 0.1038. For the effect of random forests with different specifications, the effect of using Gini index is the best, and the average value of its $F1$ evaluation is 0.1033; for the effect analysis of random forests with different numbers of trees, 7 to 10 decision trees are the best, and its average value is 0.1033. The mean value of the $F1$ assessment was 0.10175.

At present, for the application of random forest in the recommendation system, more in-depth and broader research can be carried out. In terms of feature selection, this paper only studies the feature selection scheme of random selection and linear combination. The removal of information redundancy and information overlap between trees is also a direction that can be researched; for the balanced classification problem, this paper combines the algorithms of

balanced random forest and weighted random forest, and for the distribution difference of data feature space level data can also be research in depth. In addition, there are many research directions related to the optimization and application of random forests, such as optimization of decision tree node splitting algorithm and optimization of parallelization implementation.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] J. Xu and S. Mu, "Research on the construction of crossborder e-commerce logistics service system based on machine learning algorithms," *Discrete Dynamics in Nature and Society*, vol. 2022, Article ID 3943869, 12 pages, 2022.

[2] X. Zhou, M. Su, G. Feng, and X. Zhou, "Intelligent tourism recommendation algorithm based on text mining and MP nerve cell model of multivariate transportation modes," *IEEE Access*, vol. 9, no. 99, p. 1, 2020.

[3] H. Luo and Z. Li, "Research on construction of recommendation system on account of CNN and PMF model," *Journal of Physics: Conference Series*, vol. 1992, no. 3, 5 pages, Article ID 032078, 2021.

[4] C. Pang, "Construction and analysis of macroeconomic forecasting model based on biclustering algorithm," *Journal of Mathematics*, vol. 2022, Article ID 7768949, 10 pages, 2022.

[5] X. Zheng, S. Yi, and X. De ng, "Evaluation model construction of automobile appearance design based on random forest algorithm," *Journal of Physics: Conference Series*, vol. 1941, no. 1, 9 pages, Article ID 012072, 2021.

[6] C. Iwendi, E. Ibeke, H. Eggoni, S. Velagala, and G. Srivastava, "Pointer-based item-to-item collaborative filtering recommendation system using a machine learning model," *International Journal of Information Technology and Decision Making*, vol. 21, no. 01, pp. 463–484, 2022.

[7] R. C. Rumín, "Digital marketing attribution: understanding the user path," *Journal of Electronics*, vol. 9, no. 11, 2020.

[8] S. Liu, B. Wang, M. Xu, and L. T. Yang, "Evolving graph construction for successive recommendation in event-based social networks," *Future Generation Computer Systems*, vol. 7, no. 2, pp. 48–55, 2019.

[9] M. W. Hasanat, A. Hoque, M. Hassan, B. l. Mou, and A. B. A. Hamid, "The lack of digital marketing skills: developing a digital marketer model for the retail industries," *Xi'An Jianzhu Keji Daxue Xuebao/Journal of Xi'an University of Architecture & Technology*, vol. 12, no. 3, pp. 2673–2680, 2020.

[10] S. Ohtake, T. Kodaka, and Y. Seki, "Steiner tree based recommendation system for combination of APIs and IoT devices," *ASIA PAC SOFWR ENG*, vol. 4, no. 1, pp. 66–72, 2017.

[11] J. J. Prabhu, "Digital marketing techniques, innovation & recommendation for SMEs business," *Journal of Social Sciences*, vol. 23, no. 2, pp. 404–414, 2020.

[12] S. Mathai and S. Jeswani, "Effectiveness of print media marketing in digital age: a study on Indian telecommunication industry," *FIIB Business Review*, vol. 10, no. 3, pp. 242–254, 2021.

[13] L. Xiang, Z. Wang, L. Wang, Q. Zhu, and R. Hu, "A multidimensional context-aware recommendation approach based on improved random forest algorithm," *IEEE Access*, vol. 6, pp. 45071–45085, 2018.

[14] Z. Feng, M. Hua, P. Lei, and L. Zhang, "Recommendation algorithm of cloud computing system based on random walk algorithm and collaborative filtering model," *International Technology Management*, vol. 3, no. 3, 2017.

[15] A. D. Rewade, S. W. Mohod, and S. P. Bargat, "Content based alternate medicine recommendation by using random forest algorithm," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 4, pp. 1163–1168, 2019.

[16] B. A. Hammou, A. A. Lahcen, and S. Mouline, "An effective distributed predictive model with Matrix factorization and random forest for Big Data recommendation systems," *Expert Systems with Applications*, vol. 137, pp. 253–265, 2019.

[17] S. Belkacem, K. Boukhalfa, and O. Boussaid, "Expertise-aware news feed updates recommendation: a random forest approach," *Cluster Computing*, vol. 23, no. 3, pp. 2375–2388, 2020.

[18] R. Abraham, M. E. Samad, A. M. Bakhach et al., "Forecasting a stock trend using genetic algorithm and random forest," *JRFM*, vol. 15, no. 188, 2022.

[19] C. Li, Y. Li, C. Wang et al., "The Multimedia Recommendation Algorithm Based on Probability Graphical model," *Multimedia Tools and Applications*, vol. 81, pp. 1–16, 2020.

[20] H. Zhao, Z. Li, and F. Guo, "Precision marketing based on K2 algorithm," *Computer Applications and Software*, vol. 4, no. 3, pp. 26–33, 2019.

[21] B. Lin and J. Xiao, "Multiple criteria recommendation algorithm based on matrix factorization and random forest," *Journal of South China Normal University (Social Science Edition)*, vol. 8, no. 13, pp. 88–93, 2019.