Prediction of User Consumption Behavior Data Based on the Combined Model of TF-IDF and Logistic Regression

Shuwei Xiao, Weiqin Tong *
School of Computer Engineering and Science, Shanghai University, No.99, Shangda Road, Shanghai 266555, China

*1216210029@qq.com

Abstract. In the era of the rapid development of computers and the Internet, e-commerce has become a part of the economy of many countries. Therefore, how to use historical data of user consumption behavior to predict user shopping intentions accurately and subsequent personalized recommendations has turned into research hotspots in the field of e-commerce. This paper conducts a basic analysis of JD’s e-commerce data based on machine learning. Specifically, this article constructs user-product matrix and product and user clustering by means of text processing and clustering, as well as implements a logistic regression classifier to predict the user's purchase intention of products in a certain target category in the next 5 days. Based on the JD competition data set, this article has a prediction accuracy rate of 98%. This can help e-commerce companies make better decisions.

Keywords: e-commerce; text processing; clustering; logistic regression; classifier

1. Introduction
In recent years, with the rapid development of computer Internet technology, the mobile Internet has followed closely, and online shopping has become one of the main ways people consume in life. E-commerce refers to the commercial behavior based on computer network technology, centered on the purchase of goods [1], to realize the electronic and informatization of traditional commercial behavior. Therefore, how to better explore the shopping behavior of customers and promote the development of the industry has proved to be an important development strategy for related companies. However, a major problem at present is how to extract the value of a large amount of user behavior consumption data. This is still a problem faced by many companies, especially how to analyze and predict user behavior accurately and perform user portraits and classifications based on related behaviors for personalized marketing and product recommendation. In response to the above-mentioned problems faced by e-commerce, this article uses the historical data of JD users’ consumer behavior to conduct visual analysis and predicts the user’s purchase behavior by logistic regression, to comprehensively explore the impact
of different influencing factors on consumer purchase decisions. It is of practical significance to make personalized recommendations for relevant companies and increase their operating profits.

2. Related research
As far as e-commerce data is concerned, how to analyze e-commerce data better to obtain more potential value has become the key research content of many researchers. Many scholars focus on research from the perspective of machine learning. Qian Ma used machine learning algorithms such as logistic regression and decision trees to analyze e-commerce data to predict customers who are likely to buy the same things repeatedly [2]; Kai Peng and others used factor analysis and K-means to classify users and achieved remarkable results [3]; Qingguang Chen built an E-Commerce Data Mining and Visualization Model (EDVM) by combining data mining technology, data visualization technology with online analysis and processing technology [4]; Kui Zhao used convolutional neural networks in deep learning to mine e-commerce data. He designed a series of effective algorithms and optimization methods for the two aspects of product matching recommendation and product sales forecasting [5]; Boqing Lv combined crawler technology and corresponding data mining methods to collect e-commerce web page information and conduct exploratory analysis [6].

3. Research methods and analysis

3.1. Data set description
The source of experimental data in this article is the JD competition data set, which includes user data, user behavior data, user evaluation data, and commodity data. The attributes of these four data are shown in Tab.1 - Tab.4 below:

| sku_id  | a1 | a2 | a3 | cate | brand |
|---------|----|----|----|------|-------|
| 10      | 3  | 1  | 1  | 8    | 489   |
| 100002  | 2  | 2  | 1  | 8    | 489   |
| 100003  | 1  | -1 | -1 | 8    | 30    |
| 100006  | 1  | 2  | 1  | 8    | 545   |
| 10001   | -1 | 1  | 2  | 8    | 244   |
| 100016  | 3  | -1 | 2  | 8    | 214   |
| 100029  | 3  | 2  | 2  | 8    | 214   |
| 10003   | 3  | 1  | 2  | 8    | 214   |
| 100045  | 2  | 2  | 2  | 8    | 124   |

Tab.1 Commodity data
In the above table, “sku_id” represents product ID, and represent the product attributes (unknown), “cate” represents the category of the product (data has been desensitized), and “brand” represents the brand of the product (data has been desensitized).
In Tab.2, “sku_id” represents product ID (data has been desensitized), “dt” represents the deadline (granularity is days), “comment_num” represents the segment of the cumulative number of comments (0: no comment, 1: only one comment, 2: 2-10 comments, 3: 11-50 comments, 4: more than 50 comments), “has_bad_comment” indicates whether the product has bad reviews (0: no bad reviews, 1: have bad reviews), and “bad_comment_rate” indicates the rate of bad reviews (the ratio of the number of bad reviews to the total number of reviews).

| user_id   | age | sex | Use lv_id | user_reg_tm |
|-----------|-----|-----|-----------|-------------|
| 200002    | -1  | 0   | 1         | 2016/1/26   |
| 200003    | 4   | 1   | 4         | 2016/1/26   |
| 200004    | -1  | 2   | 1         | 2016/1/26   |
| 200005    | 2   | 0   | 4         | 2016/1/26   |
| 200006    | 4   | 2   | 2         | 2013/4/10   |
| 200007    | 4   | 2   | 3         | 2016/1/26   |
| 200008    | -1  | 2   | 3         | 2016/1/26   |
| 200009    | 4   | 2   | 2         | 2016/1/26   |
| 200010    | 4   | 2   | 3         | 2016/1/26   |

Tab.3 User data

In Tab.3, “user_id” represents user ID (data has been desensitized), “age” represents age (-1 represents unknown), “sex” represents gender (0 represents male, 1 represents female, and 2 represents confidentiality), “user_lv_cd” represents the user level (ordered enumeration of levels, the higher the level, the higher the number), and “user_reg_tm” represents the registration date (granularity is days).

| user_id | sku_id | time       | model_id | type | cate | brand |
|---------|--------|------------|----------|------|------|-------|
| 287842  | 75018  | 2016/3/31 23:59 | 14       | 6    | 9    | 630   |
| 208266  | 31662  | 2016/3/31 23:59 | 1        | 8    | 545  |
| 209390  | 118799 | 2016/3/31 23:59 | 111      | 6    | 8    | 244   |
| 237311  | 5825   | 2016/3/31 23:59 | 1        | 8    | 885  |
| 257651  | 128104 | 2016/3/31 23:59 | 1        | 4    | 300  |
| 297325  | 128747 | 2016/3/31 23:59 | 0        | 6    | 8    | 489   |
| 257651  | 128104 | 2016/3/31 23:59 | 0        | 6    | 4    | 300   |
| 213380  | 81163  | 2016/3/31 23:59 | 0        | 6    | 6    | 306   |

Tab.4 User behavior data

In Tab.4, “user_id” represents user ID, “sku_id” represents product ID, “time” represents the time when this behavior occurred, “model_id” represents the click module ID, type represents the behavior of the user to purchase the product (1: browse the product, 2: add to the shopping cart, 3: delete from the...
shopping cart, 4: place an order, 5: follow, 6: click), “cate” means product category, and “brand” means product brand.

In order to better show the user, product, and product name purchased by the user, this paper conducts a visual analysis of the data, as shown in Fig.1:

![A Week Purchase Table](image)

**Fig.1** User purchases in a week

3.2. Data preprocessing

3.2.1. Word segmentation

This article uses the Chinese word segmentation tool, “Jieba” (Chinese for “to stutter”) Chinese text segmentation, to segment the text. It splits the sentence into words with language semantic meaning to improve the predictive ability of the model.

3.2.2. Remove stop words

There are many unnecessary words or phrases in the text content. Some unnecessary words and phrases need to be filtered in text processing in order to save storage space and improve search efficiency. The stop word list used is a Chinese stop word database.

3.2.3. Keyword extraction

Keyword extraction is a key step in text processing. This article uses TF-IDF to extract keywords. The calculation formula is as follows:

$$TF = \frac{\text{The number of times the current word appears in the document}}{\text{The total number of words in the document}}$$  \hspace{1cm} (1)

$$IDF = \log \left( \frac{\text{Total number of documents}}{\text{The number of documents in which the current word appears}} \right)$$  \hspace{1cm} (2)

3.2.4. Similarity calculation
After the keyword extraction, the similarity calculation of the text is also required, so that we can process the text data by using word frequency to represent text features. The similarity calculation formula used in this article is cosine similarity, and the calculation formula is as follows:

\[ \text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{||A|| ||B||} \]  

(3)

3.3. Algorithm and model introduction

3.3.1. TF-IDF model

TF-IDF (term frequency-inverse document frequency) is a commonly used weighting technique for information retrieval and text mining. TF-IDF is a statistical method used to evaluate the importance of a word or phrase to a document set or one of the documents in a corpus. The importance of a word or phrase increases in proportion to the number of times it appears in the document, but at the same time, it decreases in inverse proportion to the frequency of its appearance in the corpus. The main idea of TF-IDF is: if a word appears in an article with a high frequency of TF and rarely appears in other articles, it is considered that the word or phrase has good classification ability and is suitable for classification. The main idea of TF-IDF is: if a word or phrase appears in an article with a high frequency and rarely appears in other articles, it is considered that the word or phrase has good classification ability and is suitable for classifying.

A high word frequency in a particular document and a low document frequency of the word in the entire document collection can produce a high-weight TF-IDF. Therefore, TF-IDF tends to filter out common words and keep important words.

\[ TF - IDF = TF \times IDF \]  

(4)

The TF-IDF algorithm is very easy to understand and easy to implement, but its simple structure does not consider the semantic information of words, so it cannot handle the situation of polysemous and righteous words.

3.3.2. Logistic regression model

Logistic regression and multiple linear regression actually have many similarities, the biggest difference is that their dependent variables are different, and the others are basically the same. Because of this, these two regressions can be attributed to the same family, namely the generalized linear model. The logistic regression algorithm is a supervised learning algorithm for classification commonly used in the field of machine learning. The LR model can predict the probability of a certain situation under different independent variables, and then use the probability to complete the binary classification process. In terms of model training and recognition time, LR has great advantages over support vector machines and artificial neural networks [7]. The logistic regression model is:

\[ h(w^T z) = \frac{1}{1 + e^{-w^T z}} \]  

(5)
In the formula, $w$ is the regression coefficient, $z$ is the input, and the value of $h(w^Tz)$ represents the probability of predicting the dependent variable to 1.

The core of the construction of the LR model is to use the known sample data to obtain the best regression coefficients through limited-time training. We should let the LR model have better generalization capabilities on the basis of ensuring a better learning effect. For solving the regression coefficients in the LR model, the gradient descent method is a very effective method. The gradient descent method obtains the final regression coefficient based on the deviation between the actual result and the predicted result and the learning rate (setting parameter), and after multiple iterations of adjustment on the basis of the set initial value of the regression coefficient. The activation function used by logistic regression is the sigmoid function, and the function image is shown in Fig.2 below:

![Sigmoid activation function](image)

**Fig.2** Sigmoid activation function

4. Result analysis

First, this article uses TF-IDF to extract text keywords, and then vectorizes them and uses them as input features together with other attributes. If the purchase intention is greater than 0.75, then we determine that the user will purchase such products in the next week. Subsequently, this article uses logistic regression to train input features. In order to see more vividly whether the classification result is correct, this article uses a confusion matrix for visualization, and the confusion matrix is obtained as shown in the following Fig.3:

![Confusion matrix](image)

**Fig.3** Confusion matrix
It can be seen from Fig.3 that the correct rate is 98%, so the combined model used in this article is experimentally convincing.

5. Conclusion
This paper uses a composite model combined TI-IDF with logistic regression to analyze and predict JD’s user behavior and consumption data, and achieves good results. However, this article still has certain limitations. First, the number of data sets is far from enough; second, it is necessary to use big data frameworks, such as Hadoop and spark due to a large amount of data in e-commerce; third, the data obtained is not complete enough, so more multi-dimensional data are needed for experimentation. Therefore, in the later experiments, the big data framework will be used for data modeling to maximize the value of data.

References
[1] Y M Zhang 2020 Application of Computer Network Security Technology in Electronic Commerce China Computer & Communication pp 198-200
[2] M Qian 2017 Repeat Buyer Prediction for E-Commerce Based on Machine Learning Lanzhou University
[3] K Peng, Y B Qin and D Y Xu 2011 Customer Segmentation Modeling on Factor Analysis and K_MEANS Clustering Computer Science pp 154-158
[4] Q G Chen 2007 Research and application about the E-Commerce data mining and visualization analysis system model Computer Engineering and Applications
[5] K Zhao 2017 Deep Data Mining in E-commerce Based on Convolutional Neural Networks Zhejiang University
[6] B Q Lv 2018 Analysis of E-commerce Information Based on Crawler and Data Mining Lanzhou University
[7] A H Hou, W Gao and L Wang 2017 Research on Traffic Anomaly Detection Method Based on The Logistic Regression Model J. Chinese Journal of Engineering Mathematic pp 479-489