Personalized Recommendation Method based on User Behavior Analysis

Yu Wang, Jin Shang, Xiaofang Wu, Maofu Liu

College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, 430065, China

email: 674637291@qq.com

Keywords: User Behavior; Personalized Recommendation; Two Classification; Artificial Rules

Abstract. The characteristics of user's behavior in the real scene are analyzed, and a personalized recommendation method based on user behavior analysis is put forward. In the electronic commerce user behavior can be divided into clicking, purchasing, collecting, plussing shopping cart, etc. The current mainstream algorithm collaborative filtering algorithm can not deal with other acts in addition to the purchase behavior. Take the method based on the artificial rule and the improved hierarchical fusion model based on bagging, converting the problem to a two classification problem for predicting whether or not to buy and recommending to users. Experimental results show that the proposed method makes full use of the user's behavior information, avoiding the limitations of the traditional methods, so that the recommended effect is significantly improved.

Introduction

The electronic commerce recommendation has been a hot issue in the field of data mining, the most widely used and more mature method is the collaborative filtering algorithm［1-4］. However, on the recommendation of e-commerce, collaborative filtering algorithm still has some shortcomings in many details. (1) When the user purchased a number of items, how to determine the weights of these items in the calculation, is take the "1" appropriate, whether bring different effects on the recommended results or not. (2) Similarity matrix symmetry is not suitable, such as the user who bought a mobile phone is recommended for the mobile phone shell, it is an ideal recommendation results, but when the user bought the phone shell, recommend it for the phone is unreasonable. (3) When calculating the similarity matrix, how to deal users' click behavior and collect behavior? Only using purchase records can cause huge waste of information. (4) When the matrix is extremely sparse, the effect of the algorithm is low.

At home and abroad, a series of recommendation algorithm contest, collaborative filtering algorithm and a series of improved algorithms have not entered the final, the recommended effect is poor. It also shows that collaborative filtering algorithm is not suitable for e-commerce scenarios.

On the above consideration, we find that the most important index of measuring recommendation algorithm in e-commerce is whether users purchase after recommended or not, whether let users because of the recommendation system find what they really want to buy. We will try to convert the recommendation problem to predict what items the user will buy, and convert complex problems into two classification models, so that a variety of behavioral information of users can be made full use of. Multiple items behavior information of users can get good treatment, and avoid the problem of asymmetric similarity matrix. Considering that there are many purchase of users have a certain law, but the algorithm can not predict and measure, the two level model fusion [5] make data samples more real, more accurate. Therefore, this paper proposes an improved Bagging two level fusion model [6], and joined out prediction according to certain rules of the results, union the two sets, to improve the recall rate of prediction results and accuracy, and ensure the performance of the recommendation system.

Personalized recommendation process based on rule and level fusion model

Recommendation based on rule and level fusion model.

Extract feature of data sets, screened by artificial rules, add User Goods Pair to the predictset1 which meet the rules, use data does not conform to the rules as the input of the model training algorithm, logistic regression model, then use trainset as training set, use logistic regression model to obtain the prediction results, the prediction results are added to trainset, which is as the second level of the input of the model.

Use trainset to train single model, KNN model, random forest model, logistic regression model, GBDT model etc., take top 3 to form the second level model. Finally, use the Bagging algorithm to vote to get the predictset2 set.

Union the predictset1 set and the predictset2 set, form the final set of User Goods Pair. According To recommend according to this collection for the user.

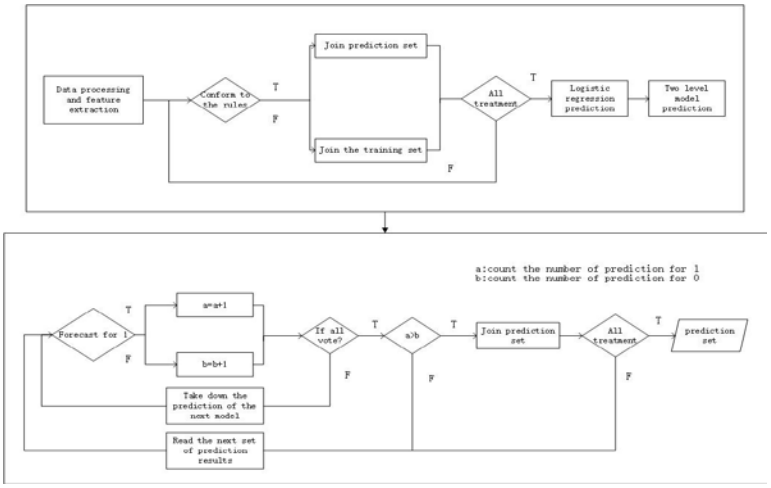

Fig .a Personalized recommendation process

Data preprocessing and feature extraction.

In order to avoid the user clicks on or buys an item too many times and bring deviation to the model, so normalize the data. Through the comparison of the experimental data, select the better user and the goods cross characteristics.

Chart a Feature extraction description

Extracted cross features	Feature description
The times of clicking	Count the times of clicking on an item
The times of purchasing	Count the times of buying on an item
The times of collecting	Count the times of collecting on an item
The times of joining the shopping cart	Count the times of joining the shopping cart on an item
Interaction between users and items	Calculate the times of user behavior on item a/ total times of user behavior, reflecting user preferences for this item
The user of the purchase conversion rate	the times of buying behavior on item a/the times of clicking behavior on item a，reflecting the user's desire to purchase this item
Percentage of user purchases of items	the times of buying behavior on item a/ total times of buying behavior，reflecting whether the item is often needed for the user

Improved algorithm strategy.

Integrated learning is not a perfect method for combination based classification model [7], and most of the research is to improve the base classifier selection, this paper adopts the two level

fusion model and the first stage used logistic regression model, use Sigmoid function

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

to classify. The training set train is predicted, and the prediction result pre1 is added to the training set to form the training set train1, which is used as the training set of the second level model, use this new training set as the training set of the second level model. The estimated regression coefficients are shown in the table below, the results can be seen that the results have better interpretability in the second level training. The probability of the final prediction results of the users who was predicted purchase increased by 6 times, making the data closer to the real situation of the model.

Chart b regression coefficient

The times of clicking	The times of purchase	……	pre1
1.0496465	1.0226648		6.408171

According to the previous theorem of generalization error, for any two classification problem, the probability of any function is δ at least, making

$$R(f) \leq \hat{R}(f) + \mathcal{E}(d, N, \delta) \tag{2}$$

and among of it $\mathcal{E}(d, N, \delta) = \sqrt{\frac{1}{2N}(\log d + \log\frac{1}{\delta})}$ (3)

It can be seen that the smaller the training error, the smaller the generalization error, the better the effect of the recommendation system。The empirical risk minimization principle shows that the empirical risk minimization model is the optimal model［8］。Suppose the classification function is

$$f: R^n \rightarrow \{c_1, c_2\} \tag{4}$$

the probability of error classification is

$$P(Y \neq f(X) = 1 - P(Y = f(X)) \tag{5}$$

Misclassification rate of prediction results for multiple models

$$\frac{1}{K}\sum_{x_i \epsilon N_k(x)} I(y_i \neq c_j) = 1 - \frac{1}{K}\sum_{x_i \epsilon N_k(x)} I(y_i = c_j) \tag{6}$$

Voting is equivalent to empirical risk minimization. When the amount of data increases, the effect of empirical risk minimization can be gradually improved. So the second level uses the Bagging method, avoiding a single weak prediction effect for learning, reducing the training error by Bagging.

Select KNN, random forest, logistic regression model to sample randomly, train the model1, model2, model3 separately. Three models are fused to get the final fusion model. By using the two level model, the data can be more realistic and reliable through the first level model.

Algorithm block diagram

Input: User Behaviour Data Set
Output: Prediction set
Step 1: Extract feature
Normalize processing data
Step 2: Logistic algorithm results are added to the training set

$$P(Y = 1|x) = \frac{\exp(w*x)}{1+\exp(w*x)}$$

w*x is linear function for x classification
Step 3: for i in n
Random sampling
Step 4: Using the weak learning algorithm A (i) to get Weak classifier H(i)
for i in all
for j in n
H(j),i=1,2…n， Vote
Step 5: The final model is

$$W(x_i) = \sum_{i=1}^{n} H_i(x_i)$$

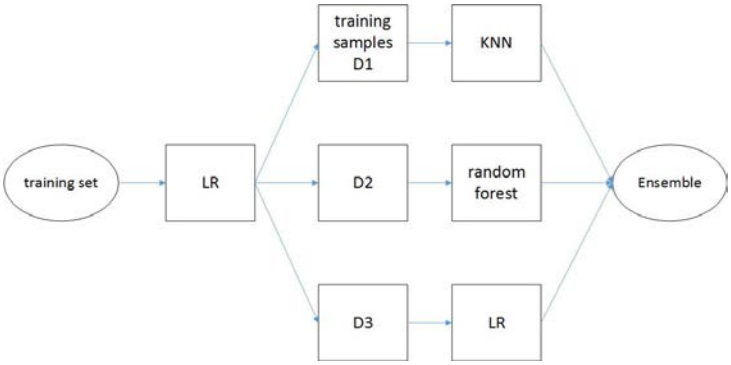

Fig.b Two level fusion model

The use of artificial rules.

Through a large number of experiments and comparative studies, we find that although we convert the problem into a two classification problem, there are still a large number of items in the predicted user non purchase list， but the user eventually bought, which is the reason for weak prediction. Therefore, according to the artificial rules, the rules of the data sets are selected and the user items are predicted according to the rules，and union the set predicted by voting by the Bagging method . Namely join the set predicted according to artificial prediction rules.

In addition, by using the knowledge of conditional probability, the range of artificial prediction data set is reduced. Through the experimental verification, it can be seen that the closer to the forecast day, the more valuable the user behavior. As the figure, the horizontal axis represents the user from the prediction of the length of time, the vertical axis represents the probability of purchase.

The manual rules used in this paper are:

if(The user u put the B items into the shopping cart before the prediction day) then recommend B to u

if(The user u in the previous period of time usually buys goods B, the number of times more than 10 times) then recommend B to u

if(The user u almost every day in the first two weeks to click on goods B before the prediction

day) then recommend B to u

if(The last item to visit of the user before the prediction day) then recommend B to u

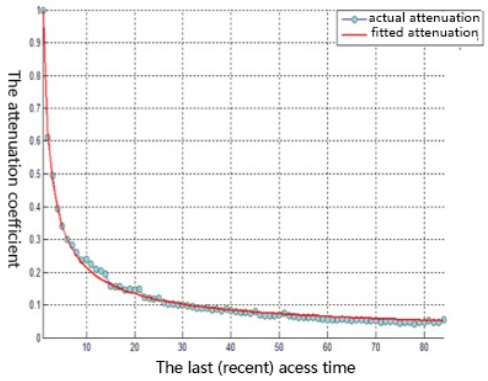

Fig.c Conditional probability curve

The blue point refer to real point, the red line refer to simulating line.

Chart c data set description

field	Field description
user_id	User identification
brand_id	Commodity brand identity
time	User behavior time
behavior_type	User behavior type

Experimental verification

Selection of data sets.

The experimental data with large data platform provided by the Alibaba Tianchi Taobao 4 months (April 15th August 15) the user behavior to verify the recommended methods in this paper, including user ID, item ID, behavior, behavior and other types of information, such as the field table. The data set includes 181882 items of user behavior information, after a variety of behavior statistics, a total of 78873 records. This paper selects the data before June 15th as the training set, a total of 30402 records, from June 16th to July 31st data as a test set, to verify the prediction in August 1st to buy users in August 15th, a total of 316 real purchase records.

Verification method.

For the evaluation of the personalized recommendation system, we can not simply look at the correct rate of prediction, this paper uses the international customary way, using the classical accuracy , recall and F1 as the evaluation index. Specific calculation formula is as follows, the PredictionSet is the prediction of the purchase , ReferenceSet is for the true answer of buying data sets. We use F1 as the sole criterion for evaluation.

$$\text{Precision} = \frac{|\cap(\text{PredictionSet},\text{ReferenceSet})|}{|\text{PredictionSet}|} \qquad \text{Recall} = \frac{|\cap(\text{PredictionSet},\text{ReferenceSet})|}{|\text{ReferenceSet}|}$$

$$\text{F1} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

Evaluation formula

Comparison and analysis of experimental results.

1) The recommendation effect based on the Collaborative filtering algorithm

In this recommendation, the similarity of user behavior is used to calculate the similarity of interest, for example, the user a and user B, with N (a) means that the user has purchased the collection of a, N (b) means that the user B purchased goods collection, using cosine similarity formula to calculate the similarity of interest. In order to avoid the number of users over two users of the similarity calculation takes too long so take the investigations to establish table, for each item had a list of user purchase behavior, first calculate the User Goods Pair {a, b} whose behavior intersection is not 0 . Ultimately through the adjustment of K (the most interested in a commodity

brand K users), m (recommended to the user before the M commodity brand) value, the final results have been recommended.

Chart d results of collaborative filtering algorithm

k	m	Predicted number	Correct number	accuracy	recall	F1
4	4	2464	7	0.284%	2.215%	0.503%
5	5	3180	10	0.314%	3.164%	0.572%
7	7	4611	13	0.281%	4.113%	0.527%
7	14	8669	21	0.242%	6.645%	0.467%

$$W_{ab} = \frac{|N(a) \cap N(b)|}{\sqrt{|N(a)| * |N(b)|}}$$

Formula b cosine similarity formula

2) Fusion model recommendation effect based on Adaboost algorithm

Some domestic scholars have proposed a fusion model based on Adaboost [9], which can predict the user's purchase by clicking rate prediction[10-11]. Use of the experimental data set to verify the recommendation effect, the selection of the basic model uses the same model and different models of the two strategies. Different positive and negative samples in logistic regression model, different random spanning tree in the forest tree. The fusion of three models: logistic, random forest, GBDT. The results are as follows:

Chart e recommendation effect of fusion model based on Adaboost algorithm

Predicted number	Correct number	accuracy	recall	F1
325	22	8.615%	8.860%	8.736%

3) The recommendation effect based on the rule and level fusion model

On the test set, the user item pairs are selected by artificial rules, combined with the results of the two level fusion model. The logistic model with positive and negative samples than 1:11, KNN nearest neighbor model number 8, random forest model generation tree is 30, results are as follows.

Chart f recommendation results based on rules and hierarchy models

Predicted number	Correct number	accuracy	recall	F1
1280	100	7.812%	31.646%	12.531%

4) Result analysis

From the above results, we can see that the collaborative filtering algorithm is not suitable for such tasks, in any case to adjust the K, m value impact little on the results, the overall effect is poor; Other scholars put forward the fusion algorithm based on Adaboosting model method has greatly improved the effect, but the weight of each single model in the boosting algorithm in the real-time change of each prediction function can only be sequentially generated, the computation time cost greatly; the proposed hierarchical model based on the rules of fusion, due to the improvement of Bagging algorithm based on the prediction of the function parallel generation time, less costly, and due to the addition of artificial rules, effect is significantly improved.

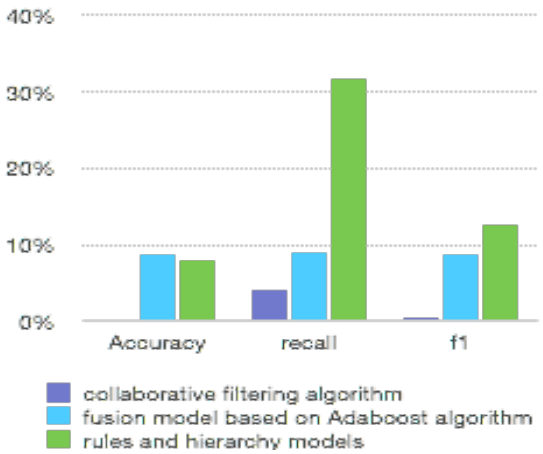

Fig. d Comparison of recommended results

Conclusion

In this paper, based on the user's behavior information, the advantages and disadvantages of the traditional recommendation algorithm are fully considered. On the basis of previous studies, the recommendation problem is converted into two classification problems. The main contributions of this paper are：1、use a variety of user behavior information to model recommendation, rather than just use the purchase behavior, make full use of the data.2、the problem is converted into two classification problem, based on the Bagging algorithm, the method of the fusion of the two level model is proposed, which solves the problem of low recommendation effect of a single model, and has a certain improvement compared with the previous results.3、This paper puts forward the use of certain artificial rules to predict, predicts some of the algorithm can not explain the purchase situation, so that the results of the recall rate is greatly improved, the overall effect is more significant. The experimental results show that the artificial rule recommendation results and recommendation method combining improved Bagging fusion model based on, effectively improve the effectiveness of the recommendation, and solve the collaborative filtering algorithm is not suitable for the electronic commerce recommendation problem.

Acknowledgement

In this paper, the research was sponsored by the innovation fund of Wuhan University of Science and Technology(Project No. 16Z2C079) .

Reference

[1] JIANGUO LIU, BINGHONG WANG, QIANG GUO. IMPROVED COLLABORATIVE FILTERING

[2] Lee H C, Lee S J, Chung Y J. A Study on the Improved Collaborative Filtering Algorithm for Recommender System[C]// Acis International Conference on Software Engineering Research, Management & Applications. IEEE Computer Society, 2007:297-304.

[3] YANHONG GUO. Research on collaborative filtering algorithm and application of recommendation system[D]. Dalian University of Technology, 2008.

[4] ZHONGJUN LI, QIHAI ZHOU, QINGHONG SHUAI. A recommendation system model based on the integration of content and collaborative filtering [J]. computer science, 2009, 36(12):142-145.

[5] Maclin R, Opitz D. Popular Ensemble Methods: An Empirical Study[J]. Journal of Artificial Intelligence Research, 2011, 11:169-198.

[6] ZHONGLIANG FU. Research on the effectiveness and optimal combination of linear combination of classifiers [J]. Computer research and development, 2009, 46(7):1206-1216.

[7] GUOQIANG LIU. Research and application of ensemble learning algorithm based on combined sampling [D]. Ocean University of China, 2011.

[8] Koltchinskii V. Oracle inequalities in empirical risk minimization and sparse recovery problems :[M]. Springer Verlag, 2011.

[9]YING CAO , QIGUANG MIAO, JIAZHEN LIU. Research progress and Prospect of AdaBoost algorithm [J]. Journal of automation, 2013, 39(6):745-758.

[10] LING YAN. Research on the prediction of click through rate in display advertising [D]. Shanghai Jiao Tong University, 2015.

[11] ZHUO WANG. User brand purchase forecast [D]. Shanghai Jiao Tong University, 2015.