Accounting and Corporate Management (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acccm.2024.060505
ISSN 2523-5788 Vol. 6 Num. 5

Research on Customer Traffic Value Recognition Model Based on Improved Random Forest Algorithm

Wei Wu

AWS EKS Team, Amazon, Seattle, Washington, 98121, United States
weiwupaper@outlook.com

Keywords: Customer traffic, value recognition, random forest, machine learning, algorithm optimization

Abstract: In today's data-driven business environment, user behavior analysis and customer traffic value evaluation have become an important basis for enterprises to formulate marketing strategies. The customer traffic data of friends contains potential market insight. How to effectively identify and evaluate the traffic value of friends customers has become the key for enterprises to gain advantages in the fierce market competition. This paper presents a customer traffic value recognition model based on improved random forest algorithm. Through the optimization of the algorithm, we not only improve the prediction accuracy of the model, but also achieve remarkable results in analyzing the value characteristics of customer traffic. The experimental results show that the improved random forest model is superior to the traditional method in accuracy, recall rate, F1 score and AUC, and has higher recognition accuracy and application potential. The model proposed in this paper provides a reliable technical support for enterprises in the fierce market competition to mine the value of business traffic, which is helpful to improve the scientific and accurate degree of marketing decision.

1. Introduction

Driven by the popularity of the Internet and the rapid development of digital technology, the richness and complexity of user behavior data have reached an unprecedented level. Enterprises pay more and more attention to the value of user data in data-driven marketing decisions, and user behavior analysis and customer traffic value evaluation have become important bases for enterprises to formulate marketing strategies. Customer traffic value evaluation is the core content of enterprise marketing and strategy optimization[1-2], and the traffic data of friends customers can reflect the potential demand and trend in the competitive market. The value identification of customer traffic can not only help enterprises to allocate resources more accurately, but also provide support for formulating personalized marketing programs and optimizing customer management strategies. How to effectively identify and evaluate the traffic value of friends and customers becomes the key for enterprises to gain advantages in the fierce market competition. The unstructured characteristics of the traffic data, the uneven distribution of data and the redundant information brought by the multi-dimensional characteristics increase the difficulty of data analysis, and bring certain challenges to the accuracy and efficiency of the model. In this paper, a value recognition model

based on improved random forest algorithm is proposed to improve the recognition accuracy while maintaining the generalization ability of the algorithm. In this paper, Bagging and Boosting techniques are combined to further optimize the integration strategy and improve the applicability and prediction accuracy of the model when dealing with high-dimensional and unbalanced data sets. The experimental part is based on the real customer traffic data set of an e-commerce platform to verify the effectiveness of the proposed model. The experiment also reveals that the model can capture high-potential customers more accurately in the identification of customer traffic value, and has strong commercial application value. Future research will focus on the application of multi-source data fusion and deep learning technology in traffic value recognition to further improve the applicability and scalability of the model.

2. Related Work

In the field of customer traffic value analysis and user behavior prediction, a large number of studies have explored how to tap the potential value of users through data mining and machine learning algorithms from different perspectives[3-4]. Among them, the traditional analysis methods are mostly based on statistical models or simple clustering methods, and mostly focus on the user behavior analysis of a single platform. However, with the diversification of data acquisition channels and the continuous improvement of model algorithms, the comprehensive analysis and model optimization of multi-source heterogeneous data have become a research hotspot[5-6].

2.1 User Behavior Analysis

User behavior analysis research attempts to extract user preferences and potential demands by modeling user behavior characteristics such as browsing, clicking and length of stay on multiple channels such as e-commerce and social platforms. The collaborative filtering recommendation model proposed by Liu et al. (2018) is a typical method, which constructs user interest characteristics by analyzing user access behaviors on different platforms and makes recommendations to potential customers. This method has a high utilization rate of existing user data, but it is not universal when dealing with cross-platform and cross-domain friend data, so it is only applicable to scenarios with relatively simple data sources and structures.

2.2 Customer Traffic Valuation

In order to evaluate the value of customer traffic, researchers divide users into different subgroups and evaluate the value based on the behavioral characteristics of different groups. Wang et al. (2020) used K-means clustering method to classify users to achieve subdivision management of customer traffic. K-means clustering is suitable for processing a large number of user data, and has the advantages of high computational efficiency and simple implementation. However, the distribution form of data has high requirements, which is easy to be affected by noisy data. Therefore, the effect is not as good as expected in complex multidimensional data. In order to deal with these problems, in recent years, researches have begun to explore algorithms based on ensemble learning, such as random forest, gradient lifting decision tree (GBDT), etc., in an attempt to achieve effective processing of multidimensional data while improving computational efficiency.

2.3 Application of Ensemble Learning in Customer Traffic Identification

Since the random forest algorithm was proposed by Breiman (2001), it has been applied to classification and regression tasks because it can improve the prediction accuracy of the model

through the combination of multiple decision trees. In particular, the performance of random forest is more stable than that of traditional single model when processing tasks with high data dimension and complex feature association. The multi-tree structure based on ensemble learning can effectively reduce the risk of overfitting and improve the tolerance of the model to noisy data. However, the random forest algorithm still has the problems of low computational efficiency and overfitting when dealing with high dimensional complex data. To solve these problems, many scholars have proposed a series of improvement methods, such as introducing feature selection mechanism, optimizing splitting criteria, etc., to improve the performance of the model in traffic recognition.

With the successful application of deep learning in image recognition, natural language processing and other fields, the application of neural network-based models in customer value recognition has attracted attention. Models such as LSTM (Long Short-term Memory Network) and GRU (Gated cycle Unit) are widely used for time series data analysis and are suitable for user behavior data containing time series. These deep learning models have higher requirements for data volume and computing resources, and are not as explanatory as traditional models such as random forests, so their application in customer traffic value assessment still needs further research and verification[7-8].

Although significant progress has been made in user behavior analysis and customer traffic value recognition, there is still room for improvement in cross-platform multi-source data fusion, heterogeneous data feature extraction, and computational efficiency of algorithms. Based on the improved random forest algorithm, this paper will improve the recognition accuracy of the model through feature selection and split criteria optimization, so as to achieve accurate recognition of the value of customer traffic of friends, and provide decision support for marketing strategy and customer resource management.

3. Model Design Based on Improved Random Forest Algorithm

3.1 Model Frame

The traffic value recognition model framework proposed in this paper based on improved random forest algorithm mainly includes four modules: data preprocessing, feature engineering, model construction and result evaluation. Each module realizes accurate prediction of traffic value through gradual optimization. In the whole framework design, data preprocessing and feature engineering laid the data foundation for model construction. The improved random forest algorithm improved the prediction performance and computational efficiency of the model by optimizing the feature selection and splitting criteria, and finally verified the model performance through a variety of evaluation indicators. The first three modules are closely related to the following steps, and the last module results evaluation refers to the multidimensional evaluation of model performance by using accuracy, recall rate, F1 score and AUC and other indicators in the result evaluation stage of the model. Accuracy is used to measure the accuracy of the overall prediction of the model, while recall rate and F1 score pay more attention to the actual application effect of the model in traffic value recognition, especially when dealing with unbalanced data. AUC can better reflect the prediction performance of the model under different thresholds. Multiple tests on the improved model show that it is superior to the traditional random forest model in various evaluation indexes, especially in the recall rate and AUC value, which further verifies the effectiveness and practical value of the improved model.

3.2 Data Preprocessing

In the customer traffic value recognition model based on improved random forest algorithm, data preprocessing is an important part of model design. Due to the complexity and variety of data sources, data preprocessing is not only the basis to ensure the performance of the model, but also directly affects the prediction accuracy and generalization ability of the model. The data preprocessing process in this paper includes three parts: data cleaning, feature coding and sample balance processing.

One is data cleaning. First, the missing values and outliers in the data are processed. Missing values can be caused by users not fully accessing a particular page or by loss during data collection, and common approaches include mean filling, interpolation, or deleting incomplete samples. The detection methods of outliers include standard deviation based method or boxplot analysis, and delete or correct these outliers to reduce their interference to the model.

The second is feature coding. Friend customer traffic data contains a large number of non-numerical characteristics (. In order for the data to be used effectively by the random forest model, it is necessary to encode the non-numerical features. In this paper, One-Hot Encoding is used to deal with categorical variables. By converting categorical features into binary features, the model can distinguish different categories of traffic data more easily.

Finally, sample balance treatment. In the task of customer traffic identification, the number of samples of different traffic categories is uneven, especially in the identification of high-value and low-value traffic. In order to avoid the bias of the model to the majority class, SMOTE (Synthetic Minority Over-sampling Technique) algorithm is introduced in this paper to generate synthetic data of minority class samples. SMOTE solved the problem of sample imbalance effectively by interpolating between a few class samples and increasing the number of minority class samples, and improved the prediction ability of the model on a few class samples.

3.3 Feature Engineering

Feature engineering is also a key step in the customer traffic value recognition model, which improves the predictive performance and generalization ability of the model by constructing, selecting and optimizing features. Feature engineering mainly includes three parts: feature construction, feature selection and feature dimension reduction.

Start with feature building. In order to describe the value characteristics of the customer traffic of friends, the behavior characteristics, interaction characteristics and business characteristics are designed on the basis of data preprocessing. Behavioral characteristics include user visit frequency, page click rate, average stay time, etc., which can reflect the user's attention and interest in friends brand; Interactive features describe the changes of users' behavior in different time periods, such as the fluctuation of peak access time and access frequency; Business characteristics include the conversion rate of users, estimated consumption potential, etc., which is used to measure the potential value of customers. Through these three types of characteristics, the model can more accurately capture the user's value trends and behavioral intentions.

Then there is feature selection. In order to avoid the computational complexity and overfitting risk caused by high-dimensional features, Recursive Feature Elimination (RFE) method was used to select features. RFE gradually removes low correlation features based on the importance of the features, and retains variables that have a significant impact on traffic value. Combined with the built-in feature importance evaluation mechanism in the random forest algorithm, a set of key features highly related to customer traffic value are retained, which improves the training efficiency and prediction accuracy of the model.

Finally, feature dimension reduction. Since customer traffic data may contain redundant or

highly correlated features, this paper uses principal component analysis and other methods to reduce the dimensionality of some features. PCA can reduce the number of features while retaining the main information, reduce the complexity and dimension of the data, and improve the generalization ability of the model.

3.4 Improved Random Forest Model Construction

The random forest algorithm is the core of the customer traffic value recognition model proposed in this paper. Although the traditional random forest algorithm has strong ensemble learning ability and high robustness, it will encounter problems such as low computational efficiency and overfitting when dealing with high-dimensional complex data. In order to improve the predictive performance and adaptability of the model, the random forest algorithm is optimized and improved, including feature selection optimization, splitting criteria improvement and integration learning strategy enhancement.

Feature selection optimization refers to the Recursive Feature Elimination (RFE) method integrated in the random forest algorithm to reduce the computational complexity of the model for feature selection. RFE recursively removes unimportant features step by step, selects key features by combining feature importance scores within the random forest, and retains variables that are highly correlated with traffic value. Different from traditional feature selection methods, RFE can dynamically optimize feature combinations during model construction to better adapt to multi-dimensional features of complex data sets. The data set after feature screening not only reduces the burden of the model, but also reduces the risk of overfitting.

Improvement of split criteria: In the decision tree construction of random forest algorithm, node split criteria directly affect the accuracy and computational efficiency of the model. Traditional splitting criteria may have limited effectiveness when dealing with highly nonlinear and heterogeneous data. In this paper, Information Gain Ratio (IGR) is introduced into the splitting criteria, which is more sensitive in measuring the correlation between features and target variables, and is especially suitable for data with uneven sample distribution. Regularization terms are added to the model to punish the overly complex tree structure in the splitting process and avoid overfitting problems. This improvement makes the construction of decision tree more optimized, and also improves the ability of the model to deal with multi-source heterogeneous features.

Ensemble learning strategy enhancement: Traditional random forest realizes the combination of multiple weak classifiers through Bagging (self-sampling) technology, but classification bias may occur on unbalanced data sets. In order to improve the generalization ability and adaptability of the model, Boosting mechanism is integrated on the basis of Bagging strategy to enhance the sensitivity of the model to a few class samples. In terms of concrete implementation, the model first generates several decision tree basic classifiers (using Bagging sampling), then reweights a few class samples, and gradually increases the weight of hard-to-classify samples, so that subsequent decision trees pay more attention to these hard-to-classify samples. With this combination of Bagging and Boosting, the model not only maintains the efficient computing power of random forest, but also achieves better accuracy in customer traffic value recognition tasks.

Perform hyperparameter optimization. In order to further improve the performance of the model, the hyperparameters are optimized using Grid Search and Cross-Validation techniques. The key parameters of the random forest algorithm, such as the number of decision trees, the maximum depth and the minimum number of split samples, directly affect the complexity and generalization ability of the model. Through grid search, multiple parameter combinations are traversed, and the optimal parameter configuration is selected with cross-validation to balance the accuracy of the model and the computational cost.

4. Conclusion and Prospect

In this paper, a recognition model based on improved random forest algorithm is proposed by deeply studying the problem of customer traffic value recognition. With the wide application of multi-source heterogeneous data, customer traffic value analysis is becoming more and more important in enterprise marketing decision-making, but the traditional algorithm has the bottleneck of accuracy and computing efficiency when dealing with complex features and cross-platform data. By combining feature selection mechanism and split criteria optimization, we have effectively improved the random forest algorithm, which not only improves the prediction accuracy of the model, but also significantly improves the applicability and stability of the model under high-dimensional data and diversified features. Experimental results show that the improved model is superior to the traditional model in accuracy, recall rate, F1 score, AUC and other evaluation indicators, especially in the multidimensional complex data, which verifies the superiority of the proposed method in the task of customer traffic value recognition.

From the perspective of practical application, the improved model proposed in this paper provides a new technical means for enterprises to identify and explore the potential value of the traffic of friends and merchants, and has strong application and promotion value. It can help enterprises more accurately identify high-value traffic customer groups, optimize marketing resource allocation, and effectively improve customer conversion rates and overall marketing effects. The optimized random forest model based on feature selection and split criteria can achieve a good balance between model complexity and accuracy, which lays a foundation for the subsequent application in larger scale and multi-source data environment.

There are also some limitations in this study, such as the adaptability of the model in the face of real-time data flows, and the potential application of deep learning models in traffic value recognition has not been deeply discussed. In the future, further research can be carried out in two aspects: On the one hand, combining the capability of feature extraction and time series prediction in deep learning to explore a hybrid model more suitable for real-time data analysis; On the other hand, the extended model is applied to more different types of traffic data sources to further enhance the universality and promotion value of the model, and provide more comprehensive support for customer value management and precision marketing.

References

[1] Bensheng Yun, Yulu Sun, Kebin Fang, et al. Large-scale network Burst traffic identification model, method and model training method :CN202011516886.7[P].CN112633475A[2024-11-10].

[2] Jing Jing, Yongzh Zhui. Spark platform weighted hierarchical subspace random forest algorithm study [J]. Journal of software Tribune, 2020, 19 (3): 5. DOI: 10.11907 / RJDK. 191691.

[3] Lili Zhu. Design of following table-level sorting learning recommendation system for random forest algorithm [J]. Journal of Huaiyin Institute of Technology, 2023, 32(5):62-68.

[4] Xiaobin Lu, Yangyi Zhang, Guancan Yang, et al. Research on disequilibrium classification in emerging technology recognition: a cost-sensitive random forest algorithm [J]. Journal of Information Technology, 2022, 41(10):1059-1070.

[5] Balaram A, Vasundra S.Prediction of software fault-prone classes usingensemble random forest withadaptive synthetic sampling algorithm[J]. Automated Software Engineering, 2022, 29(1):1-21.DOI:10.1007/s 10515-021-00311-z.

[6] Onwuka U C. Hybrid Machine Learning Algorithm for Arrhythmia Classification Using Stacking Ensemble[J]. Random Forest and J. 48 Algorithm. 2021, 6(11): 536-543.

[7] Zhang X, Huang W, Lin X,et al. Complex image recognition algorithm based on immune random forest model (Retraction of Vol 24, Pg 12643, 2020)[J]. Soft computing: A fusion of foundations, methodologies and applications, 2023.

[8] Wang Z, Zhao G, Zhao Y,et al. New Word Discovery Algorithm for Power Marketing Audit Combining Adjacency Entropy and Random Forest[J]. 2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT), 2021:214-217. DOI: 10.1109 / ISCIPT53667.2021.00050.