DOI: http://dx.doi.org/10.26483/ijarcs.v16i6.7372

Volume 16, No. 6, November-December 2025

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

CUSTOMER CLASSIFICATION MODELING USING LDA CONSIDERING SPATIAL AREA STRUCTURE

Hikari Ishihara
Graduate School of Science and Engineering
University of the Ryukyus
1 Senbaru Nishihara,Okinawa, Japan

Takeo Okazaki
Faculty of Engineering
University of the Ryukyus
1 Senbaru Nishihara,Okinawa, Japan

Abstract: To improve operational efficiency in the driver service industry, it is necessary to understand customer behavioural trends and characteristics. In this study, we propose a customer clustering method using customer data accumulated in a chauffeur service dispatch app, based on Latent Dirichlet Allocation (LDA), which incorporates spatial area structure as a feature. Specifically, we combine DBSCAN and GMM to zone areas with a high concentration of order locations and generate features that represent the area. Furthermore, we design a dataset that combines basic information such as customer age and time of use and extract potential customer groups using LDA. Finally, we evaluate the validity of the classification results and consider their applicability to service improvements.

Keywords: Customer Classification, Latent Dirichlet Allocation (LDA), Spatial area structure, DBSCAN, Gaussian Mixture Model (GMM)

1. INTRODUCTION

The designated driver service has long suffered from inefficient business practices and intense price competition. These challenges are particularly evident in dispatch management. For example, the time between a customer requesting a service and the designated driver vehicle arriving is often excessively long, creating a mismatch between supply and demand. A key underlying issue is the lack of sufficient customer analysis based on past usage history. This results in an inadequate understanding of demand in terms of both region and time-of-day. Consequently, optimal dispatch strategies tailored to customer characteristics or effective promotional measures are difficult to devise. This study aims to address these challenges by constructing a clustering model that comprehensively captures customer usage patterns and characteristics using actual customer data from the designated driver service. Specifically, the study proposes performing more realistic customer segmentation by considering not only attributes such as customer gender, age group, and order time, as well as the spatial distribution of order locations. To achieve this, the study proposes using Latent Dirichlet Allocation (LDA) for customer classification. This approach captures the characteristics of each customer segment derived from the LDA analysis.

2. CUSTOMER CLASSIFICATION USING LDA

LDA is a probabilistic generative model developed for purposes such as document classification in natural language processing. However, its structure is highly versatile as a model for extracting underlying “factors” or “tendencies” from observed data, making it applicable to non-linguistic data as well. The customer data from the designated driver service, which is the subject of this research, lacks explicit confirmation of purpose or preference. Therefore, LDA is considered an effective method for estimating the underlying potential usage purposes and behavioral patterns.

Furthermore, as LDA is a probabilistic method, it allows for “soft clustering”, accounting for the possibility of each customer belonging to multiple groups. Compared to methods like k-means or hierarchical clustering which assign customers to a single fixed cluster, LDA can flexibly capture customers' multifaceted characteristics and usage patterns. Furthermore, while probabilistic latent semantic analysis (PLSA) is another method for handling latent structures, PLSA is prone to overfitting by excessively adapting to the training data, necessitating retraining when applied
to new observational data. In contrast, LDA assumes a Dirichlet distribution as its prior distribution, specifically as a consensus prior, making it superior for applying to new data and for model extensibility. Tsukai and Tsukano[2] aimed to demonstrate the usefulness of the topic model LDA for geographic information at sub-municipal aggregation levels (i.e. detailed geographic information). They extracted characteristics of this information at the Level 3 mesh level and conducted a comparative validation with the existing factor analysis method. The analysis revealed that while factor analysis yielded only three types of geographic characteristics, LDA analysis revealed eight. Furthermore, the spatial distribution observed for each mesh was confirmed to accurately capture the structure of actual cities. Kamiya, Fuse et al. [3] proposed a method for understanding regional population characteristics using LDA and its extension, HDP-LDA. Specifically, they applied a topic model to 500m mesh mobile spatial statistics data, treating each mesh as a document and visitors’ residences as words, in order to extract population characteristics by region. Thus, LDA is advancing beyond document analysis. When combined with spatial information and statistical data, it is increasingly applied as an analytical method to capture the latent characteristics of regions and groups from multiple perspectives. Similarly, this study attempts to capture regional, temporal, and attribute-based characteristics comprehensively by applying LDA to customer data related to designated driver services. When applying LDA to customer data, each customer is treated as a single

"document," and attribute information such as gender, age group, and order time is treated as "words," aiming to extract latent customer groups. This study analyzes customer data from Okinawa Prefecture (159,277 cases) collected between January 2024 and August 2025. For each attribute below, the occurrence frequency of attribute values per customer order is counted. This data is then represented as a Bag-of-Words (BoW) vector format before being applied to LDA.

Table 1: Overview of customer data set

Item	Details
Gender	Male、Female
Age	Ages 10 to 80 (10-year increments)
Order time	0:00 to 23:00 (in one-hour intervals)
Weekdays/Holidays	Weekdays / Weekends & holidays / Day before weekends & holidays
Departure Point	Departure Location
Distance(km)	The distance from the departure point to the destination.
Weather	Weather information for the day

The preprocessing of each variable in the above Table 1 was conducted according to the following procedures.
First, Usage Time was extracted directly from the Order Time present in the order data. To determine the Age Group, the age was calculated by computing the duration between the customer's Date of Birth (from customer data) and the Order Date (from order data). The distinction for Weekday/Holiday was made by judging whether the Order Date fell on a weekday, a Weekend/Public Holiday, or the Day before a Weekend/Public Holiday. For the Origin variable, the region was specified by converting the provided Latitude and Longitude data into an Area Mesh. Regarding Distance, the distance value in kilometers (km) was first calculated using the latitude and longitude data of the origin and destination. This continuous variable was subsequently categorized into 2km intervals, starting from a 3km baseline (e.g. 3-5km, 5-7km). Finally, Weather information was obtained by extracting the corresponding meteorological data for the Order Date from the publicly available records of the Japan Meteorological Agency, Ministry of Land, Infrastructure,
Transport and Tourism.
Furthermore, we adopt collapsed Gibbs sampling [4] as the method for estimating the posterior distribution of latent variables during LDA execution. While variational Bayesian methods are also available for LDA estimation, collapsed Gibbs sampling is chosen because it is relatively easy to implement and is known to provide high estimation accuracy.
Below is the equation for customer classification using LDA with collapsed Gibbs sampling.

$$P\left(z_j = k \middle| Z_{\bar{j}}, W\right) \propto \frac{N_{kd\bar{j}} + \alpha_k}{N_{d\bar{j}} + \sum_{k=1}^{K} \alpha_k} \cdot \frac{N_{kv\bar{j}} + \beta_v}{N_{k\bar{j}} + \sum_{v=1}^{V} \beta_v} \quad (6)$$

N_{kd} : Number of attributes assigned to group k of customer d.
N_{kv} : Number of occurrences of attribute v in group k.
N_k : Number of times group k appeared in group set z.
N_d : Number of attributes contained in customer d.

From the sampling probability formula (6), the parameters are efficiently estimated as follows.

$$\hat{\theta} = \frac{N_{kd} + \alpha}{N_d + K\alpha} \quad (7)$$

$$\hat{\phi} = \frac{N_{kv} + \beta}{N_k + V\beta} \quad (8)$$

3. HOW TO CHOOSE THE OPTIMAL NUMBER OF CLASSES

When performing customer classification using LDA, it is necessary to predefine the number of groups K. This study adopts Perplexity [1], one of LDA's evaluation metrics, as a method for determining the optimal number of groups. Generally, Perplexity tends to decrease as the number of groups increases. A lower value indicates higher model predictive performance and better representation of the data structure. However, an excessively large number of groups can cause overfitting, necessitating careful selection. This study selects the optimal number of classes using Perplexity through the following procedure.

Step1. Training and Perplexity Calculation per Topic Count
Set the number of groups incrementally from 2 to 20. Train the LDA model for each group count and calculate the corresponding perplexity value.
Step2. Smoothing Perplexity
Apply moving average smoothing to the calculated perplexity series. This process suppresses the influence of local noise and clarifies the trend in the relationship between topic count and perplexity.
Step3. Calculate Change Rate After Smoothing
Calculate the change rate between adjacent topic counts for the smoothed perplexity. Determine the optimal group count based on the values before and after the change rate.

4. INTERPRETATION OF EACH CUSTOMER GROUP PROFILE

Using the ϕ distributions obtained via LDA, we interpret the profiles of each group. However, due to differences in the number of attributes contained within each item, interpretation based solely on simple ϕ distributions risks overestimating the influence of certain categories.
Therefore, in this study, we performed weighting to correct for the imbalance in the number of attributes across categories. This was achieved by multiplying the ϕ distribution value for each attribute by a weight representing the number of attributes in the category to which that attribute belongs. After this weighting adjustment, the top 15 highest-scoring attributes within each group are identified as that group's defining characteristics and interpreted as the group profile.
Traditionally, various representation methods have been used when analyzing point data. Tomita et al. [5] attempted to extract representative tourist behavior patterns by performing asymmetric clustering based on the frequency of movement between predefined tourist areas, using GPS data from visitors to Yokohama. However, this method fixes the analysis areas in advance, making it difficult to fully account for the actual spatial density distribution of people flow or the geographical proximity between areas. Consequently, it becomes challenging

to capture continuous and flexible spatial behavior patterns, potentially failing to fully represent the diversity of actual behavior.

Additionally, regional meshes are commonly used as a simplified representation method for point data. This technique discretizes spatial data by converting it into grid-like areas formed by dividing latitude and longitude into fixed-size cells, making the data easier to handle.

In this study, we attempted customer classification using LDA by employing a regional mesh-based method for identifying departure points. Part of the customer profiles obtained from these results are shown below.

Table 2 Group Profile Example (Place names)

Group 1	Group 3
Under 3km	Okinawa City Moromizato 3
Female	Okinawa City Uechi 3
10:00 PM	Male
Ginowan City Futenma 2	Okinawa City Misato 2
9:00 PM	Okinawa City Goya 2
Cloudy	Okinawa City Hiyagon 2
Weekends & Holidays	Under 3km
11:00 PM	Cloudy
HaebaruTown Tsukayama	10:00 PM
Ginowan City Ganeko 4	Weekends & Holidays

Based on the above profile results, certain patterns were observed in attributes such as gender, age group, and usage time periods. On the other hand, regarding profiles related to departure locations, a challenge arises because the data is expressed as place names, making it difficult to grasp the proximity relationships between locations.

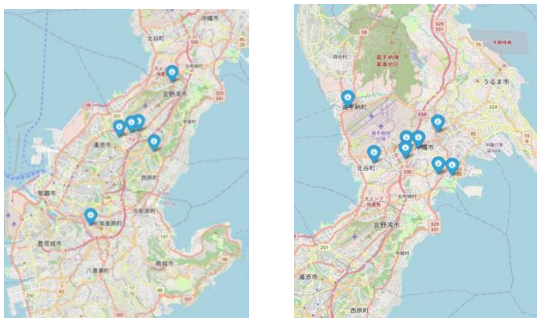

Fig. 1 Departure points for each group profile
(Left : Group 1, Right : Group 3)

5. CLASSIFICATION CONSIDERING SPATIAL AREA

Regarding departure points, methods such as regional meshes where the scope is fixed and granularity increases after division may make spatial interpretation of classification results difficult. Considering these challenges, a solution involves converting point data into “Areas” that account for spatial density and distribution spread.

Otokita [6] attempted to classify the purpose of stays based on mobile GPS data using a Gaussian Mixture Model (GMM). The GMM classification results confirmed that each cluster captured spatial spread and reflected differences in behavioral purpose.

Building on this approach, this study defines location data as flexible areas captured spatially, aiming to grasp more natural customer behavior patterns.

Generally, customers using designated driver services tend to concentrate in specific popular spots or high-frequency areas. For example, customers cluster in specific zones like major transit stations, downtown districts, business areas, and residential centers. In such areas, a natural central point exists, with customer activity spreading out from it. Therefore, this study examines a method for constructing broad areas based on a central point, assuming each area has one, when forming areas from customer order location data. To identify representative points (central points) in areas where customers concentrate, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [7] is adopted. DBSCAN is a clustering method that can automatically extract clusters based on the spatial density of data.

Each area extracted by DBSCAN contains multiple core points. In this study, the center point of area C_k is defined as the centroid c_k of the core point set C_k^{core} as follows.

$$c_k = \frac{1}{|C_k^{core}|} \sum_{x_i \in C_k^{core}} x_i \quad (3)$$

- $x_i = (x_i^{(1)}, x_i^{(2)})$: Coordinate vector for each core point

Next, we model the areas using a Gaussian Mixture Model (GMM) based on the center points c_k of each area obtained by DBSCAN. GMM is a method that represents data as a mixture of multiple normal distributions (Gaussian distributions), enabling the probabilistic definition of each area. This allows consideration of ambiguous boundaries between areas and the possibility of customers belonging to multiple areas, facilitating a flexible capture of customer behavior patterns within each area.

Below is the definition formula for the Gaussian distribution of each area using the center point c_k.

$$N\left(x \middle| c_k, \Sigma_k\right) = \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_k\right|}} \exp\left(-\frac{1}{2}(x - c_k)^T \Sigma_k^{-1} (x - c_k)\right) (4)$$

- c_k : Center point of Area C_k (Mean vector)
- $\sum_k$: Covariance matrix of Area C_k
- d : Number of dimensions of the data

Finally, for each area obtained via GMM, we quantitatively evaluate the spatial position of each order location within that area. Specifically, by considering spatial relationships—such as whether a location is “near the center of the area” or “close to the boundary”—we can more precisely capture user behavior patterns, including their spatial distribution.

To evaluate the spatial relationships among order locations within each area, we employ Mahalanobis distance. Mahalanobis distance considers the variance structure of each cluster and can measure, in a standardized manner, how far each point is from the area's center within a Gaussian distribution with elliptical contour lines. This enables a more accurate assessment of whether each order location falls within the “Center” , “Edge” or “Outside”.

Furthermore, it is known that the squared Mahalanobis distance $D_M(x)^2$ follows a chi-squared distribution with 2 degrees of freedom. Utilizing this property, the spatial relationships of each order location within the area are statistically classified as follows.

- Center : $D_M(x)^2 \leq \chi^2_{2,0.68}$ $(\sim 68\%)$
- Edge : $\chi^2_{2,0.68} < D_M(x)^2 \leq \chi^2_{2,0.95} (68 \sim 95\%)$
- Outside : $\chi^2_{2,0.95} < D_M(x)^2$ $(95\% \sim)$

In this way, by setting a statistical threshold for Mahalanobis distance that leverages the properties of the Gaussian distribution via GMM, we quantitatively evaluate whether each order location belongs to the “Center”, “Edge” or “Outside” of the area.

6. VERIFICATION OF APPLICATION EFFECT

Area zoning was used to create a distribution map of departure points, assigning corresponding area information to each order data point. Based on this area information, customer classification was performed using LDA. We verified whether this approach enables the extraction of more meaningful customer groups compared to the conventional mesh-based method. Fig. 2 visualizes the results of the area zoning performed using DBSCAN and GMM on a map. Each area is represented as a circle based on the mean position and variance of its constituent points. For DBSCAN parameters, a grid search was performed within the ranges $\varepsilon \in [0.01, 0.05]$ and MinPts $\in [10, 50]$. GMM was applied to the cluster results obtained for each parameter setting, selecting the combination yielding the minimum BIC value
($\varepsilon = 0.01$, MinPts = 23). The result is 24 Areas.

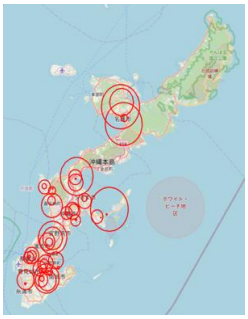

Fig. 2 Area Zoning Results (24 areas)

Fig. 3 shows the rate of change in the Perplexity value for each case when applying the customer data set to the LDA model and varying the number of groups from 2 to 20.

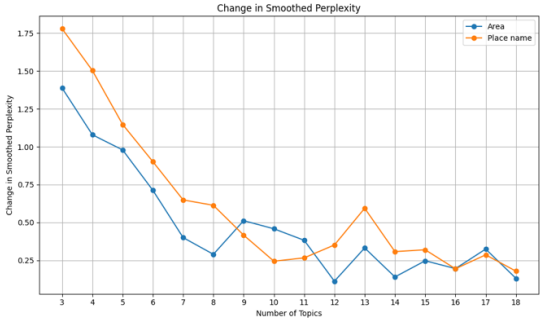

Fig. 3 Rate of change in the Perplexity

Figure 3 shows that the change rate value drops sharply when the number of groups is 10 for place name expressions and 8 for area expressions. Therefore, this study sets the number of groups to classify as 10 and 8, respectively.
The procedure for comparatively verifying the effectiveness of the two methods—the proposed "Area expression" and the "Conventional regional place name expression" is described below.

First, we conduct an "Evaluation of Group Profile Quality”.

1. Creation of Datasets Including Spatial Expression Data and Classification by LDA.
2. Creation of Profile Distributions from the Top 15 Attributes of Each Group (ϕ distribution)

Based on the calculated ϕ distribution for each group, extract the top 15 attributes with the highest probability. Use these attributes to conduct a qualitative evaluation of each group profile, focusing on aspects such as "spatial interpretability" and "balance with other features".
Calculate the proportion of spatial attributes within the entire profile distribution for each classification result.

Table 3 shows the proportion of spatial attributes among all attributes obtained by each representation method and the proportion of spatial attributes included in the profiles.

Table 3 Comparison Table of Spatial Attribute Ratios

D-point	Spatial Attribute Ratio (All Attributes)	Average Spatial Attribute Ratio (Profile)
Place name	0.91	0.42
Area	0.48	0.31

Table 3 confirms that when using place names, the number of spatial attributes is very high, and spatial attributes account for a large proportion within the group profile.

Conversely, when using area expressions, the reduced number of spatial attributes made it easier to extract non-spatial attributes such as era and time of day within the group profile.

Next, to compare the grouping results, we will examine representative group profiles.

Table 4 Example Group Profiles for Each Expression

Location expression	Group profile
Place names	Cloudy holiday, 9-11 PM: Male group using short-distance rides from Okinawa City (Moromizato3, Uechi3, Misato 3, Goya 2, Hiyagon 2, Hiyagon 6), Chatan Town (Kuwae), Kadena Town (Mizugama)
Area	Sunny or Cloudy holiday 10 PM – midnight: Male group using short-distance rides from Okinawa City area to Chatan Town area.

Based on the above results, when using place names, the large number of spatial attributes caused group profiles to skew toward spatial information, resulting in a tendency for non-spatial attributes such as age and usage time periods to be relatively underrepresented.

Conversely, when using area expressions, the departure points for each group were aggregated as geographically contiguous areas, ensuring spatial cohesion. Furthermore, the reduced number of spatial attributes facilitated the reflection of non-spatial attributes like age and usage time periods, resulting in more multifaceted and balanced group profiles.

These findings indicate that classifying customers by representing spatial information as areas achieves a balance in overall feature characteristics. Moreover, it allows for a broader and more flexible understanding of the approximate regions where each user group originates, thereby enhancing the informational value of the profiles for customer classification.

7. VALIDATION OF CUSTOMER GROUP PROFILES

Next, we quantitatively evaluate the descriptive validity of the profiles obtained via LDA to reflect the characteristics of the respective groups in the actual customer data.

1. Creation of Hypothetical Profiles

Based on the profile distributions obtained in the first step (Qualitative Evaluation), one attribute per category is selected to create multiple hypothetical profiles using all possible combinations. For example, if the profile distribution of Group 1 dictates selecting two attributes from the "Usage Time" category (e.g., "11 PM" and "10 PM") and one attribute each from the "Area," "Gender," and "Weekdays/Holidays" categories (e.g., "Naha City," "Male," and "Weekday"), we can create two hypothetical profiles with different usage times: "Male group using the Naha City area on weekdays at 11 PM" and "Male group using the Naha City area on weekdays at 10 PM." Through this process, we define a set of hypothetical profiles based on all major attribute combinations for each group.

2. Extraction of Customers by Matching with Actual Data

The actual customer dataset is compared against the hypothetical profiles defined in step 1. Customer IDs whose data matches four or more of the specified attribute categories in each hypothetical profile are extracted. This extracted customer set is defined as the Profile-Matching Customer Set (L_n).

3. Calculation of Match Rate

The Profile Match Rate is calculated between the extracted Profile-Matching Customer Set (L_n) and the original group member set classified by LDA (P_n) to quantitatively evaluate profile validity. The formula for the match rate is defined as follows

$$Match\ Rate = \frac{|L_n \cap P_n|}{L_n} \qquad (5)$$

Table 5 below shows the calculated agreement rates for each group of location expressions.

Table 5 Match rate between classified group members and actual members

Group	Match Rate
1	71%
2	59%
3	53%
4	73%
5	84%
6	43%
7	50%
8	45%
9	77%
10	67%
Avg	62%

Group	Match Rate
1	89%
2	93%
3	89%
4	79%
5	74%
6	80%
7	84%
8	86%
Avg	84%

Table 5 results show that the average customer match rate when expressed by area exceeds 80%, indicating that the group profiles derived from classification appropriately capture each group's characteristics. The high match rate is likely attributable to the emergence of features beyond spatial information as profile elements, facilitating the incorporation of diverse behavioral patterns observed in actual customer data into the classification.

In contrast, the location-based representation yielded an average match rate of 60%. Groups with lower match rates tended to have numerous spatial features representing departure points within their profiles. This suggests that these groups have fewer other features, and their profile distributions are skewed toward spatial attributes, leading to mismatches with actual usage histories.

8. CONCLUSION

This study compared customer classification results for designated driver services using LDA, employing both conventional place name representations and the proposed density-based area representations for spatial expression of departure points. The area representation extracted departure points in spatially cohesive clusters. Furthermore, the reduction in the number of spatial attributes facilitated the extraction of non-spatial attributes such as usage time periods, customer gender, and age group as features, demonstrating an improvement in profile quality. This makes it easier to grasp diverse customer behavior patterns, revealing that the approach to capturing spatial attributes also influences the extraction of non-spatial behavioral characteristics.

Future challenges include adopting area zoning methods that consider actual road networks, such as topography and traffic conditions.

Furthermore, in addition to the Okinawa Prefecture data used in this study, we possess data from other regions like Kyushu and Kanto. Therefore, applying the same methodology to other regions is necessary to verify whether regional dependencies exist.

9. ACKNOWLEDGEMENTS

We are indebted to everyone at Alpaca. Lab Inc. for their cooperation in providing data for the execution and writing of this research.

10. REFERENCES

[1] David~M. Blei, Andrew~Y. Ng, and Michael~I. Jordan, "Latent dirichlet allocation", Journal of Machine Learning Research 3 993-1022(2003)

[2] Makoto Tsukai, Yuta Tukano, "An analysis on fine-scale geographical data by using topic model", Japanese Journal of JSCE, D3 Vol.74,No.2,111-124 (2018).

[3] Keita Kamiya, Takashi Fuse, "Proposal of a method to grasp regional population characteristics using topic models",Proceedings of the 55th Annual Conference of the Japan Society of civil Engineers, Vol. 55, 42-10 (2017).

[4] T.L. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences, Vol.101, No.Suppl. 1, pp.5228-5235,April 2004.

[5] Yuya Tomita, Satoru Yokoyama, Takayuki Arima, "Analysis of GPS data using asymmetric cluster analysis method – understanding tourist behavior in the Yokohama tourist area", Bulletin of Data Analysis of Japanese Classification Society, Vol12 No.1,17-31 (2023).

[6] Shunpei Otokita, Hiroki Sakaji, Itsuki Noda, "Classification of reasons for visits using multivariate Gaussian Mixture Models on mobile GPS data", The Japanese society for artificial intelligence, SIG-SAI-051-03.

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, A Density-

[8] Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, pp.226-231, 1996.