Data of E-Commerce Users Based on Data Mining Technology

Haohua Qing¹*, Jiali Zhang², Die Fu³

¹Department of Economic And Trade, Guangzhou College of Technology and Business, Guangzhou, 510850, Guangdong, China
²Department of Computer Science and Engineering, Guangzhou College of Technology and Business, Guangzhou, 510850, Guangdong, China
³Sichuan University of Arts and Science, Dazhou, 635002, Sichuan, China

*Corresponding author e-mail: 13986536@asau.edu.cn

Abstract. With the rapid development of information technology, the rapid development of e-commerce Internet has been involved in every corner, resulting in the growth of the data volume of e-commerce users. This paper mainly studies the data analysis of e-commerce users based on data mining technology. This paper introduces the research background and significance of e-commerce user behavior. On this basis, relevant technologies of e-commerce user behavior analysis are studied, including components of Spark platform and classification algorithm of relevant user behavior analysis model. Then, the advantages and feasibility of the improved model are verified through comparative experiments between the improved Spark XGBoost model and traditional machine learning method. The study in this paper provides a parallel method for the prediction of e-commerce user behavior, which can be applied to daily life as an effective method to predict e-commerce behavior.

Keywords: Data Mining, TV Platform, User data Analysis, Spark Platform

1. Introduction

With the development of the industry in recent years, the Internet, especially the mobile Internet industry on e-commerce platforms, is developing steadily and rapidly, thus continuously improving the living standard of the common people. Nowadays, the population of e-commerce platforms is no longer limited by the young people in the cities, and even the middle-aged and old people in the villages have joined the army of e-commerce platforms, which leads to a relatively stable growth rate of China's online shopping market. At the same time, people leave a large amount of information about users' consumption behavior when browsing, clicking, collecting and buying goods on shopping websites, which makes it possible to study and analyze users' consumption behavior. There are more and more e-commerce behaviors on the Internet, which promotes the rapid development of Internet users on e-commerce platforms and brings the whole e-commerce industry into the era of big data. It is precisely because of the advent of the era of e-commerce big data that the analysis of e-commerce user...
behavior is of great practical significance for social development, whether economically, politically or culturally. For example, the number and types of commodities on the websites of some large e-commerce users are huge. How to show the products suitable for users and consumers, and recommend the products they are likely to consume, is a major problem that e-commerce platforms mainly solve. At the same time, accurate marketing recommendation also needs to send commodities and other relevant information to e-commerce users, which is more accurate and real-time than the traditional mode of marketing [1]. Therefore, to solve the above problems, we must study the behaviors of e-commerce users.

At the end of last century, researchers of various platforms have started offline or real-time user behavior data analysis and built a series of algorithm modeling. Dongre use of social network users in a platform release millions of social network content for natural language processing and analysis, so as to reflect the current social network content and the theme of the semantic analysis of solid content, and then look for external users behavior associated with the topic content, enhances the semantic understanding of the contents of this social, finally using the semantic enhance depicting the portrait of social network users, and use the training model of social network users personalized recommendation of various relevant content [2]. Mikkonen constructs the characteristics of e-commerce users and commodities, constructs a set of relevant training models using deep learning method, and conducts comparison experiments with traditional machine learning integrated learning methods such as decision tree and random forest, and finds that deep learning has a better prediction effect [3].

In this paper, from the perspective of the electric business platform, user behavior data of a electric business platform for real-time analysis, to ensure that the actual reliability of the data, introduces some related technologies, after comparing the choose to use the framework of the Spark for data processing, also some algorithms are used to support to complete the analysis of the large data, and the results have been obtained.

2. Data Mining is Used to Process the Data of E-Commerce users

2.1 User Behavior Data Processing Technology

(1) Introduction to Spark
Spark is a big data processing framework. After Spark cluster deployment, the Master process needs to be started on the Master node of Spark and the Work process needs to be started from the slave node, so that the entire Spark cluster can be operated and controlled [4]. The Driver application is the starting place of the Spark framework logic execution. It mainly dispatches the tasks of Spark jobs and transmits them to various workers. It can also process these tasks in parallel, that is, create multiple partitions to process these tasks [5].

The differences between Spark and Hadoop are mainly as follows:
1) Hadoop mainly has two important functions, the first is the distributed file system HDFS, the second is the distributed computing system MapReduce, these two functions are one of the advantages of Hadoop, because Spark only provides computing function, there is no distributed file system function [6]. However, one of the advantages of Spark is that it can be used in combination with Hadoop's components, and one of its running modules, Spark on Yarn mode, is a prime example;
2) Both frameworks have computing advantages, but the latter is faster than the former, that is, Spark is many times faster as a running framework. In addition, Spark can take advantage of more performance.

(2) User Behavior Data Analysis Model
1) Decision tree is a common machine learning algorithm in both the e-commerce field and other fields. As the name suggests, it is a tree model, and the usual linear model of machine learning
algorithms, such as machine learning algorithms in common logistic regression, it will be all of the features after the last step is converted to a percentage probability, by comparing the size of the threshold value to judge the classification of the prediction, less than a certain threshold, divide it into a class, on the contrary, it is another kind of [6]. Furthermore, the decision tree can segment nonlinear characteristics. Thus, it can be seen that the machine learning algorithm based on tree model is more in line with the thinking characteristics of human beings and has the explanatory ability for the machine learning model generated by feature division.

2) XGBoost, like GBDT, is one of the ensemble Learning Boosting algorithms. Relative to the decision tree and GBDT, XBGoost algorithm steps with the two basic same, is first put a constant process, unlike GBDT, GBDT only using machine learning L1 regularization, XGBoost is using the machine learning L1 and L2 regularization, then slowly formed several weak learning, slowly learning residual update weight [7]. Also, XGBoost does a second-order Taylor expansion of the loss function, followed by the addition of the L1 and L2 regularization described above to the XGBoost final target function, which balances the overall complexity of the training model.

The core idea of the improved XGBoost algorithm is as follows:

Add regularization to the target function:

\[ L(\Phi) = \sum_i L(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (1) \]

where \( \Omega(f) = \gamma T + \frac{1}{2} \lambda \| \omega \|^2 \quad (2) \)

Use the \( \lambda \), \( \gamma \) criteria for splitter points and associated with the regularization term:

\[ L_{\text{split}} = \frac{1}{2} \left( \frac{(\sum_{i \in L} g_i)^2}{\sum_{i \in L} h_i + \lambda} + \frac{(\sum_{i \in R} g_i)^2}{\sum_{i \in R} h_i + \lambda} - (\sum_{i \in L} g_i)^2 - \sum_{i \in R} g_i^2 \right) - \gamma \quad (3) \]

Through the above improvements, XGBoost algorithm not only has the advantages of decision tree and GBDT algorithm, but also makes up for the shortcoming that GBDT algorithm may fall into local fitting, and makes the gradient descent direction more accurate, which improves the overall accuracy of the model.

2.2 Process of Analyzing the Behavior Data of E-Commerce Users

(1) Data Preprocessing

1) Data cleansing

After obtaining the behavior data of e-commerce users, we should process these data, remove those redundant data or some data with no research value, delete these data, and then operate through the deleted data [8].

2) User identification

Users to identify ways to determine whether the user of this electric business platform of the customer, determined by the identity of the visit to some, if the user is the first time to visit, you will need to provide in the platform and register your personal basic information, such as setting up an ID, or a name, convenient access to the first time after get to the user's information, and through each of the IP address of the access platform to determine whether the same user at login, ensure the user's uniqueness.

3) Session recognition

Session recognition is all the actions a user takes between logging in to a platform and logging off from that platform. The task of session identification is to obtain the session process data of a user, the session has a length of time, in some short time, do not do too much operation, but once the session is too long, we need to use some methods to determine the login access operation of each user.
(2) Data Mining Stage

The data mining stage is an important part in the field of big data, which is widely applicable to some recommendation algorithms, machine learning, artificial intelligence and other popular fields [9]. This paper analyzes the behavior data of e-commerce users through mining.

Understanding the business: This paper studies the business perspective of e-commerce platform, through the platform to further understand the behavior of users in the industry, and translate these into a series of relevant definitions of data mining, so as to make a preliminary plan.

Understanding data: Through the plan, we will make a brief analysis of the collected data to understand each user's use of the e-commerce platform. According to these preliminary analysis, we will complete the exploration of data and the authenticity of data, etc.

Prepare data: the original obtained data will be processed by some algorithms or some technologies, and the provided data will be transformed into a data class cluster, including the user's login data, user's access data, user's order data and some user's history data, etc.

Modeling: Select an appropriate mathematical model for further manipulation of the data.

Model evaluation: Conduct a preliminary evaluation of the selected mathematical models, and determine the order of each operation by monitoring these models to ensure that it is truly applicable to the e-commerce field.

Model deployment: After the mathematical model is determined, it does not represent the end of the experiment. Although the significance of establishing the mathematical model is to improve the analysis of data, it should also be reflected through the acquired knowledge system through the user's behavior.

(3) Data Analysis Stage

Data analysis phase is to use the way of a variety of data analysis and data model of data to be processed for analysis and research, this article is based on the electric business platform of data to be processed in the user behavior data analysis and research, through this stage can be integrating data, and to research the internal relation between data and some within the law to [10]. Through understanding the characteristics of these data can help us more conveniently to accurately analyze the specific relationship between the data.

3. User Behavior Data Experiment

3.1 Experimental Holistic Approach

This paper mainly focuses on supervised learning method, which is a common binary classification problem in machine learning user behavior model. Therefore, this paper takes the goods purchased by the last user as the label value, and this paper focuses on whether the classification of the last purchased goods is accurate. There are many common standards for measuring the second classification, such as accuracy, recall rate, F1, ROC curve, etc. In this paper, F1 is mainly used as a unified evaluation standard. Formula of F1 is as follows:

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$  \hspace{1cm} (4)$$

The Precision of the above formulas refers to the proportion of the correct samples predicted in the training model to the total number of classified predicted samples. Recall refers to the ratio between the number of positive samples in the training model and the number of positive samples in the original training set. In order to obtain the confusion matrix as shown in Table 1, the predicted final positive sample can be set to label value 1 and the negative sample to be 0.
Table 1. Confusion matrix

| Predicted results | Actual results | 1     | 0     |
|-------------------|----------------|-------|-------|
| 1                 | TP             |       |       |
| 0                 | FN             |       |       |
| 0                 | TN             |       |       |

As can be seen above, if the predicted result is a positive sample and the actual result is also a positive sample, it is classified as true positive (TP). If the predicted result is a positive sample but the actual result is a negative sample, then it is classified as false positive (FP). In conclusion, it can be seen that the accuracy rate and recall rate can be expressed as:

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{P}$$

4. Experimental Results of User Behavior Data

4.1 XGBoost Model F1 Mean

![Figure 1. The F1 mean of the training model](image)

As shown in Figure 1, the above diagram shows that the decision tree training model is the worst overall, while the XGBoost training model is the best overall. The next step is to compare the improved XGBoost model to the previous single XGBoost model, but unlike the previous single XGBoost model, instead of combining all the training sets, each sliding window uses an XGBoost and outputs the final prediction by weighting. Similarly, the improved XGBoost model still uses a ten-fold cross-validation approach, using nine of the sliding Windows as training data sets and the last one as test data sets.

4.2 Improved XGBoost Model F1 Mean
Table 2. A comparison of the single and hybrid XGBoost models

|                | Training set 1 | Training set 3 | Training set 5 | Training set 7 | Training set 9 |
|----------------|----------------|----------------|----------------|----------------|----------------|
| Single model   | 15.47%         | 15.45%         | 15.61%         | 15.66%         | 15.12%         |
| Hybrid model   | 15.84%         | 15.79%         | 15.57%         | 15.71%         | 15.39%         |

As shown in Table 2, the F1 mean of the improved XGBoost model is slightly higher than that of the XGBoost single model, indicating that the improved model on this dataset does not have a very high advantage, but it does have a significant advantage over other single models, showing the importance of the XGBoost model.

4.3 Variance Comparison of Each Model

As shown in Figure 2, the improved XGBoost model had the smallest variance value of the predicted results, followed by a single XGBoost, while the F1 variance value of traditional machine learning methods was generally higher, indicating that traditional machine learning methods were less stable than the XGBoost model in terms of tree building and pruning. In conclusion, it can be concluded that the improved XGBoost model is a good model in terms of both accuracy and stability of prediction results.

4.4 Summary of This Chapter

The experiment of user behavior analysis mainly uses the method of ten-fold cross validation to iterate the experiment, no matter it is a single model or a mixed model. As predicted by the comparison between the single models, the XGBoost model performed better overall than the other single models. As you can see from the comparison between the single model and the hybrid model, the hybrid XGBoost as a whole is not that different from the single-model XGBoost, but it ends up being slightly better than the single-model XGBoost. From the standpoint of F1 mean variance, the improved XGBoost is more stable than other models.

5. Conclusions
This article begins with an introduction to the relevant technologies of user behavior, including the Spark platform and components, as well as the commonly used machine learning classification model algorithm for user behavior analysis. Then, the data source of user behavior and the key task of this paper are described, that is, the dichotomies of predicting whether users will buy or not, and the unsupervised learning of user behavior is transformed into the supervised learning of user behavior. Then, a series of user behavior preprocessing is carried out, including the elimination of outliers, the sharding strategy of time series, feature screening and the balance of positive and negative samples, etc. After the pre-processing, the improved Spark XGBoost fusion design is compared with other traditional algorithms, and it is verified that the improved algorithm can simplify the complexity of classification model to some extent, and improve the accuracy of prediction classification standards.

References
[1] Akter S, Wamba S F. Big data analytics in E-commerce: a systematic review and agenda for future research[J]. Electronic Markets, 2016, 26(2):173-194.
[2] Dongre V, Raikwal J. An Improved User Browsing Behavior Prediction using Regression Analysis on Web Logs[J]. International Journal of Computer Applications, 2015, 120(19):19-23.
[3] Mikkonen T, Klamma R, Hernández, Juan. [Lecture Notes in Computer Science] Web Engineering Volume 10845 || Leveraging Analysis of User Behavior from Web Usage Extraction over DOM-tree Structure[J]. 2018, 10.1007/978-3-319-91662-0(Chapter 14):185-192.
[4] Longhui W, Yong W, Yudong X. Implementation of a Parallel Algorithm Based on a Spark Cloud Computing Platform[J]. Algorithms, 2015, 8(3):407-414.
[5] Yang Y, Yu J, Yang M, et al. Probabilistic modeling of renewable energy source based on Spark platform with large-scale sample data[J]. European Transactions on Electrical Power, 2019, 29(3):e2759.1-e2759.13.
[6] Decaestecker C, Velthoven R V, Petein M, et al. The use of the decision tree technique and image cytometry to characterize aggressiveness in World Health Organization (WHO) grade II superficial transitional cell carcinomas of the bladder.[J]. Journal of Pathology, 2015, 178(3):274-283.
[7] Chen W, Fu K, Zuo J, et al. Radar emitter classification for large data set based on weighted-xgboost[J]. IET Radar Sonar ? Navigation, 2017, 11(8):1203-1207.
[8] Wu S L. Analysis of Food E-commerce Consuming Behavior Based on Web Data Mining Theory[J]. Advance Journal of Food ence and Technology, 2016, 10(11):874-878.
[9] Tang M, Wu Z. Research on the mechanisms of big data on consumer behavior using the models of C2C e-commerce and countermeasures[J]. African journal of business management, 2015, 9(1):18-34.
[10] Baumann A, Haupt J, Gebert F, et al. The Price of Privacy: An Evaluation of the Economic Value of Collecting Clickstream Data[J]. Wirtschaftsinformatik, 2019, 61(4):413-431.