Comparison with Recommendation Algorithm Based on Random Forest Model

Yu Jiang[1,2], Lili He[2(✉)], Yan Gao[2], Kai Wang[2], and Chengquan Hu[2]

[1] Key Laboratory of Information System Security of Ministry
of Education, TNLIST, School of Software,
Tsinghua University, Beijing 100084, China
[2] College of Computer Science and Technology,
Jilin University, Changchun 130012, China
helili@jlu.edu.cn

Abstract. Product recommendation based on user behavior is a hot research topic In the Internet era in the same data set, the features that the results of the various classifications are a greater difference were handled with random forest model. This paper compares the mainstream classification algorithm C4.5 and CART and analyzes 578,906,480 user behavior records on the results of actual transaction in Alibaba. The results show that CART decision tree algorithm is more suitable for large e-commerce data mining.

Keywords: User behavior · Random forest model · Decision tree · C4.5 · CART

1 Introduction

User implicit demand excavated from the mass of information on user behaviors is essential for service providers. Currently, the recommended system [1] has been preliminarily applied in business, but how to construct a highly efficient and intelligent recommendation algorithm is still a hot topic. Random Forests model that a classification prediction model [2] is proposed by Leo Breiman, it has many advantages, such as learning faster, less parameters and fault tolerance, since it was proposed in many fields received applications. Guo Yingjie et al. used random forest classification to identifies plant resistance gene [3]; Li Jiangeng et al. analyze gene pathways of cancer microarray data based on random forest [4] and Fang Kuangnan predicts fund yields direction used random forests model [5].

In this paper, the dataset is massive amounts of user behavior in the Alibaba website real deal. We defined user behavior attribute set and compared with classification algorithm C4.5 and CART based on random forest model to provide evidence for better user recommendation.

J.J. (Jong Hyuk) Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 421, DOI 10.1007/978-981-10-3023-9_72

2 Basic Theory

2.1 Random Forests Model

Random Forests is classifier made more decision independent trees [6, 7]. The generation of decision tree is generally controlled by the property division and pruning, but when a large number of features, it may be over-fitting problems. Random forests use boosting [8, 9] resampling method to extract plurality of samples from the original data set, and to construct the decision tree for each sample, through the plural the of decision tree, it can forecast the final prediction results (Fig. 1).

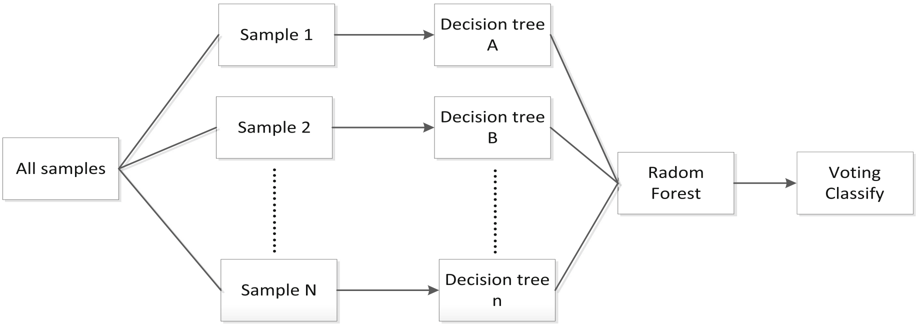

Fig. 1. Random forests model

2.2 C4.5 Algorithm

C4.5 algorithm [10] starting from the root node assigned the best properties. The value of each attribute will generate the corresponding branch, and generate new nodes on each branch. Best attribute selection criteria is based on the definition of information entropy gain ratio to select test properties of the node, entropy characterizes the purity of any sample set. There are four steps to establish C4.5:

(1) Handling the data source to convert the continuous data into discrete;
(2) Calculating its information gain and information gain ratio for each attribute;
(3) The possibility value of each attribute corresponds to a subset, it is start from the root node; Second step is performed recursively until each subset of data gets the same value attributes and generates decision trees.
(4) Extraction of classification rules based on the decision trees can classify new data set.

2.3 CART Algorithm

Classification and Regression Tree (CART) algorithm [11, 12] is a very effective non-parametric classification and regression algorithm, it achieves the purpose of prediction by constructing a binary tree. Binary Tree is not easy to generate data fragments and its accuracy will be higher than often multi-tree, so we choose binary tree in the CART algorithm. CART with Gini index as the division standard. CART is established by the following three steps:

(1) Creating binary trees used data sets, then disrupting each attribute node until all samples of leaf nodes are classified into the same category or disrupted attribute sets are empty;
(2) Pruning, pruning algorithm is continuously used to get smaller trees and form an ordered sequence of sub-tree;
(3) Selecting the final result, the final decision tree is chosen the best sub-tree from the subset of sub-tree sequence according to the validation.

3 Comparative Analysis

3.1 Experimental Design

Experimental dataset is 578,906,480 recorded data provided by large data competition of Alibaba, The data spans a period of 4 months. the data of each record is by user_id, brand_id, type, visit_time four fields, which user_id uniquely identifies the user; brand_id uniquely identifies brand; type is user's behavior, for example 0 indicate clicks, 1 indicate purchase, 2 indicate collections, 3 indicate adding to Shopping cart; visit_time is constituted by month and day. Finally, the form of forecast results is user_id, brand_id1, brand_id2….. and comparing with the actual result of the purchase. The assessment indicators are as follows:

$$\text{precision} = \frac{\sum_i^N hitBrands_i}{\sum_i^N pBrands_i} \tag{1}$$

N is the number of users predicted, pBrandsi is the number of the predicted brand list for the user i, hitBrandsi is the number of intersection between the predicted brand list and really bought brands list for the user i.

$$\text{Recall} = \frac{\sum_i^M hitBrands_i}{\sum_i^M bBrands_i} \tag{2}$$

M is the number of users actually generated transactions, bBrandsi is the number of really bought brands for the user i, hitBrandsi is the number of intersection between the predicted brand list and really bought brands list for the user i.

Finally, F1-Score is used to fit the precision and recall rate.

$$\text{F1} = \frac{2 * \text{P} * \text{R}}{\text{P} + \text{R}} \tag{3}$$

3.1.1 Attribute Selection

This paper constructs 50 property values based on user behavior and date, such as interaction attributes, user attributes, brand attributes and complex attributes. Selection of the property's value play a very important role for the classification recommended of the mass user behavior, good properties can get a better classification results. In the experiment, we divided data set into two parts, the first part is the data set of the first three months as a training set, the other part used as a prediction set, each of which is nearly 90 days of data.

Interaction attribute: interaction attribute that is summarized based on user behavior attributes. As it is a set of training data to predict the final month of the user's purchasing behavior, the closer to the right of the last day of the user's behavior, the more significant. We also make the number of the user clicks, purchase, collection and add to cart behavior with respect to time decay. The coefficient of attenuation is 1 / ((days-1) / 30 + 1), where days is the number of distance from the last day.

Brand attributes: it mainly generate based on the number of this brand's user clicks, purchase, collection and add to the cart.

User attribute: it mainly generate based on the user's own clicks, purchase, collection and add to the cart number.

Composite attribute: it mainly composite interaction attribute, user attributes or brand attributes together (Table 1).

Table 1. The classification of attributes

Attributes	Type	Description
The hits of last 1, 3, 6 days	Interaction attribute	It is mainly based on user clicks and purchases during a period. The selection method is similar with dichotomy period.
The hits of last 7–15, 16–30, 31–60, 61–90 days		
The purchases of last 6, 7–15, 15–30, 31–60, 61–90 days		
The hits of this brand in the last 15, 16–30, 31–60, 60–90 days		It is mainly get hits, the more frequently click this brand, the more interest.
The total numbers of purchase this brand in last 90 days		It is mainly makes statistical sampling based on the number of collection and adding to cart in the last a month. The more number, the more interest.
The number of adding to cart in the last 3, 7, 7–15 days		
The number of collection in the last 7, 15, 30 days		
The days of click this brand in the last 30 days * the days of click this brand in the last 31-60 days		It is mainly used to determine whether the user continued attention or purchase to the brand.
The hits of last 6 days * the hits of last 7–15 days		
(The hits of last 6 days + the hits of last 7–15 days) *(the hits of last 16–30 days)		
(The hits of last 6 days + the hits of last 7–15 days + the hits of last 16–30 days) * sqrt (sqrt(the hits of last 31–60 days))		

(*continued*)

Table 1. (*continued*)

Attributes	Type	Description
(The hits of last 6 days + the hits of last 7–15 days + sqrt(the hits of last 16–30 days)) * (the hits of last 31–60 days) * (the hits of 61–90 days)		
The purchases of last 6 days * the purchases of last 7–15 days		
(The purchases of last 6 days + the purchases of last 7–15 days) * sqrt(the purchases of last 16–30 days))		
(The purchases of last 6 days + the purchases of last 7–15 days + sqrt(the purchases of last 16–30 days)) * sqrt(sqrt(the purchases of last 31–60 days))		
(The purchases of last 6 days + the purchases of last 7–15 days + sqrt(the purchases of last 16–30 days)) * (the purchases of last 31–60 days) * (the purchases of last 61–90 days)		
(The hits of last 6 days + the hits of last 7–15 days) * (the days of click in the last 15 days −1)		
The hits of this brand in the last 1, 3, 7, 15 days/the hits of all		It mainly is a percentage of between attention in the last a pried and total attention. The higher the percentage, the more attention.
The number of knowing and purchase this brand/the total number of knowing this brand (TaoBao conversion rate)	Brand attribute	It is mainly represent the popularity of this brand, smoothly pop or rapidly popular brand should be recommended.
The re-purchase rate of this brand		
The tendency of brand hot (according to the number of purchases)		
Average number of purchase this brand every month		
The purchases of this brand in the last 7 days		
The on-line days of last 10, 20 days	User attribute	It is mainly represent activity status in the near future. The more frequently, the more likely re-purchase.
The days of purchase in the 90 days (frequency)		
The numbers of purchase brand/the numbers of knowing brand		
The purchases of last 3, 7, 15 days/the total purchases		
The days of purchase this brand in the last 30 days * the re-purchases rate of this brand	Composite attribute	It is mainly represented whether users will re-purchase this brand in a month, if users re-purchase this brand, it is likely to purchase this brand more time.

3.1.2 Parameter Configuration

Due to continuous property values, so we can use C4.5 and CART algorithms, in the parameters configuration, the other parameters are the same except the decision tree algorithm. Here, the number of random forest trees is 1000 (range from 10 to 1000), the number of each step algorithm divided attributes is log (N), the maximum number of records per tree is 1,000,000 (range from 1000 to 1000000).

3.2 Training Results

3.2.1 Confusion Matrix

See Table 2.

Table 2. Confusion matrix

	Random forest model based on C4.5	Random forest model based on CART
The Negative examples of correct prediction (TN)	16,444,969	16,449,164
The number of negative examples mistaken positive (FP)	159,915	155,720
The number of positive examples mistaken negative (FN)	1,186,845	1,090,448
The positive examples of correct prediction (TP)	393,039	489,436
The number of actual negative examples	16,604,884	16,604,884
The number of actual positive examples	1,579,884	1,579,884
The number of predicted negative examples	17,631,814	17,539,612
The number of predicted positive examples	552,954	645,156
Total number	18,184,768	18,184,768

3.2.2 ROC Curve

In the Fig. 2, the left is ROC curve of random forest model based on C4.5 decision tree algorithm, the right is ROC curve of random forest model based on CART decision tree algorithm.

The closer to the upper left corner ROC curve, the higher the accuracy of the test. The point closest to the upper left corner of the ROC curve is a minimum fault of best threshold, the total number of false positive and false negative is minimum [13]. As shown in Fig. 2, the ROC curve of random forest model based on CART algorithm is closer to the upper left corner and more accurate.

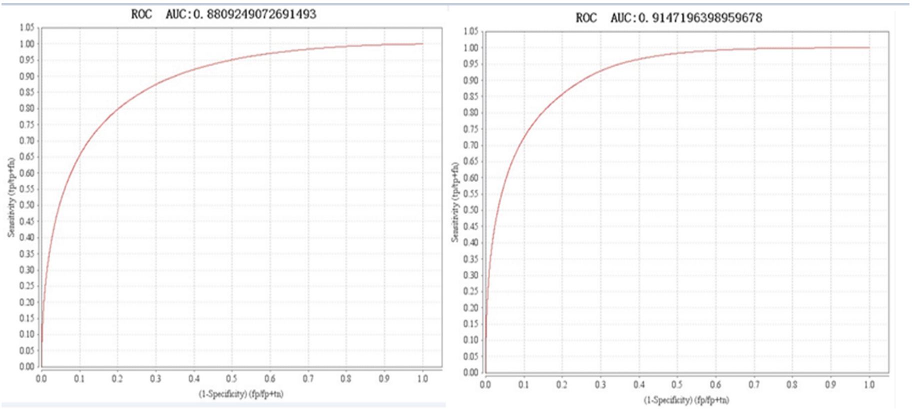

Fig. 2. Curve analysis

3.3 Prediction Results

Based on these two trained models, we predict the brand that the user is about to buy in the next month. Predicted table is mainly constitute by user_id, brand_id, probability where probability is a decimal from 0 to 1 to show the probability users may purchase in next month. As the final validation set is about 2.5 million, we take forecast result that probability are more than 0.4 about 2.8 million data to verification (Table 3).

Table 3. Comparison results of classification

	Precision (P)	Recall rate (R)	F1
Random forest classification result of C4.5	5.66 %	5.64 %	5.65 %
Random forest classification result of CART	5.83 %	5.82 %	5.82 %

The results showed that classification results of random forest based on CART decision tree algorithm is superior to classification results of random forest based on C4.5 decision tree algorithm. Both model evaluation and actual results showed that CART algorithm is better than C4.5 algorithms in user behavior classification decisions.

4 Conclusion

The paper used Random Forest model to classify and compared the results of C4.5 and CART based on the massive actual user data. The results show that CART algorithm is superior to C4.5 algorithm on actual user transactions and difficult attributes will affect the classification results.

References

1. Resnick, P., Varian, H.R.: Recommender systems. Commun. ACM **40**(3), 56–58 (1997)
2. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
3. Guo, Y., Liu, X., Guo, M., et al.: Identification of plant resistance gene with random forest. Jisuanji Kexue yu Tansuo **6**(1), 67–77 (2012)
4. Jiangeng, L., Zhikun, G., Xiaogang, R.: Random forest based gene pathway analysis of gastric cancer microarray data. J. Biol. **27**(2), 1–4 (2010)
5. Fang, K., Zhu, J., Xie, B.: A research into the forecasting of fund return rate direction and trading strategies based on the random forest method. Econ. Surv. **2**(9), 61–66 (2010)
6. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Elsevier, New York (2011)
7. Srivastava, J., Cooley, R., Deshpande, M., et al.: Web usage mining: discovery and applications of usage patterns from web data. ACM SIGKDD Explor. Newsl. **1**(2), 12–23 (2000)
8. Schapire, R.E.: The strength of weak learnability. Mach. Learn. **5**(2), 197–227 (1990)
9. Freund, Y.: Boosting a weak learning algorithm by majority. Inf. Comput. **121**(2), 256–285 (1995)
10. Quinlan, J.R.: C4.5: Programs for Machine Learning. Elsevier, New York (2014)
11. Breiman, L., Friedman, J., Stone, C.J., et al.: Classification and Regression Trees. CRC Press, Boca Raton (1984)
12. Fan, Y., Chen, L., Qifeng, Z.: Random forest based potential k nearest neighbor classifier and its application in gene expression data. Syst. Eng. Theory Pract. **32**(4), 815–825 (2012)
13. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143**(1), 29–36 (1982)