Predict Consumers' Automobile Purchase Behavior Based on the Preferences of Explicit Features and Implicit Topic

Du Yanan, Qian Yang
School of Management, HeFei University of Technology, HeFei 230009, China
Duyanan2046@163.com

Abstract. In order to know the role of online reviews in predicting consumers' automobile purchasing behavior in the social network environment, this paper uses online reviews to construct a consumer interest preference model to obtain consumers' explicit feature preferences and implicit topic preferences. Based on the preferences of consumers, introduce the tree structure to construct hierarchical prediction model to predict consumers' purchase behavior. Meanwhile, consider the effect of online interactive comment on consumers' automobile purchases. Several results are obtained in this paper: (1) The proposed hierarchical model introducing more information could achieve multi-granularity purchase-decision predicting; (2) the predictive effect based on implicit topic preference is better than that based on explicit feature preference; (3) integrating online interactive comments have positive effect in analyzing and predicting consumers' purchasing behavior.

1. Introduction
With the rapid development of the automobile industry, more and more automobile brands and types have sprung up, which result in more and more diverse in customer demands and preferences. Getting insight of individual customer’s preferences for the automobile has been important issue that automobile companies and marketers focus on. Consumer buying behavior is the embodiment of heterogeneity of consumer preferences [1-3]. Predicting the purchase choice of consumers based on the heterogeneity of consumer preferences is a hot research topic. Online reviews contain the characteristics which consumers’ concern and imply consumers' interest. So that they provide an opportunity to explore consumers' heterogeneous preferences and predict individual consumers' buying behavior.

The previous work about purchase behavior mainly focuses on consumers' characteristics, product attributes and marketing strategies. On consumers' characteristics, [4-6] analyses how individual characteristics and family characteristics, such as gender, age, income, education level, family population, number of children, affect consumer demand and purchase decisions. On product attributes, literature [7] analyzed the influence of five automobile attributes, namely fuel consumption, safety, comfort, style and quality, on consumer preference and choice behavior by using survey data of new automobile consumers. On marketing strategies, literature [8] analyzes the influence of brand advertisements, price promotion and the interaction between them on consumers' purchase of automobile. These studies are mainly based on the survey data of the offline market, which is costly to obtain and limited in data quantity.

Massive online reviews provide a rich pool of data resources. These data could reflect individual’s characteristics such as consumers' buying interest, attitude and personality traits [9-11]. In addition, online reviews contain consumers' perception and understanding of product features, price promotion,
macro policies, etc. [12, 13]. The differences in review content among different individuals reflect the preferences of different consumers [14]. While in on-line community, users interact with each other, share common hobbies, and discuss issues of common interest. Unconsciously, individual preferences would be affected by other person's behavior [15]. For example, when intending to buy an automobile, consumers may look over others' reviews, seek advices from others, discuss with others about a certain brand or a certain type on the social platform, so the final decision-making will be influenced in this process.

In this paper, utilize online reviews to obtain explicit features and implicit topics which reflect consumers' heterogeneous preferences and introduce the tree structure to construct a hierarchical prediction model to predicting consumers' purchase behavior. In order to better predict consumers' automobile purchasing behavior, consider online comments that interact with consumers while acquiring consumers' preferences of explicit features preferences and implicit topic preferences.

2. Model Building

2.1 Multivariate Discrete Selection Model Considering Hierarchical Relations

When making a purchase decision, consumers will judge the utility of the candidate products and make a rational decision based on the principle of utility maximization. Discrete selection model is to handle the decision-making problem from the perspective of utility maximization for consumers.

The observable data set contains $n$ consumer samples $\{(x^{(i)}, y^{(i)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(n)}, y^{(n)})\}$, where $x^{(i)} = (x_{1}^{(i)}, x_{2}^{(i)}, \cdots, x_{L}^{(i)})$ is an L-dimensional covariant, represents explicit features preferences and implicit topic preferences which influence purchase decision; $y_{i}$ is response variable, represents the car line selected by consumers. $U_{j} = x_{j} \beta_{j}$ ($j = 1, 2, \cdots, J$) are the utility function when the consumer chooses the $j$th car line. Then the probability that consumers choose to buy $j$th car line can be expressed as (1):

$$P(y = j | x, \beta) = \frac{\exp(U_{j})}{\sum_{j'} \exp(U_{j'})} = \frac{\exp(x_{j} \beta_{j})}{\sum_{j'} \exp(x_{j'} \beta_{j'})}$$

$$\beta_{j} | \Sigma \sim N(\mu_{j}, \Sigma^{2})$$

$$P(\mu, \Sigma^{2}) = N(\mu | \mu_{0}, \Sigma^{2}/\rho)IW(\Sigma^{2} | v_{0}, \Lambda_{0})$$

Parameter $\beta$ obeys the multivariate normal distribution with the mean is $\mu$ and the covariance is $\Sigma^{2}$; $\mu$ is multivariate normal distribution with the mean is $\mu_{0}$ and the covariance is $\Sigma^{2}/\rho$ ; covariance $\Sigma^{2}$ is Inverse-Wishart distribution with the degree of freedom is $v_{0}$ and the scale parameter is $\Lambda_{0}$.

In MNL model, there is no correlation between classes. But in this problem, there are some similarities in consumers' preferences. Use hierarchical model to connect classes with each other by constructing a tree structure classifier. $m$ is a parent-node in hierarchical model, it has $c_{m}$ child nodes $S_{t}$ ($t = 1, 2, \cdots, c_{m}$), which represent the buying candidate set of consumers. The probability that the consumer selects the parent node $m$ can calculate by formula (2). And the probability for any nodes is the continuous product of intermediate nodes’ probabilities, as (3).

$$P(y \in S_{t} | x, \beta_{m}) = \frac{\exp(x_{m} \beta_{m})}{\sum_{m'} \exp(x_{m'} \beta_{m'})}$$

$$\beta_{m} | \Sigma_{m} \sim N(\mu_{m}, \Sigma^{2}_{m})$$

$$P(\mu_{m}, \Sigma^{2}_{m}) = N(\mu_{m} | \mu_{0}, \Sigma^{2}_{m}/\rho)IW(\Sigma^{2}_{m} | v_{0}, \Lambda_{0})$$
2.2 Consumer’s explicit features extraction

Employ Targeted topic model (TTM) automatically catch terms related to a given term, then filter out worthless words, this is the process of building fine-grained product feature knowledge by human-computer interaction. Next, map product features to multi-dimensional coarse-grained feature classes, and calculate consumers’ attention on each explicit feature. \( F = \{ F_1, F_2, \ldots, F_n \} \) is a set of feature classes, each \( F_i = \{ f_1, f_2, \ldots, f_f \} \) is also a set, represent fine-grained feature in feature class. \( \text{Num}(F_i, f_j) \) is the number of fine-grained features corresponding to \( F_i \) appearing in consumers’ review. It is used to reflect consumer attention on coarse-grained features and the formula is as (4).

\[
\text{AttentionRate} = \frac{\sum \sum \text{Num}(F_i, f_j)}{\sum \sum \text{Num}(F_i, f_j)}
\]

2.3 Consumer’s implicit topic preference measurement

LDA is used to explore the distribution of implicit topic preferences in consumers' online reviews which is used to depict consumers' purchase motivation. In Fig.1. \( p \) represents the topic distribution of online reviews posted by users, \( z_w \) represents the topic corresponding to the word \( w \), \( \phi_k \) is word distribution corresponding to topic \( k \). \( \alpha \) and \( \beta \) are hyper parameter. Given the number of topics \( K \), the implicit topic preference distribution of consumers can be directly obtained, as shown in equation (5).

\[
\text{Preference}_v = \left\{ \left( \text{Topic}_1, p_{1v} \right), \left( \text{Topic}_2, p_{2v} \right), \ldots, \left( \text{Topic}_k, p_{kv} \right) \right\}
\]

2.4 Preference measure combining interactive comments

In order to study the influence of interacting between consumers on predicting consumers’ automobile purchase decision, in this step need to integrate the interactive comments. The steps are similar to the previous steps: based on the method described in 2.2, calculate explicit features preference which integrating interactive comments, marked as \( \text{AttentionRate}_v \); based on the method described in 2.3, calculate implicit topic preference which integrating interactive comments, marked as \( \text{Preference}_v \). \( \text{Preference}_v \) is used to characterize consumers' purchase motivation.

2.5 Parameter learning

Four utility functions (6), (7), (8), (9) were constructed based on explicit features preference, implicit topic preference, preference integrated with social comment and related demographic variables.

\[
\begin{align*}
U_{1i} &= \alpha + \text{Age}_{\beta_i} + \text{Province}_{\beta_i} + \text{AttentionRate}_{\beta_i} + \cdots + \text{AttentionRate}_{L-2\beta_i} \\
U_{2i} &= \alpha + \text{Age}_{\beta_i} + \text{Province}_{\beta_i} + \text{Topic}_{\beta_i} + \text{Topic}_{\beta_i} + \cdots + \text{Topic}_{L-2\beta_i} \\
U_{3i} &= \alpha + \text{Age}_{\beta_i} + \text{Province}_{\beta_i} + \text{AttentionRate}_{\beta_i} + \cdots + \text{AttentionRate}_{(L-2)\beta_i} \\
U_{4i} &= \alpha + \text{Age}_{\beta_i} + \text{Province}_{\beta_i} + \text{Topic}_{\beta_i} + \text{Topic}_{\beta_i} + \cdots + \text{Topic}_{(L-2)\beta_i}
\end{align*}
\]
In these formulas, $Age$ represents the age of the user, $Province$ represents the province of the user. Hamiltonian Monte Carlo (HMC) sampling is used to estimate the unknown parameters $\beta$. Define Hamiltonian function $H(\beta, r)$ as (10):

$$H(\beta, r) = E(\beta) + \frac{1}{2} r^T M^{-1} r$$  (10)

Where $E(\beta)$ is potential energy function, $\frac{1}{2} r^T M^{-1} r$ is kinetic energy function, $r$ is momentum (auxiliary parameter), $M$ is a mass matrix (usually set as a unit matrix). The target distribution for sampling in this scenario is $P(\beta | Data) \propto \exp[E(\beta)]$, potential energy function $E(\beta)$ was expressed as (11), $LL(\beta)$ is the logarithm of likelihood function and $P(\beta)$ is the prior function.

$$E(\beta) = -LL(\beta) - \log P(\beta)$$

$$= - \sum_{n=1}^{N} \sum_{j=1}^{J} 1\{y^{(n)} = j\} \ln \frac{\exp(x^{(n)} \beta_j)}{\sum_{j=1}^{J} \exp(x^{(n)} \beta_j)} - \frac{1}{2} \sum_{j=1}^{J} (\beta_j - \mu_j)^T (\Sigma_j)^{-1} (\beta_j - \mu_j)$$  (11)

The posterior distribution of mean $\mu_j$ and covariance $\Sigma_j$ in (11) obeys the normal Inverse-Wishart distribution (12).

$$P(\mu_j, \Sigma_j | \beta_j) = N(\mu_j | \mu'_0, \rho' \Lambda') IW(\Sigma_j | v', \Lambda')$$  (12)

Where

$$\mu'_0 = \frac{\rho}{\rho + J} \mu_0 + \frac{J}{\rho + J} \bar{\beta}$$

$$\Lambda' = \Lambda_0 + \frac{J}{\rho + J} (\beta_j - \bar{\beta}) (\beta_j - \bar{\beta})^T + \frac{\rho J}{\rho + J} (\bar{\beta} - \mu_0) (\bar{\beta} - \mu_0)^T$$

$$\rho' = \rho + J$$

$$v' = v + J$$

$$\bar{\beta} = \sum_{j=1}^{J} \beta_j$$

Compute partial derivative about $E$ with respect to $\beta$ in order to update $\beta$, the formula is as (13).

$$\nabla E(\beta) = \sum_{j=1}^{J} \left( (\Sigma_j)^{-1} (\beta_j - \mu_j) + (\beta_j - \mu_j)^T (\Sigma_j)^{-1} \right) - \sum_{n=1}^{N} \sum_{j=1}^{J} 1\{y^{(n)} = j\} \frac{\exp(x^{(n)} \beta_j)}{\sum_{j=1}^{J} \exp(x^{(n)} \beta_j)}$$  (13)

Parametric update pseudocode as following table 1.

| Algorithm: Hamiltonian Monte Carlo |
|-----------------------------------|
| **Input:** Initial value $\beta^{(0)}$ and step size $\varepsilon$ |
| **Output:** Updated parameters $\beta^{(t+1)}$ |
| For $t = 1, 2, \ldots$, do |
| $(\beta_t, r_t) = (\beta^{(t)}, r^{(t)})$; $r_t \leftarrow r_t - \frac{\varepsilon}{2} \nabla E(\beta_t)$ |
| For $i = 1$ to $m$, do |
| // The parameters are iteratively updated using formula (13) |
| $\beta_i \leftarrow \beta_{i-1} + \varepsilon M^{-1} r_{i-1}$ |
\[ r_i \leftarrow r_{i+1} + e \nabla U(\beta_i) \]

End for

\[ r_m \leftarrow r_m - \frac{e}{2} \nabla E(\beta_m) ; \left( \hat{\beta}, \hat{r} \right) = (\beta_m, r_m) \]

\[ \psi \sim \text{Uniform}(0,1) \]

//Metropolis-Hastings amendment; decide whether to accept or not based on formula (10)

if \[ \psi < \min \left( 1, \exp \left( H(\hat{\beta}, \hat{r}) - H(\beta^{(i)}, r^{(i)}) \right) \right) \], then \[ \beta^{(i+1)} = \hat{\beta} \]

else \[ \beta^{(i+1)} = \beta^{(i)} \]

end for

end for

3. Experiment and discussion

3.1 Data

The data in this paper come from “Auto Home”, the most valuable Internet automobile marketing platform in China. The data include demographic information of consumers (such as age, region, etc.), online reviews, purchase information (such as brand, vehicle type, car line, purchase time, etc.), and interactive comment and friends. The data set includes 7021 consumers, 37 brands, 10 vehicle type and 145 car line.

3.2 Experimental result and analysis

Obtain fine-grained product feature knowledge based on human-computer interaction, divide fine-grained product features into 10 feature classes. Table 2 shows the 10 feature classes constructed in this paper and some corresponding product feature knowledge. For example, in feature class “price/performance ratio” include “price”, “cost price”, “preferential”, “cheap”, and etc. terms to describe the fine-grained characteristics of product.

| Feature Classes          | Terms Intra-class                                           |
|--------------------------|------------------------------------------------------------|
| Price/Performance Ratio  | price, cost price, preferential, level, cheap, car tax, quality, practicality, economy, etc. |
| Appearance               | color, glass, wheel hub, windshield sticker, wiper, door handle, taillights, headlights, etc. |
| Trim                     | dashboard, camera, disc box, navigator, cigarette lighter, steering wheel, backrest, etc. |
| Space                    | trunk, vehicle volume, capacity, vehicle width, lateral space, head space, internal space, etc. |
| Control                  | gear shift, steering gear, remote control, player, control line, automatic gear, manual gear, etc. |
| Power                    | engine, acceleration, start, horsepower, torque, clutch, spark plug, idle speed, speed, etc. |
| Fuel                     | fuel, average fuel consumption, instantaneous fuel consumption, fuel saving, urban fuel consumption, etc. |
| Safety                   | safety belt, air bag, anti-theft, ABS, anti-lock system, tire pressure monitoring system, etc. |
| Customer                 | customer service, maintenance, repair, 4S, the first insurance, insurance, time cost, etc. |
| Service                  | disassembly, etc.                                        |
| Comfort                  | comfortable, shock absorption, sound insulation, noise, sound, wind drying, wind drying, etc. |

LDA model is used to obtain consumers' implicit topic preferences, and to depict their purchase motivation. The comparison of attention and implicit topic preference between two users is shown in Fig. 2. It can be seen from subgraph (a), user 1 pays more attention on power and cost performance, while user 2 pays more attention on control, comfortable and power. These reflect the different preferences of the two consumers. Subgraph (b) shows the comparison of two users' implicit topic preferences in the case of topics number is 5. User 3 pays more attention on Topic1 and Topic5, while
user 4 pays more attention on Topic2 and Topic3, these reflect the different purchase motivations of the two consumers.

3.3 Prediction and Analysis of Consumer’s Car Purchase Behavior

In order for convenience of expression, simplify the names of different methods. Discrete selection model is recorded as MNL, hierarchical discrete selection model is recorded as HMNL. The benchmark algorithm for comparison in this paper is Maximum Likelihood Estimation Multivariate Discrete Selection Model, which is recorded as MLEMNL. U_MLEMNL, U_MNL, U_HMNL are denoted the models constructed using consumers' explicit features preferences respectively. UF_MLEMNL, UF_MNL, UF_HMNL are represent the models constructed using features preferences which engaging interactive comments. Table 3 shows the results in case of topics number is 5.

| Model                  | Explicit features preferences | Implicit topic preference (5 Topic) |
|------------------------|-------------------------------|------------------------------------|
|                         | Accuracy | Precision | F_Score | Accuracy | Precision | F_Score |
| U_MLEMNL                | 0.0652   | 0.1462    | 0.0153  | 0.1182   | 0.2041    | 0.031   |
| U_MNL                  | 0.0805   | 0.1645    | 0.0161  | 0.1275   | 0.2217    | 0.0252  |
| U_HMNL                 | 0.0954   | 0.1778    | 0.0239  | 0.1417   | 0.2392    | 0.0382  |
| UF_MLEMNL              | 0.0786   | 0.1579    | 0.0193  | 0.1296   | 0.2163    | 0.0321  |
| UF_MNL                | 0.0959   | 0.1726    | 0.0231  | 0.1225   | 0.2425    | 0.0303  |
| UF_HMNL              | 0.1132   | 0.1982    | 0.0339  | 0.1475   | 0.2753    | 0.0412  |

It could be found that hierarchical models performed better, all 3 indexes (Accuracy, Precision, F_Score) were improved. This is because the hierarchical model introduces more information. Using implicit topic preference in predict consumers’ automobile purchase performed better. Integrating interactive comments has a positive effect on predicting consumer’ purchase behavior. In Fig.3 it can be seen clearly, integrating interactive comments on hierarchical model is better on both preference model, due to more consumers information was added.
4. Conclusion

Based on the results and discussions presented above, the conclusions are obtained as below:

1. The hierarchical MNL model considers multi-granularity information such as brand, vehicle type and car line etc., which can achieve multi-granularity purchase decision predicting.

2. Implicit topic preference is more useful than explicit features preference, as it can better reflect consumers' purchase motivation.

3. Integrating interactive comments can increase the insight on consumers, thereby have positive effect in analyzing and predicting consumers' purchasing behavior.

Acknowledgments

This work was financially supported by the Post-doctoral Research Fund of Hefei University of Technology (JZ2018HGBH0281).

References

[1] Allenby, G.M. and P.E. Rossi, Marketing models of consumer heterogeneity. Journal of econometrics, 1998. 89(1): p. 57-78.

[2] Boxall, P.C. and W.L. Adamowicz, Understanding heterogeneous preferences in random utility models: a latent class approach. Environmental and resource economics, 2002. 23(4): p. 421-446.

[3] Rossi, P.E., R.E. McCulloch, and G.M. Allenby, The value of purchase history data in target marketing. Marketing Science, 1996. 15(4): p. 321-340.

[4] Hackbarth, A. and R. Madlener, Consumer preferences for alternative fuel vehicles: A discrete choice analysis. Transportation Research Part D: Transport and Environment, 2013. 25: p. 5-17.

[5] Qian, L. and D. Soopramanien, Heterogeneous consumer preferences for alternative fuel cars in China. Transportation Research Part D: Transport and Environment, 2011. 16(8): p. 607-613.

[6] Sanko, N., et al., Household car and motorcycle ownership in Bangkok and Kuala Lumpur in comparison with Nagoya. Transportmetrica A: Transport Science, 2014. 10(3): p. 187-213.

[7] Wang, M., et al., Analyzing and Predicting Heterogeneous Customer Preferences in China's Auto Market Using Choice Modeling and Network Analysis. SAE International Journal of Materials and Manufacturing, 2015. 8(2015-01-0468): p. 668-677.

[8] Xu, L., et al., Price advertising by manufacturers and dealers. Management Science, 2014. 60(11): p. 2816-2834

[9] Darley, W.K., C. Blankson, and D.J. Luethge, Toward an integrated framework for online consumer behavior and decision making process: A review. Psychology & marketing, 2010. 27(2): p. 94-116.

[10] Liu, Y., J. Wang, and Y. Jiang, PT-LDA: A latent variable model to predict personality traits of social network users. Neurocomputing, 2016. 210: p. 155-163.

[11] Vermeulen, I.E. and D. Seegers, Tried and tested: The impact of online hotel reviews on consumer consideration. Tourism management, 2009. 30(1): p. 123-127
[12] Kim, Y. and R. Krishnan, On product-level uncertainty and online purchase behavior: An empirical analysis. Management Science, 2015. 61(10): p. 2449-2467.
[13] Zhao, Y., et al., Modeling consumer learning from online product reviews. Marketing Science, 2013. 32(1): p. 153-169.
[14] Moon, S., et al., Consumers’ pre-launch awareness and preference on movie sales. European Journal of Marketing, 2016. 50(5/6): p. 1024-1046
[15] Li, X., M. Wang, and T.-P. Liang, A multi-theoretical kernel-based approach to social network-based recommendation. Decision Support Systems, 2014. 65: p. 95-104.