Research on Consumer Behavior Prediction Based on E-commerce Data Analysis

He Li

School of Management, Liao Ning Communication University, Shenyang 110136, China

176866986@qq.com

Abstract. The user behavior log of e-commerce platform contains a wide range of data, which can be roughly divided into three types: user basic information, merchant information and user behavior information. Behavioral information includes clicks, favorites, shopping carts and purchases. Because of the asymmetry of online shopping information, the virtuality of trading environment, the incomplete perception of goods and the unsynchronization of trading process, consumers are faced with many uncertainties in the process of online shopping. According to the specific research problems, this paper applies the real consumer behavior data in the e-commerce data platform to build a model to predict consumer behavior. The consumer behavior sequence and other characteristics are analyzed by two machine learning methods respectively. First, the consumer behavior sequence based on time sequence is analyzed. Add the obtained probability value as a new feature to other feature sets as the input of NB (Naive Bayes) model, and train the model to get the probability of whether consumers will buy, and then get the final judgment result of the model. The results show that the fusion model has a balanced performance for consumer behavior sequences with different lengths, and its accuracy can be kept at around 0.9. The research on the problems raised in this paper can also be extended to other fields.

Keywords: E-commerce, Consumer Behavior, Prediction

1. Introduction

At present, it is an era of Internet information technology. The construction and development of social e-commerce platform in China should keep pace with the times and keep up with the pace of the times. Due to the asymmetry of online shopping information, the virtuality of trading environment, the incomplete perception of goods and the unsynchronization of trading process, consumers are faced with many uncertainties in the online shopping process [1-2].

If modern social e-commerce platform wants to stand out in the fierce competition e-commerce market, it must attach great importance to the research on the behavior characteristics of platform consumers, highlight the main position of users on the platform, and carry out marketing services around the daily behavior characteristics of users. At this time, it is a better choice to predict the next behavior of consumers through the implicit feedback data generated by consumers, so as to understand the consumer's purchase behavior bias of a certain commodity. According to the specific research problems, this paper applies the real consumer behavior data in the e-commerce data platform to build a model to predict consumer behavior.

If merchants lack understanding of consumers' purchasing intentions, the push of product information will lose focus and timeliness, increase the cost of pre purchase services, and reduce the return rate. This method analyzes user behavior to determine their purchasing intention, and then uses the recommendation system to push personalized products for users, making it a better choice. A personalized recommendation system effectively filters and extracts massive product information based on consumers' own taste and purchasing needs, providing consumers with more targeted product recommendations, thereby improving their satisfaction and loyalty.

2. Research method

2.1 Establishment of consumer behavior prediction model

When consumers buy goods, they often get recommendations from friends, relatives, classmates and colleagues around them. The process of purchase and consumption requires highly correlated individual trust, and also requires insight and analysis of consumers' user behavior [3-4]. Compared with the consumption and shopping behavior on the traditional e-commerce platform, they have more shopping consumption patterns on the social e-commerce platform. The social e-commerce platform has a stronger social communication effect and the rapid response of the fan community, so that the value of its own marketing planning scheme can be maximized.

The user behavior log of e-commerce platform contains a wide range of data, which can be roughly divided into three types: user basic information, merchant information and user behavior information. Behavioral information includes clicks, favorites, shopping carts and purchases. If we directly use the original e-commerce user behavior log data, we will face various challenges, such as noise and deviation [5-6]. There is considerable uncertainty in the prediction of e-commerce users' revisiting, which is mainly due to two reasons: first, there are various factors of users' revisiting, that is, the characteristic space of data is scattered; Second, the user's access records are unevenly distributed in the data set. In this chapter, our question is about users' accurate revisit prediction to businesses.

In this paper, the consumer behavior sequence and other characteristics are analyzed by two machine learning methods respectively. First, the consumer behavior sequence based on time sequence is analyzed. Combining the obtained probability with other feature sets, and taking it as the input of NB (Naive Bayes) model, the probability of whether customers will buy or not is obtained by training this model, so as to obtain the final judgment of this model, which is expressed as 0,1. Figure 1 shows the construction of the fusion model.

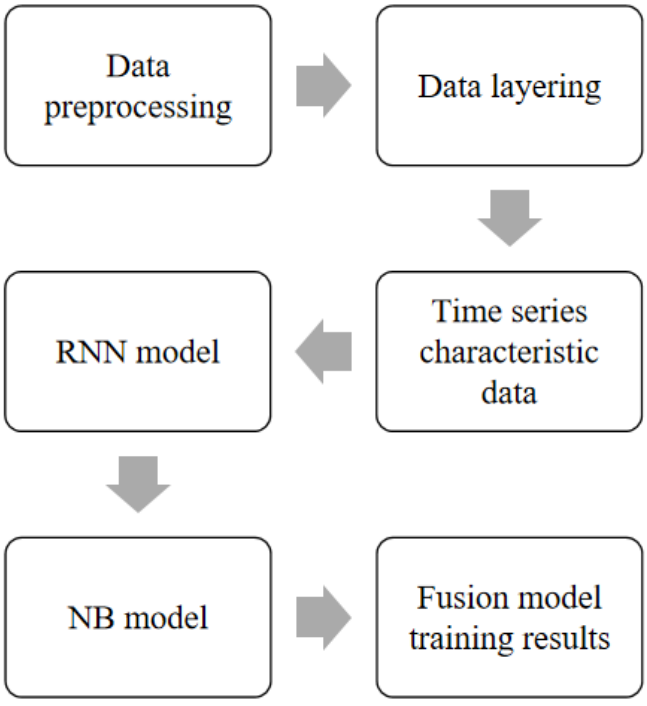

Figure 1 Fusion model structure diagram

In this paper, we use RNN (Recurrent neural network) to classify behavior data. In this process, an output value will be generated to judge the next Consumer behaviour. RNN classification label C is shown in the following formula:

$$C = \begin{cases} Y_0 = 0, \ Consumers\ will\ not\ buy \\ Y_1, \ Consumers\ will\ buy \end{cases} \quad (1)$$

Because the problem studied in this paper is a classification problem, the Softmax function is used, and finally Score and c are calculated as the last hidden layer state.

$$\text{Score} = \text{Softmax}(\text{Vh}_\text{l} + \text{c}) \quad (2)$$

In the NB model, we will use two possible classification labels $Y = \{C_1, C_2\}$ to describe customers' buying and not buying behaviors. Where X is the sample set to be classified, and the score of the initial classification result obtained by RNN model is considered as an attribute of the sample set to be classified. Suppose there are $\text{d} + 1$ kinds of attributes, among which $\text{d} + 1$ kinds are the result of RNN classification.

2.2 Data preprocessing

Real data is the basic element of data mining. In order to obtain the real data set, this paper uses the web crawler technology to crawl the consumer behavior data of an e-commerce company for one month. In addition, in the use of tools, because Python has a good support of third-party modules, and Python syntax is simple, so the work of this paper mainly adopts Python. After crawling the webpage data to the local area, Python's beautifulsoup module is used to parse the webpage data in html format, and generate the item detail file and user file.

Variables related to whether consumers buy a product, such as user activity, product popularity, user's preferred brand, etc., need to be extracted from the original user behavior log, and the variables are screened and analyzed through the model. Because the original data contains all the information of the user's behavior every time, the variables needed to judge the user's purchase can be constructed on this basis [7-8]. It can be said that the key of this paper is to use theoretical knowledge and business experience in the field of e-commerce to extract and screen variables from the original data for the prediction of user consumption. The choice of variables determines the upper limit of the prediction ability of the model, and different algorithms just keep approaching this upper limit.

There are always several problems in the original data: data inconsistency, data repetition, data loss and information redundancy. The data quality determines the upper limit of the accuracy of the prediction algorithm, so some data processing should be done before the main processing. For some quantitative attributes, the effective information is interval division, such as age. Age can be divided into several intervals according to age, such as "1" for those under 18 years old and "2" for those aged 18-25 years old.

Occasionally, the data will be inconsistent, such as the gender of a user may be different in different items [9]. If the gender of the user can be judged, the gender of the user can be corrected. Otherwise, you can give a gender or delete the user. Missing data is supplemented or deleted according to other data. After the screening and sorting of the above process, 8369 pieces of data were obtained. Because the abnormal values contained in the data may affect the actual effect of clustering, this paper adopts the way of data visualization to clean and preprocess the data [10].

3. Result analysis

The data used in this paper has been processed by Mysql5.6 and saved in the form of csv file. This experiment was completed in Python3.5, which is used for programming, graphic analysis and demonstration. TensorFlow is a deep learning method open to Google. TensorFlow has a good structure, and can be easily run on different platforms, such as mobile devices, servers, and one or more CPUs. It is very flexible, portable and can be used in many languages.

This topic takes the operation sequence of consumers as the breakthrough point. In terms of consumer behavior, RNN is used to analyze and learn the behavior sequence of consumers, and the training results of RNN are added to the remaining feature sets, and then they are input into NB model, so as to obtain the prediction results. A comparison of the predictive effects of the two models is shown in Figure 2.

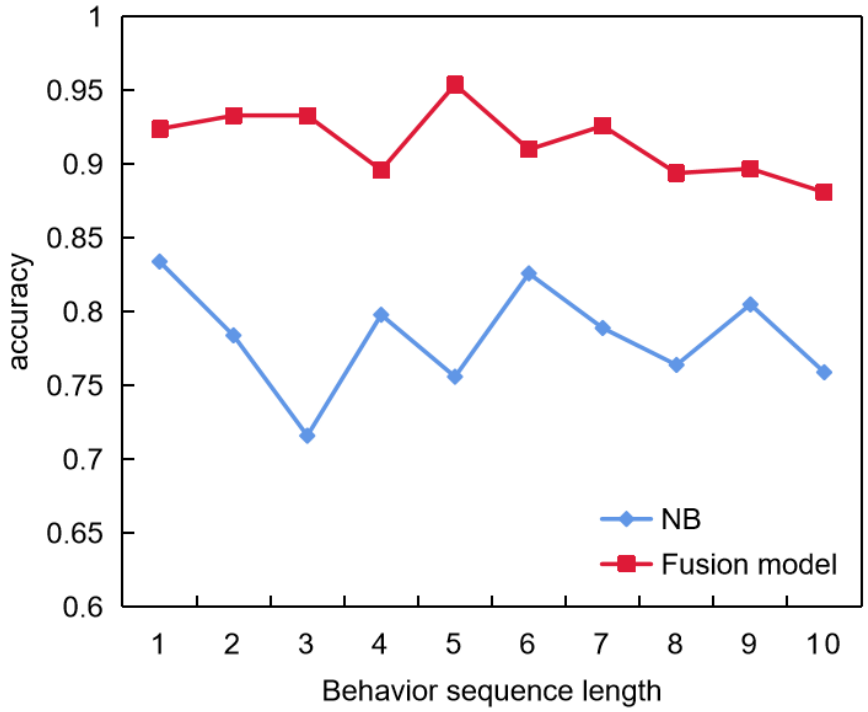

Figure 2 Comparison of prediction effects of models

The fusion model has a balanced performance for consumer behavior sequences with different lengths, and its accuracy can be kept at around 0.9. Similar to NB, the prediction accuracy also improves with the increase of the length of the behavior sequence. Among them, the ability to recognize short sequences is slightly stronger, and it can better handle consumer behavior information with different lengths. When the number of consumer behaviors reaches six, the model shows the best effect.

In the problem studied in this paper, if the model can perform well in predicting short consumer behavior, it can meet the purpose of predicting buying behavior from a qualitative point of view. From the overall index of model training, the fusion model has a better classification effect in the prediction of short behavior series, which is more effective than NB single model.

Accurately predicting users' re-visit to merchants on e-commerce platform will have unpredictable value for both platforms and merchants. If it helps the e-commerce platform to provide personalized services for users, and establish a close relationship between merchants and users, it will promote the long-term, stable and high-speed development of the e-commerce platform; It is helpful for merchants to accurately market users and establish a stable loyal user base. The research on the problems raised in this paper can also be extended to other fields, such as not only e-commerce platforms, but also music websites, search engines, payment websites and take-away platforms, so the research on this issue has great practical significance and value.

4. Conclusions

If modern social e-commerce platform wants to stand out in the fierce competition e-commerce market, it must attach great importance to the research on the behavior characteristics of platform consumers, highlight the main position of users on the platform, and carry out marketing services around the daily behavior characteristics of users. According to the specific research problems, this paper applies the real consumer behavior data in the e-commerce data platform to build a model to predict consumer behavior. In the fusion model, firstly, RNN is used to analyze and learn the consumer behavior sequence, and then, RNN training results are added to the remaining feature sets as new feature items, and the final prediction results are obtained by inputting NB model. The fusion model has a balanced performance for consumer behavior sequences with different lengths, and its

accuracy can be kept at around 0.9. Similar to NB, the prediction accuracy also improves with the increase of the length of the behavior sequence. The research on the problems raised in this paper can also be extended to other fields.

References

[1] Prinzie, A. , & Poel, D. V. D. (2011). Modeling complex longitudinal consumer behavior with dynamic bayesian networks: an acquisition pattern analysis application. Journal of Intelligent Information Systems, 36(3), 283-304.

[2] Joachims, T. , Granka, L. , Pan, B. , Hembrooke, H. , Radlinski, F. , & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. Acm Transactions on Information Systems, 25(2), 7.

[3] Ishigaki, T. , Takenaka, T. , & Motomura, Y. (2011). Customer behavior prediction system by large scale data fusion in a retail service. Transactions of the Japanese Society for Artificial Intelligence, 26(6), 670-681.

[4] Mu, W. (2019). A big data-based prediction model for purchase decisions of consumers on cross-border e-commerce platforms. Journal Europeen des Systemes Automatises, 2019(4), 52.

[5] Meso, P. , Musa, P. F. , & Mbarika, V. W. A. (2010). Towards a model of consumer use of mobile information and communication technology in ldcs: the case of sub-saharan africa. Information Systems Journal, 15(2), 119-146.

[6] Ren, Z. , Wan, J. , Shi, W. , Xu, X. , & Zhou, M. (2014). Workload analysis, implications, and optimization on a production hadoop cluster: a case study on taobao. IEEE Transactions on Services Computing, 7(2), 307-321.

[7] Wen, Y. , Kong, L. , & Liu, G. (2021). Big data analysis of e-commerce efficiency and its influencing factors of agricultural products in china. Mobile Information Systems, 2021(1), 1-8.

[8] Goh, K. Y. , Heng, C. S. , & Lin, Z. (2013). Social media brand community and consumer behavior: quantifying the relative impact of user- and marketer-generated content. Information Systems Research, 24(1), 88-107.

[9] Zhang, H. , Wang, M. , Yang, L. , & Zhu, H. (2019). A novel user behavior analysis and prediction algorithm based on mobile social environment. Wireless Networks, 25(2), 791-803.

[10] Duan, J. , Liu, H. , & Zeng, J. (2013). Posterior probability model for stock return prediction based on analyst's recommendation behavior. Knowledge-Based Systems, 50(10), 151-158.