Research on User Behavior Analysis and Personalized Recommendation Based on CatBoost Algorithm

Shiyang Li

Beijing Sport University, Beijing, Haidian, 100080, China

2020010037@bsu.edu.cn

Keywords: User behavior analysis; Personalized recommendation; CatBoost algorithm; Classification feature processing; Ordered boosting

Abstract: Considering the problems of data sparsity, cold start, and classification feature processing in user behavior analysis and personalized recommendation in Internet applications, this paper proposes a research method for user behavior analysis and personalized recommendation based on the CatBoost algorithm. First, based on the user's historical behavior data, a user profile and interest model are established to obtain information such as the user's primary attributes, behavioral characteristics, interest preferences, and consumption capabilities. Second, in order to eliminate the prediction bias in the gradient-boosting decision tree, an ordered boosting strategy is proposed. Thirdly, based on the classification feature processing capability of the CatBoost algorithm, a personalized recommendation model that comprehensively considers the influence of user features and item features is constructed to describe the matching degree between users and items accurately. Finally, a user behavior analysis and personalized recommendation system based on the CatBoost algorithm is built. In addition, the effectiveness of the method proposed in this paper is proved through simulation and experimental verification.

1. Introduction

User behavior analysis and personalized recommendations are widely used in e-commerce, social networking, entertainment, and other fields due to their excellent user experience, high conversion rate, and fast feedback response [1]. In practical Internet applications, users often hide partial information. If the user profile is not accurate enough, the recommendation will fail. In addition, the user's interest quality affects the feedback response to the recommendation, which is also influenced by a cold start. Therefore, user behavior analysis and personalized recommendations have essential research value for improving user satisfaction and business efficiency.

The traditional personalized recommendation algorithm mainly makes recommendations based on the collaborative filtering model, divided into user- and item-based methods. Collaborative filtering algorithms require a large amount of user behavior data. Although the collaborative filtering model is the most commonly used in the personalized recommendation, its description of classification features is not accurate enough, leading to problems such as data sparsity and overfitting in traditional recommendation algorithms [2]. Therefore, it is necessary to study the classification features of user behavior to establish a more accurate recommendation model.

To solve the problems of the inaccurate user profile, matching deviations between users and items in recommendation results, and low recommendation accuracy in current recommendations, this paper proposes a research method for user behavior analysis and personalized recommendation based on the CatBoost algorithm. First, the user's historical behavior data is preprocessed to be the input of the CatBoost algorithm to obtain accurate user profile information. Secondly, the parameters of the CatBoost algorithm are optimized to obtain the optimal model. And then, an ordered boosting strategy is proposed to use the optimal model to complete the ranking of personalized recommendations. Next, careful consideration of user and item features' impact on recommendations is integrated with experimental phenomena for more accurate modeling. In addition, the CatBoost algorithm is used to build a personalized recommendation model. Finally, the

  DOI: 10.25236/icceme.2023.012

method proposed in this paper is verified as effective by simulation and experiment.

2. Research Background of User Behavior Analysis and Personalized Recommendation Based on the CatBoost Algorithm

2.1 Significance and Challenges of User Behavior Analysis and Personalized Recommendation

The concept of user behavior analysis and the personalized recommendation was popularized with the development of the Internet. It is "soaked" with the concept of big data, showing the user's personalized orientation and reflecting the intelligent business strategy since the information age [3]. However, it is hard to have consistency when certain statistical standards are used to construct the definition and essence of user behavior analysis and personalized recommendation.

2.2 Basic Principles and Advantages of the CatBoost Algorithm

This paper proposes a user behavior analysis and personalized recommendation method based on the CatBoost algorithm. This method uses the classification feature processing capability of the CatBoost algorithm to transform the recommendation issue into a ranking issue and to achieve accurate matching of users and items by training the gradient boosting tree model. To verify the effectiveness of this method, this paper uses MovieLens as a simulation platform to generate user behavior data of different types, locations, and interests as the input of the gradient boosting tree model. The simulation results show that the method can realize personalized recommendations in the data sparsity and cold start environment with good accuracy and generalization ability. In the process, the CatBoost algorithm is a gradient-boosting algorithm used to optimize the loss function. The function of the CatBoost algorithm is as follows:

$$L(\theta)=\sum_{i=1}^{n} l\left(y_i, f\left(x_i;\theta\right)\right)+\Omega(\theta) \tag{1}$$

In the function, $L(\theta)$ is the loss function, $l\left(y_i, f\left(x_i;\theta\right)\right)$ is the loss item of the sample i, y_i is the true label of the sample i, $f\left(x_i;\theta\right)$ is the predicted value of the sample i, θ is the model parameter, and $\Omega(\theta)$ is the regularization item. The logic of the CatBoost algorithm is shown in Figure 1.

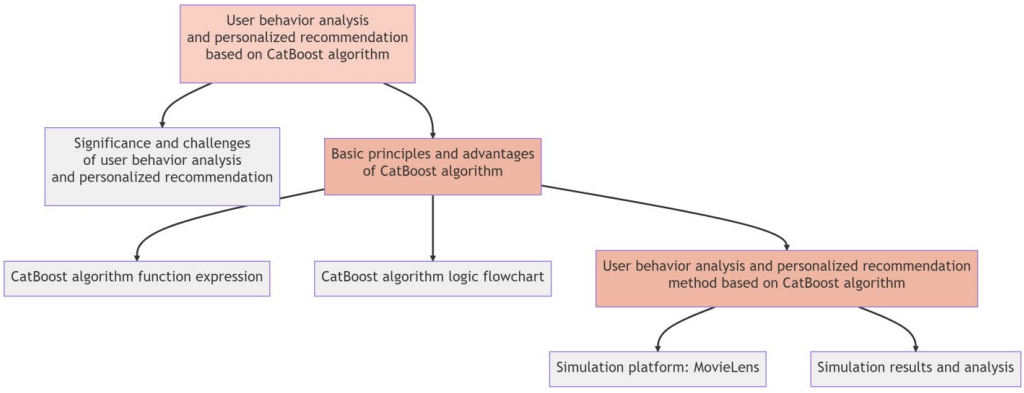

Figure 1 Logic of CatBoost algorithm

3. Research Basis and Key Technologies of User Behavior Analysis and Personalized Recommendation Based on CatBoost Algorithm

3.1 Data Preprocessing and Feature Engineering

Data quality is an essential criterion for data analysis, and it is a manifestation of data completeness, accuracy, consistency, availability, and timeliness [4]. Data quality and cleaning discuss different definitions of data from the perspectives of data source, structure, and content.

Since data quality is more objective to some extent and belongs to applied science aimed at improving the value of data, some scholars believe that data quality is the degree to which data meets the needs of users or the applicability of data. Notably, the research history of data quality can even be traced back to the development of statistics, which mainly includes data collection, verification, repair, and evaluation. In the meantime, the concepts and methods of data cleaning are closely related to the development of databases. Through data cleansing, data quality becomes an important indicator of database management. The main contribution of data cleaning theory in the relational database era is the data quality measurement based on the standard attributes of relational algebra.

3.2 Parameter Tuning and Model Evaluation of the CatBoost Algorithm

In the research on user behavior analysis and personalized recommendation, the static features of users and items were mainly considered. In contrast, the dynamic interaction between users and items was ignored, leading to inaccurate, imprecise, and poor optimization recommendation results [5]. To solve the above problems, parameter tuning and model evaluation were carried out for the CatBoost algorithm by mainly combining grid search technology and cross-validation model to deeply optimize for the traditional gradient boosting tree framework structure to achieve a better recommendation effect of the CatBoost algorithm. Specifically, firstly, the main parameters of the CatBoost algorithm were systematically searched and compared using grid search technology to find the optimal parameter combination. Secondly, the training results of the CatBoost algorithm were evaluated and analyzed using a cross-validation model to verify the generalization ability and stability of the CatBoost algorithm. Then, the CatBoost algorithm was improved by using the ordered boosting strategy, which solved the problem of prediction bias in the gradient boosting tree algorithm. Finally, the CatBoost algorithm and other recommendation algorithms were compared to prove the superiority of the CatBoost algorithm. In the process, the grid search technology is a parameter optimization technique used to find the optimal parameter combination. The function of the grid search technique is as follows:

$$\theta^* = \underset{\theta \in \theta}{\operatorname{argmin}} L(\theta) \tag{2}$$

In the function, θ^* is the optimal parameter combination, θ is the parameter space, and $L(\theta)$ is the loss function.

The parameter tuning process of the CatBoost algorithm is shown in Figure 2.

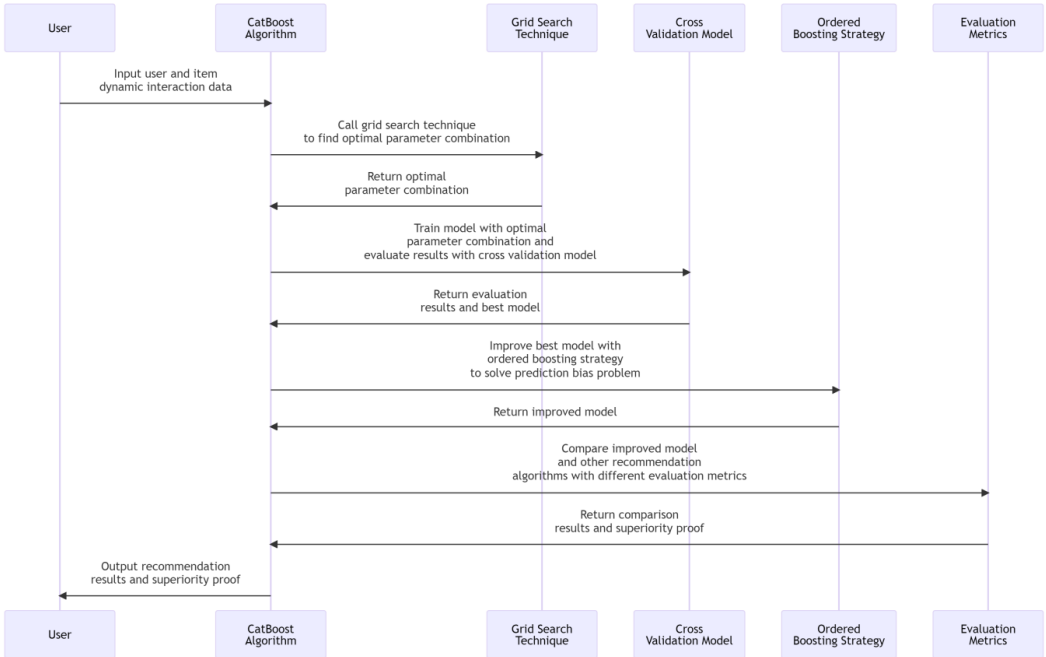

Figure 2 The parameter tuning process of the CatBoost algorithm

3.3 Improvement and Extension of CatBoost Algorithm

This section proposes an improved and extended version of the CatBoost algorithm based on deep neural networks. This method uses the deep neural network to encode the features of users and items. And it extracts the hidden vectors related to the matching degree of users and items as the input of the CatBoost algorithm to train the gradient boosting tree to identify the preferences and ratings of users and items. Therefore, this method can effectively mine the nonlinear and complex patterns existing in user behavior with high accuracy and generalization. In the process, the deep neural network is a multi-layer nonlinear model, which is used to learn the high-level abstract features of the data. The function of the deep neural network is as follows:

$$h(x) = f_n\left(f_{n-1}\left(\ldots f_1\left(x;W_1\right);W_{n-1}\right);W_n\right) \tag{3}$$

In the function, $h(x)$ is the output of the deep neural network, x is the input, f_i is the activation function of the i th layer, and W_i is the weight matrix of the i th layer. The expansion logic is shown in Figure 3.

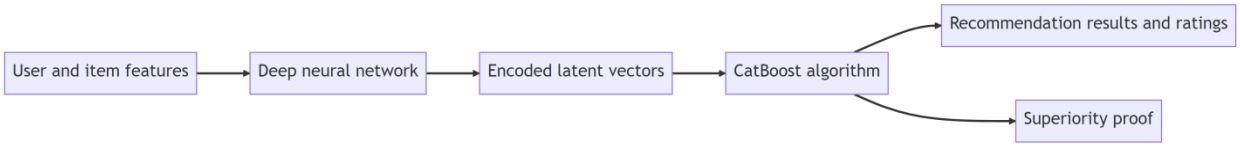

Figure 3 The improvement and extension logic of the CatBoost algorithm

4. User Behavior Analysis and Personalized Recommendation Modeling and Application Based on CatBoost Algorithm

4.1 Construction of User Behavior Analysis Model

User behavior analysis aims to provide users with recommendations in line with interest preferences and to improve user satisfaction and business benefits by building models and optimizing algorithms. Although user behavior analysis is not a new topic, the CatBoost algorithm outlines the classification feature dimension of user behavior analysis, expands the technical connotation of user behavior analysis, and endows user behavior analysis with the value of integrating data mining, machine learning, and artificial intelligence by integrating user behavior analysis with gradient boosting trees and other elements in an orderly manner [6]. Therefore, the CatBoost algorithm has been successfully practiced and explored in e-commerce, social networking, entertainment, etc., providing experience and reference for user behavior analysis. However, compared with the requirements of the theoretical construction and mechanism design of user behavior analysis and the current Internet requirements, the CatBoost algorithm still needs to be further optimized by closely matching the dynamic interaction between users and items to generate personalized recommendations. The user behavior analysis model is shown in Figure 4.

This paper proposes a user behavior analysis method based on user profiles. This method uses the classification feature processing ability of the CatBoost algorithm to transform the user behavior analysis issue into the user profile construction issue to realize accurate profiling of users by training the gradient boosting tree model. To verify the effectiveness of the method, this paper uses MovieLens as a simulation platform to generate user behavior data of different types, locations, and interests as the input of the gradient boosting tree model. The simulation results show that the method can realize user behavior analysis in the data sparsity and cold start environment with good accuracy and generalization ability. In the process, the user profile is a description of user features. And its function is to reflect information such as the user's essential attributes and behavioral features, interest preferences, and consumption capabilities. The function of the user profile is as follows:

$$u(x) = f(x;\theta) \tag{4}$$

In the function, $u(x)$ is the user profile, x is the user behavior data, $f(x;\theta)$ is the gradient boosting tree model, and θ is the model parameter.

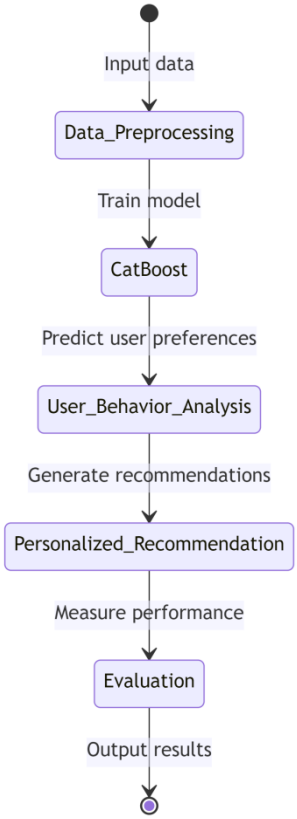

Figure 4 User Behavior Analysis Model

4.2 Personalized Recommendation Model Construction

In the related research on personalized recommendations, the static features of users and items were mainly considered before. In contrast, the dynamic interaction between users and items was ignored, leading to the problems of inaccurate, imprecise, and poor optimization recommendation results. To solve the above problems, the CatBoost algorithm is improved and extended by mainly integrating the multi-layer fusion feature network technology and the ordered boosting strategy to deeply optimize the traditional gradient boosting tree frame structure to achieve a better recommendation effect of the CatBoost algorithm. Specifically, firstly, the multi-layer fusion feature network is used to fuse the features of users and items to extract hidden vectors related to the matching degree of users and items. Secondly, the CatBoost algorithm is improved by using the ordered boosting strategy to solve the problem of prediction bias in the boosting tree algorithm. Then, the CatBoost algorithm is used to train the gradient boosting tree model to sort users and items automatically. Finally, different evaluation indicators are used to compare the CatBoost algorithm with other recommendation algorithms, proving the superiority of the CatBoost algorithm. In the process, the multi-layer fusion feature network is a feature learning model to fuse multiple user and item features. The function of the multi-layer fusion feature network is as follows:

$$h(x_u, x_i) = f_n\left(f_{n-1}\left(\ldots f_1\left(x_w, x_i; W_1\right); W_{n-1}\right); W_n\right) \tag{5}$$

In the function, $h(x_u, x_i)$ is the output of the multi-layer fusion feature network, x_u is the user feature, x_i is the item feature, f_i is the fusion function of the ith layer, and w_i is the weight matrix of the ith layer. The operation process of the personalized recommendation model is shown in Figure 5.

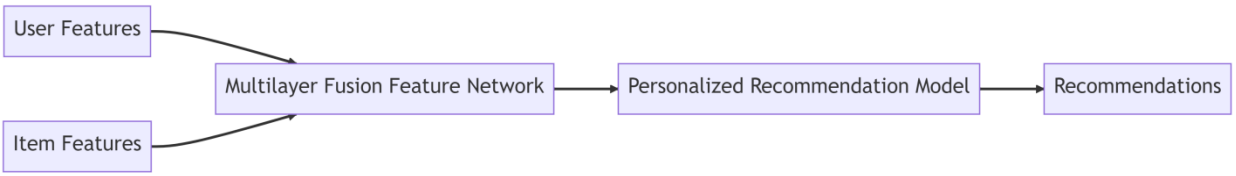

Figure 5 Operation process of the personalized recommendation model

4.3 Integration of User Behavior Analysis and Personalized Recommendation Model

Evidently, both user behavior analysis and personalized recommendation cannot avoid the "classification features" of data in terms of data mining. In recommendation mechanisms, classification features are standard and effective feature tools that are important in matching users and items. This also makes classification features not only a technical concept but also a value concept. Therefore, the CatBoost algorithm based on "classification features" has become the core user behavior analysis and personalized recommendation mechanism. Notably, the practice of the CatBoost algorithm is generally an optimization path gradually formed on the basis of the gradient boosting tree. However, this path includes the attempt at a deep neural network. From user profiles to personalized recommendations, the CatBoost algorithm always revolves around classification features. Therefore, the CatBoost algorithm should be dedicated to improving the recommendation effect to meet Internet requirements. However, in the case of data sparsity and cold start being amplified, this also brings a dilemma called the phenomenon of prediction bias. Overall, the CatBoost algorithm still needs to be improved regarding data processing and model construction. In addition, its recommendation performance needs further improvement, which is also an essential task for user behavior analysis and personalized recommendation. The proportion of each link of model fusion is shown in Figure 6.

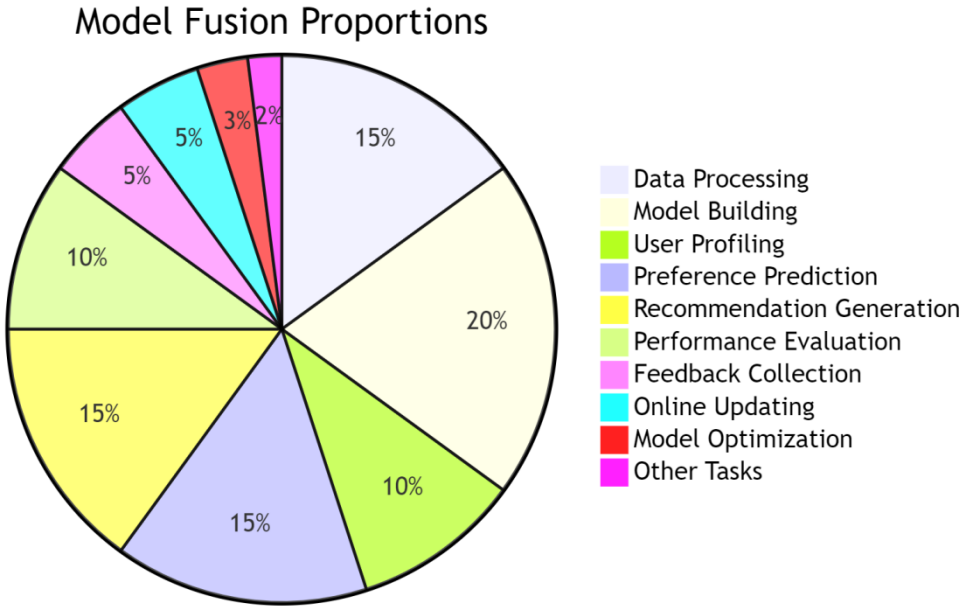

Figure 6 The proportion of each link of model fusion

4.4 Application Cases of User Behavior Analysis and Personalized Recommendation Model

From the user experience perspective, data sparsity and cold start have limited the capability of personalized recommendations for a long time. Since the 21st century, the CatBoost algorithm, which integrates data mining, machine learning, and artificial intelligence, has reshaped user behavior analysis and personalized recommendation through classification feature processing. However, the disadvantages of traditional gradient-boosting trees still restrict the recommendation effect. Due to prediction bias and the impact of dynamic interactions between users and items, the CatBoost algorithm still needs to be perfected. In the context of the Internet, deep neural networks are seen as a straightforward way to optimize recommendation capability. However, the practical

effect of static feature-based deep neural networks on the matching degree of users and items still needs to be determined. At the same time, deep neural networks lack classification feature processing due to the difficulty in the data processing. Therefore, deep neural networks also do not consistently achieve the goal of optimizing recommendations. Therefore, behavior analysis and personalized recommendation are not only technical problems but also problems of value.

5. Conclusion

This paper proposes a novel solution for user behavior analysis and personalized recommendation based on the CatBoost algorithm. First, the user profile model is reconstructed to take advantage of the classification feature information in the CatBoost algorithm model. After further tuning the CatBoost algorithm parameters, an ordered boosting strategy is used to sort users and items completely. Then, the personalized recommendation model is constructed by integrating deep neural network technology based on the experiments. Theoretical analysis, simulation, and experimental results show that the CatBoost algorithm can effectively solve data sparsity and cold start, improve recommendation accuracy and generalization ability, and optimize the user experience and business benefits.

References

[1] Sharma S, Rana V, Kumar V. Deep learning based semantic personalized recommendation system[J]. International Journal of Information Management Data Insights, 2021, 1(2): 100028.

[2] Celik E, Omurca S. Comparative Analysis of Offline Recommendation Systems with Machine Learning Algorithms[J]. PROCEEDINGS BOOK, 2021.

[3] Cao W, Wang K, Gan H, et al. User online purchase behavior prediction based on fusion model of CatBoost and Logit[C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 2003(1): 012011.

[4] Chabane N, Bouaoune A, Tighilt R, et al. Intelligent, personalized shopping recommendations using clustering and supervised machine learning algorithms[J]. Plos one, 2022, 17(12): e0278364.

[5] Chen S, Huang L, Lei Z, et al. Research on personalized recommendation hybrid algorithm for interactive experience equipment[J]. Computational Intelligence, 2020, 36(3): 1348-1373.

[6] Dou X. Online purchase behavior prediction and analysis using ensemble learning[C]//2020 IEEE 5th International conference on cloud computing and big data analytics (ICCCBDA). IEEE, 2020: 532-536.