2014

A PREDICTIVE MODEL FOR CUSTOMER PURCHASE BEHAVIOR IN E-COMMERCE CONTEXT

Jiangtao Qiu
Southwestern University of Finance and Economics, Jiangtaoqiu@gmail.com

A PREDICTIVE MODEL FOR CUSTOMER PURCHASE BEHAVIOR IN E-COMMERCE CONTEXT

Jiangtao Qiu, School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, Sichuan, China, Jiangtaoqiu@gmail.com

Abstract

Predicting customer purchase behaviour is an interesting and challenging task. In e-commerce context, to tackle the challenge will confront a lot of new problems different from those in traditional business. This study investigates three factors that affect purchasing decision-making of customers in online shopping: the needs of customers, the popularity of products and the preference of the customers. Furthermore, exploiting purchase data and ratings of products in the e-commerce website, we propose methods to quantify the strength of these factors: (1) using associations between products to predict the needs of customers; (2) combining collaborative filtering and a hierarchical Bayesian discrete choice model to learn preference of customers; (3) building a support vector regression based model, called Heat model, to calculate the popularity of products; (4) developing a crowdsourcing approach based experimental platform to generate train set for learning Heat model. Combining these factors, a model, called COREL, is proposed to make purchase behaviour prediction for customers. Submitted a purchased product of a customer, the model can return top n the most possible purchased products of the customer in future. Experiments show that these factors play key roles in predictive model and COREL can greatly outperform the baseline methods.

Keywords: predictive model, purchase behavior, E-commerce

1 INTRODUCTION

If a firm is able to predict customers purchase behavior, the firm will really benefit much from this ability, such as improving success rate of acquiring customer, increasing sales and establishing competitiveness. There have been researchers from marketing and customer relationship management (CRM) making their contributions for the purchase behavior prediction. Market basket analysis [Raymond et.al. 2005, Shu et.al. 2011] examines purchase lists of supermarket or shop to identify purchase pattern. With the pattern, the needs of one customer can be predicted. Scholars in CRM employ techniques of data mining to evaluate customers' value [E. H. Suh et.al. 1999, J. R. Bult et. al. 1995, J. A. McCarty et. al.2007], helping firms to acquire customers or improve customer retention.

Different from traditional business, enterprises in e-commerce have no way to acquire information about customer demography, geography and family background. Instead, it is convenient to obtain reviews, ratings of product and visiting tracks. Therefore, those methods and algorithms for customer purchase behavior prediction in traditional business are not suitable for e-commerce context.

Is it probable to predict customer purchase behavior in e-commerce context?

Anand [Anand 2008] suggests that purchases are viewed as two distinct types in online shopping: firm-initiated purchase, which is a consequence of the firm making recommendations, and others apart from the firm-initiated purchase. In our opinion, the second type may be further divided into self-initiated purchase and association-initiated purchase. The self-initiated purchase indicates that current purchase is not related with past purchase behavior while the association-initiated purchase considers it to be associated with previous experience. Let's envisage a scenario: a beginner photographer bought a camera; with improvement of his photography skills, a tripod can possibly emerge on his schedule, and then a remote shutter will also probably follow it.

As a tool for firm-initiated purchase, recommender system is a technology adopted widely in e-commerce websites. Recommender system [Linyuan et.al. 2012, D. Jannach et.al. 2011] predicts which items a customer will most probably like or be interested in via either exploiting ratings made by customers with similar taste (collaborative filtering) or using ratings of the customer for other products in past time (content-based recommendation). However, recommender system generally predicts a rating for a candidate product that only represents what impression the customer will have for the product. The rating is far from predicting purchase behavior of customers. According to our experiments, using collaborative filtering as a tool for predicting customer purchase behavior leads to a very poor performance.

To predict the self-initiated purchase is an extremely difficult task since too many factors can activate a customer's purchase desire while many of them cannot be acquired in the online context. For example, a man may go to buy a TV set due to old one was out of order. The desire of customer is almost impossible to be predicted in e-commerce context.

Referring to the association-initiated purchase, however, past purchased products of customers may reveal their needs. Hence, it is viable for predicting customer needs via exploring associations among products. From our point of view, customers' purchase behaviors generally are motivated by their needs. Additionally, the popularity of products has an impact on the purchasing decision-making of a customer. For example, a product in which no one has shown any interest is difficult to activate a customer's desire. Also, customers' preference for products plays key roles for their purchasing decision-making.

This study focuses on the association-initiated purchase prediction in e-commerce context. We propose a predictive model incorporating customer needs, the popularity of products and customer preference, which are usually ignored in previous works. When a customer submits one purchased product into this model, it can return top *n* the most probably purchased products by the customer in future.

We explore associations between products, exploiting them to predict customer needs. The popularity of a product is subject to many factors such as on-shelf date, recent review date, ratings and the number of reviews. We develop a Support Vector Regression (SVR) based model, called Heat model, to calculate popularity of products. However, we fail to find out train dataset for Heat model from e-commerce websites. According to our experience, customers are generally able to make a correct judgment on the popularity of a product when they are doing an online shopping. Relying on crowdsourcing approach, we develop an experimental platform in which participants give a comparison on the popularity of a pair of products. The system converts the choices of participants to instances of train data for learning the parameters of Heat model. The discrete choice analysis is usually used to build the behavioral model for customers' decision-making. We also combine collaborative filtering and a hierarchical Bayesian discrete choice model to learn customers' preference for products.

2 RELATED WORK

Researchers in marketing and retailing fields have paid their efforts on predicting customers purchase behavior, which may help firms to implement cross selling and up selling campaigns. Market basket analysis or Association Rule is a main technique for the prediction. They find customer purchasing patterns by extracting associations or co-occurrences from stores' transactional databases. [Raymond et.al. 2005, Shu et.al. 2011].

CRM system of a firm generally needs to maintain a large amount of customer data, such as age, gender, income and purchasing lists. Data mining techniques are often employed in analytical CRM to transform this large amount of data into valuable knowledge that can be used to support marketing decision making. Based on such data mining techniques, customers may be segmented into clusters with internally homogenous and mutually heterogeneous characteristics [Chihli et.al. 2008]. Besides segmentation, customers can also be ranked on their probability to behave in a certain way (e.g. buying a specific product or responding to a certain marketing campaign). With helps of these segmentation schemes and rankings, a firm is able to approach carefully selected customers, resulting in a higher success rate of their marketing campaigns [E. H. Suh et.al. 1999]. Consequently, researchers often try to improve CRM by enhancing the data mining techniques themselves. Researches in CRM have evolved from RFM models (i.e., recency, frequency and monetary value of customer purchases) to classification techniques such as chi-square automatic interaction detection (CHAID) and regression models [J. R. Bult et.al. 1995, J. A. McCarty et. al. 2007]. Recently, researchers try to outperform these primitive techniques by introducing more advanced machine learning algorithms, like support vector machines, neural networks and random forests [H. Shin et.al. 2006, J. Zahavi et. al. 1997].

In general, it is difficult to acquire customer demography and privacy information such as income, age and gender in e-commerce context. Instead, it is easier to acquire web accessing data such as product reviews and ratings, which provide richer information than traditional CRM. Therefore, those prediction techniques in CRM based on customer data cannot be well transferred to e-commerce domain. There have been few researches focusing on predicting purchasing behavior in e-commerce context or online shopping. [Poel et.al. 2005] employ logit modeling to predict whether or not a purchase will be made during the next visit to the website. The model uses click steam behavior, customer demography and historic purchase behavior as variables. Their study shows that click stream behavior is important when determining the tendency to buy.

E-commerce firms generally use recommender system for predicting customer purchase behavior. However, as discussed in Section 1, recommender system that generally calculate a rating for a candidate product only predict what impression the customer will have for the product, which is far different from predicting purchase behaviors of the customer. There have been many excellent surveys [Linyuan et.al. 2012, D. Jannach et.al. 2011] about recommendation system. We will not give a detailed discussion about it in this paper.

3 CUSTOMER PURCHASE BEHAVIOR PREDICTIVE MODEL

This section proposes a customer purchase behavior prediction model COREL (**C**ust**O**mer purchase p**RE**diction mode**L**). Let c_k be a customer; d_i and d_j be products. When c_k purchased d_i in time t, COREL can return top n the most likely purchased products by c_k after t time.

As c_k bought d_i in t time, the probability that c_k will also purchase d_j after t time may be

$$p(d_j \mid c_k, d_i) = \frac{p(d_i \mid c_k, d_j)\, p(d_j, c_k)}{p(d_i, c_k)}$$

Suppose c_k and d_i are independent of each other, i.e. c_k may purchase any product in time t, the probability may be

$$p(d_j \mid c_k, d_i) = \frac{p(d_j \mid d_i)\, p(d_j \mid c_k)}{p(d_j)}$$

where $p(d_j|c_k)$ is the probability that c_k will purchase d_j. $p(d_j|d_i)$ is the probability that a customer bought d_i will also purchase d_j. $p(d_j)$ is prior probability of d_j.

Let $\omega=\{d_1,\ldots,d_{i-1},d_{i+1},\ldots,d_m\}$ be a collection of candidate products. We calculate $p(d_j|c_k, d_i)$ for each product $d_j \in \omega$, and then rank them. When the prior probability of a product $p(d_j)$ is assumed to be uniform across all products, $p(d_j)$ can be ignored. Therefore COREL may be

$$p(d_j \mid c_k, d_i) \propto p(d_j \mid c_k)\, p(d_j \mid d_i)$$

COREL can be understood as a two-stage approach: using $p(d_j|d_i)$ to build a product collection ω where products are associated with d_i; using $p(d_j|c_k)$ to pick up the most likely being purchased candidates from ω. Therefore, the most important task for building a model COREL is estimating both parameters $p(d_j|d_i)$ and $p(d_j|c_k)$.

3.1 Estimating p($d_j|d_i$)

Parameter $p(d_j|d_i)$ represents the probability that d_j will also be purchased by the same customer after d_i is purchased. The parameter can be estimated by exploring association between d_i and d_j, which may be calculated via market basket analysis. When both products occur in same one market basket, it is generally thought that there exists an association between both products. Using maximum likelihood estimation,

$$p(d_j \mid d_i) = \frac{|d_i \cap d_j|}{|d_i|} \qquad (1)$$

$|d_i|$ denotes the number that product d_i is purchased; $|d_i \cap d_j|$is the frequency of both products d_i and d_j co-occurring in the one market basket. However, experiment shows that the collection of candidates built using formula (1) is so small that COREL fail to achieve a good performance.

Therefore, we suggest building association of category, and then picking up candidates from associated category of a product. Generally, e-commerce websites assign their products the multi-levels categories. For instance, jingdong (www.jd.com) has three level categories for their products. For one item “EPSON LQ-630k Printer”, its categories from first level to third level is “Computer or Office Equipment->Printing related Office Equipment->Printer”. We generate categories association in third level of categories. $Thr(d_i)$ denotes the third category of product d_i. So,

$$p(d_j \mid d_i) = \frac{|Thr(d_i) \cap Thr(d_j)|}{|Thr(d_i)|} \qquad (2)$$

Experiments reported about in section 4 demonstrate that associations of categories can extend the candidate collection. COREL achieve a better performance by picking up top *n* associated categories than that of using formula (1).

3.2 Estimating p(d_j|c_k)

Parameter p(d_j|c_k) indicates the probability that customer c_k will purchase product d_j. However, it is almost impossible to accurately estimate the parameter. According to our experiences of online-shopping, we may find that the parameter is subject to two factors: the popularity of d_j and preference of c_k. Suppose both factors are independent of each other, formally, p(d_j|c_k) is approximately computed with formula (3)

$$p(d_j \mid c_k) \approx Hot\left(d_j\right) * Preference\left(c_k, d_j\right) \quad (3)$$

It is believed that a product that is purchased much more frequently and has higher ratings than others will be more popular in customers' purchase decision-making. Hence we develop a model, called Heat model Hot(d_j), to calculate the popularity of products.

Customer's preference also plays an important factor during his/her purchase decision- making. We propose an approach to learn customers' preference of products Preference(c_k,d_j), which is presented in Section 3.2.2.

3.2.1 Heat Model

Besides the needs of a customer, the number of reviews and ratings of a product also play import roles on purchase decision-making of the customer. If a product is given a lot of low ratings or sparse reviews, it is believed that a customer will hesitate to purchase the product even if he has a strong need for it. We exploit reviews and sale information of a product to calculate its popularity. They include (1) Qr, the number of reviews; (2) Qs, the average of ratings; (3) Qa, the number of days since on-shelf; (4) Qu, the number of days since recent review. We use a vector to represent the product d_i in which there are four elements (Qr, Qs, Qa, Qu).

Previous researches have shown that SVR (Support Vector Regression) is an excellent tool for predictive tasks. We develop a SVR based model to calculate popularity of products, called Heat model Hot(d_i). Given a product with its four attributes Qr, Qs, Qa, Qu, Heat model can calculate a score for its' popularity. The train set is a necessary component for learning a Heat model. As far as we know, none of the e-commerce websites provides labeled data about popularity of products. However, we can observer that a visitor can possibly decide which one of two products is more popular in online shopping. Based on the observation, we use following four steps to generate a Heat Model.

Step 1: Relied on crowdsourcing approach, we develop a platform in which participants need to pick up a more popular product from a pair of products displayed in web page. The interface of the system shows in Figure 1.

The system works with following steps. Picks up any two products A and B from products database; displays factors Qr, Qs, Qa and Qu of both products in web page; a participant chooses a more popular one from them; If Hot(A)>Hot(B), generates two instances of train set.

Err_Qr	Err_Qs	Err_Qa	Err_Qu	label
Qr(A)- Qr(B)	Qs(A) -Qs(B)	Qa(A)- Qa(B)	Qu(A)- Qu(B)	1
Qr(B)- Qr(A)	Qs(B) -Qs(A)	Qa(B)- Qa(A)	Qu(B)- Qu(A)	-1

In the obtained train set, one instance includes five fields: Err_Qr, Err_Qs, Err_Qa, Err_Qu and label. Qr(A) denotes the element Qr of vector A.

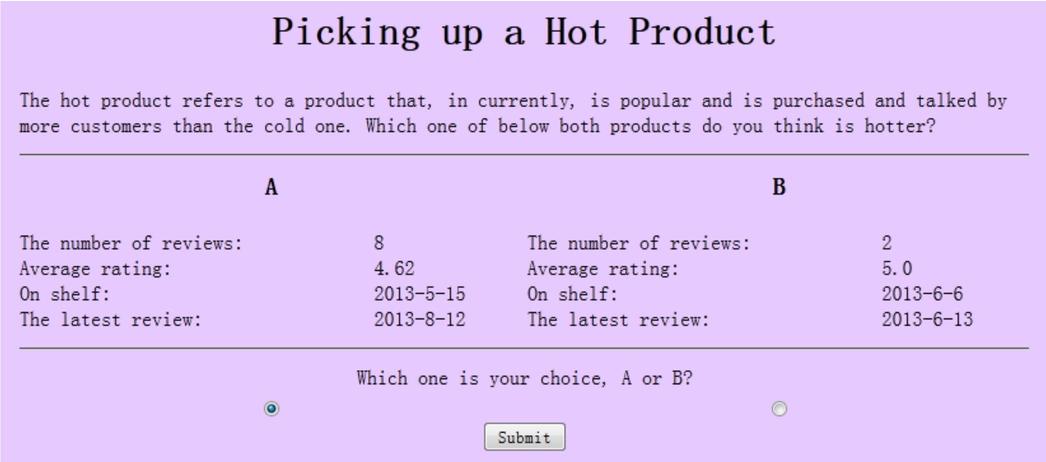

Figure 1. A crowdsourceing approach based experimental platform

Step 2: We build a logistic regression model $f(\varphi)$ that may be able to compare the popularity of two products. φ in the model is a vector where the elements, denoted as Err_Qr, Err_Qs, Err_Qa, Err_Qu, represent difference between elements of both compared product vectors.

$$f(\varphi) = \frac{\exp(\pi(\varphi))}{1 + \exp(\pi(\varphi))}$$

Where

$$\pi(\varphi) = \beta_0 + \beta_1 \times Err\_Qr + \beta_2 \times Err\_Qs + \beta_3 \times Err\_Qa + \beta_4 \times Err\_Qu$$

We employ train set obtained in step 1 to train the logistic regression model.

Step 3: An algorithm using logistic regression model f(φ) to calculate popularity of products is described as follows.

Algorithm 1. Calculate Popularity of Products

Input: a collection of products ω, logistic regression model f(φ)
Output: the popularity of products in ω
steps：

1. P ←[]
2. For each pair <a,b>, a,b∈ω, a≠b
3. φ =V(a)-V(b)
4. score=f(φ)
6. P[a]= P[a]+score-0.5; P[b]= P[b]+0.5-score
9. End
10. normalize P to range [0,1]
11. Return P

In the algorithm 1, the array P stores the calculated popularity of all products in ω in range of [0, 1].

Step 4: Using algorithm 1, we calculate popularity for each products in set ω, and further generate a train set for SVR model. Two instances in the train set are shown in following table where score refers to popularity of a product and Ln(Qr) is natural log of Qr attribute.

Ln(Qr)	Qs	Ln(Qa)	Ln(Qu)	score
1.0986	4	5.9054	5.8833	0.23539
0.69315	5	6.0497	5.9636	0.32821

In this study, we explore ε-SVR and μ-SVR combining with the polynomial kernel and the radial basis function that are used as the kernel function of SVR respectively. Since there is few general guidance to determine the parameters of SVR, this study varies the parameters to select optimal values for the best prediction performance. Experimental results show that ε-SVR with radial basis function can reach best performance in our study. This study uses LIBSVM software system [Chih et.al. 2011] to perform experiments.

Given a set of data points, $\{(X_1,z_1),\ldots,(X_m,z_m)\}$, such that $X_i \in R^n$ is an input and $z_i \in \mathrm{R}^1$ is a target, the standard form of ε-SVR is

$$\min_{w,b,\xi,\xi^*} \quad \frac{1}{2} w^T w + C\sum_{i=1}^{l} \xi_i + C\sum_{i=1}^{l} \xi_i^*$$

Subject to

$$\begin{aligned} & W^T \phi(X_i) + b - z_i \le \varepsilon + \xi_i \\ & z_i - W^T \phi(X_i) - b \le \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \ge 0, i = 1,\ldots,l \end{aligned}$$

When Heat model reaches the best performance, the parameters of ε-SVR are C=1 and ε=0.3.

3.3 Learning Customer Preference

Economic models of choice typically assume that an individual's latent utility is a function of brand and attribute preference [Sha et.al. 2003]. Collaborative filtering (CF) may be used to estimate a customer's rating for one product in e-commerce via exploiting the product ratings made by customers with similar taste. However, CF doesn't consider a customer's preference for price and brand of products that play an important role in customer's purchase decision making. We predict the c_k's rating for d_j using collaborative filtering, CF(c_k,d_j), and then propose a hierarchical Bayesian discrete choice model to learning preference of customers to price and brand, DC(c_k,d_j). CF(c_k,d_j)*DC(c_k,d_j) refer to preference of customer c_k to product d_j.

CF(c_k,d_j) is calculated in formula(4).

$$CF(c_k, d_j) = \frac{\sum_{s \in S} Sim(c_k, s) \times rating(s, d_j)}{|S|} \qquad (4)$$

where S denotes a set of customers that consists of top 10 the most similar customers with c_k; rating(s,d_j) refers to a rating that customer s make for product d_j. The possible rating values are defined on a numerical scale from 0 (strongly dislike) to 5 (strongly like). Sim(c_k, s) indicates the similarity between customers c_k and s, which can be calculated by using cosine measure. A customer feature vector is defined as a set of ratings of products. For example, the feature vector of c_k, V(c_k)=(0, 4, 1, 0, 5) represents that c_k did not purchase product d_1 (or he/she give a 0 rating value) and gave d_2 a rating value 4, etc.

$$Sim(c_k, c_l) = \frac{V(c_k)V(c_l)}{|V(c_k)||V(c_l)|} \qquad (5)$$

Experiments reported about in section 4.3 analyze how CF impact on performance of COREL. Experimental results show that a model combining p(di|dj) with CF can outperform the basic models using only either p(d_j|d_i) or CF on predicting customer purchasing behavior.

We propose a hierarchical Bayesian discrete choice model to learn customer c_k's preference for price and brand. Employing the model, we can calculate to what extent customer c_k prefer a product d_j, DC(c_k, d_j).

We divide price and brand of every product into three levels respectively: high, medium and low price; large, moderate and small brand. In this way, the feature vector x of a product d has six binary value features x = (p_hi, p_me, p_lo, b_la, b_mo, b_sm) corresponding to three price levels and three brand levels, respectively. Only one of three price levels in the feature vector has a value 1 while others is 0. For example, (p_hi=1, p_me=0, p_lo=0) indicates the price of a product is in high level. Brand features are also subjected to the rule. For example, (b_la=0, b_mo =0, b_sm=1) means that a product belongs to the small brand.

$$DC(c_k, d_j) = P(y_j = 1) = \frac{1}{1+\exp(-V(d_j, c_k))} \quad (6)$$

Utility function

$$U\left(d_j, c_k\right) = V\left(d_j, c_k\right) + e_{jk}$$

$$V\left(d_j, c_k\right) = \beta_1 \times p\_hi + \beta_2 \times p\_me + \beta_3 \times p\_lo + \beta_4 \times b\_la + \beta_5 \times b\_mo + \beta_6 \times b\_sm$$

$P(y_j=1)$ denotes the probability of selecting product d_j. Six coefficients $\beta_1 \sim \beta_6$ in the utility function are decided by features of customer. It means that each customer can face the utility function with different coefficients. We use following features to construct a customer feature vector.

- R (Recency): the number of months that have passed since the customer last purchased
- F (Frequency): number of purchases in the last 12 months.
- M (Monetary): the amount of value from the customer in the last 12 months.
- Sd, which is standard deviation of prices of total purchase products of the customer. The value reveals the habit of online shopping of the customer. The smaller Sd means that the customer like purchase fixed variety of product while larger Sd indicates that the customer don’t mind price of products in online shopping.
- Age, which denotes time interval in year from current date to date when the customer first purchase product in the website.

Every customer may have his preference for products’ price and brand. For example, someone prefers large brand products while other ones do not care a products’ brand in condition that it is cheap. In the hierarchical Bayesian model, the coefficients in utility function are decided by customers’ features.

Use B denotes $\beta_1 \sim \beta_6$.

$$B = Z\Delta + U; \qquad u_i \sim N(0, V_\beta)$$

The matrix Z contains features of customers. The coefficient matrix Δ has a normal distribution with means vec($\bar{\Delta}$) and covariance matrices given by Kronecker product of A^{-1} and V_β.

$$\beta_n \sim N(\Delta' Z_n, V_\beta)$$

$$vec(\Delta) \sim N(vec(\bar{\Delta}), A^{-1} \otimes V_\beta)$$

$$V_\beta \sim IW(v, V)$$

The *vec* operator creates a column vector from a matrix by stacking the column vectors of $\bar{\Delta}$ [Jan et.al. 2007]. Hyperparameter V_β has an Inverted Wishart prior. We set noninformative prior v, V, $\bar{\Delta}$ and A to v=m+3, $V = v \cdot I$, $\bar{\Delta}$=0 and A=0.01 where *m* is the number of coefficients in utility function.

Parameters in the hierarchical Bayesian model can be described using a DAG (Directed Acyclic Graph) in Figure 2.

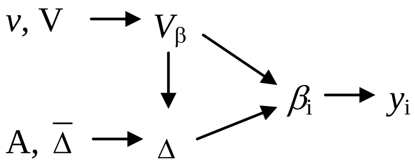

Figure 2 DAG of parameters in the hierarchical Bayesian model

We employ MCMC-metropolis hasting algorithm to estimate parameters in the hierarchical Bayesian model, using a normal distribution as the proposal distribution for the MCMC algorithm. The log-likelihood function is

$$L(X,Y,B)=\sum_i \log(p(x_i)*y_i+(1-p(x_i))*(1-y_i))$$

$$p(x_i)=\frac{\exp(x_i*B)}{1+\exp(x_i*B)}$$

The steps for estimating parameters of model are as follows.

Algorithm 2. Use MCMC-metropolis hasting algorithm to estimate parameters

steps：

1. initiating β_{old}
2. draw from $V_\beta|v,V \sim IW(v+n,V+S)$
3. draw from $vec(\Delta)\,|\,\bar{\Delta},A,V_\beta \sim N(vec(\bar{\Delta}),A^{-1}\otimes V_\beta)$
4. draw $\beta_{\text{new}} \sim N(\beta_{\text{old}}, V_\beta)$
5. Compute

$$\alpha(\beta_{old},\beta_{new}) \sim \min(1, p(\beta_{new})q(\beta_{new},\beta_{old})/p(\beta_{old})q(\beta_{old},\beta_{new}))$$

Where

$$p(\beta_{new})/p(\beta_{old})=\exp(L(X,Y,\beta_{new})-L(X,Y,\beta_{old})),$$

$$q(\beta_{new},\beta_{old})/q(\beta_{old},\beta_{new})=\exp\{(\beta'_{new}-Z\Delta)*V_\beta*(\beta_{new}-(Z\Delta)')-(\beta'_{old}-Z\Delta)*V_\beta*(\beta_{old}-(Z\Delta)')\}$$

6. If $\alpha<1$ then
7. $\beta_{\text{old}}=\beta_{\text{new}}$ with probability α
8. else
9. $\beta_{\text{old}}=\beta_{\text{new}}$
10. End
11. Goto step (2) until loop end

Using saved draws, we can plot posterior distribution of coefficients. Figure 3 illustrates posterior distributions of three coefficient p_hi, p_me and p_lo for one customer. It can be observed that means of three distribution is about -2.7, 0.8 and 0.5, respectively. From point estimate of three coefficients, we can conclude that the customer generally rejects high price product and tends to prefer medium price products to low price one. The trained model reveals more information about customers.

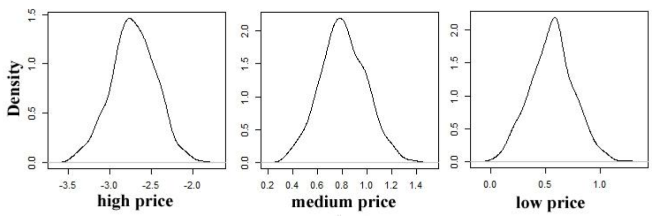

Figure3. Posterior distribution of coefficients of a customer

To learn a hierarchical Bayesian discrete choice model, it is necessary to know the choices of customers in a finite alternative set. In e-commerce context, however, we can only know what customers purchased, not to know what customer gave up in their choice. When training a hierarchical

Bayesian discrete choice model, both positive and negative samples are necessary components. Regarding the purchased products as positive data, we develop a technique to generate one negative instance from the positive one. One instance in train dataset is a feature vector of a purchased product combined with a label. Six features p_hi, p_me, p_lo, b_la, b_mo, b_sm in an instance represent price level and brand level of a product, respectively.

p_hi	p_me	p_lo	b_la	b_mo	b_sm	label
1	0	0	1	0	0	1

When each feature in the positive instance is inverted, we can derive a negative instance.

p_hi	p_me	p_lo	b_la	b_mo	b_sm	label
0	1	1	0	1	1	0

4 EXPERIMENTS

4.1 Dataset

Jingdong is a well-known B2C e-commerce website in P.R.China. We collected customer information and reviews of products from the website. The collected data contain 727,878 product items, 342,451 customers and 14,634,059 reviews from 2004 to 31 January, 2013. The products in Jingdong are assigned three level categories. There are 19 first-level categories, 124 second-level categories and 1078 third-level categories.

It is difficult to directly collect purchase data of customers from e-commerce website as they are generally regarded as privacy. Our study is based on an assumption: if a customer frequently writes reviews in an e-commerce website, his reviews can almost totally exhibit purchased products of the customer (in Jingdong website, only customers who purchased a product is authorized to write review for the product). Therefore, we pick up customers with high reviewing frequency to generate purchase data from their reviews. In addition, there are 55 participants using our crowdsourcing platform to generate train data that have 1351*2 instances.

The collected data are processed as follows.

(1)Dividing dataset: Divide the dataset into three sections by date. A section: before 30th June, 2012; B section : from 30th June, 2012 to 31th July,2012; C section: after 31th July, 2012.

(2)Picking up customers: We pick up customers who wrote reviews in all three sections and the number of reviews in A section is more than 30. There are total of 2770 customers meeting the requirement.

(3)Test set: those products that are reviewed in C section by the picked customers.

(4)Train set: purchased data of the picked customer in A section.

(5)Target set: C section

(6)Setting price level for every product: Let *d* be a product, thr be its third-level category. If the price of *d* is up to 75th percentile of price of all products in thr, we assign features of d to p_hi=1, p_me=0 and p_lo=0. If the price of *d* is less than 25th percentile of price, the features is p_hi=0, p_me=0 and p_lo=1. Otherwise, p_hi=0, p_me=1 and p_lo=0

(7)Setting brand level for every product: we examine the distribution of the number of product items for all brands. If the number of product items of a brand is larger than 75th percentile of the distribution, we set features all products belonging to the brand as b_la=1, b_mo=0 and b_sm=0. If the brand is less than 25th percentile of the distribution, the feature is b_hi=0, b_mo=0 and b_sm=1. Otherwise, b_hi=0, b_mo=1 and b_sm=0

We call the processed data as JD dataset.

4.2 Experimental Results

This section investigates performance of COREL on predicting customer purchase behavior. We compare COREL to several baseline methods. These models are shown in Table 1.

name	model	description
M1	p(d_i|d_j)	Using formula (1) to pick up associated products
M2	p(d_i|d_j)*CF	Using formula(2) to build candidate set and employing collaborative filtering to pick up products from candidates
M3	p(d_i|d_j)*CF*Hot	A model combing M2 with Heat model
M4	COREL	Incorporating customers' price and brand preference

Table 1 Prediction model

Four models in Table 1 make predictions on JD data set. These models use the latest purchased product d_i in B section by customer c_k to predict purchased ones by c_k in C section. The candidates are generated by picking up ten most associated categories of d_i. These models calculate a prediction score for all candidates, and pick up top n candidates to build a product subset ω. If any product in ω occurs in test set Φ, we call customer c_k is successfully predicted. Using precision as a measure, we present the experimental result in Table 2.

n	M1	M2	M3	M4
1	4.7%	9.9%	11.4%	15.8%
3	6.8%	13.1%	15.7%	17.5%
5	8.9%	19.9%	23.5%	23.5%
10	9.9%	27.7%	32.9%	32.2%

Table 2 Evaluate prediction performance for four models

Table 2 shows that market basket analysis (M1) has a poorest performance on predicting purchasing behavior of customers. Combing collaborative filtering and market basket analysis (M2), predictive model can dramatically improve its performance. When combining M2 and Heat model, the predictive performance can be further increased. The model M4 incorporates price and brand preference of customers. We can see brand and price preference do not make a significant improvement for prediction performance when n=10. But M4 outperform other three models when n=3 and n=1. It means that the proposed model COREL in this study is feasible and effective on predicting customers' purchase behavior.

5 CONCLUSION

Researchers from marketing and CRM fields make a lot of significant contributions on customer purchase behavior prediction for traditional business. However, in e-commerce context, new methods and techniques need to be developed to deal with the predictive problem.

In this study, we investigate several key factors that have an impact on purchasing decision-making of customers in e-commerce context, including the needs of customers, the popularity of products and the preference of customers. Furthermore, exploiting purchase data and ratings of products, we propose methods to quantify the strength of these factors.

We believe that there exist association-initiated purchase in on-line shopping and it can be exploited to predict the needs of customers. Experiments in this study favor our point of view. It is reasonable that we divide online shopping to three types: firm-initiated purchase, self-initiated purchase and association-initiated purchase. Accordingly, experiments also show that associations between categories of products can significantly improve the predictive performance.

We develop a SVR based model, called Heat model, to calculate the popularity of products. However, none of the e-commerce website provides the labeled data for training the model. Relying on crowdsourcing approach, we develop a system that can generate train set via collecting participant's judgment on popularity of products. Experiments prove that our approach is feasible and popularity of products is also a key factor impacting on-line shopping. We also combine CF and a hierarchical Bayesian discrete choice model to learn preference of customers. Experiments demonstrate that customers' preference play an important role on the purchase decision-making of customer.

The model COREL, which combines products associations, the popularity, and customer preference, may be applied to predict the most possible purchased products of a customer. Experiments show that COREL obviously outperform the baseline methods.

Acknowledgements

This work was supported by Major Program of National Natural Science Foundation of China (No. 91218301) and the Fundamental Research Funds for the Central University of China (No. JBK120505).

References

Anand V. Bodapati (2008). Recommendation Systems with Purchases Data, journal of marketing research, 45 (1),77-93.

Chihli Hung and Chih-Fong Tsai (2008). Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand, Expert System with Application. 34(1), 780–787.

Chih-Chung Chang and Chih-Jen Lin (2011). LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 27(2), 1-27.

D. Jannach, M. Zanker, A. Felfernig, G. Friedrich, (2011). Recommender Systems: An Introduction. Cambridge University Press, New York.

E. H. Suh, K. C. Noh, and C. K. Suh (1999). Customer list segmentation using the combined response model, Expert System with Application, 17(1), 89–97.

H. Shin, and S. Cho, (2006). Response modeling with support vector machines, Expert System with Application, 30(4), 746-760.

Jan R. Magnus, Heinz N. (2007). Matrix Differential Calculus with Applications in Statistics and Econometrics, JOHN WILEY & SONS, New York.

J. A. McCarty, and M. Hastak (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression, Journal of Business Research, 60(3), 656–662.

J. R. Bult, and T. Wansbeek (1995). Optimal selection for direct mail. Mark. Sci., 14, 378–394.

J. Zahavi, and N. Levin, (1997). Applying neural computing to target marketing. Journal of Direct Marketing, 11(2), 5–23.

Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, Tao Zhou (2012). Recommender systems, Physics Reports, 519(1), 1–49.

Poel, D. V. D., & Buckinx, W. (2005). Predicting online-purchasing behavior. European Journal of Operational Research, 166, 557–575.

Raymond Chi-Wing Wong, Ada Wai-Chee Fu, Ke Wang, (2005). Data Mining for Inventory Item Selection with Cross-Selling Considerations. Data Mining and Knowledge Discovery, 11 (1), 81-112

Sha Yang, Greg M. Allenby (2003). Modeling Interdependent Consumer Preferences. Journal of Marketing Research, 40(3), 282-294.

Shu-hsien Liao, Yin-ju Chen, Hsin-hua Hsieh (2011). Mining customer knowledge for direct selling and marketing, Expert Systems with Applications, 38 (5), 6059–6069