A-025

空間的エリア構造を考慮した LDA による顧客分類モデルの提案

Proposal of a customer classification model using LDA considering spatial area structure

石原 光竜† 岡崎 威生‡
Hikari Ishihara Takeo Okazaki

1. はじめに

運転代行業界は、非効率的な業務形態や過度な価格競争に悩まされてきた。特に、配車管理の面で課題が顕在化している。例えば、利用者がサービスを依頼してから代行車が到着するまでの時間が長くなるケースが多く、需要と供給のミスマッチが生じている。こうした問題の背景には、過去の利用履歴を活用した顧客分析が十分に行われておらず、地域別・時間帯別の需要の把握が不十分なことがある。その結果、顧客特性に応じた最適な配車戦略や効果的なプロモーション施策の立案が難しい状況にある。

本研究では、こうした課題を解決するために、運転代行業の実際の顧客データを用いて、顧客の利用傾向や特性を多角的に捉えるクラスタリングモデルの構築を目指す。特に、顧客の性別や年代、注文時間帯などの属性情報に加え、注文地点の空間的な分布を考慮し、より実態に即した客層分類を行うことを目的とする。具体的には、顧客分類を行う手法として、潜在クラス分析手法の Latent Dirichlet Allocation[1]（以下、LDA と記述）を用い、LDA による分析を通じて得られる各客層の特徴を捉える方法を提案する。

2. LDA による顧客分類

LDA は、自然言語処理における文書分類などを目的として開発された確率的生成モデルである。しかし、その構造は、観測データに潜在する「要因」や「傾向」を抽出するモデルとして汎用性が高く、非言語的なデータにも応用が可能である。本研究の対象となる運転代行業の顧客データは、明示的な目的や嗜好が確認されていないことから、その背後にある潜在的な利用目的や行動パターンを推定する手法として LDA は有効であると考えられる。

また、LDA は確率的手法であるため、各顧客が複数のグループに所属する可能性を考慮した「ソフトクラスタリング」が可能であり、K-means 法や階層クラスタリングのような 1 つのクラスタに固定される手法と比較して、顧客の多面的な性質や利用傾向を柔軟に捉えることができる。なお、潜在構造を扱う手法として、確率的潜在意味解析（以下、PLSA と記述）も存在するが、PLSA は学習データに過剰に適合することで過学習が生じやすく、新たな観測データにおいて再学習が必要となる。一方、LDA は事前分布としてディリクレ分布を協約事前分布として仮定しており、新規データへの適用やモデルの拡張性に優れている。

塚井・塚野ら[2]は、市町村以下の集計単位の地理情報(詳細地理情報)に関するトピックモデル LDA の有用性を示すことを目的とし、3 次メッシュレベルの詳細地理情報の特性を抽出して、既存手法である因子分析との比較検証を行った。分析の結果、因子分析からは 3 種のみの地理特性しか得られないのに対して、LDA による分析により 8 種類の地理特性を得られるとともに、得られたメッシュ別の空間分布が実都市の構造を適切に捉えることが確認された。

神谷・布施ら[3]は、LDA とその拡張である HDP-LDA を用いて、地域別人口特性の把握手法を提案している。具体的には、500m メッシュのモバイル空間統計データを対象に、メッシュを文書、滞在者の居住地を単語と見立ててトピックモデルを適用し、地域ごとの人口特性を抽出している。

このように、LDA は文書データの分析にとどまらず、空間情報や統計データと組み合わせることで、地域や集団の潜在的な特徴を多面的に捉える分析手法としての応用が進んでいる。本研究でも同様に、運転代行サービスに関する顧客データを LDA に適用することで、地域的・時間的・属性的な特徴を統合的に捉えることを試みる。顧客データを LDA に応用するにあたり、各顧客を 1 つの「文書」、そして性別や年代、利用時間帯などの属性情報を「単語」として扱うことで、潜在的な顧客グループの抽出を試みる。本研究では、2024 年 1 月から 2025 年 3 月までの沖縄県の顧客データ（54093 件）を分析対象とする。なお、以下の各属性データについて、各顧客の注文毎に発生する属性値の出現頻度を集計し、Bag-of-Words（BoW）形式のベクトルとして表現した上で、LDA に適用する。

項目	詳細
性別	男性、女性
年代	10 代区切りで 10 代～80 代
利用時間帯	1 時間区切りで 0 時～23 時
平日/休日	平日、休日・祝日、休日・祝日の前
出発地点	注文された出発地の緯度経度
距離(km)	出発地から目的地までの距離
天気情報	注文日の天気情報

表 1　運転代行業の顧客データ　概要

なお、LDA 実行に伴う潜在変数の事後分布推定手法として、「崩壊型ギブスサンプリング」を採用する。LDA における推定手法としては「変分ベイズ法」も挙げられるが、崩壊型ギブスサンプリングは実装が比較的容易で、推定精度にも優れていることが知られているため、採用する。

以下に、崩壊型ギブスサンプリングを推定手法とした LDA による顧客分類を実行するための定義式を示す。

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

- N_{kd}：顧客 d のグループ k が割り当てられた属性数
- N_{kv}：グループ k における属性 v の出現回数
- N_k：グループ集合 z においてグループ k が表れた回数
- N_d：顧客 d に含まれる属性の数
- N_{dj}：顧客 d の n 番目の属性を除いたときの属性数

上記のサンプリング確率(6)より、それぞれ以下のようにパラメータを効率的に推定する。

† 琉球大学大学院理工学研究科　Graduate School of Science and Engineering University of the Ryukyus
‡ 琉球大学工学部工学科　Faculty of Engineering University of the Ryukyus

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

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

3. 最適なクラス数の選択方法

LDA による顧客分類を実行する際、事前にグループ数 K を設定する必要がある。本研究では、最適なグループ数を決定するための手法として、LDA の評価指標の一つである **Perplexity**[4]を採用する。一般的に Perplexity はグループ数の増加に応じて値が低くなる傾向があり、値が低いほどモデルの予測性能が高く、データ構造をより良く表現するとされる。しかし、過度に大きなグループ数は過学習の原因となるため、適切な選択が求められる。本研究では、以下の手順により Perplexity による最適なクラス数の選定を行う。

Step1.トピック数ごとのモデル学習と Perplexity の算出

グループ数を 2 から 30 までの範囲で 1 刻みに設定し、各グループ数に対して LDA モデルの学習を行い、対応する perplexity の値を算出する。

Step2. Perplexity の平滑化

計算された perplexity の系列に対し、移動平均による平滑化を行う。この処理により、局所的なノイズの影響を抑え、トピック数と perplexity の関係の傾向を明確化する。

Step3. 平滑化後の変化率の算出

平滑化後の perplexity に対し、隣接するトピック数間の変化率を計算する。変化率の前後の値に基づき、最適なグループ数を決定する。

4. 各顧客グループのプロフィール解釈

LDA によって得られるφ分布を用いて、各グループがどのような特徴を持つのかプロフィールの解釈を行う。ただし、各データ項目に含まれる属性数の違いにより、単純なφ分布に基づく解釈では一部のカテゴリの影響が過小評価または過大評価される可能性がある。そのため本研究では、属性ごとのφ分布値に対し、その属性が属するカテゴリの属性数を重みとして乗じることで、カテゴリ間の属性数の偏りを補正する重み付けを行った。重み付けによる調整後、各グループにおけるスコアが高い上位 15 属性を、そのグループの特徴として捉え、プロフィールとして解釈を行う。

従来、地点データを分析対象として扱う場合において、様々な表現手法が使用されている。冨田ら[7]は、横浜来訪者の GPS データをもとに、あらかじめ設定された観光地エリア間の移動頻度に基づいた非対称クラスタリングを実施し、代表的な観光行動パターンの抽出を試みている。しかしこの手法では、分析対象となるエリアが事前に固定されており、実際の人流の空間的な密度分布やエリア間の地理的近接性を十分に考慮することができない。そのため、連続的かつ柔軟な空間的行動パターンの把握が困難となり、実際の行動の多様性を捉えきれない可能性がある。

また、地点データの簡易な表現手法として、地域メッシュを用いる方法も一般的に用いられる。これは緯度経度を一定のサイズで区切った格子状の領域に変換することで、空間データを離散化し、扱いやすくする手法である。

本研究では、はじめに従来の簡易的な表現手法である地域メッシュによる出発地点の特定手法を用いて、LDA による顧客分類を試みた。その結果得られた顧客プロフィールの一部を以下に示す。

Group	プロフィール　解釈結果
1	曇りの 21 時～23 時に 那覇市{東町 泊　沖縄市{比屋根 諸見里宜野湾市大山か 胡屋 ら長距離利用の 40 代男性
2	曇りの 21 時～23 時に 那覇市{屋富祖 当山　沖縄市{東 照屋　宜野湾市我如古 から短距離利用の 20 代女性

表 2　グループプロフィール例（地域メッシュ）

上記のプロフィール結果より、性別や年代、利用時間帯などの属性においては一定のまとまりが確認された。一方で、出発地点に関するプロフィールにおいては、同一グループ内において、下記の図 1 のように那覇市や沖縄市など地理的に離れた地点が含まれるケースが多く、行動パターンの直感的な解釈や施策への応用が難しいという課題が見られた。

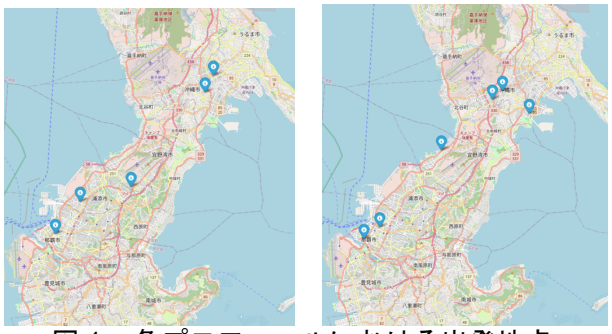

図 1　各プロフィールにおける出発地点
（左:Group1, 右 Group2）

5. 空間的エリア構造を考慮した顧客分類

注文地点について、地域メッシュのように事前に範囲が固定され、かつ粒度の細かい手法では、分類結果の空間的解釈が困難になる場合がある。これら課題を踏まえ、地点データの空間的な密度や分布の広がりを考慮したエリアとして柔軟に捉える手法が解決策として挙げられる。音喜多ら[6]は、携帯 GPS データを基に多次元ガウスモデル(Gaussian Mixture Model, 以降 GMM と記述)による行動の滞在目的の分類を試みている。GMM による分類の結果、クラスタごとに空間的な広がりを捉え、行動目的の違いを反映することが確認されている。

本研究でもこの考え方を踏まえ、地点データを空間的に捉えた柔軟なエリアとして定義し、より自然な顧客の行動パターンの把握を目指す。

一般的に、運転代行業の利用顧客は特定の人気スポットや利用頻度の高いエリアに集中しやすい傾向があると考え

る。例えば、主要な交通機関の駅や繁華街、ビジネスエリア、住宅地の中心など、特定のエリアに顧客が集中することとなる。このようなエリアでは、自然に中心点が存在し、その周辺に顧客の行動が広がっていくと考えられる。

そのため本研究では、顧客の注文地点データからエリアを形成するにあたり、各エリアに中心点があると仮定し、その中心点を基とした広範囲なエリアを構築する手法を検討する。顧客が集中するエリアの代表点（中心点）を特定するために、**DBSCAN**（Density-Based Spatial Clustering of Applications with Noise）[7]を採用する。DBSCAN は、データの空間的な密度に基づいてクラスタを自動的に抽出できるクラスタリング手法である。

DBSCAN により抽出された各エリアには、複数のコアポイントが存在する。本研究では、エリアC_kの中心点として、コアポイント集合C_k^{core}の重心c_kを以下のように定義する。

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

- $x_i = (x_i^{(1)}, x_i^{(2)})$：各コアポイントの座標ベクトル

次に、DBSCAN で得られた各エリアの中心点c_kを基に、GMM (Gaussian Mixture Model) によるエリアのモデル化を行う。GMM は、データを複数の正規分布（ガウス分布）の混合として表現する手法であり、各エリアを確率的に定義することが可能である。これにより、エリア間の曖昧な境界や、顧客が複数のエリアに属する可能性も考慮でき、エリア内での顧客の行動パターンを柔軟に捉えることが期待される。

以下に、中心点c_kを用いた各エリアのガウス分布の定義式を示す。

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

- c_k：エリアC_kの中心点（平均ベクトル）
- $\sum_k$ ：エリアC_kの共分散行列
- d：データの次元数

最後に、GMM により得られた各エリアに対して、各注文地点がそのエリア内でどのような空間的位置にあるかを定量的に評価する。具体的には、各地点が「エリアの中心付近か」、または「境界に近い位置にいるのか」といった空間的な関係性を考慮することで、ユーザの行動パターンを空間的な分布も含めてより精緻に捉えることができると考えられる。

各エリア内の各注文地点の位置関係を評価するために、**マハラノビス距離**を採用する。マハラノビス距離は、各クラスタの分散構造を考慮し、楕円状の等高線を持つガウス分布において、各地点がエリアの中心からどれだけ離れているかを標準化して測定できる特徴を持つ。そのため、各注文地点がエリアの「中心付近」「端」「エリア外」のいずれに該当するかをより正確に評価することが可能となる。そして、マハラノビス距離の二乗$D_M(x)^2$は自由度 2 のカイ二乗分布に従う性質が知られている。これを利用し、各注文地点のエリア内の位置関係を統計的に以下のように分類する。

- 中心付近：$D_M(x)^2 \leq \chi^2_{2,0.68}$　(~ 68%)
- エリア周辺：$\chi^2_{2,0.68} < D_M(x)^2 \leq \chi^2_{2,0.95}$(68 ~ 95%)
- エリア外：$\chi^2_{2,0.95} < D_M(x)^2$　(95% ~)

このように、GMM によるガウス分布の性質を活かしてマハラノビス距離の統計的閾値を設定することで、各注文地点がエリアの「中心付近」「周辺」「外部」のどの位置に属するのかを定量的に評価を行う。

6. 適用効果の検証

上述のゾーニング手法により、出発地点のエリア分布を生成し、各顧客の注文データに対応するエリア情報を付与した。このエリア情報をもとに LDA による顧客分類を実施し、従来のメッシュベースの方法と比較して、より意味のある顧客グループの抽出が可能であるかを検証する。

図 2 は、DBSCAN および GMM を用いて実施したエリアゾーニングの結果を地図上に可視化したものである。各エリアは、所属ポイントの平均位置と分散をもとに円形で表現している。なお、DBSCAN のパラメータは、ε ∈ [0.01, 0.05]、MinPts ∈ [1, 20] の範囲でグリッドサーチを行い、各パラメータ設定で得られたクラスタ結果に対して GMM を適用し、BIC の値が最小となる組み合わせ（ε = 0.011, MinPts = 5）を選択した。その結果、得られたエリア数は 44 となる。

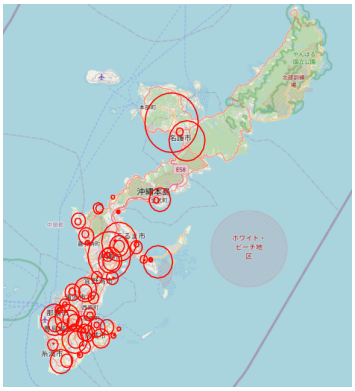

図 2　エリアゾーニング結果（エリア数 44）

図 3 は、顧客データセットを LDA モデルに適用し、グループ数を 2 から 30 まで変化させた時のそれぞれの Perplexity の変化率を表したものである。

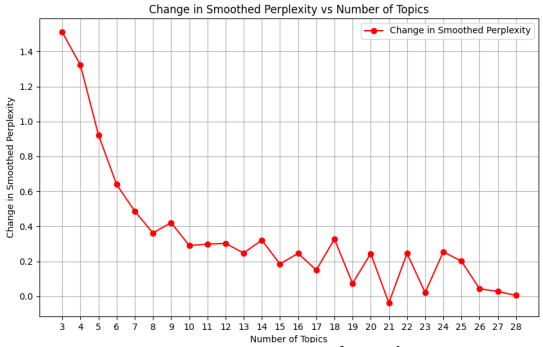

図 3　Perplexity 変化率

グループ数 10 以降より Perplexity 値が安定し、解釈可能なクラスタ構造となると考えられる。よって本研究では、分類するグループ数を 10 と設定する。

LDA に適用して得られた φ 分布の上位 15 属性をもとに、各グループのプロフィールを作成する。そして、本研究で提案したエリア表現を考慮し顧客分類した場合と、従来の地域メッシュ表現による顧客分類した場合のプロフィールを比較し、表現手法の違いからグループのクオリティにどう影響するのか検証する。

各表現手法で得られた全属性における空間属性の割合、及びプロフィールに含まれる空間属性の割合を示す。

出発地 表現手法	全属性中の 空間属性割合	プロフィールにおける 空間属性の割合（平均）
メッシュ	0.93	0.48
エリア	0.63	0.36

表 3　空間属性割合の比較

表 3 より、地域メッシュ表現では空間属性の種類が非常に多く、プロフィールにおいても空間情報の比重が大きい傾向が確認された。一方、エリア表現では空間属性が一定の地域に集約され、属性数も抑えられることで、年代や時間帯などの非空間属性が相対的に抽出されやすくなった。

次に、それぞれの分類結果の比較にあたり、代表的なグループプロフィールを以下に取り上げる。

地点表現	グループプロフィール
地域 メッシュ	曇りの休日 0 時に ・沖縄市 (上地, 諸見里) ・那覇市 (松尾, 泊, おもろまち, 奥武山町) ・うるま市 (安慶名) から中距離利用する 30 代女性
エリア	晴れの休日 23 時〜0 時に沖縄市エリア から短距離利用する若年女性層

表 4　表現手法別のグループプロフィール例

上記の結果より、地域メッシュ表現を用いた顧客分類では、同一グループ内に「沖縄市」、「那覇市」、「うるま市」など地理的に離れた地点が混在し、空間的なまとまりに欠ける傾向が見られた。また、空間属性の種類が多いため、プロフィールが空間情報に偏重し、年代や利用時間帯などの非空間属性の相対的な存在感が薄れる傾向も確認された。

一方、エリア表現を用いた分類では、各グループの出発地点が地理的に連続したエリアとして自然に集約され、空間的なまとまりが確保された。さらに、空間属性数が抑制されたことで、年代や利用時間帯などの非空間属性が反映されやすくなり、得られるグループプロフィールが多面的かつバランスの取れたものとなった。

この結果から、エリア表現を用いた顧客分類を行うことで、全体的な特徴におけるバランスが保たれ、また各グループの出発地点がどのあたりであるのかを広範囲かつ柔軟に捉えられることで、顧客分類におけるプロファイルとしての情報価値が高まったと考えられる。

7. おわりに

本研究では、LDA による運転代行業の顧客分類を行うにあたり、出発地点の空間表現として、従来手法の地域メッシュ表現と、本研究で提案した密度を考慮したエリア表現を用いて、それぞれの分類結果の違いを明らかにした。結果として、エリア表現を用いることで、出発地点が空間的にまとまりのある形で抽出され、空間属性が適度に集約されることで、非空間属性が相対的に際立ち、プロフィールとしての質が向上することを示した。これにより、顧客の利用範囲や属性に応じた多様な行動パターンの把握が可能となり、空間属性の捉え方が非空間的な行動特徴の抽出にも影響を与えることが明らかとなった。

今後の課題としては、地形や交通状況など、実際の移動に影響を与える要因を考慮した分析手法の導入や、提案手法のモデルの安定性や頑健性などについての定量的な評価方法が挙げられる。また、本研究では推定手法として崩壊型ギブスサンプリングを用いた LDA を採用したが、変分ベイズ法などの他の推定手法や分類手法との比較検証の検討の余地があると考える。

さらに、本研究で対象とした沖縄県のデータに加え、九州・関東地方など他地域のデータも保有している。そのため、適用範囲を拡大することで、それぞれの地域特性の違いを踏まえた比較分析や、より汎用的な顧客分類手法の確立を今後の課題とする。

謝辞

本研究の遂行および執筆にあたり、顧客データ提供にご協力いただいた株式会社 Alpaca.Lab の皆様に、心より御礼申し上げます。

参考文献

[1] David~M. Blei, Andrew~Y. Ng, and Michael~I. Jordan, “Latent dirichlet allocation”, Journal of Machine Learning Research 3 993-1022(2003)
[2] 塚井誠人, 塚野裕太, “トピックモデルによる詳細地理情報分析”, 土木学会論文集 D3 Vol.74,No.2,111-124 (2018).
[3] 神谷啓太, 布施孝志, “トピックモデルを利用した地域別人口特性の把握手法の提案”, 第 55 回土木計画学研究発表会・講演集, Vol. 55, 42-10 (2017).
[4] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, Vol. 3, pp. 993–1022, 2003.
[5] 冨田裕也, 横山暁, 有馬貴之, “非対称クラスター分析法を用いた GPS データの分析 横浜観光エリアにおける観光行動の把握”, 日本分類学会 データ分析の理論と応用 Vol12 No.1,17-31 (2023).
[6] 音喜多俊平, 坂地泰紀, 野田五十樹, “多次元ガウスモデルによる携帯 GPS データの滞在目的分類”,人工知能学会第二種研究会資料 社会における AI 研究会, SIG-SAI-051-03
[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, A Density-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.