人の流れデータを用いた人々の人口統計学的属性の推定

Estimation of Human Demographic Attributes from Person Flow Data

学籍番号　47-136760
氏　　名　西村隆宏 （Nishimura,Takahiro）
指導教員　柴崎 亮介 教授

1. 概要

近年，災害発生時の緊急対応や，交通状況の詳細な把握，アドテクノロジーに代表されるデジタルアドバタイジングの観点から，「人の流れ」に関する非集計の位置情報に対する産学官の関心は非常に高い．しかし，位置情報発信機能を持つ機器からは緯度経度情報と時間情報程度しか取得できないため，詳細な分析を行うには様々な属性情報を付与する必要がある．特に人口統計学的属性と呼ばれる「性別」，「年齢」，「職業」の 3 属性は分析の基本となる切り口であるため特に重要な属性である．しかし，属性つき人の流れデータは機密性が非常に高いため様々な人が使用できるようになるには困難を多数乗り越えなくてはならない．

この困難を解決する手段として，属性をユーザーから直接取得するのではなく，行動履歴を元にして推定する手段が考えられる．しかし推定モデルを構築するには大量の属性付き行動履歴データが必要になるため，推定モデルを構築することは非常に困難である．松尾ほか[1]はセンサを用いて屋内における人の行動特徴から人々の属性を推定したが，GPS を代表する屋外測位手法には適用が難しい．パーソントリップ調査（PT）は大規模，広域の人の流れの調査で最も利活用が進んでいる調査だが，この調査で取得できる人の位置情報は起終点のおおまかなエリアのみである．そのため一人ひとりを細かく観察する分析には何らかの方法で位置情報を確率的に再配分する必要がある．

本研究は，ユーザーの移動履歴を元にユーザーの人口統計学的属性を推定するモデルを構築する．属性推定モデルを構築するに際し，PT と GPS データという，特徴が異なる 2 つのデータでそれぞれ属性推定モデルを構築し，データの特徴の違いによる識別性能の違いについて考察する．さらに特徴の異なる両データを組み合わせ構築した属性推定モデルについて考察し，多種多様な移動履歴情報への適用について検討する．

表 1 使用したデータ

データ名	人数	調査期間
PT調査	約60万人	平日1日
GPSデータ	159人	1ヶ月

2. 研究手法

2.1 使用データと特徴量

本研究では空間配分版東京都市圏 PT と KDDI 研究所提供のユーザーの利用許諾済 GPS データを用いて人口統計学的属性の推定モデルを構築する．人の移動に着目したとき，得られる特徴は，大きく 3 つが考えられる．

- 1 日の移動履歴
- 滞在場所の特徴
- 行動パターン

この特徴のうち, PT は調査期間が平日 1 日のみという特徴があるため, 1 日の移動履歴と滞在場所の特徴をユーザーの行動から抽出し, GPS データは調査期間が 1 ヶ月のため上記 3 つの特徴を平休日別に抽出した. 抽出した特徴は以下の通りである.

- {自宅/勤務地}出発時間
- {自宅/勤務地}到着時間
- {自宅/勤務地}訪問回数
- 1 日の総行動距離
- 時間別行動距離
- 1 日の総滞留点数
- 居住地から勤務地までの距離
- グルーピングした地域メッシュの総滞在時間

GPS データは移動履歴情報が複数日あるため, これら特徴量を平休日別に全ての計測日において取得し, 取得した特徴量の平均を 1 日の平均移動履歴に設定した. また, 取得した特徴量の分散を行動パターンとみなし, 分析に使用する特徴量に加えた.

滞在場所の特徴量抽出は, 事業所・企業統計調査という政府が実施する, 1/2 地域メッシュ内の業種別事業所数のデータと, 株式会社マイクロベース提供の 1/2 地域メッシュ内の 5 歳階級別人口データを使用した. ２つの地域情報を元にメッシュをクラスタ数を自動で決定する手法の一つである, xmeans 法を用いてクラスタリングし, 各クラスタへの滞留時間を滞在に関する特徴抽出とした.

2.2 使用したパターン認識手法と検証

人口統計学的属性推定モデルの構築には

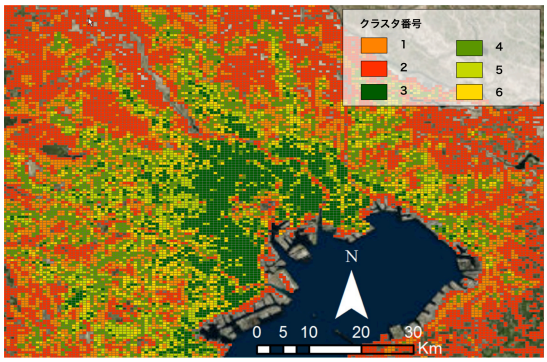

図 1 構築した地域特徴データ

主流のパターン認識手法である, Neural Network, Random Forest, Support Vector Machine を用いた. 設定する必要のあるハイパーパラメータは未知であるためグリッドサーチにより決定した状態で試行し, 前処理に次元圧縮や特徴空間の直交化に用いられる主成分分析の使用有無, 滞在場所の特徴考慮, 滞留時間の考慮有無の計 18 パターンを適用し, 最も識別性能が高い属性推定モデルを採用した. GPS データを元にした属性推定モデルの構築では, さらに滞留場所の特徴集計期間の違い, パーソントリップ調査で構築したモデルの適用による 30 パターンを試行した.

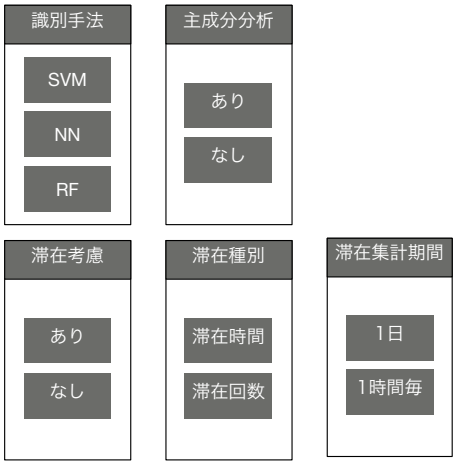

図 2 本研究で試行したモデル構築パターン

属性推定モデルの性能評価は, 事前にデータ集合を学習データ集合と検証データ集

合にランダムに分配し，学習データ集合で構築されたモデルを検証データ集合に適用した時の，F-measure を評価指標とした．また，使用した人口統計学的属性情報は事前学習によりクラスが非常に細かい事による学習困難性が判明したため，年代属性は 5 歳階級別から 10 歳階級別，職業に関してはサラリーマン，学生，主婦，その他の 4 つにクラスを集約して属性推定モデルを構築した．

3．属性推定モデルの構築結果

3.1 パーソントリップ調査を元に，構築した属性推定モデルの結果

PT を元に属性推定モデルを構築した結果，前処理で主成分分析を行わない Random Forest が，安定的に識別性能が良い事がわかった．これは使用した特徴量間の相関がほとんどないため，特徴空間の軸に対し垂直方向の識別境界を描く Random Forest が効果的に作用したと考えられる．表 2 は各属性推定モデル構築の変数重要度を計算した結果である．これより，自宅勤務地の距離，行動距離，滞留時間のばらつき，早朝の行動，幹線道路隣接商業地域への滞留が性別の違いに，自宅や勤務地の出発時間や行動距離が年代の違いに，滞在開始時間のばらつきや自宅や勤務地の発着時間が職業の違いに関係していることがわかった．また，どの人口統計学的属性においても深夜帯の行動特徴は属性に関係なく，性別や年代は滞在する場所と関係がある事がわかった．

3.2 GPS データを元に構築した属性推定モデルの結果

GPS データを元に属性推定モデルを構築した結果，サンプル数<使用パラメータ数であったため，SVM を用いた推定モデルは全て識別性能が低かった．性別推定モデルにおいては PT から構築した推定モデルより良い識別性能を持つモデルが構築できた．これは休日 7 時の行動傾向に男女差が大きい事が原因であり，PT ではこの特徴を抽出できない．よって GPS データから構築した推定モデルが高い識別性能であったと考えられる．また，場所特徴の集計期間はサンプル数が少ないため，1 時間おきの集計より 1 日全体の集計の方がモデルの安定性が高い．しかし，1 時間おきの集計は 1 日全体の集計と比べ情報量が多いため，サンプル数が増えることで良い識別性能を生む要因になると考えられる．年代や職業推定モ

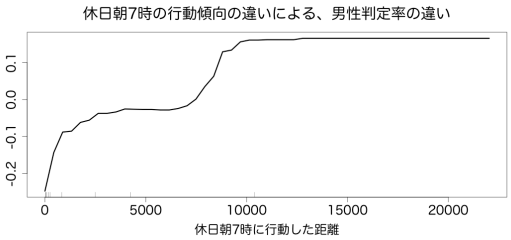

図 4 休日朝 7 時の行動傾向の違いによる男性判定率の違い

表 2 PT から構築した各属性推定モデルの変数重要度

	性別		年代		職業	
	特徴量	変数重要度	特徴量	変数重要度	特徴量	変数重要度
1	行動距離	3194.921	行動距離	4452.572	滞留時間の分散	6148.948
2	滞留時間の分散	2941.177	自宅勤務地距離	4066.363	退社時間	5648.507
3	自宅勤務地距離	2695.569	帰宅時間	3787.804	帰宅時間	5588.521
4	7 時の行動距離	2050.485	退社時間	3697.906	行動距離	4204.337
5	幹線道路隣接商業地域	1696.148	滞留時間の分散	3626.717	自宅出発時間	4137.308

デルはクラス数が多いため, Random Forest の使用に十分なサンプル数ではなかったと考えられる．そのため本研究で試行した識別手法の中で最も単純な Neural Network が他の手法より高い識別性能であることが多いと考えられる．また，パーソントリップ調査で用いた属性推定モデルを GPS データに対し適用した結果，年代推定において，データが異なるものの識別性能の差が GPS データから構築したモデルとほとんど変わらないため，年代推定に対しては適用可能である．

表 3 PT から構築したモデルを GPS データに適用した時の結果

Attribute	Accuracy	Recall	Precision	F-measure
性別	0.532	0.475	0.25	0.176
年代	0.587	0.299	0.115	0.25
職業	0.545	0.075	0.704	0.207

表 3 属性別の最良モデル

属性	データ取得期間	パターン認識手法	PCA使用有無	滞在時間考慮有無	滞在特徴量集計期間	F-measure
性別	1日	Random Forest	なし	あり	1日	0.61
性別	1ヶ月	Random Forest	あり	なし	1時間	0.8
年代	1日	Random Forest	なし	あり	1日	0.48
年代	1ヶ月	Neural Network	なし	あり	1日	0.29
職業	1日	Neural Network	なし	なし	1日	0.61
職業	1ヶ月	Neural Network	なし	あり	1時間	0.89

3　まとめ

本研究は人の行動に着目して，人の行動履歴を元にその人の性別，年代，職業といった人口統計学的属性の推定モデルの構築を試みた．そこでパーソントリップ調査と GPS データといった 2 つの異なるデータを用い，特徴の抽出可能性，特徴選択，手法や前処理，集計方法の違いによる計 51 パターンの推定モデルを各属性に対して構築し，属性推定モデルの識別性能について検証した．

その結果，性別，職業の 2 つのモデルに関しては識別性能の高い属性推定モデルが構築でき，属性の違いが行動距離や特定の時間の行動パターンに影響を与えている事がわかった．一方で年代推定モデルに関しては本研究で属性推定モデルを構築するために用いた特徴量や場所の特徴データではうまく分類できない事がわかった．

PT から構築したモデルを GPS データに適用した結果，類似した異なるデータで属性推定モデルを構築した場合でも年代推定に関しては識別が行える事がわかったほか，属性内の各クラスに属するサンプル数が十分に多い場合 Random Forest，少ない場合 Neural Network を使用することで識別性能が高い推定モデルを構築することができることがわかった

今後の展望は抽出する特徴量を増やしモデルを改善する他, Tr-adaboost に代表される Inductive Transfer Learning と Laplacian Support Vector Machine 等の Online-Semi-Supervised Learning を組み合わせ頑健なモデルを構築することである．

本研究では東京都市圏を対象に人口統計学的属性推定モデルを構築したため，今後は別の都市を対象に同様の属性推定モデル構築し，推定モデルの差による都市の比較を行う事である．

参考文献

[1] 松尾豊，岡崎直観，中村嘉志，西村拓一，橋田浩一，中島秀之．位置履歴か らのユーザ属性の推定．情報処理学会論文誌 48(6), 2106-2117, 2007

[2] 斎藤参郎，椿井昌邦，中嶋貴昭．都心商業空間における商業施設への消費 者来街者数と回遊パタンの同時推定逆問題について．地域学研究 30(1), 213-229, 1999