1F4-OS-09a-2

教師なし学習を用いた移動軌跡データからの意味情報推定

Estimating Semantic Information from Trajectory Data by Unsupervised Learning

田中 優子 *1　　上原 邦昭 *1
Yuko Tanaka　　Kuniaki Uehara

*1神戸大学大学院システム情報学研究科計算科学専攻
Graduate School of System Informatics, Kobe University

Because of the large amount of trajectory data produced by mobile devices, behavioral analysis using trajectory data are widely studied. However, raw GPS data consists of time series data of the coordinates, and does not have any semantic information. Furthermore, because of the problem of private protection, the personal attributes are also covered by the data. This paper estimates semantic information of trajectory data using multiple unsupervised learning methods. It is useful as a technique of the privacy-protection data mining by using data without the meaning information. This method estimates the personal attributes of trajectory data, by clustering characteristics such as behavior time. Then, to improve the estimation accuracy, we introduce a cluster ensemble. We perform an experiment using trajectory data in Tokyo metropolitan area and estimate 5 job attributes. In addition, we have also estimated the means of transportation and the semantic places such as office and school.

1. はじめに

人の行動推定や行動予測では，機械学習の手法が使われている．GPS データのような人の移動軌跡の場合，データは座標と時間の時系列データであり，行動分析に必要な移動手段や滞在場所などの意味情報は付随されていない．[田中 14] では，移動軌跡データから行動分析を行うために，機械学習を用いてオントロジーを構築し，データに意味付けをする研究を提案している．しかし，意味付けの過程では教師あり学習を用いている．すなわち，あらかじめ用意された意味情報付きデータを用いる必要がある．現実的には，移動軌跡データに意味情報が付加されたデータを用意することは困難である．

さらにプライバシーの問題があるために，移動軌跡データには人の属性情報も付けられていないことがある．そのため，GPS データを用いた災害避難行動分析 [宋 12] では，個人属性は使わず，規則正しく通勤する人かどうか，その土地に良く訪問するかどうかなどの情報を推定し，分析することができるとしている．しかしながら，通常時には，人々は属性の違いによって行動の特徴にも違いがあるはずであり，行動の特徴を見れば個人属性が推定できると考えられる．そこで本研究では，移動軌跡データに意味情報を推定する手法を提案する．具体的には，データの個人属性を用いずに，教師なし学習の枠組みのなかで推定を行う，プライバシー保護データマイニングとして行動推定を行う仕組みを提案する．

提案手法では，教師なしデータ分類手法であるクラスタリングを用いて人の属性推定を行う．学生や社会人など，人の属性によって行動時間に違いがあると仮定し，行動時間を特徴量として人を分類する．クラスタリング手法には初期値に依存しやすいなどの欠点があり，精度が低くなるため，クラスタアンサンブルを導入する．クラスタアンサンブルは，複数のクラスタリング結果を統合して，より頑健性と安定度を高めた結果を得ることができる手法である．複数のクラスタリング結果を用いる際には，多様なクラスタを用意した方が高精度な結果が得られることが知られている．そのため，本研究では 2 種類の異なるクラスタリング手法を用いて，初期クラスタリングを行う．それぞれの手法で複数回クラスタリングを行った結果から，クラスタアンサンブルを実行する．

連絡先: 田中 優子，神戸大学大学院システム情報学研究科，神戸市灘区六甲台町 1-1，tanaka@ai.cs.kobe-u.ac.jp

2. クラスタリングの精度向上

2.1 クラスタリング

人間の行動分析には，人の属性，つまり年齢や職業といった情報が重要である．そこで，移動軌跡データから人の属性を推定することを考える．人は学生や社会人など属性によって行動時間に違いがあると考えられる．帰宅時間や移動時間などの時間を特徴量としてクラスタリングを行えば，人の属性別のクラスタが発見できる．

本研究では，クラスタアンサンブルを導入するため，複数のクラスタリングを行う．多様なクラスタリングを行うことが望ましいため，後に説明する 2 つの手法，PLSI と GMM を利用する．PLSI は，特徴量をカテゴリカルデータとして扱い，特徴量の共起性をもとに，データをクラスタリングする手法である．一方，GMM は特徴量を数値として扱っている．データは複数の正規分布から生成されたものであると仮定し，それぞれの分布を求めてクラスタリングを行うものである．

人の属性を分析するクラスタリング手法として，文書分類や購買データ分析などで使われる確率的潜在意味解析（PLSI）[Hofmann 99] を用いる．PLSI は潜在クラス分析の手法である．潜在クラス分析とは，観測変数の背後にカテゴリカルな潜在変数があると仮定して，潜在構造を説明するモデルである．潜在クラス分析は，確率的なクラスタリング手法としてみなすことができる．

軌跡データの利用の際には，以下の 3 つの仮定を PLSI でモデル化している．まず，人は潜在クラスで分類することができる．また，一日の行動は帰宅時間の早い遅いや，通勤時間の長短など，行動時間で分類することができる．そして，各潜在クラスは特定の行動時間の傾向がある．これを図に示したのが図 1 である．

各潜在クラスが職業にあたるものだと仮定すれば，人の属性を推定することができることになる．具体的な特徴として，主婦や小学生は家に帰る時間は早く，社会人は遅いと思われる．

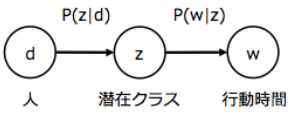

図 1: PLSI で構築するモデル

通学時間に関しては，小中学生は家から近い学校に通うため短く，高校生や大学生は通学時間が長い可能性がある．これらの行動知識を利用し，PLSI でクラスタリングを行い，人の属性を分類する．

混合正規分布（GMM）は，複数の正規分布を混ぜ合わせて表される確率モデルである．人の属性ごとの行動時間を考えた場合，正規分布になると仮定する．例えば，帰宅時間別の人数を考えると，主婦や小中学生は午後の早い時間に帰る人が多く，社会人では夜に帰る人が多いだろう．各属性ごとに正規分布があるとすれば，全データでは混合正規分布としてモデル化できると考えられる．さらに，他の特徴量として，通勤にかかる時間や外出時間を考えても，人の属性ごとに異なるピークがあると仮定できる．それぞれの特徴量について，GMM によって異なる正規分布を推定すれば，属性ごとのクラスタとして捉えることができる．

2.2 クラスタアンサンブル

上記のクラスタリング手法のような教師なし学習は，初期値・パラメータに依存しやすいという欠点があり，教師あり学習に比べて結果が不安定である．このため，クラスタアンサンブルに基づく教師なし学習を用いて，推定をより洗練化して精度を上げる手法を提案する．

クラスタアンサンブルは，異なる特徴量を使うなどして複数のクラスタリングを行い，それらの結果を統合して，より識別能力の高いクラスタを生成することである．手順としては，最初にデータの特徴量を用いて複数回クラスタリングを行い，複数のクラスタ（弱クラスタ）を得る．このとき，異なる特徴量を使う・異なる初期値やパラメータを使う・異なるクラスタリング手法を使うなどを適用して，多様性のある弱クラスタを用意する．そして，弱クラスタをもとに，データ間やクラスタ間の類似度を求め，最終的なクラスタ（強クラスタ）を求める．

弱クラスタは図 2 のように行列で表される．ここでは x が n 個のデータを表し，λ が m 回の初期クラスタリングを表す．各クラスタリングでは，k_i $(i \in \{1, ..., m\})$ 個のクラスタを生成し，$\sum_{i=0}^{m} k_i$ 個の弱クラスタを得る．この弱クラスタをもとに，最終的に k 個の強クラスタを得る．クラスタを統合する考え方には，行を統合するものと，列を統合するものの 2 種類に分けられる．

	λ_1	λ_2	⋯	λ_m
x_1	1	2	⋯	2
x_2	2	3	⋯	2
⋮	⋮	⋮		⋮
x_n	1	1	⋯	3

図 2: 弱クラスタの行列表現

行を統合する，つまりデータを統合する方法は，弱クラスタを元に全データ間の類似度を計算する．類似度をもとに，階層的クラスタリングなどの方法でデータを再度クラスタリングして，新たな k 個の強クラスタを得る．このような方法は，全データ間の類似度計算を行う必要があるため，大規模なデータでは計算量が膨大になる可能性がある．一方，列を統合する方法は，弱クラスタ間の類似度を求めて，類似しているクラスタの集まりを強クラスタとする．この手法であれば，計算量が少なく，大規模データにも適応することができる．本稿では，これら 2 種類の統合方法について，1 つずつクラスタアンサンブル手法を紹介する．

クラスタアンサンブルの手法として，Meta-Clustering Algorithm（MCLA）[Strehl 02] を利用する．MCLA は列を統合する，つまりクラスタを統合する考え方の手法であり，計算量が少なく，大規模データに適応できることが特徴である．MCLA では，弱クラスタをメタグラフとして扱う．各弱クラスタをノードと見なし，クラスタ間の類似度を辺の重みとしたグラフを生成する．クラスタ A，B 間の類似度はジャッカール係数を用いて以下の式で定義される．

$$Similarity(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

ここで，$|A|$ はクラスタ A 中のデータ数を表す．生成したメタグラフを，グラフ分割問題として解けば，k 個の弱クラスタの集まりに分割できる．これを強クラスタとする．グラフ分割問題には，オープンソースの実装ライブラリ METIS を用いる．そして最後に，各データの属する強クラスタを決定する．弱クラスタを強クラスタに変換し，最も属する数の多い強クラスタを選択する．ここで，属する数が同じ強クラスタが複数ある場合は，その中からランダムで選ぶ．

MCLA の計算量は $O(nk^2m^2)$ である．MCLA のようなクラスタ統合を行う方法を用いれば，計算量が少ないというメリットがある．しかし，MCLA ではデータがどの強クラスタに属するか判断できない場合があり，弱クラスタの精度や数によっては，ランダムで振り分けられるものが多くなる可能性がある．

もう一つのクラスタアンサンブルの手法として，Link-based Cluster Ensemble（LCE）[Iam-On 12] を紹介する．LCE はクラスタ間類似度を求めた後に，行（データ）を統合して強クラスタを求める手法である．図 2 で示した初期クラスタは，図 3 のような 2 値関連行列に変換することができる．各データがクラスタに属するか属さないかを 1 or 0 で表現している．LCE では，この 2 値行列中で値が 0 となっている所を，クラスタ間の類似度の値に変更した改良行列 RM を用いる．RM 中の値は，次の式で表される．

$$RM(x_i, cl) = \begin{cases} 1 & if cl = C^t(x_i) \\ sim(cl, C^t(x_i)) & otherwise \end{cases} \tag{2}$$

ここで，$C^t(x_i)$ はデータ x_i が属するクラスタを指す．

クラスタ間類似度は，Weighted Triple-Quality（WTQ）という方法で求められる．WTQ では，MCLA の時と同様に，ノードを弱クラスタ，ジャッカール係数を重みとしたグラフを考える．あるクラスタ C_x と C_y の類似度を求めるには，C_x・C_y の両方と辺で結ばれているクラスタ C_k を考える（この関係は triple と呼ばれる）．これら 3 つのノード間の WTQ スコアは，次の式で表される．

$$WTQ^k_{x,y} = \frac{1}{W_k} \tag{3}$$

	$\lambda_{1,1}$	$\lambda_{1,2}$	$\lambda_{2,1}$	$\lambda_{2,2}$	$\lambda_{2,3}$	⋯
x_1	1	0	0	1	0	⋯
x_2	0	1	0	0	1	⋯
⋮			⋮			
x_n	1	0	1	0	0	⋯

図 3: 弱クラスタの 2 値関連行列での表現

ここで，W_k はクラスタ C_k に隣接する辺の重みの合計である．そして，クラスタ C_x と C_y 間の WTQ スコアは，全ての triple なクラスタ（1...q）に対する WTQ スコアの合計として次の式で表される．

$$WTQ_{x,y} = \sum_{k=1}^{q} WTQ_{x,y}^{k} \tag{4}$$

最終的に，クラスタ C_x と C_y 間の類似度は次の式で表される．

$$sim(x,y) = \frac{WTQ_{x,y}}{WTQ_{max}} \tag{5}$$

RM の値が求まった後，グラフ分割手法を元にして最終的クラスタを得る．RM は重み付き 2 部グラフに変換される．グラフは全データと全弱クラスタをノードとし，RM の値を辺の重みとする．データ間，クラスタ間は辺がないため，2 部グラフとなる．このグラフを対称行列で表し，k 個の固有ベクトルを求める．固有ベクトルの値を特徴量として，各行を k-means 法でクラスタリングし，最終的な強クラスタを得る．

LCE 手法では，弱クラスタでデータが属さないクラスタに対しても，属するクラスタと似ていれば値を与えることにより，より精度の高い強クラスタを得ることができる．一方で，$(n+m)$ 次の正方行列の計算をする必要があるため，大規模データではメモリ不足などの懸念がある．

3. 評価

実験データには，東京大学空間情報科学研究センター（CSIS）が提供している平成 20 年度東京都市圏の「人の流れデータ」[人の流れプロジェクト 13] を用いる．このデータは，パーソントリップ調査（PT 調査）データに対し，GPS データのような座標の時系列データとなるように，ジオコーディング処理を施したものである．PT 調査とは，どのような人が，どのような目的で，どのような移動手段で，どこからどこへ移動したかを把握するために，都市圏で行われているアンケート調査である．そのため，人の流れデータには，時間と座標以外にも性別や年齢，職業など人の属性の情報や，移動手段・移動目的などの情報が含まれている．今回の実験では，これらを正解ラベルとして用い，社会人・主婦・小中学生・高校生・大学生の 5 つの職業に分けて推定を行った．使用したデータは一日分の移動軌跡 257,575 件である．一日中座標が変化しないものは属性を推定できないため，あらかじめ排除している．PLSI の実装には，オープンソースソフトウェアである plsi-0.03 を利用している．GMM には統計解析ツール R を用いている．MCLA 中で用いるグラフ分割は，METIS-5.1.0 を用いている．

3.1 実験結果

クラスタリング手法 PLSI・GMM・クラスタアンサンブル手法 MCLA を用いて，人の属性推定を行った．PLSI で指定する温度パラメータは 0.75 としている．MCLA の実験では，PLSI でクラスタ数 $k = 3,5$ を指定した結果と，GMM で帰宅時間・移動時間（通勤時間）・外出時間の 3 つのデータ特徴量を用いた場合の結果を弱クラスタとして利用している．推定の結果，PLSI では 42.7 %，GMM では 59.1 %，MCLA では 66.3 %の精度となった．精度を見れば，各クラスタリング手法のみを使う場合に比べて，クラスタアンサンブルの導入により推定の精度が上がっていることが分かる．

MCLA では，まず弱クラスタを統合した強クラスタが生成され，弱クラスタごとにラベルが変更される．どのような弱クラスタが統合されたかを見るため，統合によるラベル変化を表 1 に示す．3 つの GMM では，社会人クラスタが統合されて，社会人の強クラスタとなっている．また，帰宅時間を特徴量とした GMM では，弱クラスタと強クラスタのラベルが全て一致している．しかし，他の手法で得られた弱クラスタのうち，小中学生・高校生・主婦のクラスタは，他の強クラスタとして統合されるものが多い．理想的には，各手法において同じラベルの弱クラスタが統合されることが望ましいが，社会人以外のクラスタの区別は難しいことが分かる．

表 1: 弱クラスタの統合によるラベルの変化

GMM（帰宅時間）	前	社会人	小中生	高校生	大学生	主婦
	後	社会人	小中生	高校生	大学生	主婦
GMM（移動時間 ）	前	社会人	小中生	高校生	大学生	
	後	社会人	大学生	主婦	高校生	
GMM（外出時間）	前	社会人	高校生			
	後	社会人	主婦			
PLSI（$k=5$）	前	社会人	小中生	高校生	大学生	主婦
	後	高校生	小中生	大学生	大学生	小中生
PLSI（$k=3$）	前	社会人	高校生	主婦		
	後	高校生	主婦	小中生		

クラスタアンサンブルによって，ラベルが修正されたかどうかについて，データ数とラベルの正誤数を表 2 に示す．1 列目は弱クラスタで正しいラベルとなり，アンサンブルでも修正されなかったデータである．GMM でどの特徴量を用いた場合でも，正答であったデータ中の約 15 万データは修正されていない．3, 4 列目の修正されたデータ数を見ると，PLSI では初期の精度が悪いため，修正された数も多くなっている．GMM では，修正されたことにより正答となったものよりも，誤答となったものの方が多いことが分かる．

表 2: クラスタアンサンブルでラベル修正されたデータ数

	修正されない		修正された	
	正	誤	正	誤
GMM（帰宅時間）	154,591	30,279	16,171	56,534
GMM（移動時間）	143,125	30,094	27,637	56,719
GMM（外出時間）	156,974	37,668	13,788	49,145
PLSI（$k=5$）	54,284	32,356	116,478	54,457
PLSI（$k=3$）	75,140	27,496	95,622	59,317

クラスタリングでは，各手法によって異なった間違いが起こる．PLSI では，異なる特徴量間の関係性が似ているデータがクラスタとなる．しかし，数値的な特徴は考慮されないため，

値が近くても別のクラスタになる可能性がある．一方，GMM は数値的に似ているものが同じクラスタとなる．例えば，アルバイトをしている学生の場合，単に帰宅時間が遅いだけで，他の特徴量を考慮せずに，データが社会人の弱クラスタに分類される．クラスタアンサンブル手法によって，それぞれの間違いが修正されるために，推定精度が向上したと考えられる．

しかしながら，今回の MCLA の実験では，データ中の 1 割のデータ 26,420 件が，強クラスタ決定の際にランダムで振り分けられている．異なる手法による初期クラスタリングの結果の違いが大きければ，強クラスタを一意に決めることができない可能性が高い．そのため，より精度の高い弱クラスタを利用することや，多くの弱クラスタを用意することが有用であると考えられる．今後は，LCE などのデータ統合を行う手法を利用することを検討しているが，大規模データに対応できるような実装方法が必要となる．現在は，スーパーコンピュータ FX-10 上で並列処理を行うフレームワークである，京 MapReduce（KMR）を用いて実装を行っている段階である．

3.2 滞在場所の推定

人の属性を推定することができれば，さらに移動軌跡データから意味的情報を推定することができる．人の属性ごとに集まる場所を求めれば滞在場所の意味が分かる．例えば，大学生が昼間に集まる場所は大学であるし，社会人が集まる場所は会社である．このように，属性別に人の集まる場所が分かれば，行動知識にもとづいて意味付けを行うことができる．人の集まる場所は，密度に基づくクラスタリングによって検出できる．クラスタリング手法としては DenStream がある．DenStream とは，密度に基づくクラスタリングアルゴリズム DBSCAN を，ストリームデータに適用したものである．

属性別に滞在場所の推定を行った結果が表 3 である．軌跡データ中で，移動目的が通勤または通学となっているものの到着点を正解座標とし，Denstream で求めた座標の適合率を計算している．上は人の属性の正解ラベルを使用し，正しい属性別のデータごとに滞在場所を求めた結果である．社会人や小中学生は，MCLA の結果を用いても精度が高く求められるが，大学生や高校生では，大きく精度が下がっている．これは，MCLA による属性推定の精度が影響していることによる．

表 3: 滞在場所推定の精度

	会社	高校	小中学	大学
正解ラベル	96.1 %	96.3 %	96.2 %	77.9 %
属性推定	94.9 %	68.3 %	93.7 %	57.7 %

3.3 移動手段の推定

移動軌跡の意味的情報として，滞在場所の他に移動手段がある．移動手段についても，教師なし学習を用いて推定することを考える．都市部でよく使われる移動手段として，徒歩・自転車・バス・車・電車の 5 つがある．データからは移動の速度を求めることができるため，速度の時系列データに対する移動手段のラベルを推定する，系列ラベリング問題として捉えることができる．そこで，系列ラベリング問題を解く手法として，隠れマルコフモデル（HMM）を用いる．HMM は，Baum-Welch アルゴリズムを利用した場合，教師なし学習として用いることができる．HMM では，現在の状態は一つ前の状態に依存するという仮定のもとに，観測データ（速度）に対応する状態ラベル（移動手段）を推定することになる．状態遷移の確率を求めることにより，前後の移動手段の関係性を考慮することができる．例えば，バスと車は速度だけでは区別がつきにくいが，電車移動の後は車よりバスの可能性が高い，といったことを考慮して推定を行うことができる．

5 分ごとの移動平均速度を入力データとし，HMM で移動手段の推定を行った結果，推定精度は 50.0 %となった．比較手法として，教師あり学習で系列ラベリング問題を解く手法である CRF を用いた場合，推定精度は 75.4 %である．CRF では，前後 2 つの速度を特徴量とし，10,000 個の軌跡軌跡を訓練データとして移動手段の推定を行っている．CRF に比べて，HMM の精度が低いことが分かる．HMM による推定は，人の属性推定のときと同様に，初期値への依存度が高いなどの問題がある．そこで，移動手段の推定についても，クラスタアンサンブルを導入することで精度向上が期待できると考えている．

4. おわりに

本稿では，プライバシー保護データマイニングの観点から，意味的情報が付加されていない移動軌跡データに対して，教師なし学習手法を用いて意味情報を推定する方法を提案した．具体的には，クラスタリング手法を用いて，人の行動時間の特徴から人の属性を推定を行った．1 つのクラスタリング手法を用いるだけでは，初期値への依存などで精度が低くなるという問題点があるため，クラスタアンサンブルを導入した．さらに，人の属性を推定した移動軌跡に対して，移動手段や滞在場所などの意味的情報の推定を行った．

人の属性推定については，精度としてはまだ十分ではないため，弱クラスタの検討や他のクラスタアンサンブル手法を用いることを検討している．さらに，移動手段の推定においても，クラスタアンサンブルを導入し，異なる初期値や異なる時間間隔で推定した結果から，最終的な結果を得ることによって，より精度の高い推定を行う予定である．

参考文献

[Hofmann 99] Hofmann, T.: Probabilistic Latent Semantic Indexing., in *Proc. of the 22nd Annual International SIGIR Conference*, Vol. 3, pp. 583-617 (1999).

[Iam-On 12] Iam-On, N., Boongeon, T., Garrett, S. and Price, C.: A Link-Based Cluster Ensemble Approach for Categorical Data Clustering, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No.3, pp. 413-425 (2012).

[Strehl 02] Strehl, A. and Ghosh, J.: Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions, *Machine Learning Research*, Vol. 3, pp. 583-617 (2002).

[宋 12] 宋 軒・関本 義秀, 160 万人の長期 GPS 移動データに基づく災害避難行動の分析とシミュレーションモデル構築に関する研究, 平成 24 年度国土政策関係研究支援事業研究成果報告書 (2012).

[人の流れプロジェクト 13] 東京大学 空間情報科学研究センター，人の流れプロジェクト：http://pflow.csis.u-tokyo.ac.jp/index-j.html (2013).

[田中 14] 田中優子・関和広・上原邦昭: 人間の行動知識を用いた移動軌跡データからの固有行動検出, 2014 年度人工知能学会全国大会論文集 (2014).