人の流れデータセットを用いたデモグラフィック属性の推定及び GPS データへの適用可能性に関する研究

西村隆宏・秋山祐樹・金杉洋・Teerayut Horanont・柴崎亮介・関本義秀

Study of Estimate Human Demographic Attributes Using Person Flow Datasets and Application It for GPS Log Data

Takahiro NISHIMURA, Yuki AKIYAMA, Hiroshi, KANASUGI, Teerayut HORANONT, Ryosuke SHIBASAKI and Yoshihide SEKIMOTO

Abstract:In recent years, we have high-function mobile device called smartphone. As a result, we can get various information; location, time, gyro and so on. To analyze mobile phone GPS log data is specially focused on but it is hard to get validation data. We focus on Person Flow dataset (PFlow dataset) made from Person Trip survey (PT) that has label of demographic attributes, transfer history and develop model from it is applied for GPS log data using Transductive Transfer Learning. We developed gender, age and work type classifier model from PFlow dataset. As a result, we can classify high accuracy for these demographic attributes.

Keywords: 「GPS データ(GPS log data)」,「人流データ(Pflow data)」, 「トランスダクティブ学習(Transductive Transfer Learning)」

1. はじめに

近年，消費者の嗜好が多様化しマスマーケティングから個人へ最適化されたマーケティングへと移行している．そのため，企業は消費者のログデータを解析し，個人へ最適化されたマーケティング活動を行っている．特に近年はスマートフォンの普及により携帯電話に内蔵されたジャイロやGPSによる位置情報などからユーザーの行動を反映した様々な情報を取得できる．しかし，これらの情報からは緯度経度情報や角速度情報しか取得できず，マーケティング活動に利用するには何らかの方法でユーザーのデモグラフィック属性を推定する必要がある．さらにこの問題はデモグラフィック属性付き位置情報データの取得困難性により，単一データの利用ではデモグラフィック属性分類モデルを構築する事は困難である.本研究は携帯電話から取得される位置情報履歴データと類似性が高い人の流れデータセットを用いて分類モデルを構築し，位置情報履歴に適用する事を目的とする．人の流れデータセットは政府によるパーソントリップ調査（PT 調査）を高精細化したデータであり，PT 調査で付与されたデモグラフィック属性と緯度経度情報の両方が付与されている．さらにトランスダクティブ転移学習の適用可能性について，携帯電話の位置情報履歴に構築モデルを適用できるか考察する.

西村隆宏　〒181-0004　東京都目黒区駒場 4-6-1
東京大学生産技術研究所 Cw-503
Phone: 03-5452-6412
E-mail: nishimura@csis.u-tokyo.ac.jp

2. 手法の概略

2.1　人の流れデータセット

本研究では人の流れデータセットを用いてデ

モグラフィック属性の推定を行う．本データはPT調査を高精細化したデータであり，1分間隔でユーザーID，緯度経度，移動目的，性別，年代，職業が付与されている．表1は付与されている3つのデモグラフィック属性の一覧である．

表1　デモグラフィック属性の一覧

属性番号	性別	年代	職業
1	男性	-5歳	農林水産業従事者
2	女性	5-10歳	生産工程／労務作業者
3		10-15歳	販売従事者
4		15-20歳	サービス職業従事者
5		20-25歳	運輸・通信従事者
6		25-30歳	保安職業従事者
7		30-35歳	事務従事者
8		35-40歳	専門的・技術的職業従事者
9		40-45歳	管理的職業従事者
10		45-50歳	その他職業
11		50-55歳	園児・小学生・中学生
12		55-60歳	高校生
13		60-65歳	大学生・短大生・各種専門学校生
14		65-70歳	主婦・主夫
15		70-75歳	無職
16		75-80歳	その他
17		80-85歳	不明
		85歳-	

2.2　特徴量抽出及び学習，検証

本研究はデモグラフィック属性の学習器としてSVMを採用した．SVMは機械学習手法の一つであり、カーネルトリックと呼ばれる非線形写像を内部的に行うことで非常に優れた分類性能を示す手法である．使用したカーネルにはガウシアンカーネルであり、SVM で最も使用されるカーネル関数を利用した．表2は抽出した特徴量であり，性別，年代，職業の3つに関して同一の特徴量を使用し，それぞれ SVM を用いて学習器を作成した．

表2　抽出した特徴量

特徴量	単位
居住地出発時刻	時
居住地到着時刻	時
勤務地出発時刻	時
勤務地到着時刻	時
居住地訪問回数	
勤務地訪問回数	
移動距離総計	m/日
時間あたり移動距離	m/時
滞留時刻の分散	
滞留点数	

また，学習に用いるサンプル数を決定した．nをサンプル数，Tを処理時間，Eを汎化誤差とするとサンプル数の決定式は式(1)となる．

$$argmin(T(n)/E(n)) \quad (1)$$

また滞留点の決定には人の流れデータセットに付与されているトリップ ID を元に構築し，滞留点の集約閾値は羽田野(2012)の手法を用いた．

モデル検証は学習に利用するサンプル数と同数のサンプルを別途用意し，推定値の正解率を検証結果として採用した．

2.3　トランスダクティブ転移学習への適用考察

転移学習とは学習に用いるソースデータと適用するターゲットデータが異なる学習手法であり，ソースデータとターゲットデータのラベルの有無に4タイプに分類される．本研究ではラベルつきソースデータとして人の流れデータセットを利用し，ターゲットデータとしてラベルなしGPSログデータを利用する．表3は転移学習の種類の一覧であり，本研究の手法はトランスダクティブ転移学習に分類される．

表3 転移学習の種類

		Does Target have Labels?	
		Yes	No
Does Source have Lebals?	Yes	Inductive Transfer Learning	Transductive Transfer Learning
	No	Self-Taught Learning	Unsupervised Transfer Learning

また，トランスダクティブ転移学習はソースデータとターゲットデータ間で以下の式（2）を満たす必要がある．

$$P[Y^{(s)}|X^{(s)}] = P[Y^{(t)}|X^{(t)}] \quad (2)$$

大数の法則によりGPSデータの場合は2.2.1式を満たすことが期待されているが，人の流れデータセットの場合はサンプル数によって母集団の分布と異なる可能性が生じる．そこで定量的に母集団との同一性を検証するために

Kullback-Leibler divergence(KL divergence)を計算し，母集団のヒストグラムと人の流れデータセットからサンプリングしたヒストグラムの類似度を計算する．式3はKL divergenceの導出式でありP,Qは離散確率分布である．
KL divergenceの値が0の時，２つの分布は等しいといえる．

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \quad (3)$$

3. 適用例

3.1　サンプル数決定問題の結果

図1はサンプル数を変えつつ学習を行った結果である．

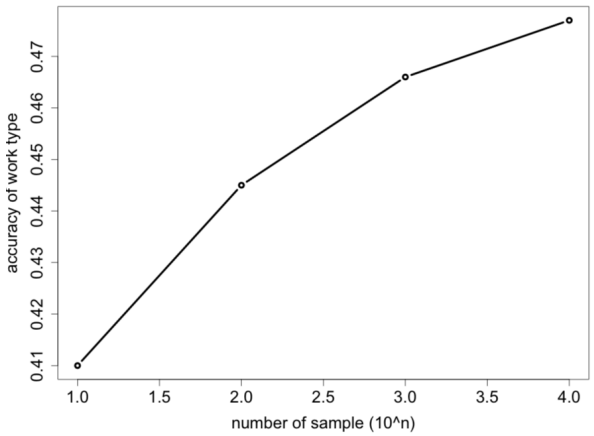

図1 サンプル数の違いによる学習効率の結果

図1の通り，学習に用いるサンプル数は5000が適当であると思われる．よって以降の学習では5000人をランダムサンプリングして学習を行う．

3.2　各デモグラフィック属性の推定結果

3.2.1 付与されている属性ラベルを用いた場合

表4は表2で示した特徴量を用いて性別，年代，職業の3属性を学習した結果である．

表4 3属性の学習結果

属性	accuracy
性別	0.67
年代	0.25
職業	0.4

表4の通り，全ての属性において学習がうまく進んでいないことがわかる．これは性別に関しては使用した特徴量では学習が難しいことが原因と考えられる．また，年代や職業はクラス数が多いため分類性能が低くなってしまうものと考えられる．そこで 学習前に年代と職業のクラスを集約したクラスを学習対象として再度学習を行った．表5は年代，職業のクラス集約を示した表である．

表5 年代と職業に関するクラス集約表

年代	属性集約後年代	職業	属性集約後職業
-5歳	10歳未満	農林水産業従事者	会社員
5-10歳		生産工程／労務作業者	
10-15歳	10代	販売従事者	
15-20歳		サービス職業従事者	
20-25歳	20代	運輸・通信従事者	
25-30歳		保安職業従事者	
30-35歳	30代	事務従事者	
35-40歳		専門的・技術的職業従事者	
40-45歳	40代	管理的職業従事者	
45-50歳		その他職業	
50-55歳	50代	園児・小学生・中学生	学生
55-60歳		高校生	
60-65歳	60代	大学生・短大生・各種専門学校生	
65-70歳		主婦・主夫	主夫
70-75歳	70代	無職	無職・その他
75-80歳		その他	
80-85歳	80歳以上	不明	
85歳-			

3.2.2 クラス集約後の推定結果

3.2.1で集約したクラスを用いて学習を行った．表6はクラス集約後の学習結果である．

表6 クラス集約後の学習結果

属性	accuracy	recall	precision	f-measure
性別	0.67	0.644	0.64	0.64
年代	0.37	0.33	0.29	0.244
職業	0.82	0.74	0.69	0.711

クラスを集約し学習を行った結果，職業の推定に関しては高いF-measureを検出した。一方で年代に関してはクラス集約前と比較するとややaccuracyが上昇した．

3.3　転移学習の適用可能性結果

まず KL divergence に用いる分布を示す．図2.1は人の流れデータセットに含まれるユーザーの人口ピラミッドで図2.2は東京都の人口ピラミッドである．

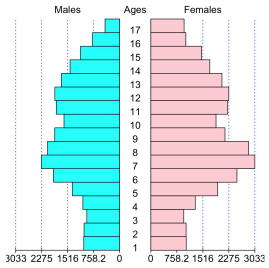

図 2.1 人の流れデータセットの人口ピラミッド

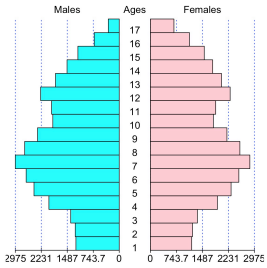

図 2.2 東京都の人口ピラミッド

また，表7は性別に計算したKL divergenceの平均値である．表 7 の値が非常に小さいことより両者の分布はほぼ一致しているみなせるため，人の流れデータセットは転移学習に利用可能であると考えられる．

表 7 KL divergence の計算結果

性別	KL Divergence
男性	0.0168
女性	0.0116

4. おわりに

本研究はデモグラフィック属性と緯度経度が付与されている人の流れデータセットを用いて，ユーザーの移動情報を元にユーザーのデモグラフィック属性を推定した．また，トランスダクティブ転移学習を適用するために人の流れデータセットと携帯電話の位置情報履歴との類似性を考察した．その結果，プレトレーニングでは人の流れデータセットに含まれる全ユーザーのうち，5000 ユーザーをサンプリングすることで十分な精度を保った状態で学習が行われることがわかった．またSVMを用いて学習した結果，高精度で職業の推定が行えることが分かった一方で，性別，年齢ラベルの推定精度は低かった．さらにKL divergenceの値より，人の流れデータセットを用いてGPSデータのクラス分類問題を解くトランスダクティブ転移学習の手法を適用できることが分かった．

本研究では移動情報のみ使用してデモグラフィック属性の推定を行ったため，構築したモデルには滞留地点の特徴が含まれていない．今後は滞留地点の特徴をモデルに適用し，GPS データへ適用後の精度検証を行う他，他地域の人の流れデータセットを用いて推定モデルを構築したい．

謝辞

本研究を進めるにあたり東京大学空間情報科学研究センターより人の流れデータセットを提供していただきました．ここに感謝の意を申し上げます．

参考文献

人の流れプロジェクト
http://pflow.csis.u-tokyo.ac.jp/index-j.html.

Akiyama, Y., Takada, T. and Shibasaki, R. (2013): "Development of Micropopulation Census through Disaggregation of National Population Census", CUPUM2013 conference papers, 110.

羽田野真由美・上山智士・秋山祐樹・Horanont Teerayut・柴崎亮介 (2012):GPSデータを用いた商業集積地来訪者の行動パターン抽出方法の検討，第２１回地理情報システム学会講演論文集（CD-ROM, F-3-4）

Arnold A., Ramesh N., and William W. C..(2007): "A comparative study of methods for transductive transfer learning". Data Mining Workshops, ICDM Workshops 2007. Seventh IEEE International Conference