Latent Class Analysis and Application for O2O Service User Classification: Postprint

Authors: Pingfeng Liu, Wang Bei, Lei Jie

Date: 2017-10-11T00:00:00+00:00

Abstract

Objective: To achieve more objective and accurate user classification in the O2O (Online to Offline) model and formulate corresponding service strategies for different user groups. Method: An O2O user classification model based on Latent Class Analysis (LCA) was designed, the LCA method was employed to classify users, and food and beverage group-buying O2O was used as a case study to verify the simplicity and efficiency of applying LCA to O2O user classification. Results: Users were divided into four categories: potential, loyal-curious, cautious, and discriminating types. For each user type, latent characteristics and group types were analyzed, and corresponding marketing strategies were proposed. Limitations: The study employed a binary classification method for user features, manual processing of source data, and the selection of boundary thresholds in binary classification is subjective. Conclusion: LCA can achieve O2O user classification and precision marketing, expanding the application scope of latent class models.

Full Text

Preamble

Latent Class Analysis and Application for O2O Service User Classification

Liu Pingfeng, Wang Bei, Lei Jie
(School of Economics, Wuhan University of Technology, Wuhan 430070, China)

Abstract

[**Objective**] This study aims to achieve more objective and precise user classification under the Online to Offline (O2O) model, enabling the development of tailored service strategies for different user groups. [**Methods**] We designed an O2O user classification model based on Latent Class Analysis (LCA), using

LCA to categorize users. Taking catering group-buying O2O as an example, we validated the simplicity and efficiency of applying LCA to O2O user classification. [**Results**] Users were classified into four types: Potential, Loyal-Curious, Cautious, and Picky. For each user type, we analyzed their latent characteristics and group profiles, and proposed relevant marketing strategies accordingly. [**Limitations**] The study employed a binary classification method for user features, requiring manual preprocessing of source data, where the selection of cutoff boundaries involved subjectivity. [**Conclusions**] LCA can effectively achieve O2O user classification and precision marketing, expanding the application scope of latent class models.

Keywords: O2O; Latent Class Analysis; User Classification; Marketing
Classification Numbers: F713.36; G35

With the development of mobile internet, O2O services can provide customers with enhanced experiences through online-offline collaboration, achieving rapid growth in catering, tourism, entertainment, retail, wedding, and automotive industries. Since real-time acquisition of user transaction and behavioral data enables one-to-one precision marketing in e-commerce, this approach has been widely adopted online. However, obtaining real-time offline information in O2O services remains challenging, making one-to-one precision marketing difficult for O2O users. A feasible alternative is to classify O2O users based on their online and offline characteristics, then develop different marketing strategies for different user categories.

Although numerous studies have addressed internet user classification, most employ methods such as K-means [**?**], K-medoids [**?**], and KNN [**?**], where the K value and number of clusters are subjectively set based on experience without accurate justification. Moreover, domestic researchers have focused primarily on behavioral and demographic dimensions [**?**] for user classification. For instance, Guan et al. [**?**] established classification models based on user behavioral attributes and basic characteristics, but lacked analysis of psychological features, resulting in incomplete user profiling. Some studies have used structural equation modeling [**?**] and factor analysis [**?**] for statistical analysis of user characteristics. However, O2O user attribute variables (such as psychological features) are categorical, which structural equation modeling and factor analysis cannot handle effectively. Therefore, methods capable of processing categorical variables are needed for O2O user classification. Furthermore, O2O services constitute an emerging research field, with limited domestic research on O2O user classification that fails to meet classification demands.

Latent Class Analysis (LCA) can handle categorical variables without requiring pre-specification of the number of categories, thus avoiding subjectivity in determining classification numbers. LCA has been widely applied to diagnose psychological, emotional, and behavioral issues [**?**] to study group heterogeneity. This paper draws on LCA' s advantages and applies them to O2O user classification, establishing an optimal latent class model based on O2O users' psychological and behavioral characteristics. This approach achieves precise user classification

based on objective data and proposes corresponding marketing strategies according to classification results, addressing the challenge of integrating online and offline resources. Using real survey data, this study validates the effectiveness of applying LCA to O2O user classification.

2. O2O User Feature Analysis and Attribute Extraction

The O2O (Online to Offline) concept was proposed by Rampell in August 2010, and no unified definition has yet emerged. In 2013, Zhang Bo [**?**] suggested that O2O represents a new business model in the mobile internet era where online and offline interactions occur in lifestyle consumption domains, extending to various industries. The O2O model involves online transactions or reservations for offline consumption experiences, with the specific consumption process illustrated in Figure 1 [Figure 1: see original paper]. For example, a user searches for hotpot restaurant information on Meituan, purchases a group-buying voucher, completes payment online, receives a verification code, visits the physical store within the validity period, presents the code for consumption, and may optionally leave a review after dining. Unlike traditional online shopping, O2O includes the crucial offline consumption component, which represents the core experience that cannot be controlled online.

O2O services have transformed people' s lifestyles, user characteristics [**?**], and demands. O2O service users exhibit the following features:

(1) **Fragmentation.** In the mobile internet era, technological advances such as smartphones, 4G networks, QR codes, and electronic vouchers have led to fragmented information dissemination, acquisition methods, and individual consumption behaviors. Users rapidly switch between various information sources, resulting in fragmented marketing channels (Weibo, WeChat, forums, video sites, subways, billboards, flyers, etc.) to capitalize on users' fragmented time. Future O2O will integrate offline micro-traffic, delivering appropriate products (services) in appropriate locations, at appropriate times, in appropriate quantities, and at appropriate prices to appropriate users. Based on time fragmentation and channel fragmentation characteristics, we selected users' consumption time and product information acquisition channels as user attributes.

(2) **Online-Offline Integration.** O2O marketing channels and payment methods involve both online and offline components. User experience, sharing, and O2O loop completion result from combined online-offline interactions. Therefore, we selected payment methods and user source channels as user attributes.

(3) **Convenience Orientation.** Advances in mobile payment and location technologies have made payments and electronic voucher acquisition convenient and fast, promoting O2O development and establishing "point-to-point" relationships between merchants and users. This enables users to make reservations or purchases anytime and anywhere, while allow-

ing service providers to update product or activity information and conduct low-cost marketing through multiple channels. This convenience in consumption and information acquisition increases user consumption frequency, making consumption frequency a relevant user attribute.

(4) **Emotional Tendencies.** Under the new O2O business model, individual behaviors tend to trust subjective feelings and friends' recommendations rather than rational thinking. For instance, users with strong curiosity may consume upon receiving push notifications about new offerings; users feeling tired or hungry during shopping may place orders when receiving restaurant promotional offers; and users may consume based on friends' recommendations. In these scenarios, product distance, price, personal income, attitudes toward new things, and attitudes toward others' evaluations all influence consumption decisions. Therefore, we selected these influencing factors as user attributes.

Based on O2O user characteristics and existing research [**?**], this study extracted corresponding user attributes across three dimensions: basic attributes, behavioral attributes, and psychological attributes, as detailed in Table 1 .

3.1. User Classification Process

The LCA-based O2O user classification primarily involves selecting relevant attributes based on O2O user characteristics to establish a Latent Class Model (LCM), fitting the LCM using model fitting methods in Mplus, selecting the optimal model, and conducting parameterization and classification analysis. Figure 2 [Figure 2: see original paper] illustrates the classification flowchart.

3.2. O2O User Classification Modeling Based on LCA

Based on the selected O2O user attributes, we established an O2O user classification model using LCA, as shown in Figure 3 [Figure 3: see original paper]. LCA operates under the fundamental assumption of local independence among manifest variables (user attributes). The LCM starts from a zero model that assumes complete independence among manifest variables, gradually increasing the number of latent categories, comparing model evaluation indices to determine the optimal model. The latent class variable is then used to explain associations among manifest variables, thereby maintaining local independence.

The LCA mathematical model is shown in formula (1) [**?**]:

$$\pi_{ijk}^{ABC} = \sum_{t=1}^{T} \pi_t^X \pi_{it}^{A|X} \pi_{jt}^{B|X} \pi_{kt}^{C|X}$$

where X is the latent variable with $t(t = 1, 2, \ldots, T)$ categories; A, B, C are three user attributes with values i, j, k respectively. π_{ijk}^{ABC} represents the joint probability of the latent class model. π_t^X is the probability that latent variable

X belongs to category t, with $\sum_{t=1}^{T} \pi_t^X = 1$. $\pi_{it}^{A|X}$ represents the conditional probability of $A = i$ given $X = t$, i.e., $P(A = i|X = t)$, and similarly for others.

3.3. Model Fitting and Parameter Estimation Methods

For exploratory LCA, the number of latent categories is typically determined based on statistical indices such as Pearson-$\chi^2, likelihoodratiochi - square, AIC(AkaikeInformationCriterion), andBIC(BayesianInformationCriterion)$[?].$AICandBICaret$ $SizeAdjustedBIC), Pearson- ^{2, andEntropyasfittingevaluationindices.AIC, BIC, SSBIC, andPearson-}$ 2$ values are better when smaller. Entropy measures the uncertainty of an event, with larger values indicating greater uncertainty.

Parameter estimation for latent class models primarily uses Maximum Likelihood (ML) methods, with various algorithms for the iterative process. Mplus software defaults to MLR (Maximum Likelihood Estimator with Robust Standard Errors) for dichotomous variables, which is a modified marginal maximum likelihood estimation. MLR uses Sandwich estimation for standard errors, making it suitable for non-normal data and small sample estimation. Specific explanations of MLR can be found in reference [?] and the Mplus User' s Manual [?].

Based on parameter estimation results, O2O service users are classified to provide appropriate O2O social marketing and classification management strategies for enterprises. The specific classification steps are: (1) Calculate the joint probability of the user latent class model according to formula (1); (2) Conduct latent classification for each individual based on Bayesian posterior probability theory, calculated as follows: π_{tijk} is the latent class probability, with $\sum_{t=1}^{T} \pi_{tijk} = 1$, where π_{tijk} represents the probability that X belongs to latent category t and each user attribute is at levels i, j, k respectively. Formula (2) determines which category a user belongs to.

4.1. Source Data Collection and Preprocessing

This study utilized survey data on catering group-buying O2O user behavior. Since catering O2O involves a broad population and income attributes have minimal impact, this attribute was removed during questionnaire design. We employed an online questionnaire survey method, designing and generating the questionnaire through the "Questionnaire Star"platform and distributing the survey link via QQ groups and WeChat Moments to classmates, teachers, friends, and relatives. The survey was conducted from April 22 to April 26, 2015, collecting 338 responses, of which 327 were valid (96.75% validity rate).

The collected data were preprocessed according to selected user attributes. Based on value distributions, relevant options were merged. For example, in Question 3 of the survey, options "1-2 times per month," "rarely used," and "never used" were merged into "less than 2 times," while "more than 2 times" was changed to "2 times or more," forming Question B in Table 2 . Question 6

options were similarly processed. For convenience in statistical analysis, further preprocessing was conducted: gender "male" was coded as "0" and "female" as "1" ; monthly usage frequency "2 times or more" as "0" and "less than 2 times" as "1" ; and Questions C, D, E, F, G, H, I were similarly coded.

4.2. Model Fitting and Selection

Mplus is the mainstream software for latent class analysis, so we selected Mplus [**?**] to conduct LCA on O2O user data. Table 3 presents the relevant fitting indices.

Since all indices are better when smaller, and following Lin et al.'s [**?**] recommendation that AIC is preferable for samples below 1,000, we selected the four-class solution based on minimum AIC. The SSBIC value was also smallest for four classes, while Pearson-χ^2 did not improve significantly with five classes. Considering model parsimony, the four-class model was selected, with the Entropy value also within an acceptable range. Based on comprehensive evaluation of all indices, the four-class model was chosen as optimal.

4.3. Model Parameter Estimation

Following the four-class latent class analysis model, we ran Mplus software to estimate latent class probabilities and conditional probabilities for each item, with results shown in Table 4 . The results indicate that in Question G, result 1 represents price-conscious users, who account for high proportions across all four classes, suggesting surveyed users are relatively price-sensitive. In Question C, result 0 represents online payment, with 87.4%, 78.7%, and 82.8% of Class1, Class2, and Class4 respectively choosing online payment. In Question B, Class2 and Class4 show 82.6% and 62.8% respectively with monthly consumption frequency greater than two times. Meanwhile, in Question A, Class2 and Class4 show female proportions of 79.5% and 56.9% respectively. Correspondingly, Class1 and Class3, which have higher proportions of monthly consumption frequency less than two times, show higher male proportions in Question A, indicating that females have higher monthly consumption frequency than males.

4.4. Latent Classification Results

After parameterizing the optimal model, we calculated posterior probabilities for classification into each category based on the obtained latent class probabilities and conditional probabilities for each item. Using formulas (1) and (2), classification probabilities were computed, and classification results were obtained based on the principle of maximum posterior probability, as shown in Table 5 . Class1-Class4 contained 119, 55, 24, and 129 individuals respectively.

4.5. Results Discussion

Specifically, 36.4% of respondents belong to Class1, which is predominantly male (61.5%), with 60% having monthly consumption frequency below two times, 89.1% obtaining product or service information through promotional offers, 93.3% consuming on non-workdays, and most not paying attention to distance or reviews but with 63.4% being price-conscious. This group' s characteristics include low consumption frequency but price sensitivity, with information sources being promotions and offers. With proper marketing, this group shows significant consumption potential and can be defined as "Potential" users. 16.8% of respondents belong to Class2, which differs from Class1 in being predominantly female (79.5%), with 82.6% having monthly consumption frequency greater than two times. This group shows 100% price consciousness, high consumption frequency, and willingness to accept new things, and can be defined as "Loyal-Curious" users. 7.3% of respondents belong to Class3, whose distinguishing features include 79.4% consuming on workdays, 76.7% obtaining product or service information through friends' recommendations, 55.1% choosing offline payment (significantly higher than other classes), and 85.4% paying attention to reviews to avoid risks. This group can be defined as "Cautious" users. 39.5% of respondents belong to Class4, which differs little from Class2 in most attributes but notably shows high attention to service/product price, distance, and reviews, making them more "Picky" users.

Table 5 presents the classification results based on the model, which have underlying classification reasons. According to the results in Table 5, we analyzed the latent characteristics and group types for different categories and proposed relevant marketing strategies, as detailed in Table 6 .

This study classified O2O users incorporating psychological and behavioral characteristics using LCA, employing objective probabilities as judgment criteria to analyze latent group heterogeneity. This approach compensates for the arbitrariness of cluster analysis, yielding more objective and realistic classification results that address the challenge of integrating online and offline resources. The results provide objective foundations for enterprise marketing and management, help improve user experience, and enhance service provider operational efficiency. Moreover, Mplus software offers simplicity, rapid processing, and comprehensive capabilities for handling latent class models. However, LCA assumes local independence among manifest variables, which is difficult to achieve in practice and requires model constraints, representing a technical area for improvement in LCM. Future work should focus on selecting the most objective indices to determine optimal models for studying O2O user classification.

References

[**?**] Rahman M A, Islam M Z. A Hybrid Clustering Technique Combining a Novel Genetic Algorithm with K-Means [**?**]. Knowledge-Based Systems, 2014, 71: 345-365.

[?] Zadegan S M R, Mirzaie M, Sadoughi F. Ranked K-medoids: A Fast and Accurate Rank-based Partitioning Algorithm for Clustering Large Datasets [?]. Knowledge-Based Systems, 2013, 39: 133-143.

[?] Ma Qing, Xie Juanying. New K-medoids Clustering Algorithm Based on Granular Computing [?]. Journal of Computer Applications, 2012, 32(7): 1973-1977.

[?] Li Wei. The Research on Mobile Internet User Behavior [?]. Beijing: Beijing University of Posts and Telecommunications, 2013.

[?] Hu Muhai, Cai Shuqin. The Model of Mobile Customer Segmentation Based on Hypergraph of Context Preference Knowledge [?]. Chinese Journal of Management, 2011, 8(10): 1509-1516.

[?] Guan J F, Dai Y, Zhang M, et al. Evaluation Research for Internet Users/Services Attributes Extraction and Classification [?]. The Journal of China Universities of Posts and Telecommunications, 2013, 20(S1): 81-85.

[?] MacCallum R C, Austin J T. Applications of Structural Equation Modeling in Psychological Research [?]. Annual Review of Psychology, 2000, 51: 201-226.

[?] Bartholomew D J, Knott M. Latent Variable Models and Factor Analysis [?]. The 2nd Edition. Edward Arnold, 1999.

[?] Kang J, Ciecierski C C, Malin E L, et al. A Latent Class Analysis of Cancer Risk Behaviors Among US College Students [?]. Preventive Medicine, 2014, 64: 121-125.

[?] Li Zhihua, Yin Xiayun, Cai Taisheng, et al. Latent Class Analysis of the Characteristics of Left-Behind Children' s Emotional and Behavioral Problems: The Person-Centered Perspectives [?]. Journal of Psychological Science, 2014, 37(2): 329-334.

[?] Zhang Jieting, Jiao Can, Zhang Minqiang. Application of Latent Class Analysis in Psychological Research [?]. Advances in Psychological Science, 2010, 18(12): 1991-1998.

[?] Jackson N, Denny S, Sheridan J, et al. Predictors of Drinking Patterns in Adolescence: A Latent Class Analysis [?]. Drug and Alcohol Dependence, 2014, 135: 133-139.

[?] Ward R M, Cleveland M J, Messman-Moore T L. Latent Class Analysis of College Women' s Thursday Drinking [?]. Addictive Behaviors, 2013, 38(1): 1407-1413.

[?] Zhang Bo. O2O-Commercial Revolution of the Mobile Internet Era [?]. Beijing: China Machine Press, 2013: 11.

[?] Tai C L, Hong J Y, Chang C M, et al. Determinants of Consumer' s Intention to Participate in Group Buying [?]. Procedia-Social and Behavioral Sciences, 2012, 57: 396-403.

[?] Ahn T, Ryu S, Han I. The Impact of the Online and Offline Features on the User Acceptance of Internet Shopping Malls [?]. Electronic Commerce Research and Applications, 2005, 3(4): 405-420.

[?] Zhang Yufeng, Zhou Lei, Yang Wei, et al. Research on Customer Perceived Risks by Customers in E-Commerce Group Buying [?]. Information Science, 2011, 29(10): 1505-1508.

[?] Zhang Chunxia. Research on Consumption Characteristics and Psychology of Group-Buying2.0 [?]. Beijing: Beijing University of Posts and Telecommunications, 2012.

[?] Kumar N, Benbasat I. Research Note: The Influence of Recommendations and Consumer Reviews on Evaluations of Websites [?]. Information System Research, 2006, 7(4): 425-439.

[?] Liu Yingzi, Wu Hao. A Summarization of Customer Segmentation Methods [?]. Journal of Industrial Engineering and Engineering Management, 2006, 20(1): 53-57.

[?] Hagenaars J A, McCutcheon A L. Applied Latent Class Analysis [?]. Cambridge University Press, 2002: 56-58.

[?] Lin T H, Dayton C M. Model Selection Information Criteria for Non-Nested Latent Class Models [?]. Journal of Educational and Behavioral Statistics, 1997, 22(3): 249-264.

[?] Liu Hongyun, Luo Fang, Wang Yue, et al. Item Parameter Estimation for Multidimensional Measurement: Comparisons of SEM and MIRT Based Methods [?]. Acta Psychologica Sinica, 2012, 44(1): 121-132.

[?] Mplus User' s Guide [?]. http://www.statmodel.com/.

Author Contributions: Liu Pingfeng: Conceived research ideas, designed research framework, revised final manuscript; Wang Bei: Drafted manuscript, conducted experiments; Wang Bei, Lei Jie: Collected, cleaned, and analyzed data.

Conflict of Interest Statement: All authors declare no conflict of interest.

Supporting Data: Supporting data is available in the journal' s online version at http://www.infotech.ac.cn.
[?] Wang Bei. 338_{338}2.xls. Survey data on catering O2O user behavior.
[?] Wang Bei. Attachment 1: Catering O2O User Behavior Survey.doc. Catering O2O user behavior questionnaire.

Received: August 3, 2015
Revised: January 15, 2016

Note: Figure translations are in progress. See original paper for figures.