Research on Feature Engineering Training Model System for Deep Optimization of Merchant Transaction Data

Fenwei Guo*

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

*Corresponding author email: 19722090@bjtu.edu.cn

Keywords: Merchant Transaction Data, Fake Transaction, Deep Optimization, Feature Extraction, Stacking Fusion Model.

Abstract: With the continuous increase of the number of enterprises, especially the continuous expansion of the scale of small and micro merchants, there are more and more risky merchants with illegal transactions and fraudulent behaviors. Based on the sales records of commodities and the basic information of merchants, this paper proposes a false transaction identification method combining the Stacking fusion model and multi-layer perceptron, and identifies false transactions by identifying commodities that can increase sales by brushing orders. First, the deep belief network is used to learn the transaction features to obtain higher-level abstract features; then the multi-layer perceptron is used to perform the classification task to identify fake transactions. The transaction records and comment data of products are crawled from Taobao for experimental verification. Compared with the experimental results of other machine learning models, its performance has been significantly improved.

1. Introduction

With the rapid development of e-commerce, there are more and more e-commerce retail platforms, and more and more merchants have settled on each platform. Traditional marketing methods and operation methods have been unable to effectively improve the operational efficiency of e-commerce retail platforms. E-commerce retail platforms generate massive amounts of operational data every day. Big data is an inherent feature of the Internet industry. Digital operation is a unique "artifact" for e-commerce companies [1]. The core competitive advantage of e-commerce in the future comes from the ability to interpret data and the ability to respond quickly to changes in data. Data is the core resource in the e-commerce retail platform. Deeply excavate and analyze the deeper value of the data, build a set of keys, hierarchical, and realistic operation monitoring index system, and comprehensively manage and control various business risks, which can improve the management level of the platform. Operational efficiency.

The current mainstream risk prevention and control methods mainly include blacklist system, expert rule system and machine learning feature model. However, the blacklist system relies heavily on the information of the blacklist database and external data, and is less effective in identifying new risk cases; the expert rule system mainly relies on the experience accumulation of business personnel of financial institutions, and the advantage is that it is an expert who has been iteratively verified [2]. The general effect of rules is good, but the migration of new business is poor, the accumulation of rules requires a long time period and high labor costs, and the feature dimension monitored by expert rules is limited and the generalization ability is weak; machine learning features Model is currently a popular risk control research field, and has become one of the main means of financial anti-fraud. It has the advantages of wide coverage of features, strong data processing capabilities, and relatively weak requirements for business capabilities.

2. The process of data mining in the data operation management of e-commerce retail platform

Regardless of the application, the data mining process in the data operation management of the e-commerce retail platform is generally divided into three stages: data preparation stage, data mining stage and data display stage as shown in Figure 1 (the picture is quoted from Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA).

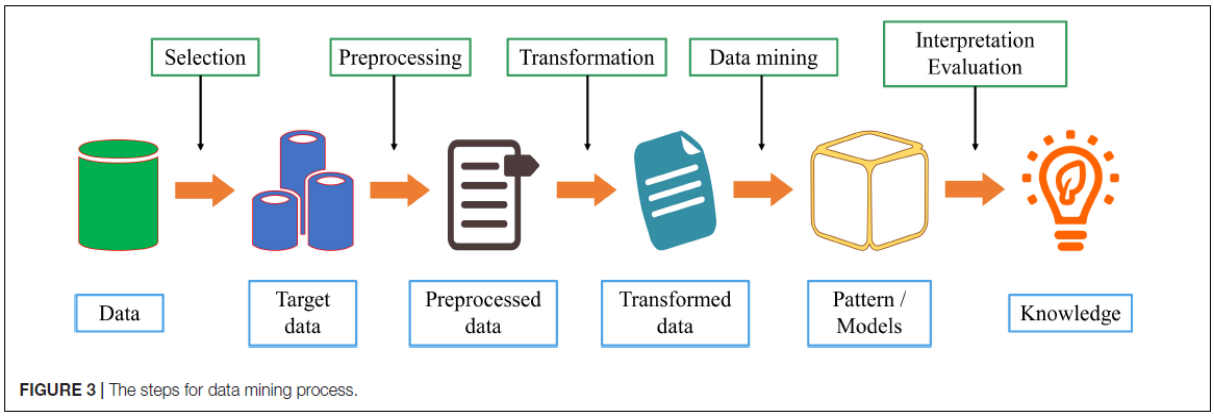

Figure 1. The process of data mining

2.1 Data Preparation

The data preparation stage mainly includes three steps: data selection, data cleaning, and data conversion. The choice of data is to collect relevant platform data according to the purpose of mining [3]. Data cleaning includes noise and missing value processing in the data, discretization of continuous attributes, etc. Data transformation is to convert the data format into a description form suitable for data mining.

2.2 Data Mining

According to the characteristics of the data and the analysis target, select the appropriate data mining algorithm and parameters and implement it with the tools. Commonly used data mining methods include association rules, clustering, classification and so on. Commonly used data mining tools are WEKA, SPSS, RapidMiner, etc.

2.3 Data display

Display the results of data mining to users with visualization techniques such as graphics and reports, and convert them into content that users can understand.

3. Merchant transaction data optimization algorithm design

3.1 Data description

The data set in this article comes from the real behavior data of users on the Tmall platform published by Ali. As one of the well-known e-commerce platforms in China, Tmall has 800 million active users and has accumulated a large amount of user information and behavior data [4]. Therefore, the selection of Tmall platform users as the research object is representative and has certain practical significance. The dataset contains user information and behavior data of 3,986 anonymous merchants and 260,864 anonymous buyers corresponding to these merchants. The rules and characteristics contained in this information have great research value. For example, users can be obtained through data visualization analysis. The basic characteristics of, deep mining can explore the user's purchasing rules and so on. There are three tables in the data set used in this paper: the information table containing the user's gender and age, the user's behavior data table and training data set generated on the Tmall e-commerce platform. The age or gender information of some users is missing in the user information table. The attributes of the user personal information table are shown in Table 1. In the User Age Range field, the value for users younger than 18 is 1; the value between 18 and 24 is 2; the

value between 25 and 29 is 3; the value between 30 and 34 is 4; and the value between 35 and 39 is A value of 5; a value of 6 between the ages of 40 and 49; and a value of 7 for the age of 50 and over.

Table 1. User Personal Information Form

Field	Describe
User_id	User's unique ID code
Age_range	User age range
gender	User gender, 0 is female, 1 is male

The user behavior information table includes various behaviors of users at the merchant, such as browsing, purchasing, and favorites, including data from May to November. The specific field descriptions are shown in Table 2. In the user behavior category field, 0 means click behavior; 1 means add to shopping cart; 2 means purchase; 3 means add to favorites [5]. This basic information can help you understand who the merchant's products are primarily intended for. The user behavior information table is the information table that this paper focuses on. Although there are only 7 features in the table, the features can be combined according to the relationship between each feature to construct more features.

Table 2. User behavior information table

Field	Describe
User_id	User's unique ID code
Item_id	Unique ID code for the item
Cat_id	The identifier of the category to which the product belongs
Merchant_id	Merchant's unique ID code
Brand_id	Product trademark
Time_stamp	Time (month, day)
Action_type	User behavior categories (including browsing, purchasing, favorites, etc.)

The original data set contains the training set table. The training data set table includes the ID code of the user, the ID code of the merchant, and the identification of whether the user has repeatedly purchased at the merchant. As shown in Table 3, there are few features in this table, and this paper will study it in to construct a new training set.

Table 3. Training set table

Field	Describe
User_id	User's unique ID code
Merchant_id	Merchant's unique ID code
label	Whether the user has a repeat purchase behavior identifier

There are different tables in the original data set that are not easy to mine features and train models. Therefore, according to the same ID attribute in the three tables, this paper merges the user personal information table, user behavior information table and training table. The combined table is shown in Table 4.

Table 4. Combined user behavior information table

Field	Describe
User_id	User's unique ID code
Merchant_id	Merchant's unique ID code
Age_range	User age range
Gender	User gender, 0 is female, 1 is male
Cat_id	The identifier of the category to which the product belongs
Brand_id	product trademark
Time_stamp	Time (month, day)
Action_type	User behavior category
Label	Whether the user has a repeat purchase behavior identifier

3.2 Data Processing Algorithms

Logistic regression, abbreviated as LR, is a type of linear model, which is simple, fast, and has strong generalization ability to new data. The function representation of logistic regression and linear model is different [6]. The function of linear model is understood as y=Xθ, while logistic regression is a transformation of the original function y=g(z), where z=Xθ. The function g here realizes the final mapping effect, for example, the final result is [0,1], which realizes the function of binary classification. The Sigmoid function is generally used to represent g, as shown in Equation 1.

$$g(z) = \frac{1}{1+e^{-z}} \tag{1}$$

The hypothesis function of logistic regression is shown in formula (2):

$$h_\theta(X) = g(X\theta) = \frac{1}{1+e^{-X\theta}} \tag{2}$$

Among them, X is the sample input, θ is the function parameter to be solved, and $h_\theta(X)$ is the model output, which in this paper indicates whether the user will repeat the purchase. The graph of the sigmoid function is shown in Figure 2.

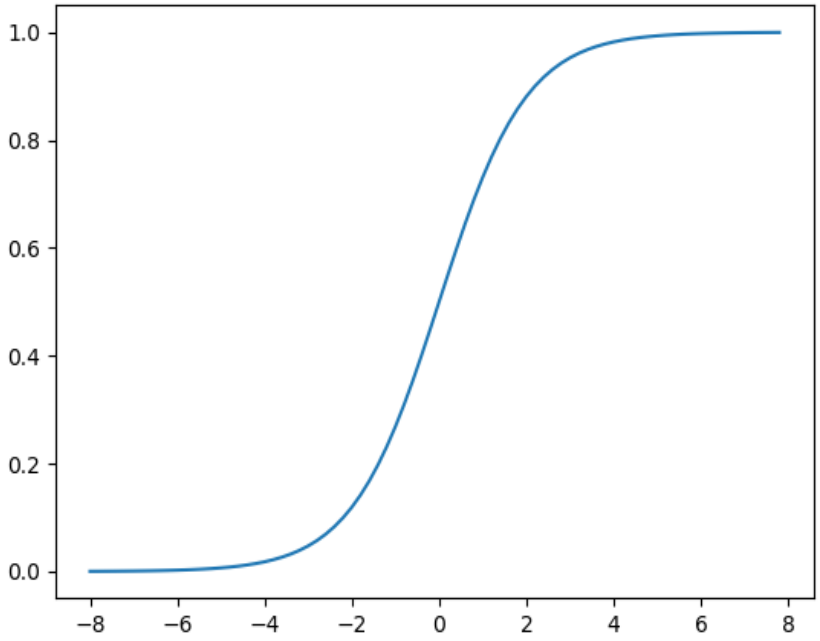

Figure 2. Sigmoid function graph

Logistic regression employs a log-likelihood loss function, as shown in Equation 3.

$$J\big(Y, P(Y|X)\big) = -\log(P(Y|X)) \tag{3}$$

The research on the repeated purchase behavior of users in this paper belongs to the binary classification problem. The final input of the sample is 1 or 0. The function expression of the correct probability of the sample prediction is shown in Equation 4, where the value range of y is 1 or 0:

$$P(y|x;\theta) = (h_\theta(x))^y(1-h_\theta(x))^{1-y} \tag{4}$$

Logistic regression uses the maximization of the log-likelihood function to solve for the model coefficients θ. The expression of the likelihood function is shown in Equation 5:

$$L(\theta) = p(y|x;\theta) = \prod_{i=1}^{n}(h_\theta(x^i))^{y^i}(1 - h_\theta(x^i)^{1-y^i} \quad (5)$$

n represents the number of samples, and the loss function of the model can be obtained by inverting the likelihood function. The loss function is shown in Equation 6:

$$J(\theta) = -\ln L(\theta) = -\sum_i^n (y^{(i)} \log(h_\theta(x^i)) + (1 - y^i) \log(1 - h_\theta(x^i))) \quad (6)$$

The gradient descent method is commonly used to solve the logistic regression loss function. It is a kind of iterative method. It is used to solve the least squares problem. By letting all the partial derivatives in the gradient drop to the lowest point, the optimal solution for each parameter is obtained [7]. The rate of gradient descent is called the learning rate, denoted by α. The ideal learning rate setting allows the function to slowly find the optimal solution without skipping the optimal point. The least square's function can be obtained for Equation 6, as shown in Equation 7:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(h_\theta(x^i) - y^i)^2 \quad (7)$$

The gradient θ_j in gradient descent can be obtained by taking the derivation of the least squares number, as shown in Equation 8:

$$\theta_j = \theta_j + \alpha\sum_{i=1}^{n}(y^i - h_\theta(x^i))x_j^i \quad (8)$$

In the previous research of this paper, logistic regression was used to predict the problem studied in this paper, and the prediction effect was not good, but logistic regression has the advantage of being simple and fast, so in this paper, logistic regression is used as a secondary learner to train and output the final prediction result. The complexity of the model can be reduced.

3.3 Data processing training model

The Relief algorithm was first proposed by Kira. It is an algorithm that assigns weights to features according to the relationship between each feature and two types of data, and filters out features with a weight greater than or equal to a certain threshold. At the same time, the Relief algorithm selects the adjacent samples participating in the weight calculation by calculating the distance between the samples, so the sample distance calculation is affected by the features, that is, the Relief algorithm combines the mutual influence and relationship between the features when screening adjacent samples [8]. The specific process is to randomly select a sample at a time, calculate the distance between the value of the sample in the feature to be tested and the adjacent similar value, and the distance between the sample and the adjacent heterogeneous value, and compare the size of the distance between the two. If the distance of the adjacent same class is smaller than the distance of the adjacent heterogeneous, it means that the feature distinguishing ability is good, otherwise the feature distinguishing ability is poor. By randomly selecting a certain number of samples for multiple evaluations, the distinguishing effect of the features is finally determined.

Table 5. Relief algorithm

Algorithm specific process
1, Set the weights of all features to 0
2, $for\ \ w_j = 1\ to\ n\ do$
3. Randomly select a sample
4. Find the nearest neighbor sample H of R from the same samples, and find the nearest neighbor sample N from different samples
5, $for\ A = 1\ to\ M\ do$
6, $w(A) = w(a) - diff(A,R,H)/N + diff(A,R,N)/N$
7, $for\ A = 1\ to\ M\ do$
8, $if\ w(A) \geq \delta$
9, Then A is the filtered feature
10, end

The Relief algorithm does not consider the problem of data imbalance when selecting features. However, in the original user behavior data, users with repeated purchase behavior only account for a small part, so it is easier to use the Relief algorithm for a class of samples that do not have repeated purchase behavior. selected to generate an offset. For merchants, they will pay more attention to users who have repeated purchase behavior. These customers are potential customers for future consumption, which is conducive to the realization of precise marketing by merchants. Considering that the traditional oversampling method will cause repeated learning of the minority class samples and amplify the noise effect of the minority class samples, and the traditional under sampling method may delete some important samples and lose information, so this paper starts from the point of feature selection. The improved Relief algorithm is used to solve the problem of data imbalance, that is, more attention is paid to the features with strong discriminating ability for a few samples.

The improvement of the Relief algorithm in this paper is mainly to improve the selection of samples. The original random selection of a sample is changed to randomly select M samples from a few samples to calculate the feature weight, and randomly select M samples from the majority of a class of samples. The sample calculates the feature weight, and finally calculates the mean of the two feature weights to represent the importance of the feature, so that the data of the two types of samples is balanced when selecting samples, and the influence on the importance of the feature is balanced. The specific algorithm process after improvement is shown in Table 6.

Table 6. Improved Relief Algorithm

The specific process of the improved algorithm
1. Set the weights of all features to 0, and set $Flag = True$
2, $for\ w_j = 1\ to\ n\ do$
3, $if\ Flag == True$
4. Randomly select a sample from the minority class samples $Flag == False$
5, $else$
6, Randomly select a sample from the majority class sample $Flag == True$
7. Find the nearest neighbor sample H of R from the same samples, and find the nearest neighbor sample N from different samples
8, $for\ A = 1\ to\ M\ do$
9, $w(A) = w(a) - diff(A,R,H)/N + diff(A,R,N)/N$
10, $for\ A = 1\ to\ M\ do$
11, $if\ w(A) \geq \delta$
12. Then A is the filtered feature
13, end

This paper uses the improved Relief algorithm to rank the importance of each feature. The results show that 18 features in the input 121-dimensional input features have an importance level of 0, and

these features have a low degree of discrimination between the two types of samples. Many of the predicted features obtained after feature extraction are useless. If these feature samples are used without any processing, the prediction effect of the model will decrease and the computational complexity of the model will be increased. Therefore, some unimportant features can be removed by feature selection. Therefore, a total of 103 features are selected in this paper, including the number of users' daily purchases, the purchase conversion rate, and the merchant's repurchase rate.

First, in the training process, the dataset X_i^m is randomly divided into five sets $D_m(m = 1,2,3,4,5)$ with the same number and mutually exclusive, so that $\overline{D_m}$ and D_m respectively represent the mth folded training set and test set, where $\overline{D_m} + D_m = D$; after five iterations, in each iteration, the base learner is the i algorithm S_i in the training set $\overline{D_m}$, each trained base learner X_i^m predicts each sample x_j in the test set D_m, and the representation method is $X_i^m(x_j)$. When the above steps are completed for the four algorithms, this paper obtains the training results of the four algorithms for all samples in D_m; when five iterations are completed, this paper obtains the training results of the four models for all samples, denoted as D'. Then use D' to train the logistic regression model to get the secondary learner $X' = S(D')$, and the secondary learner outputs the final prediction result.

Next is the testing process. Since in the above training process, each algorithm S_i can generate five learners $X_i^m(m = 1,2,3,4,5)$ in the five-fold training dataset, so in the testing process each algorithm The generated five base learners will generate five prediction results $X_i^m(x)(m = 1,2,3,4,5)$ on the test set sample x, and then take the average of these five results $\overline{X_\iota}(x)$ as the prediction result of algorithm S_i on sample x; input the sample data into four algorithms in turn to get $(\overline{X_1}(x), \overline{X_2}(x), \overline{X_3}(x), \overline{X_4}(x))$ four test results; finally, input the prediction result obtained by the basic learner into the trained secondary learner X' to obtain the final prediction result. The specific algorithm process of the Stacking fusion model is shown in Table 7.

Table 7. Algorithm design of Stacking fusion model

Algorithm specific process
Step one:
1. Divide the original data set into five sets of similar numbers and mutually exclusive$D_m(m = 1,2,3,4,5)$
2, $for\ m = 1\ to\ 5\ do$
3, $for\ i = 1\ to\ 4\ do$
4, $X_i^m = S_i(D \backslash D_m)$
5, $for\ x \in D_m$
5, Calculate$X_i^m(x)$
10, end
Step 2:
1 The prediction result of machine learning in the five-fold intersection is $(X_1^m(x), X_2^m(x), X_3^m(x), X_4^m(x)$
2, Calculate$\overline{X_\iota}(x) = (\sum_{m=1}^{5} X_i^m(x))/5$, That is, the prediction result of each base learner in the sample x, and the data set D' is obtained
3, Use the dataset D' to train the secondary learner to get $X' = S(D')$
4, Use the secondary learner to output the final prediction result

4. Data Results

The prediction effect of the Stacking fusion model constructed in this paper is better than that of each single prediction model. Due to the large difference in the number of two types of users in the original data set, there is a data imbalance problem, which affects the prediction effect of the model. Therefore, this paper uses an improved under sampling method to deal with the data imbalance

problem and build a prediction model again. In addition, this paper adds a time sliding window to dynamically update the data set based on the balanced data set. The experimental results are shown in Table 8.

Table 8. Experimental results table

Model	F_1 value	AUC value
Stacking fusion model	0.7931	0.7804
Stacking fusion model after sampling	0.8317	0.8502
Fusion model with sliding window	0.8631	0.8763

According to the experimental results, the F_1 value and AUC value of the improved Stacking fusion model are both higher than those of the model before the improvement, in which the F_1 value is increased by 5% and the AUC is increased by 9%. It can be seen that although the improved Relief algorithm is used in the feature selection process to pay more attention to the features with strong discriminating ability for a few samples, it does not completely solve the problem of data imbalance. After using the improved sampling method to obtain a balanced dataset, the predictive ability of the fusion model is further improved. Finally, this paper adds a sliding window to dynamically update the samples according to the set period, which further improves the prediction ability of the model.

5. Conclusion

This paper firstly uses the Stacking fusion model to learn the transaction features to obtain higher-level abstract features; then the multi-layer perceptron is initialized, and the multi-layer perceptron is used to perform the classification task, so as to realize the identification of false commodity transactions. According to the sales of the product, the review record and the basic information of the store, the characteristics of the product are used and quantified. In order to verify the feasibility of the method, the product information is collected from Taobao as a training and test set, and the marked product data is trained and learned. Compared with the traditional recognition method, the performance of this method is significantly improved. For the massive number of falsely traded commodities in Taobao, the experimental data in this paper is relatively small, and it is still necessary to crawl a relatively large amount of data to further verify the method in the future.

References

[1] Dong Xiaoyan. Yunshang Wuhan International Fashion Center organized merchants to participate in CHIC and held a matchmaking meeting. Textile and Apparel Weekly, Vol. 1 (2021) No.3, p. 19 - 22.

[2] Liu Zhitao, Xiao Dan. Problems and Countermeasures of Taobao Mall Merchant Management in the Era of Big Data. Journal of the School of Electronic Engineering, Vol. 9 (2020) No.2, p. 555.

[3] Sun Quan, Tang Tao, Zheng Jianbin, et al. Intelligent Fraud Detection Based on Graph Network Driven by Financial Transaction Data. Chinese Journal of Applied Science, Vol. 38 (2020) No.5, p. 120.

[4] Shi Purun, Cao Jiaying, Jia Jun. Pricing mechanism and anti-monopoly implications of monopoly online platforms based on consumer data value. Practice and Understanding of Mathematics, Vol. 51 (2021) No.22, p. 9 - 14.

[5] Meng Xuran, Bi Xiuchun, Zhang Shuguang. Research on High Frequency Algorithms and Back testing Based on Generative Adversarial Networks. Journal of University of Science and Technology of China, Vol. 50 (2020) No.6, p. 10 - 13.

[6] Zhao Bingzhen, Chen Zhiyu, Yan Longchuan, et al. Data Privacy Protection of Power Business Transactions Based on Blockchain Architecture. Automation of Electric Power Systems, Vol. 45 (2021) No.17, p. 75 - 88.

[7] Zhu Fengxia. Transaction Database Encryption Technology Based on Blockchain Technology. Electronic Design Engineering, Vol. 28 (2020) No.3, p. 93 - 97.

[8] Cheng Lisha, Wang Shijun, Tian Junfeng, et al. Research on urban network in Northeast China based on transaction data of government procurement activities. Geographical Science, Vol. 41 (2021) No.8, p. 119.