3K4-5

# スポット情報を考慮した複合データ分類モデル

## The Multi-type Data Classification Model Considered Spot Information

鈴木 優伽 *1　斉藤 和巳 *1　風間 一洋 *2
Yuka Suzuki　Kazumi Saito　Kazuhiro Kazama

*1静岡県立大学　経営情報イノベーション研究科
Graduate School of Management and Information of Innovation, University of Shizuoka

*2和歌山大学　システム工学部
Faculty of Systems Engineering, Wakayama University

We address a problem of classifying user behavior data into point-of-interests(POIs) by integrating multiple social media dataset. For this purpose, unlike a standard distance employed by conventional methods such as a nearset neighbour method, we propose a method based on the posterior probabilities obatained from Gaussian mixture models, some of whose parameters are learned from user behavior data by the EM algorithm. In our experiments using datasets of Flickr, Twitter and TripAdvisor, by showing that the proposed method could produce naturally interpretable results in comparison to those of conventional and variant methods, we confirm that our method is vial and promising.

## 1. はじめに

近年，Twiiter 等のソーシャルメディアや情報技術の発達により，個人の情報発信が容易に行われている．中でも，スマートフォン等の GPS 搭載のデバイスの普及に伴い，位置情報を伴った情報発信が盛んである．これらの情報は，個人の詳細な行動データとなりうるため，行動パターンの把握等の行動分析の研究や，重要スポット抽出等の研究に応用可能とされている [1][2][3]．例えば，[4][5] らの研究では，写真投稿サイト Flickr に投稿された写真の位置情報データを使用し，それらをクラスタリングすることで重要スポットの抽出を行っている．

これら研究の特徴として，単一のソーシャルメディアから得られるデータのみを利用していることが挙げられる．しかしながら，各個人が利用するソーシャルメディアは単一ではなく，複数のものが並行して利用されることが多い．例えば，ある場所に訪れた際に，Twitter を用いてその場所の様子や訪れた際の感情をツイートすると共に，Instagram や Flickr などの写真投稿サイトに，風景や友人との写真を投稿すると考えられる．すなわち，同じ事象に関するデータが目的別に複数のソーシャルメディアに分散して存在しており，単一ソーシャルメディアのデータを扱うだけでは，ユーザの限られた側面しか見ていない可能性がある．

そのため，情報の相互補完や情報量の充実を可能にするような，複数ソーシャルメディアのデータを突合する分類モデルが必要になる．そこで本研究では、その第一歩として，あるソーシャルメディアで得られる観光スポットデータと，他のソーシャルメディアデータで得られる行動データを突合する分類モデルを提案する．この時，最も単純に，$k$-mean クラスタリング [6] や $k$-近傍法 [7] 等の，スポットの地点とユーザの行動地点の距離に基づいた分類モデルが考えられる．しかし一般に，人気があり大規模な施設のスポット程，多くのユーザの行動データが適応されるなど，各スポットに対する人気度やスポットの面的広がりによる行動範囲の広さで，分類結果に偏りが見られると考えられる．すなわち，単純な距離を用いる k-mean クラスタリングや k-近傍法では，そのような偏りが表現されず，分類結果に限界がある可能性が示唆される．そこで本稿では，人気度等のスポット情報や行動範囲を考慮した分類モデルとして，混合ガウスモデルを土台に拡張した分類モデルを定義する．具体的には，各分布の事前確率にスポットの人気度を組み込み，各分布の標準偏差を学習させることで行動範囲を表現する．

本稿の構成は以下の通りである．まず，混合ガウスを土台にした本モデルについて述べる．次に，本モデルで扱うソーシャルメディアの詳細と，本モデルの比較モデルについて説明する．そして，評価実験において，本モデルがより自然な分類結果が得られることを示すとともに，最後にまとめを述べる．

連絡先: 鈴木優伽，静岡県立大学経営情報イノベーション研究科，静岡県静岡市駿河区谷田 52-1，054-264-5436

## 2. 分類モデル

ある時刻 $t$ での回遊者 $m$ の行動データは $\mathbf{x}_{m,t}$ となるが，本研究では，時刻や回遊者を区別しないので，一般に，$\mathbf{x}_n = (Lat, Lng)^T$ からなる行動データ集合 $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ を考える．他のソーシャルメディアから得られる観光スポット集合 $\mathcal{K} = \{1, \cdots, K\}$ のうち，$k$ 番目の観光スポットの情報として，位置を $\mathbf{y}_k = (Lat, Lng)^T$，人気度を $g(k)$ と定義する．ただし，$(Lat, Lng)$ は回遊者がソーシャルメディアを利用した位置の緯度と経度を意味し，上付き $T$ は転置を表す．

今，回遊者の行動データ $\mathbf{x}_n$ の分布は各観光スポットを中心とした混合分布で近似可能とし，混合ガウスモデルで両データを突合させるとする．各分布の中心を得られた観光スポットの位置情報 $\mathcal{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_K\}$，2 次元分散共分散行列の集合を $\mathcal{S} = \{\mathbf{S}_1, \cdots, \mathbf{S}_K\}$ で定義する．ただし本稿では，$\mathbf{S}_k = s_k\mathbf{I}$ で定義される最も単純なケースを考える．ここで，$s_k$ はスポット $k$ の標準偏差の係数に対応し，$\mathbf{I}$ は単位行列を表す．なお，以下の議論は一般の共分散行列のケースへも容易に拡張できる．この時，一般の $d$-次元ガウス分布を特殊化し，行動データ $\mathbf{x}_n$ が $k$ 番目の分布に属する確率 $N(\mathbf{x}_n; \mathbf{y}_k, \mathbf{S}_k)$ を以下で算出する．

$$N(\mathbf{x}_n; \mathbf{y}_k, \mathbf{S}_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}_n - \mathbf{y}_k)^T\mathbf{S}_k^{-1}(\mathbf{x}_n - \mathbf{y}_k)\right)}{(2\pi)^{\frac{d}{2}}|\mathbf{S}_k|^{\frac{1}{2}}},$$

$$\propto \quad \frac{1}{s_k}\exp\Big(-\frac{1}{2s_k}||\mathbf{x}_n-\mathbf{y}_k||^2\Big). \qquad (1)$$

また，一般に回遊者は口コミ等で人気のある場所に訪れる傾向があるため，回遊者が $k$ 番目の分布を滞在する確率 $\alpha_k$ はスポット情報である人気度 $g(k)$ に比例すると考えられる．後述するように，人気度 $g(k)$ はレビューサイトでのレビュー数より推定する．そのため，回遊者が各分布を滞在する確率を $\mathcal{A}=\{\alpha_1,\cdots,\alpha_k\}$ とし，行動データ $\mathbf{x}_n$ が各分布に属する確率 $p(\mathbf{x}_n)$ を以下で定義した混合ガウスモデルを，本稿におけるスポット情報を考慮した分類モデルと定義する．

$$p(\mathbf{x}_n)=\sum_{k=1}^{K}\alpha_k N(\mathbf{x}_n;\mathbf{y}_k,\mathbf{S}_k),\quad \sum_{k=1}^{K}\alpha_k=1,\quad \alpha_k\propto g(k) \qquad (2)$$

上記の混合ガウスモデルに潜在変数 $z$ を 導入し，パラメータ $\mathcal{A},\mathcal{Y}$ をスポット情報で与えられた値から不変とした上で，行動データ集合 $\mathcal{X}$ が与えられた下での尤度関数の最大化から，パラメータ $\mathcal{S}$ を求めデータを分類する．

### 2.1 パラメータ学習

導入する潜在変数 $\mathbf{z}$ は $\mathbf{z}\in\{0,1\}^K$ の $K$ 次元ベクトルであり，$p(z_{nk}=1)=\alpha_k$，$p(\mathbf{x}_n|z_{nk}=1)=N(\mathbf{x}_n;\mathbf{y}_k,\mathbf{S}_k)$ であるとする．また，データ $\mathbf{x}_n$ が与えられた下での $z_{nk}$ の事後確率 $\gamma(z_{nk})$ を以下で定義する．

$$\gamma(z_{nk})=p(z_k=1|\mathbf{x}_n)=\frac{\alpha_k N(\mathbf{x}_n;\mathbf{y}_k,\mathbf{S}_k)}{\sum_{j=1}^{K}\alpha_j N(\mathbf{x}_n;\mathbf{y}_j,\mathbf{S}_j)} \qquad (3)$$

今我々は，行動データ集合 $\mathcal{X}$ が与えられた，以下の尤度式 $\mathcal{L}$ が最大となるようにパラメータ $\mathcal{S}$ を求めていく．

$$\mathcal{L}(\mathcal{X}|\mathcal{A},\mathcal{Y},\mathcal{S})=\sum_{n=1}^{N}\ln\left(\sum_{k=1}^{K}\alpha_k N(\mathbf{x}_n;\mathbf{y}_k,\mathbf{S}_k)\right) \qquad (4)$$

式の性質上，陽に解析解が得られないため，詳細には以下のEM(期待値最大化) アルゴリズムを用いる。

A1. $i\leftarrow 0$ として， $\mathcal{A},\mathcal{S}$ を以下で初期化;

$$\alpha_k^{(i)}\leftarrow\frac{g(k)}{\sum_{k=1}^{K}g(k)},\quad s_k^{(i)}\leftarrow 1$$

A2. 現在のパラメータから事後確率 $\gamma(z_{nk})^{(i)}$ を求める;

A3. 現在の事後確率からパラメータを更新;

$$s_k^{(i+1)} \quad = \quad \frac{1}{2}\sum_{n=1}^{N}\frac{\gamma(z_{nk})^{(i)}}{\sum_{n=1}^{N}\gamma(z_{nk})^{(i)}}||\mathbf{x}_n-\mathbf{y}_k||^2$$

A4. 更新したパラメータから尤度 $\mathcal{L}^{(i+1)}$ を計算;

A5. 定数 $\epsilon=10^{-4}$ とし $(\mathcal{L}^{(i+1)}-\mathcal{L}^{(i)})/\mathcal{L}^{(i+1)}<\epsilon$ ならば終了，さもなくば，$i\leftarrow i+1$ とし，A2 に戻り再度計算;

## 3. 評価実験

### 3.1 データセット

本実験では，扱うソーシャルメディアとして，写真共有サイト Flickr，旅行レビューサイト TripAdvisor(以下，TA)，Twitter を採用した．位置情報が京都周辺と神奈川周辺を示すものを対象に，2012/1/1〜2015/1/31 の間で投稿されたデータを収集した．ただし，Twitter データはデータ量が膨大であるため，ツイート内容に「行った/行く/なう」等の行動や場所を示すと思われる単語を含むものに限定した.. Flickr から収集した写真に付随する位置情報と撮影時刻等の時間情報から得た行動履歴と，Twitter から収集した位置情報付きツイートから得た行動履歴を行動データとし，TA から得られた各観光スポットの緯度･経度をスポットの位置情報，レビュー数をスポットの人気度とみなした．京都では Flickr 行動データ数 76999，Twitter 行動データ数 70078，スポット数 649 であり，神奈川では Flickr 行動データ数 166712，Twitter 行動データ数 166712，スポット数 778 である．スポットの人気度の平均は，京都で 40，神奈川で 36 である．

### 3.2 比較モデル

比較モデルとして，スポット情報を考慮しないモデルを用いる．すなわち，パラメータ学習の際に，$\mathcal{A}$ も同時に学習するモデルを本実験の比較モデルとする．制約条件 $\sum_{k=1}^{K}\alpha_k=1$ より，次式が最大となるようなパラメータ $\mathcal{A}$ を求める．ここで，$\lambda$ はラグランジュの未定定数とする．

$$\mathcal{L}(\mathcal{X}|\mathcal{A},\mathcal{Y},\mathcal{S})+\lambda\left(\sum_{k=1}^{K}\alpha_k-1\right) \qquad (5)$$

式 5 を $\alpha_k$ で微分して 0 とおくと次式に展開できる．

$$\begin{aligned}\sum_{n=1}^{N}\frac{N(\mathbf{x}_n;\mathbf{y}_k,\mathbf{S}_k)}{\sum_{k=1}^{K}\alpha_k N(\mathbf{x}_n;\mathbf{y}_k,\mathbf{S}_k)}+\lambda &= 0\\ \sum_{n=1}^{N}\frac{\gamma(z_{nk})}{\alpha_k}+\lambda &= 0\\ \lambda=-\sum_{k=1}^{K}\sum_{n=1}^{N}\gamma(z_{nk}) &= -N\end{aligned}$$

よって，アルゴリズム A3 に以下の更新式を追加する．

$$\alpha_k^{(i+1)} \quad = \quad \frac{1}{N}\sum_{n=1}^{N}\gamma(z_{nk})^{(i)}$$

## 4. 実験結果

### 4.1 データ分類結果の考察

表 1,2 に，モデルを用いて分類後，各行動データに対する事後確率 $\gamma(z_{nk})$ が高い上位 10 スポットを示す．表 1,2 から，本モデルを用いることで,「伏見稲荷大社」,「清水寺」,「横浜中華街」などの一般的に写真が多く撮られやすい・ツイートがされやすいと考えられるスポットほど事後確率が高いことが確認でき，Flickr と Twitter の両行動データに対して自然な分類ができていると期待できる．この特徴はスポット情報を考慮した本モデルで顕著なため，本モデルの有効性が示唆される．また，両行動データに対する事後確率上位スポットに差異が見られるが，これは,「写真を撮る」「呟く」といった行動の性質の違いを反映した結果であると考えられる．

本モデルで，実際に伏見稲荷大社に分類された Flickr 行動データの例を図 1，Twitter 行動データ例を表 3 に示す．図 1 から，分類されたデータの大半が伏見稲荷大社の見どころであ

表 1: 事後確率上位 10 スポット: Flickr 行動データ

| rank | 京都:本モデル | 京都:比較モデル | 神奈川:本モデル | 神奈川:比較モデル |
|---|---|---|---|---|
| 1 | 清水寺 | 霊雲院 | 横浜中華街 | 王禅寺 |
| 2 | 伏見稲荷大社 | イオンモール Kyoto | よこはま動物園ズーラシア | 柏尾川堤の桜 |
| 3 | 天龍寺 | 新京極商店街 | 鶴岡八幡宮 | 夢見ヶ崎動物公園 |
| 4 | トロッコ列車 | 京都タワー | アンパンマンこどもミュージアム | 相模健康センター |
| 5 | 京都水族館 | 護王神社 | 寒川神社 | 神奈川県水道記念館 |
| 6 | 京都タワー | 宝泉寺禅センター | 泉の森 | よこはま動物園ズーラシア |
| 7 | 二条城 | 立命館大学国際平和ミュージアム | 大涌谷 | 日向山の森 |
| 8 | 奈良公園 | 大井神社 | 高尾山 | 四季の森公園 |
| 9 | 京都錦市場 | 竹林の道 | 電車とバスの博物館 | パシフィコ横浜臨港パーク |
| 10 | 東福寺 | 櫟谷宗像神社 | シーバス | 京浜伏見稲荷神社 |

表 2: 事後確率上位 10 スポット: Twitter 行動データ

| rank | 京都:本モデル | 京都:比較モデル | 神奈川:本モデル | 神奈川:比較モデル |
|---|---|---|---|---|
| 1 | 伏見稲荷大社 | 伏見稲荷大社 | 横浜中華街 | 桜坂 |
| 2 | 清水寺 | 百済寺跡 | 羽田空港 第二ターミナル 展望デッキ | 宝来公園 |
| 3 | 東大寺 | 交野山 | よみうりランド | 多摩川台公園古墳展示室 |
| 4 | 金閣寺 | アステイ　ロード | 藤子・F・不二雄ミュージアム | 多摩川浅間神社 |
| 5 | 二条城 | 京都まちなか交通観光案内所 | 高尾山 | 松濤園 |
| 6 | 平等院 | 宇治市源氏物語ミュージアム | サンリオピューロランド | ジョイナス彫刻の森 |
| 7 | 三十三間堂 | 京都タワー | 羽田空港 国際線ターミナル 展望デッキ | 絹の道資料館 |
| 8 | 山田池公園 | 車折神社 | 県立相模原公園 | 鹿島神社 |
| 9 | 京都タワー | イオンモール高の原 | 多摩動物公園 | 公時神社 |
| 10 | 下鴨神社 | 徳林庵 | 寒川神社 | ボーネルンド キドキド |

る鳥居や神社の写真であり，伏見稲荷観光を目的にしたデータが自然に分類されていることが確認できる．すなわち，観光スポット情報と行動データの突合という点で，本モデルが有効的であると考えられる．

図 1: 分類された行動データ例 (Flickr)

一方，Twiiter 行動データでは，伏見稲荷大社で呟かれたと予想できるデータはあるものの，友人と遊んだ等の日常生活に沿った行動データが多く含まれていることが確認できる．この理由として，Flickr 行動データの大半が「写真を撮る」という

表 3: 分類された行動データ例 (Twitter)

| |
|---|
| りんちゃんなう稲荷神社なう。商売の神様にお礼をしてきます。 |
| 友人の家の前なう |
| 明日神切りに行くか |
| 帰宅なななうう |
| うまいけど駅まで行くのめんどい |
| 明石まで行くのだりぃなぁ |
| 友達ん家なう |
| 伏見稲荷なう !!! お参りなう !!! |
| 西宮に行くか三宮に行くか…… |

性質上，観光目的のもので占められているのに対し，Twitter 行動データは日常生活等の非観光を目的としたもので占められていることが推測できる．そのため，Flickr 行動データとは異なる観点で，スポット情報との突合が可能であると考えられる．しかし，より適切なスポット情報と行動データの突合を目指すのであれば，前段階の処理として，SVM などを用いて，非観光ツイートの除外や，日常生活で利用するスポットの情報の追加等をする必要性があると考えられる．

### 4.2　スポット情報を組み込む有用性の考察

本モデルでの分類結果と比較モデルの分類結果から，スポット情報を組み込む有用性を検証する．ここでは，イオンモール Kyoto に分類されたデータを例に考察を行う．

図 2 に本モデル，図 3 に比較モデルでイオンモール Kyoto に分類されたデータの一例を示す．ただし，示すデータは本モデル・比較モデルの分類結果の内どちらかにしか現れないものに限定する．図 3 から，比較モデルでは，本願寺や京都タワー


図 2: 分類された行動データ例 (本モデル)


図 3: 分類された行動データ例 (比較モデル)

を訪れた際の行動データが分類されていることが確認できる．すなわち，イオンモール Kyoto 付近で撮影されたものの，イオンモール Kyoto に分類すべきではないデータが多く含まれている．しかし，図 2 を見ると，イオンモール Kyoto の外観や，イオンモール Kyoto 内のテナント (UNIQLO・無印良品) を示すデータが分類されており，スポット情報を組み込むことで，より自然な分類結果が得られることが示唆される．

## 5. おわりに

本研究では，情報の相互補完や情報量の充実が可能となるように，複数ソーシャルメディアのデータを突合する分類モデルの構築を目的とした．本稿では，その第一歩として，混合ガウスモデルにスポット情報を組み込んだ分類モデルを定義し，モデルの有効性を評価した．実行動データを用いた評価実験では，本モデルによってデータの自然な分類が可能であり，データ同士の突合に有効性があることを示した．しかし，Twitter などの観光ではなく日常生活中の行動（感情）がデータの大半を占める場合は，Flickr 等の観光データとは異なる分類結果が得られることが確認できた．その為，今後は，ツイートの中身を考慮する・tag を見るなどの位置情報以外の情報を考慮した分類モデルを定義し，より柔軟にデータ同士を突合させることを目指す．

**謝辞** 本研究は，総務省 SCOPE(No.142306004) 及び，科研費 (No.26330345) の補助を受けた．

## 参考文献

[1] Song C, Koren T, and Barabasi A-L Wang P. Modelling the scaling properties of human mobility. *Nature*, pp. 818–823, 2010.

[2] Y.Zheng, L.Zhang, X.Xie, and W.-Y.Ma. Mining Interesting Locations and Travel Sequences from GPS Trajectories. In *in Proc.of WWW*, pp. 791–800, 2009.

[3] M.C.Gonzales, C.A.Hidalgo, and A.L.Barabasi. Understanding Individual Human Mobility Patterns. In *Nature*, pp. 779–782, 2008.

[4] W.Chen, A.Battestini, N.Gelfand, and V.Setlur. Visual Summaries of Popular Landmarks From Community Photo Collections. pp. 1248–1255. IEEE, 2009.

[5] D.Crandall, L.Backstrom, D.Huttenlocher, and J.Kleinberg. Mapping the World's Photos. In *in Proc.of WWW*, pp. 268–288, 2005.

[6] J.MacQueen. Some methods for classification and analysis of multivariate observations. In *in Proc.of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[7] B.V.Dasarathy. *Nearest Neighbor (NN) Norms NN Pattern Classification Techniques*. IEEE Computer Society, 1991.