人の行動把握のための 教師なし学習による意味情報推定

田中 優子¹・上原 邦昭²

¹非会員 神戸大学大学院システム情報学研究科計算科学専攻 (〒657-8501 神戸市灘区六甲台町 1-1)
E-mail: tanaka@ai.cs.kobe-u.ac.jp

²非会員 神戸大学教授 大学院システム情報学研究科計算科学専攻 (〒657-8501 神戸市灘区六甲台町 1-1)
E-mail: uehara@kobe-u.ac.jp

近年、様々な分野において、都市における人の行動把握の必要性が高まっている。しかし、位置情報データを用いた行動分析では、データに移動手段や滞在場所のような軌跡の意味情報が付いていないという問題がある。また、プライバシー保護のために、人の個人属性も得ることができないことが多い。そこで本研究では、位置情報データからこれらの意味情報を推定する手法を提案する。具体的には、位置情報から得られる特徴量のみを用いて、機械学習で推定する。その際に、教師なし学習のみを用いて、正解ラベル付きデータを必要としない推定を行う。最終的に、東京都市圏の人の流れデータを用いた実験を行い、本手法の有効性を考察する。

Key Words: trajectory data mining, semantic trajectory, person trip data, probe person data

1. はじめに

近年、移動軌跡から人の行動を分析することの重要性が高まっている。例えば、観光やマーケティングの観点では、人が興味を持って訪問した場所の分析から、個人の嗜好に合わせた推薦サービスが提供されている¹⁾。また、ターミナル駅の混雑等の解消や災害時の避難計画を目的として、都市における人の行動把握が必要とされている²⁾。

これまで、都市における人の行動把握には、パーソントリップ調査 (PT 調査) が実施されており、アンケート形式でデータが収集されてきた。また、近年、携帯端末の位置情報センサや加速度センサを活用したプローブパーソン調査 (PP 調査) が開発されている。PP 調査では、詳細かつ継続的な人の行動把握が期待されている。しかし、これらのデータから行動を分析するには課題がある。

課題の一つは、位置情報データには家や学校といった滞在場所、通学や買い物といった移動目的など、軌跡の意味情報が付いていないということである。人の行動を把握するために必要な意味情報は、センサから取得できない。PP 調査では、ユーザに携帯端末から意味情報を入力してもらう仕組みがあるが、操作忘れや操作ミスが発生すること、ユーザの負担が大きいことなどが課題とされている³⁾。

また、データ活用におけるプライバシーの保護という課題もある。データ活用とプライバシー保護を両立

させるため、プライバシー保護データマイニングと呼ばれる手法⁴⁾が開発されており、データの匿名化や攪乱により個人が特定できないようにする技術が用いられている。位置情報の利用においては、空間情報のあいまい化や個人識別情報の削除などを行い、プライバシーを保護して分析されている⁵⁾。特に、年齢や職業などの個人属性は、プライバシー保護のために削除され提供されないことがある。例えば、宗・関本⁶⁾の災害避難行動分析では、個人属性を用いることができないため、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、分析するにとどまっている。しかしながら、機械学習を用いて個人属性を推定できれば、属性別の避難行動の違いなど、詳細に分析することが期待できる。

本研究では、位置情報データから得られる特徴量を用いて、移動軌跡の意味情報を機械学習で推定する手法を提案する。機械学習には、個人属性や意味情報などの正解ラベルを用いて学習する教師あり学習と、何の正解ラベルも用いない教師なし学習がある。本研究では、個人属性が利用できないことを想定した教師なし学習の枠組みで推定する。すなわち、あらかじめ意味情報が付けられたデータを用いて学習するのではなく、軌跡データから得られる行動の特徴のみから、個人属性を推定することを検討する。

また、幾何的なデータである位置情報データを stop と move の 2 つの概念に分割し、それぞれに対して意味情報を推定する。stop では滞在場所と滞在時間のパ

ターン検出, move では移動手段と移動目的を推定する. 実験では, 東京都市圏の人の流れデータを用いて意味情報を推定し, 提案手法の有効性を考察する. 最終的に, 得られた意味情報を体系的にまとめたオントロジーを構築する. オントロジーにより, 意味情報間の相互の関係性を表すことができる.

2. 関連研究

移動軌跡データの意味情報推定に関する既存研究として, 移動軌跡データとそれ以外のデータを組み合わせて, 意味情報を付与する方法がある. なかでも, 地理情報システム (GIS) 等の地図情報を利用し, 移動軌跡に対して情報を付与する研究は多く存在する. Wolf ら⁷⁾ は, 車両の軌跡データに対して GIS を利用して立ち寄った場所などを特定すれば, 移動目的が推測できることを示している. また, Yan ら⁸⁾ は, 座標と道路や線路のデータを照らし合わせて移動している場所を特定し, 速度情報などを用いて移動手段を特定している. これらの手法は, 軌跡データに対して正確な情報を付けることができるため, 高精度な推定ができる.

移動手段推定では, GPS データと加速度データを合わせて用いる手法が提案されている^{9),10)}. これらは, 位置情報端末内での処理を想定しており, リアルタイム性や省エネルギーを目的とする方向で研究が発展している. また, 矢部ら³⁾ は, PP 調査でのユーザ負担を軽減するため, 加速度波形の特徴量を用いた移動手段の自動判別手法を開発している. 松島ら¹¹⁾ は, スマートフォンによる加速度波形と移動経路の特徴量に着目し, 高精度な移動手段の推定を行っている.

位置情報データのみを用いる研究として, Zheng ら¹²⁾ は, GPS データから得られる速度と加速度から, 決定木アルゴリズムを用いて移動手段を多クラス分類すれば, 高精度で移動手段を推定できるとしている. Montini ら¹³⁾ は, GPS データから移動目的を推定する手法を提案し, 活動の時刻や期間をもとに Random Forest アルゴリズムを用いて推定している.

これらの方法はいずれも教師あり学習を用いて推定している. 教師あり学習では, 教師データとしてあらかじめ意味情報のラベル付きデータを用意し, 学習する必要がある. しかし, そもそも十分な量の教師データを用意することは困難である. そこで本研究では, ラベル付きデータを用いずに, 教師なし学習の枠組みで意味情報を推定する方法を提案する. また, プライバシー保護により削除される可能性がある個人属性についても, 移動軌跡データから推定する. 位置情報データから行動調査を行う研究として, 朝倉ら¹⁴⁾ は, PHS による移動軌跡データから移動点と滞在点を識別し, 経路

図-1 3階層の移動軌跡

を特定する手法を提案している. 本稿では, より詳細な行動分析を行うため, 滞在場所や移動手段など, 行動分析に必要な意味情報を包括的に推定する方法を提案する.

3. 意味情報推定手法

(1) 軌跡の意味情報

移動軌跡は, 観測軌跡・統語的軌跡・意味的軌跡の3種類の階層に分けられる¹⁵⁾. 図-1は3種類の軌跡を図示したものである. 観測軌跡とは, 位置情報センサから収集された生データであり, 座標と時刻から成る点列で表される. 統語的軌跡とは, 観測軌跡をstopとmoveに分割したものである. ある範囲内で一定時間以上停止している場合をstopとし, stop以外の部分列は全てmoveと定義する. これにより, 移動軌跡はstopとmoveの繰り返りで表される. 意味的軌跡とは, 統語的軌跡におけるstopとmoveに対して意味情報を付与したものである. stopには場所 (例えば, 家, 会社, 店など) の情報が付けられ, moveは移動手段 (例えば, 徒歩, バス, 電車など) の交通機関などでラベル付けされる. これらの意味情報により, 軌跡全体の意味 (例えば, 通勤, 買い物など) が解釈できる.

(2) 個人属性の推定

本稿では, 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る. そこで, まず人の行動時間をもとに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, 個人属性を推定する. クラスタリング手法だけでは, 初期値・パラメータに依存しやすいため, 結果が不安定である. このため, クラスタアンサンブルという手法を導入し, 推定をより洗練化して精度を向上させる. クラスタアンサンブルとは, 複数回クラ

図-2 PLSI のモデル図

スタリングした結果を統合して、より識別能力の高いクラスタを生成する手法である¹⁶⁾。

クラスタアンサンブルの手順としては、最初にデータの特徴量を用いて複数回クラスタリングし、複数のクラスタ（弱クラスタ）を得る。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）を得る。弱クラスタは、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して、多様性のある弱クラスタを用意するほうが良いことが知られている¹⁷⁾。そこで、本稿では2つのクラスタリング手法、確率的潜在意味解析 (PLSI)¹⁸⁾ と混合正規分布 (GMM)¹⁹⁾ を利用している。

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、潜在構造を説明するモデルであり、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。移動軌跡データにおいては、以下の3つの仮定をPLSIでモデル化する。

まず、人は潜在クラスで分類できる。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。これを図に示したものが図-2である。各潜在クラスを職業にあたるものと仮定すれば、個人属性を推定できる。具体的な特徴として、主婦や小中学生は家に帰る時間は早く、社会人は遅いと思われる。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。このような、あらかじめ知り得ている行動の特徴を利用して PLSI によりクラスタリングし、個人属性を推定する。

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。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、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。

クラスタアンサンブル手法として、Meta-Clustering Algorithm (MCLA)¹⁶⁾ を利用する。MCLA は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みなし、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クカル係数を用いて式 (1) で定義される。

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

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。これにより、より重複するデータが多い弱クラスタ同士を統合できる。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。

本稿では、初期クラスタリングとして PLSI と GMM の結果を用いる。PLSI は、特徴量を質的データとして扱い、特徴量の共起性をもとにデータをクラスタリングする。一方、GMM は特徴量を数値として扱う手法である。異なる2つの手法を用いることにより、多様性のある弱クラスタが得られる。これらの結果を用いて、クラスタアンサンブルで精度向上させる。

(3) 滞在場所の推定

都市部における人々の行動には、年齢や職業といった個人属性別のグループごとに集まる場所がある。例えば、大学生が昼間に集まる場所は大学であり、会社員が集まる場所は会社である。このように、個人属性別に人の集まる場所が分かれば、滞在場所の意味付けができる。人の集まる場所は、密度に基づくクラスタにより検出できる。このクラスタリングには DenStream²⁰⁾ を用いる（付録 III 参照）。DenStream 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手法である。密度に基づくクラスタリングは、距離のしきい値以下の点同士を隣接しているとみなし、ある密度以上に隣接する点が集まっている箇所を検出する。得られるクラスタは高い密度で人が集まる場所になる。

本研究では、前述により推定された個人属性（社会人、大学生、高校生、小中学生）別に DenStream を実行する。それぞれのグループに属するデータについて、一日の移動軌跡の始点と終点を除く stop にあたる座標を入力し、得られたクラスタの中心座標を滞在場所（会社、大学、高校、小中学校）と推定する。また、移動軌

跡の始点と終点は、座標が同じである場合、家のラベルを付ける。

(4) 滞在時間パターンの検出

stop について付け加えられるもう一つ概念として、滞在時間がある。滞在時間は stop の時刻から数値として得られるが、より詳しい情報を得るために、時間的パターンを考える。時間的パターンとは、例えば、学生は午前中に大学に滞在した後、夕方には駅周辺に滞在する傾向があるなど、複数の滞在場所と滞在時間を合わせた行動パターンとなる。本稿では、Eagle and Pentland²¹⁾のように、主成分分析により人の行動パターンを求める。滞在場所をラベル付けされた移動軌跡データを利用し、同じ属性を持つ人のグループの行動を行列で表したものを主成分分析すれば、そのグループの特徴的な行動パターンが検出される。また、その重要度も検出できる。

主成分分析で用いる行列は、人の行動時間と滞在場所を行列で表したものである。用いるデータは、前述の滞在場所推定により、既に家や会社などの滞在場所ラベルが付けられているものとする。個人の行動は H 次元のベクトル Γ で表される。ここで H は $h \times n$ 、 h はデータ数、 n はラベル数である。ベクトルの各値は、そのラベルがつけられた時刻では 1、そうでない時は 0 とする。

次に、グループの行列を $B = [\Gamma_1, \Gamma_2, \dots, \Gamma_M]$ で表す。ここで M はそのグループに属する人の数である。グループの行列の例を図-3 に示す。ここでのラベルは {Home, Office, Move, Elsewhere, Nodata} ($n = 5$) であり、データ数は 1 時間ごとの 1 日分 ($h = 24$)、グループ人数は $M = 50$ である。各行が個人の一日の行動ベクトルとなっており、色がついている部分は値が 1、つまりその時刻にそのラベルの滞在場所に居たことを表す。

この行列 B を主成分分析すれば、固有ベクトルと固有値が得られる。この固有ベクトルがグループの特徴的な行動ベクトルを表し、対応する固有値は行動ベクトルの重みとなる。行動ベクトルの各次元は時間とラベルを表すため、そのグループの人がどの時間にどのラベルの滞在場所に居るかというパターンが検出される。本稿では、個人属性をグループとみなし、属性ごとに行動パターンを検出する。

(5) 移動手段の推定

都市部でよく使われる移動手段として、徒歩、自転車、バス、乗用車および鉄道の 5 つを考える。移動軌跡データからは移動の速度が求められるため、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

図-3 社会人グループの一日の行動行列

列ラベリング問題として捉えることができる。そこで、教師なし系列ラベリング手法である隠れマルコフモデル (HMM)²²⁾ を用いる。

速度データを用いた推定では、バスと乗用車などの差が区別できない。しかし、人の行動における習慣として、「乗用車での移動の後にバスを使う可能性は低い」ことや、「電車の移動の後は乗用車よりもバスの可能性が高い」などの移動手段間の関係性が存在する。これらの移動手段の変化を考慮すれば、速度だけでは区別が困難な移動手段を推定できる。したがって、HMM を用いれば、これらの前後の関係性をモデル化できる。

移動手段推定の流れは、まず move の測位点列から、特徴量として移動速度を計算する。次に、HMM において移動手段間の遷移確率を学習しやすくするため、データを圧縮する。圧縮した速度データに対し、HMM によって移動手段のモデルを構築して推定する。

a) 速度データ圧縮

HMM により、前後の関係性を考慮して移動手段を推定するには、異なる移動手段間の遷移確率を学習する必要がある。しかし、図-4 に示す通り、一つの移動手段からは複数の値が出力される。例えば、徒歩での移動速度は [2 3 5 3] となっており、この部分列中の 3 回の遷移は、全て徒歩から徒歩への自己遷移となる。一方、考慮したい移動手段間の遷移は、徒歩から電車、電車からバスに変化する 2 回しか存在しない。そのため、単位時間ごとの速度データでは、自己遷移の割合が多くなり、異なる移動手段間の遷移確率を求めることは困難になる。そこで本稿では、まず移動手段の変化点を検出し、セグメント分割する。そして、各セグメントの平均速度を求め、各セグメントが長さ 1 となるように圧縮する。図-4 の例では、2 箇所の変化点で分割し、各セグメントの平均速度を求めると [3 34 18] となる。この圧縮されたデータを用いれば、移動手段間の遷移確率を考慮して、各セグメントに対して一つの状態 (移動手段) を推定できる。

図-4 速度データの圧縮

図-5 学生の移動パターンの例

b) 空間変換による変化点検出

変化点検出には, Lopez ら²³⁾ が提案している空間変換を用いた変化点検出手法を利用する. この手法は, あらかじめ各時刻における移動手段を推定したデータを用い, 空間変換した上で変化点を検出する. 例えば, 移動手段として walking, biking, driving を想定する場合, 入力データはあらかじめ移動手段が推定され, 各時刻 t での移動手段の確率は, 活動ベクトル $M = (m_1, m_2, m_3)$ で表されているものとする. ここでの m_1 は walking, m_2 は biking, m_3 は driving の確率を表している. そして, $w_1 = (1, 0, 0)$, $w_2 = (0, 1, 0)$, $w_3 = (0, 0, 1)$ として直交単位ベクトルを定義し, それぞれが walking, biking, driving を表す軸だと考えると, ベクトル M を 3 次元の直交空間に表せる. ベクトル M と軸 w_i の成す角度を計算すれば, 値が小さいほど軸に対応する移動手段の確率が高いことになる. これにより, 各時刻における移動手段を求め, 変化点を検出できる. 得られた変化点で速度の時系列データをセグメント分割し, 各セグメントの平均速度を新たな時系列データとすれば, 一つの移動手段による軌跡を一つ速度データで表すことになり, データ圧縮となる.

c) 系列ラベリングによる移動手段推定

圧縮した速度データを用いて, 教師なし学習の HMM により移動手段を推定する. HMM は, 現在の状態が一つ前の状態に依存するという仮定のもとに, 質的な観測値の系列データに対応する状態ラベル列を推定する手法である. ここでの状態が移動手段, 観測値が移動速度に値すると考える.

HMM で推定されるパラメータは, 初期確率 (各状態がラベル列の先頭になる確率) π , 各状態間の遷移確率 A , 出力確率 (各状態から各特徴量が出力される確率) B , の 3 つである. Baum-Welch アルゴリズム²⁴⁾ と呼ばれる学習法を用いれば, 効率的に教師なし学習できることが知られている (付録 II 参照). 学習されたパラメータをもとに, 再帰的に最尤状態ラベル列が決定できる.

(6) 移動目的の推定と分析

移動目的を推定するには, 移動の出発点と到着点を考えればよい. 例えば, 家から学校への移動の目的は通学であるし, ある場所からショッピングセンターへの移動の目的は買い物と考えられるように, 発着点から移動目的を意味付けできる. そこで, 滞在場所間の移動でよく見られるパターンを求める. ここでは Alvares ら²⁵⁾ のように, 支持度と移動パターンを定義する. 具体的には, 全軌跡のうち, 場所 A から場所 B への移動含む軌跡の割合を AB の支持度とする. 支持度がしきい値を超えた場合, 移動パターンとして検出する. このようにすれば, 支持度が高いほどよく見られる行動パターンとなる.

軌跡の集合を $T = \{t_1, t_2, \dots, t_n\}$ とし, $\overrightarrow{AB}$ は stop A から stop B へ移動する move を表すものとする. ただし A, B は stop の滞在場所ラベルである. move $\overrightarrow{AB}$ の支持度は式 (2) で求められる.

$$\overrightarrow{AB} = \frac{|\{t \in T | \overrightarrow{AB} \in t\}|}{|T|} \quad (2)$$

ここで $|X|$ は集合 X の要素数を表す. move の支持度が ς であるとき, しきい値 minsup より ς が大きければ移動パターンとする.

移動パターンの検出によって, 移動目的別のパターンが検出される. 例えば学生であれば, 学生の軌跡のうち何%が学校からショッピングセンターへ移動するか, といった移動目的ごとの情報が得られる (図-5).

4. 人の流れデータを用いた実験

(1) 実験設定

a) 実験データ

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」を用いている. このデータは PT 調査データ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, どのような交通手段で, 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。

この調査では、プライバシー保護のために、詳細な住所を特定できないようにされており、各トリップの起終点はゾーンコードで表現されている。各ゾーンは夜間人口約 15,000 人を目安とした空間解像度の粗い区画（東京都の場合は 164 区画）となっており、同じゾーンを起終点とするトリップの座標は同じ位置に集約されている。そこで、人の流れデータでは、移動目的等をもとに、起終点をゾーン範囲内の尤もらしい位置に再配分する処理が行われている。再配分処理では、住宅地図データから得られる建物の属性（住宅、学校、事業所）や延べ床面積をもとに定義した分布関数に従って確率的に位置を決定し、より実態に近い人々の移動データを再現している。これにより、各トリップの起終点は、再配分された建物の地理座標（緯度経度）で表されている。また、移動経路についても、移動手段と合致するように、道路等に沿った地理座標に変換されている。そのため、人の流れデータにおける位置情報は、全てゾーンコードではなく地理座標で表現されている。

全データ数は 587,340 人であるが、本実験では、一日中全く移動をしないデータや職業が不明であるデータを除いた 397,510 人分を使用する。PT 調査はサンプル調査であるため、現実の人口分布に近づく拡大係数が設定されているが、本稿では拡大前のデータを使用している。座標データは、ある平日 1 日の 1 分ごとの地理座標である。また、データの領域は埼玉・東京・神奈川・千葉・茨城南部である。

人の流れデータにはパーソン ID、日時、座標、個人属性（年齢、性別、職業、住所コード）、交通手段、移動目的といった情報が含まれている。本実験では時間、座標のみを用いて教師なし学習により推定するが、推定の評価ではこれらの意味情報を利用して精度を求めている。なお、データに含まれる職業ラベルは 17 種であるが、このうち、その他と不明に相当するデータはあらかじめ除去している。また、職業ラベルは業界や内容によって細かく分けられているが、これらをまとめて社会人としている。交通手段のラベルについても、細分化されているラベルを統合している。自動二輪車、タクシー、軽乗用車および貨物自動車は、乗用車のラベルに統合している。自家用バスと路線バスはバス、モノレールと鉄道は電車に統合している。船舶や航空機のラベルがつけられたデータは存在しないため、本実験では利用していない。

b) データの前処理

軌跡データには、まず stop と move に分割する処理を行う。人の流れデータは、アンケート調査から再現された位置座標であるため、一つの滞在中は必ず同じ座標がつけられている。そのため、stop は 1 分前の座

標と一致していれば stop の続きと見なし、異なれば次の move の開始とする。

個人属性推定に用いるデータは、帰宅時刻、移動時間、滞在時間の 3 種類の時間特徴量である。PLSI で用いるデータは質的データであるため、帰宅時刻・滞在時間は 1 時間単位、移動時間は 10 分単位で分割し、離散化している。GMM は量的データを扱うため、時間特徴量は全て分単位に変換して利用している。

滞在時間パターン分析におけるデータでは、各滞在時間を 30 分単位で分割し離散化する。用いるラベルは {Home, Office, Move, Elsewhere, Nodata} の 5 つとしたため、一日の行動ベクトルの次元は $5 \times 24 \times 2 = 240$ となる。移動手段推定に用いるデータは、5 分ごとの平均速度を計算した時系列データとしている。軌跡中の stop にあたる部分は全て速度 0 であり、データ圧縮や変化点検出を行う前にあらかじめ圧縮している。変化点検出で用いるデータは、圧縮前のデータであらかじめ HMM（状態数 3）を実行し、各時刻での移動手段の確率を求めている。

c) 実験環境

データの前処理と主成分分析の実装は、スーパーコンピュータ富士通 Fx10 上で動作する京 MapReduce を利用している。京 MapReduce により、軌跡データを人ごとに集約し並列計算するため、計算資源を削減できる。DenStream の実行には Massive Online Analysis (MOA) version 2013.11 を用いている。MOA はデータストリームマイニングを行うためのオープンソースフレームワークである。GMM, PLSI の実行には、統計解析のフリーソフトウェアである R version 3.1.1 を利用している。MCLA におけ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 5.1.0 を用いている。

(2) 個人属性の推定

PLSI と GMM によるクラスタリングと、MCLA によるクラスタアンサンブルを用いて、個人属性を推定し、本アプローチの精度を求める実験を行う。まず、クラスタアンサンブルの導入による精度向上について評価する。想定する属性は、社会人、大学生、高校生、小中学生および主婦の 5 つの職業である。PLSI で指定する温度パラメータは 0.75 とし、クラスタ数 $k = 3, 5$ の 2 種類で実験する。GMM は 3 種類の特徴量をそれぞれを用いて実験する。クラスタアンサンブルでは、これら 5 回分の結果をアンサンブルする。

属性推定の結果の精度を表-1 に示す。MCLA による結果が最も精度が高く、クラスタアンサンブルの導入により推定精度が向上したことが示されている。初期のクラスタリングでは、手法によって異なる間違いが

表-1 個人属性推定の精度比較

手法	適合率
PLSI ($k = 5$)	42.7 %
GMM	59.1 %
MCLA	66.3 %

表-2 クラスタと属性のクロス表

	C1	C2	C3	C4	C5	合計
社会人	50.08	14.89	10.00	7.73	17.31	100
大学生	30.89	36.54	6.90	6.76	36.96	100
高校生	31.42	36.92	13.36	2.51	16.17	100
小中学生	9.83	38.92	12.36	3.20	35.68	100
主婦	6.81	18.49	2.72	3.07	67.66	100

起こる．たとえば，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が，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．

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．MCLA によるアンサンブルは，クラスタ統合後，各データが属する強クラスタを弱クラスタの多数決で決めなおすようなものである．弱クラスタ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．

次に，アンサンブルで推定されたクラスタについて評価する．各職業の内訳を表-2 に示す．各職業のデータ総数をそれぞれ 100 % としたとき，各クラスタに含まれるデータ数の割合を示している．

クラスタ C1 と C5 には，それぞれ社会人と主婦が多く分類されており，それぞれのクラスタが得られたことが分かる．また，C2 には，高校生や小中学生が多く含まれているため，学生のクラスタを得ることができている．大学生は C1, C2, C5 に多く分類されている．大学生は行動が多様であるため，それぞれのデータでより生活パターンが似ているほうに分けられたと考えられる．小中学生については，大学生の結果と同様に，学生が主婦に大きく振り分けられた結果となっている．

表-3 各クラスタの特徴量

	C1	C2	C3	C4	C5
帰宅時間 [時刻]	21:03	18:04	20:13	17:43	18:08
移動時間 [分]	35.1	35.1	62.0	41.8	36.5
外出時間 [時間]	11.0	9.0	10.5	6.5	8.4

表-4 Denstream で検出された座標の精度

	会社	高校	小中学校	大学
適合率	94.9 %	68.3 %	93.7 %	57.7 %
再現率	44.3 %	48.0 %	35.9 %	54.5 %

一方で，クラスタ C3 と C4 はどの職業でも含まれる割合が少ない．

次に，各クラスタのデータ特徴量を表-3 に示す．C1 は社会人のクラスタであり，帰宅時間が遅く外出時間も長いことが分かる．反対に，C5 の主婦クラスタは，帰宅時間も早く外出時間も短いことが分かる．クラスタ C3 と C4 は職業別に当てはまるものが無いが，特徴的なデータが集まったクラスタである．例えば，C3 は移動時間が他のクラスタに比べて大きく，また，C4 は帰宅時間も早く外出時間も短く，家に居る時間が長い．行動時間による個人属性推定の結果として，特徴量が似ている職業は明確な区別ができなかったが，一日の生活パターンの違いによるクラスタを生成することはできている．

(3) 滞在場所の推定

属性別に DenStream によるクラスタリングを行い，会社，大学，高校および小中学校の座標を推定し，精度を評価する．DenStream におけるパラメータは，約 10 m を隣接とみなす値としている．クラスタリングによって特定した滞在場所のうち，大学の場所を地図上に示したものを図-6 に示す．図中の黒い建物アイコンがクラスタリングによって検出されたもの，緑のピンが正解の座標である．正解の座標は，データに付けられた移動目的が通勤や通学となっている軌跡の到着点から特定している．人の流れデータは正確な座標ではないため，実際にある大学の地図上の位置とは必ずしも一致しない．しかしながら，多くの点で推定した座標と正解の座標が一致しており，正しい場所が検出されていることが見て取れる．

推定で得られた座標の精度を表-4 に示す．適合率は，検出した座標のうち正確な位置であったものの割合を表す．会社と小中学校では，90 % 以上の高い精度で検出できている．一方，高校や大学は 60 % 前後の精度となっている．これは，個人属性の推定においての誤

図-6 Denstream で検出された大学の座標

分類が影響していると考えられる。また、大学生は規則的な行動をしておらず、軌跡も多様であるため、他の滞在場所に比べて精度が下がったと考えられる。また、再現率は正解座標のうち何割を検出できたかを表す。いずれも 60 % 以下の低い値であり、検出できた座標が少ないことが分かる。密度に基づくクラスタリングでは、サンプル数が少ない会社や学校は検出されないことが分かる。

(4) 滞在時間パターンの分析

主成分分析によって得られた時間的パターンを図-7に示す。これは、社会人の属性グループの行動を1時間ごとにラベル付けした行列を用いた結果である。主成分分析で得られた第一主成分ベクトルを、各ラベルに相当する箇所で区切って表示している。図中の赤い部分が、その時間にその場所にいる傾向があることを表している。図-7より、午前8時までと午後8時以降は家、9時から19時までは会社にいるといった行動が見て取れる。つまり、社会人グループの特徴的な行動を表していることになる。

主成分分析では、第二主成分・第三主成分を次々に求めれば、順に、より特徴的な行動を表すベクトルが得られる。しかし、今回の実験では、平日一日のみの移動軌跡データを用いたため、第一主成分しか行動パターンが得られていない。今後、長期間のデータや用いる滞在場所ラベルを増やせば、例えば、第一主成分には本実験と同じような通勤の行動が得られ、第二主成分には休日に多くみられる行動や買い物に行く時間帯などが検出されることが期待できる。

(5) 移動手段の推定

移動手段推定の精度を評価する。HMM による推定はパラメータの初期値にも影響されるため、いずれも3回ずつ実験し、平均精度を求めている。また、比較のために教師あり学習の系列ラベリング手法 CRF²⁶⁾を用いた推定も行っている(付録I参照)。実験結果を

図-7 社会人グループの一日の行動パターン

表-5 移動手段推定の精度比較

手法	データ	適合率
HMM	圧縮なし	43.9 %
HMM	変化点検出	55.2 %
HMM	変化点検出+MDL	57.9 %
CRF	圧縮なし	75.8 %

表-5に示す。

上3行が提案手法である教師なし学習による推定の結果である。推定精度は順に向上しており、57.9 %の適合率で推定できている。MDL²⁷⁾は、変化点検出による分割後に、同じような速度である部分列を統合する方法であり、分割数をより少なくしてデータの圧縮率を高めている。データ圧縮により推定精度は向上しており、HMM に適した学習ができていることが分かる。一方、表中の CRF は、教師あり学習による推定であり、精度は 75.8 %である。CRF は前後両方の遷移確率の考慮や、複数の特徴量が利用できるため、高精度に推定できる。しかし、十分な量の正解付きデータが必要となる。正解付きデータが利用できない場合、教師あり学習より精度は劣るものの、HMM による推定も可能であるといえる。

(6) 移動目的の推定と分析

意味的移動パターンの検出結果を図-8に示す。それぞれ社会人、大学生、高校生の移動パターンの結果である。最低支持度は $minsup = 0.1$ としている。各場所間の移動の支持度を用いて、例えば高校生の場合であれば、家から学校に登校する人は 83 %、学校から家でない他の場所へ行く人は 12 %というような移動の割合が検出できている。また、その他から学校へ行く移動は、最低支持度より低い値であり、移動パターンとして検出されていない。

移動パターンは移動目的で表せる。例えば、家→学校は通学、学校→その他は寄り道、家→その他は外出となり、移動目的で意味付けできる。今回の実験は、移

図-8 検出された属性別の移動パターン

図-9 得られた意味情報をまとめたオントロジー

動先の滞在場所ラベルから移動目的を特定する単純な方法である。しかし、同じ場所への移動でも、移動目的は人によって様々あると考えられる。例えば、移動先がショッピングセンターである場合、客としてその場所に行く人は買い物が移動目的であるが、そこで働く人にとっては移動目的が通勤となる。本実験では一日分の行動データを用いて推定しているが、数日分やそれ以上の長期間データを用いれば、移動先の場所に訪れる頻度や時間帯などから、より詳細に意味付けできると考えている。

実験で得られた意味情報をまとめたオントロジーを図-9に示す。図に示した数値は、社会人のグループから得られたものである。提案手法では、個人属性や移動軌跡の意味情報が包括的に推定できるため、図のように意味情報を体系的にまとめることができる。これらの意味情報は相互に関係があり、それぞれの結果で得られた特徴量が、他の意味情報推定に活かされる。互いの影響を考慮すれば、必要な意味情報の推定精度を相互に向上できる。

5. おわりに

本稿では、意味情報が付けら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

定する手法を提案した。具体的には、行動分析で必要と考えられる個人属性、滞在場所、滞在時間、移動手段および移動目的について、軌跡データから得られる特徴量のみを用いて推定した。個人属性の推定では、一日の行動時間を特徴量としてクラスタリングし、精度向上のためにクラスターアンサンブル手法を導入した。推定精度は66.3%となり、生活パターンが似ている人々のクラスターが得られた。滞在場所に関しては、個人属性の推定結果をもとに、密度に基づくクラスタリングを用いて会社や学校の座標を推定した。再現率は35.9%~54.5%と低いものの、会社や小中学校の適合率は94.9%、93.7%と高精度で検出された。一方、大学生のような行動が多様であると考えられる属性では、適合率が57.7%と精度が低くなることが分かった。移動手段推定では、教師なし系列ラベリング手法とデータ圧縮を用いて推定した結果、精度は57.8%であった。滞在時間と移動目的は属性別に分析を行い、社会人であれば、9時から19時は会社に滞在するような行動時間の特徴や、移動目的別に通勤の割合は0.73、寄り道の割合が0.17といった移動パターンの特徴が得られることを示した。

本研究では、平日一日のみの移動軌跡データを用いた実験を行ったが、長期的なデータを用いれば、より詳細な意味情報推定ができると考えている。個人属性の推定では、一週間ごとの周期的な行動時間を特徴量として加えれば、学生や社会人など決まった生活パターンの人々を高精度で推定できると考えられる。滞在場所の推定においても、より高精度になることが期待できる。一方、提案手法はあらかじめ行動の特徴を推定できない属性は推定が難しい。例えば、非就業者の行動把握は大きな課題であるが、一日の行動の特徴のみでは推定が難しいため、本稿では非就業者の中でも比較的让生活パターンが一定している主婦のみを取り上げた。長期的なデータを用いれば、非就業者でもいくつかの生活パターンがクラスターとして検出されることが考えられる。また、滞在場所の推定では、属性を問わず人が集まる場所を検出することはできるが、その具体的な意味付けは難しい。重要度の高い座標のみ地図情

報を参照するなどして、推定できる意味情報を増やすことは今後の課題である。滞在場所に意味付けができれば、提案手法の流れとは逆に、滞在場所の意味情報から個人属性を推定できる可能性もある。今後、意味情報の相互の関係性による推定方法も検討していきたい。移動手段の推定においては、加速度や向きなど速度以外の特徴量を導入し、精度をさらに向上させる必要がある。良い精度の推定結果が得られれば、移動手段推定においてもクラスタアンサンブルの導入により更なる精度向上と安定化が期待できると考えられる。一方、本研究で用いた人の流れデータは、人工的に座標が付けられたものであるため、GPS データにおけるノイズや移動途中の一時停止といった詳細な情報が含まれていない。今後、実データを用いて実験する際には、これらに対応する処理を行う必要がある。

謝辞：本研究は2010－2015年度CSIS共同研究「位置情報と移動時間を考慮した移動軌跡からのパターン抽出」の成果の一部に基づいている。また、大林財団平成26年度研究助成金「人の流れデータへの大規模データマイニングの適用と交通制御への応用」の援助を受けている。

付録 I CRF

CRF は、入力配列 $\mathbf{x}$ と出力ラベル $\mathbf{y}$ が共に構造を持つ条件付き確率分布 $P(\mathbf{y}|\mathbf{x})$ を表現するモデルである。入出力を共に長さ T の配列 $\mathbf{x} = (x_1, x_2, \dots, x_T)$, $\mathbf{y} = (y_1, y_2, \dots, y_T)$ とする。出力ラベルの取りうる集合を Y とし、各 y_t は Y の要素とする。特徴 $f \in F$ が、ある入出力配列 $(\mathbf{x}, \mathbf{y})$ の組において成立する数を $\phi_f(\mathbf{x}, \mathbf{y})$ とする。また、その特徴の重要度を θ_f とする。これが CRF におけるパラメータである。入力配列 $\mathbf{x}$ が与えられたときに、これに同じ長さのラベル列 $\mathbf{y}$ を割り当てる確率は、パラメータベクトル θ と特徴ベクトル ϕ の内積を用いて、以下のように定義される。

$$P(\mathbf{y}|\mathbf{x}; \theta) = \frac{\exp(\theta^T \phi(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y} \in Y} \exp(\theta^T \phi(\mathbf{x}, \mathbf{y}))}$$

ここでは、内積 $\theta^T \phi(\mathbf{x}, \mathbf{y})$ の値を、0 以上 1 以下の値に変換し、確率分布となるようにしている。これが CRF の確率モデルである。

付録 II Baum-Welch アルゴリズム

Baum-Welch アルゴリズムは、与えられた観測値の系列データから、EM アルゴリズムの考え方を用いて HMM のパラメータを推定するものである。EM アルゴリズムでは、変数の期待値を求めるステップ (E-step) と、パラメータの最尤推定解を求めるステップ (M-step)

を、パラメータが収束するまで反復する。HMM で求めるパラメータは、初期確率 π 、遷移確率 A 、出力確率 B の 3 つである。具体的な手順は次の通りである。

1. パラメータ $\pi_i, a_{ij}, b_i(k)$ をランダムな値で初期化
2. 再推定式により $\bar{\pi}_i, \bar{a}_{ij}, \bar{b}_i(k)$ を計算 (E-step)
3. $\bar{\pi}_i, \bar{a}_{ij}, \bar{b}_i(k)$ でパラメータを更新 (M-step)
4. パラメータが収束するまで、以上の操作を反復

ここで、パラメータの再推定式は以下のように表せる。

$\bar{\pi}_i$ = 時刻 1 で状態 i に滞在した確率

$$\bar{a}_{ij} = \frac{\text{状態 } i \text{ から状態 } j \text{ へ遷移する回数の期待値}}{\text{状態 } i \text{ を訪れる回数の期待値}}$$

$$\bar{b}_i(k) = \frac{\text{状態 } i \text{ から特徴量 } k \text{ を出力する回数の期待値}}{\text{状態 } i \text{ を訪れる回数の期待値}}$$

上記の再推定式により HMM のパラメータを求める方法が、Baum-Welch アルゴリズムである。

付録 III Denstream

DenStream では、オンライン処理である micro-cluster の生成と、オフライン処理である DBSCAN による micro-cluster の統合の 2 つの処理が行われる。オンライン処理部分のフローチャートを以下に示す。

オンライン処理では、データストリームから入力されたデータ p を micro-cluster に統合するマージ処理と、周期 T_p おきに micro-cluster の枝刈りを行う。マージ処理のフローチャートを次に示す。マージ処理では、micro-cluster の半径が指定したしきい値 ϵ 以下になる

ようにデータが追加されるため、位置が近いデータがクラスタリングされる。また、各 micro-cluster には、属するデータ数がより多く、より現在に近い時刻にデータが追加されるほど重くなる重み w が定義されている。 w が小さいものを o-micro-cluster と呼び、データの追加により w が大きくなれば p-micro-cluster に変わる。枝刈りの処理では、重み w が小さい micro-cluster は削除される。o-micro-cluster の枝刈りは、早すぎると p-micro-cluster に変わる前に消えてしまうため、o-micro-cluster の生成時間に合わせて単調増加する、しきい値 ξ を使用する。枝刈りの結果、近傍のデータが多い、または生成されたばかりの micro-cluster が残る。以上のオンライン処理が行われた後、オフライン処理である DBSCAN が実行される。DBSCAN では、クラスタ間のユークリッド距離とデータ数をもとに、近傍の micro-cluster を統合し、密度の高いクラスタを検出する。

参考文献

- 1) 西田京介, 戸田浩之, 倉島健, 内山匡: 確率的訪問 POI 分析: 時空間行動軌跡からのユーザモデリング, マルチメディア, 分散, 協調とモバイル シンポジウム論文集,

Vol.2013, pp.334-345, 2013.

- 2) 中村敏和, 関本義秀, 薄井智貴, 柴崎亮介: パーティクルフィルターを用いた都市圏レベルの人の流れの推定手法の構築, 土木学会論文集 D3, Vol.69, No.3, pp.227-236, 2013.
- 3) 矢部努, 若井亮太, 及川潤, 北村清洲, 廣川和希, 福沢綾乃, 牧村和彦: 生活行動調査の自動化に関する基礎的研究~プローブパーソンデータによる起終点と手段の自動判別手法の構築に向けて~, IBS Annual Report 研究活動報告 2015, pp. 39-46, 2015.
- 4) 佐久間淳, 小林重信: プライバシー保護データマイニング, 人工知能学会誌, Vol.24, No.2, pp.283-294, 2009.
- 5) 佐藤一夫: プライバシー保護を考慮したケータイ行動ログの利活用について, 情報処理, Vol.50, No.7, pp.598-602, 2009.
- 6) 宗軒, 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.
- 7) Wolf, J., Guensler, R. and Bachman, W.: Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data, *Transportation Research Record: Journal of the Transportation Research Board*, No.1768, pp.125-134, 2001.
- 8) Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S. and Aberer, K.: Semantic trajectories: Mobility data computation and annotation, *ACM Transactions on*

- Intelligent Systems and Technology*, Vol.4, No.3, 2013.
- 9) Feng, T. and Timmermans, H. J.: Transportation mode recognition using GPS and accelerometer data, *Transportation Research Part C: Emerging Technologies*, Vol.37, pp.118–130, 2013.
 - 10) Shafique, M. A. and Hato, E.: Use of acceleration data for transportation mode prediction, *Transportation*, Vol.42, No.1, pp.163–188, 2015.
 - 11) 松島敏和, 橋本浩良, 高宮進: スマートフォンによるブローパーソン調査の高度化に向けた移動手段判別手法の開発, 土木学会論文集 D3, Vol.71, No.5, pp.547–558, 2015.
 - 12) Zheng, Y., Li, Q., Chen, Y., Xie, X. and Ma, W.-Y.: Understanding mobility based on GPS data, *Proc. of the 10th International Conference on Ubiquitous Computing*, pp. 312–321, 2008.
 - 13) Montini, L., Rieser-Schüssler, N., Horni, A. and Axhausen, K.: Trip purpose identification from GPS tracks, *Transportation Research Record: Journal of the Transportation Research Board*, No.2405, pp.16–23, 2014.
 - 14) 朝倉康夫, 羽藤英二, 大藤武彦, 田名部淳: PHS による位置情報を用いた交通行動調査手法, 土木学会論文集, No.653, pp.95–104, 2000.
 - 15) Renso, C., Baglioni, M., de Macedo, J. A. F., Trasarti, R. and Wachowicz, M.: How you move reveals who you are: understanding human behavior by analyzing trajectory data, *Knowledge and Information Systems*, Vol.37, No.2, pp.331–362, 2012.
 - 16) Strehl, A. and Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions, *The Journal of Machine Learning Research*, Vol.3, pp.583–617, 2003.
 - 17) 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.
 - 18) Hofmann, T.: Probabilistic latent semantic indexing, *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57, 1999.
 - 19) Dempster, A. P., Laird, N. M. and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society Series B (Methodological)*, pp. 1–38, 1977.
 - 20) Cao, F., Ester, M., Qian, W. and Zhou, A.: Density-based clustering over an evolving data stream with noise, *Proc. of the SIAM International Conference on Data Mining*, pp. 328–339, 2006.
 - 21) Eagle, N. and Pentland, A. S.: Eigenbehaviors: Identifying structure in routine, *Behavioral Ecology and Sociobiology*, Vol.63, No.7, pp.1057–1066, 2009.
 - 22) Rabiner, L. R. and Juang, B.-H.: An introduction to hidden Markov models, *IEEE ASSP Magazine*, Vol.3, No.1, pp.4–16, 1986.
 - 23) Lopez Aguirre, A. J., Ochoa, D. and Gautama, S.: Detecting changes of transportation-mode by using classification data, *Proc. of the 18th International Conference on Information Fusion*, 2015.
 - 24) Baum, L. E., Petrie, T., Soules, G. and Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *The Annals of Mathematical Statistics*, Vol.41, No.1, pp.164–171, 1970.
 - 25) Alvares, L. O., Bogorny, V., de Macedo, J. A. F., Moelans, B. and Spaccapietra, S.: Dynamic modeling of trajectory patterns using data mining and reverse engineering, *Tutorials, Posters, Panels and Industrial Contributions at the 26th International Conference on Conceptual Modeling*, Vol. 83, pp. 149–154, 2007.
 - 26) Lafferty, J., McCallum, A. and Pereira, F. C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proc. of the 18th International Conference on Machine Learning*, pp. 282–289, 2001.
 - 27) Tanaka, Y., Iwamoto, K. and Uehara, K.: Discovery of time-series motif from multi-dimensional data based on MDL principle, *Machine Learning*, Vol.58, No.2-3, pp.269–300, 2005.

(2016. 3. 18 受付)

SEMANTIC INFORMATION ESTIMATION FOR BEHAVIOR UNDERSTANDING USING UNSUPERVISED LEARNING

Yuko TANAKA and Kuniaki UEHARA

Recently, in various fields, there is an increasing need for human behavior understanding in the city environment. However, there are problems with the behavior analysis using the position data. One of the problems is that the semantic information, such as transportation mode or trip purpose, is not attached to the data. Another problem is that, because of the privacy protection, personal attributes of people can not be obtained. In this paper, we propose a method to estimate personal attributes and semantic information from the position data. Specifically, the estimation by the machine learning uses only features obtained from the position data. By using only the unsupervised learning, it can estimate without a correct labeled data. In the experiment using people flow data in the Tokyo urban area, we consider the effectiveness of the technique.