7P-03

スポット情報組み込みモデルによる回遊行動データの分類

鈴木　優伽 †　　斉藤　和巳 †　　風間　一洋 ††
静岡県立大学経営情報学部　　和歌山大学システム工学部

1　はじめに

近年，TripAdvisor をはじめとする旅行レビューサイトの発達や GPS 機能付きスマートフォンの普及から，回遊者の行動データが容易に所得可能となり，それら行動データを利用した研究が注目を集めている．中でも，回遊者の行動分析や観光ルート推奨システムに応用可能であることから，実世界で回遊者の関心が高い場所である POI(Point of Interest) の抽出など，行動データの分類に関する研究 [1][2] が盛んに行われている．例えば，写真サイト Flickr に投稿された写真の位置情報を使用し，それらをクラスタリングすることでデータを分類している研究が存在する [3][4]．しかし，これらの研究では，単一ソーシャルメディア上の行動データのみで構成しているため，意味づけが困難である可能性がある．そのため，本研究では，より柔軟な分類を目的に，行動データと他のソーシャルメディアから得られるスポット情報を突合して分類するモデルを提案する．ソーシャルメディアに Flickr と TripAdvisor を用いて評価実験を行い，提案モデルで妥当な分類結果が得られることを示す．

2　提案手法

2.1　提案分類モデル

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を $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ を考え，他のソーシャルメディアから得られるスポット集合 $\mathcal{K} = \{1, \cdots, K\}$ のうち，k 番目のスポットの位置情報を $\mathbf{y}_k = (Lat, Lng)^T$ とし，人気度 $g(k)$ をスポット情報として定義する．ただし，(Lat, Lng) は位置データを意味し，上付き T は転置を表す．

回遊者の行動データはいくつかの中心 (ピーク) と広がりをもった分布で分類できるとする．各分布の中心を他のソーシャルメディから得られたスポットの位置情報 $\mathcal{Y} = \{\mathbf{y_1}, \cdots, \mathbf{y_K}\}$，広がりを 2 次元分散共分散行列の集合 $\mathcal{S} = \{\mathbf{S}_1, \cdots, \mathbf{S}_K\}$ 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，$\mathbf{I}$ 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 $N(\mathbf{x}_n; \mathbf{y}_k, \mathbf{S}_k)$ を以下で表す．

$$\begin{aligned} N(\mathbf{x}_n; \mathbf{y}_k, \mathbf{S}_k) &= \frac{\exp\Big(-\frac{1}{2}(\mathbf{x}_n - \mathbf{y}_k)^T\mathbf{S}_k^{-1}(\mathbf{x}_n - \mathbf{y}_k)\Big)}{(2\pi)^{\frac{d}{2}}|\mathbf{S}_k|^{\frac{1}{2}}}, \\ &\propto \frac{1}{s_k}\exp\Big(-\frac{1}{2s_k}\|\mathbf{x}_n - \mathbf{y}_k\|^2\Big). \end{aligned} \quad (1)$$

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 α_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 $\mathcal{A} = \{\alpha_1, \cdots, \alpha_k\}$ とすると，行動データ $\mathbf{x}_n$ が各分布に属する確率 $p(\mathbf{x}_n)$ は，以下の混合ガウス分布を用いた確率モデルで算出される．

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

本研究では，上記の混合ガウス分布に潜在変数 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．

2.2　パラメータ学習

導入する潜在変数 z は 1-of-K 表現で表せる K 次元ベクトルであり，$p(z_k = 1) = \alpha_k$，$p(x|z_k = 1) = N(\mathbf{x}_n; \mathbf{y}_k, \mathbf{S}_k)$ であるとする．また，データ $\mathbf{x}_n$ が与えられた下での z の事後確率を以下で定義する．

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

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各パラメータ $\mathcal{S}$ を求めていく．

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

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;

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

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;

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

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

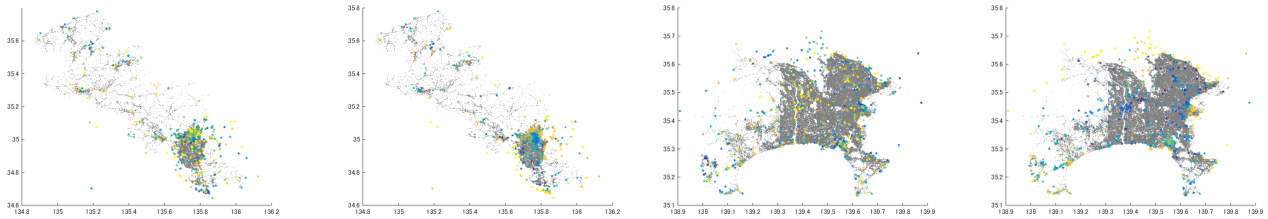
本実験では，ソーシャルメディアとして，写真共有サイト Flickr と旅行レビューサイト TripAdvisor(以下，TA) を採用し，京都と神奈川を対象にデータを収集した．Flickr から収集した写真に付する位置情報と撮影時刻等の時間情報から回遊者の行動データを構築し，TA から得られた各観光スポットの緯度・経度をスポットの位置情報，レビュー数をスポットの人気度とみなした．京都では回遊データ数 76999，スポット情報数 649，神奈川では回遊データ数 166712，スポット情報数 778 である．また，図 1 に用いたレビュー数 (人気度) の分布を示す．スポットに投稿されたレビュー数の平均は，京都で 40，神奈川で 36 である．

Classification of user behavior data based on model with spot information

†Yuka SUZUKI　†Kazumi SAITO　††Kazuhiro KAZAMA
†School of Management and Information, University of Shizuoka
††Faculty of Systems Engineering, Wakayama University

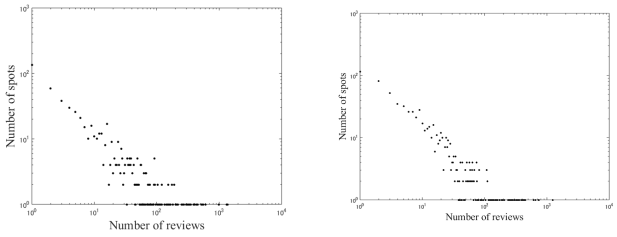
表 1: 事後確率上位 10 スポット

rank	京都 $M1$	京都 $M2$	神奈川 $M1$	神奈川 $M2$
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



(a) 京都 ($M1$)　(b) 京都 ($M2$)　(c) 神奈川 ($M1$)　(d) 神奈川 ($M2$)

図 2: 各スポットを事後確率で彩色した結果



京都　神奈川

図 1: レビュー数分布

3.1 事後確率上位スポットの比較

表 1 に，モデルを用いて分類後，各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．また，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．すなわち，比較モデルではアルゴリズム A3 に以下の更新式を追加する．

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

表中，本モデルを $M1$，比較モデルを $M2$ と表す．図 2 に，各スポットを，青色ほど事後確率が高く，黄色ほど低いように彩色して可視化する．図 2 から，$M1, M2$ では分類結果が異なることが確認できる．同様に表 1 から，$M1$ では一般的に多くの写真が撮られていると予測できるスポットにデータが分類されており，分類後の意味づけが容易であることが分かる．一方 $M2$ では，自身の知名度は高くないが周辺のスポットの知名度は高いようなスポットや，訪れる人が少ないが，確実に写真が撮られているようなスポットに分類されており，一部の回遊者の行動が全体の分類に強く影響を与えていることが分かる．これを裏付ける理由として，特に人気があるとは言えないが，その周辺には他のスポットがないような場所は事後確率が高くなることが挙げられる．我々が目指しているのは，意味づけが容易で柔軟な解釈が可能である分類モデルであり，また，一部の行動データに影響されることなく高い精度で分類可能であることが望ましい．そのため，$M1$ を用いる方が適切であると考えられる．

4 おわりに

本研究では，人気度等のスポット情報を考慮したデータ分類法を提案し，その有効性を検証した．評価実験では，スポット情報を考慮せずにパラメータ $\mathcal{A}, \mathcal{S}$ を学習するモデルを比較モデルとして用いた．その結果，スポット情報を考慮した本モデルでの分類の方が，より正確な分類が可能であり，分類後の意味づけも容易であることが確認できた．今後は，分類結果の評価法の改善を図ると共に，人気度以外の情報を組み込んだモデルの構築を試みる．

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] D.Comaniciu and P.Meer. Mean Shift: A Robust Approach Toward Feature Space Analysis. pp. 603–619. IEEE, 2002.

[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] W.Chen, A.Battestini, N.Gelfand, and V.Setlur. Visual summaries of popular landmarks from community photo collections. pp. 1248–1255. IEEE, 2009.

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