様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ２ 日現在

機関番号：１４５０１
研究種目：挑戦的萌芽研究
研究期間：2013～2014
課題番号：２５５６０１４７

研究課題名（和文）乱択アルゴリズムによる並列分散軌跡パターンマイニング

研究課題名（英文）Parallel Distributed Trajectory Pattern Mining Using Randomized Algorithm

研究代表者
上原　邦昭（UEHARA, Kuniaki）

神戸大学・システム情報学研究科・教授

研究者番号：６０１６０２０６

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：近年、GPSなどの位置情報技術の普及により、人や車など時々刻々と動的に位置が変化する、移動軌跡データが容易に入手できるようになってきている。一方、防災、交通・都市計画、マーケティングなどの分野では、移動軌跡データから現在の状況を把握したいという要求が高まっている。本研究では、ストリームに対応できるオンライン処理方式の移動軌跡マイニングを開発した。また，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，クラスタアンサンブル手法を用いて，人の行動時間の特徴から人の属性推定を行った．

研究成果の概要（英文）：With the rapid increase of the number of mobile GPS devices, it is important to develop efficient and effective algorithms to analyze massive trajectory data streams. Although there are many algorithms that can find patterns by batch processes, what we need is a new algorithm with limited resources by online processes. This research aims at developing such an algorithm and attempts to discover stay points, or the places which are becoming crowded.

Currently, 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 covered by the data. This research also 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.

研究分野：知能情報学

キーワード：地理情報システム　データマイニング　近似的アルゴリズム　GPS　クラスターアンサンブル　行動分析　移動軌跡データ

１．研究開始当初の背景

近年、GPS などの位置情報技術の普及により、人や車など時々刻々と動的に位置が変化する、移動軌跡データが容易に入手できるようになってきている。一方、防災、交通・都市計画、マーケティングなどの分野では、移動軌跡データから現在の状況を把握したいという要求が高まっている。

２．研究の目的

大規模なデータから状況を判断するために頻出パターンや相関ルールを発見する、データマイニングという研究がある。初期のデータマイニングの研究では、静的なデータ集合から頻出パターンや相関ルールを獲得することを目的としたものが多かった。このようなアプローチに基づく移動軌跡データマイニングでは、(1) 全てのデータを一括で読み込み保存するバッチ処理方式となり、メモリを大量に消費するという問題がある。さらに、(2) 人の流れのような、時間順でデータが到来し、常に新しい情報が入力される、ストリームを分析することができないという問題がある。本研究では、移動軌跡データマイニングをオンライン処理方式に拡張して、ストリームから頻出パターンを検出することを検討する。

一方、GPS データを人の行動分析に生かそうという研究が増えている。人がなぜ、どのように行動するかを分析するためには、軌跡に意味付けを行い、それらの関係性を考える必要がある。しかし、(3) GPS データには行動分析に必要な移動手段や滞在場所などの意味的情報が付けられていない。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。しかしながら、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。

３．研究の方法

(1) の問題を解決するために、注目領域の発見時に動的な使用メモリ空間の削減を可能にするアプローチを導入する。すなわち、注目領域すべてをメモリに展開することなく、小さいメモリ空間ですむように、確率的な解析を行う近似的アルゴリズムを導入して、誤差を許容したアプローチを採用する。

本研究で提案するアルゴリズムは、ストリームから頻出パターンを検出する近似的アルゴリズム Lossy Counting に基づいている。Lossy Counting は処理単位として定量的なデータを想定している。しかしながら、ストリームは逐次的に変動するため、Decaying Window と呼ぶ時間窓の概念を導入して、処理単位を時間帯として拡張している。

(2) の問題に対しては、ストリームアルゴリズムを採用する。ストリームアルゴリズムとは、過去のデータの記憶に制限を設け、過去の入力データ量に比べて非常に小さいメモリ空間を用いて計算を行う手法である。

一方、道路は混雑する前に、立ち止まる歩行者や車が増加するという現象を伴う。大量の人や車が立ち止まると、交通流が急に変化して混雑する可能性が高まる。このため、静的な密度よりも、交通流の変化に注目すべきである。本研究では、移動軌跡データストリームから交通流の変化点のみを抽出し、新しいストリームとみなして頻出パターンを検出するアプローチを提案する。

(3) の問題については、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、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クラスタリングする。

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。

４．研究成果

説明の都合上、(1)と(2)の研究をまとめてオンライン型移動軌跡マイニングと呼ぶ。また(3)の研究を移動軌跡データからの意味情報推定と呼ぶ。

[オンライン型移動軌跡マイニング]

交通分析、観光案内などでは、人々はどこでよく立ち止まるのか（stay point と呼ぶ）を発見しなければならない。最も単純な stay point の計測方法は、地図をセルに分割して、セルごとにカウンタを一つ用意して、セルに立ち止まった人数を数える方法である。しかしながら、この方法の問題点として、地図の面積が広すぎる場合、大量のカウンタを準備しなければならないが、地図全体を見ると、セルの総数に対して stay point となるセルの数は少数であることが知られている。このように、大規模なデータから出現するアイテムの頻度を求めようとすると、求めたいのは頻出アイテムであるのに、ほとんど出現しないアイテムが大多数を占め、メモリの多くを消費してしまうという問題がある。この問題を解決するための手段として、Lossy Counting という近似的アルゴリズムがある。

Lossy Counting は、長さ無限のストリームのアイテムを数えるために考案された計数アルゴリズムである。Lossy Counting のパラメータは、今までのストリームの長さ N、閾値Δとエラー率εからなる。出力はストリームでの各アイテムの出現回数である。しかも、この出力としての出現回数はΔN 以上で、実際の出現回数より最悪εN の誤差を許して頻度を数えている。このアルゴリズムの利点として、計数結果に少しの誤差εN を考慮して、カウンタ数を $e^{-1}\log(eN)$まで減らすことができる点がある。

Lossy Counting を用いて stay point を検出するには難点がある。理由として、交通流は常に変化しているということが挙げられる。たとえば、朝の出勤ラッシュ時を考えると、７時半頃ならば、多摩のような東京中心周りの駅が池袋よりも混雑している。８時を過ぎると、池袋、東京のような都心の駅が忙しくなってくる。したがって、時間を考慮せずに計数すると、直近と過去の状況が混在した結果が得られてしまうことになる。

時間を考慮した計数の基本的な考え方として、直近のデータには大きい重みを与えて、過去のデータには小さい重みを与えるという手法が考えられる。このような、時間に基づいてデータに重みを与える計数方法として Decaying Window という手法がある。Decaying Window は、各データに一つのカウンタを与えるとともに、時間間隔ごとに、各カウンタの結果に１に近い定数δ($1-10^{-5}$ など)を重みとして掛け合わす操作を行う。この結果、過去のデータは長い時間間隔を経過しているので、直近のデータよりも定数δがより多く掛け合わせられるため、徐々に重みを小さくすることができる。

以上の考え方に基づいて、Lossy Counting と Decaying Window を統合した計数アルゴリズムを開発した。具体的には、Lossy Counting は、カウンタを定期的にスキャンし、頻度が低いカウンタを削除するアルゴリズムであるが、毎回ストリームからデータを入力する時点でカウンタに重みをかける操作を追加している。この結果、時間の影響をパラメータと重みで自由に調整でき、しかも出現が少ないアイテムは数えない計数アルゴリズムとなっている。

[移動軌跡データからの意味情報推定]

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。しかしながら、教師無し学習であるクラスタリングは分類精度が悪いという問題がある。このため、複数のクラスタリングを行ったのちにクラスタアンサンブルを導入している。

本研究では、クラスタリングに PLSI と GMM 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。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。

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。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。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てクラスタリングを行い、人の属性を分類している。

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。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。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。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。

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いて、教師なし学習による推定をより洗練化して精度を上げる手法を提案する。

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。クラスタアンサンブルとして、Meta-Clustering Al-gorithm（MCLA）を利用している。

実験結果

東京大学空間情報科学研究センターの人の流れデータを利用して実験を行った。「人の流れ」データは 2008 年 10 月 1 日における東京都 60 万人の移動軌跡データである。各データはユーザーID、経緯度、時刻、性別などから構成されている。

[オンライン型移動軌跡マイニング]

DenStream との比較

まず、「人の流れ」データから stay point を抽出し、本アルゴリズムと類似したクラスタリング DenStream との比較を行う。図１は本手法（右）と DenStream（左）の実験結果を比較したものである。DenStream では、２箇所の高密度な stay point（黒い丸）が出現しているが、左は多摩センター駅、右

は走行中の電車を表している。一方、本手法の結果から、走行中の電車内では交通流が変化しないので、混雑エリアとしてみなされていないことがわかる。さらに、多摩センター駅とサンリオピューロランド付近での stay point が発見されている。

図１　本手法と DenStream の比較

メモリ消費量について

Decaying Window の有無による評価実験したところ、交通量に応じてメモリ消費量が大幅に激減していることが分かる。結果を図２に示す。横軸は時間、縦軸はカウンタ数である。紫の線はカウンタを削除せずに計数した場合の結果である。他の三つの線は閾値を変化させて実行した結果である。本手法のメモリ消費量はカウンタを削除しない場合よりも少量で済んでいることが分かる。

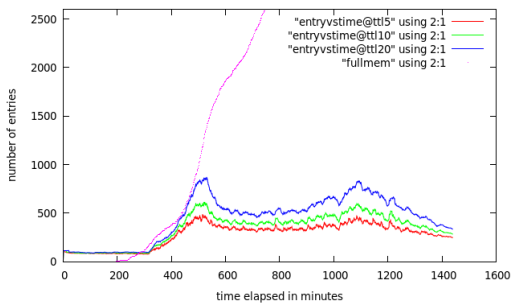

図２　メモリ消費量の比較

[移動軌跡データからの意味情報推定]

「人の流れ」データは、パーソントリップ調査（PT 調査）にジオコーディング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。このため、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。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。

人の属性推定

MCLA を用いて、人の属性推定を行った結果を示す。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。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。しかしながら、今回の実験では、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。このため、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。

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、教師なし学習である HMM を用いる。

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 37.6 %となった。HMM による推定は、人の属性推定と同様に、初期値への依存度が高いなどの問題がある。このため、移動手段の推定についても、クラスタアンサンブルを適用した。初期値を変えて３回実行した結果をアンサンブルしたところ、精度は 53.7%となり、精度向上が見られた。

今後の課題

本研究では、オンライン処理方式の移動軌跡マイニングと移動軌跡データからの意味情報推定について研究した。オンライン処理方式の移動軌跡マイニングの課題として、アルゴリズムの並列化がある。すなわち、Lossy Counting はすべてのカウンタを定期的にスキャンし、頻度が低いカウンタを削除するメカニズムを採用している。このため、すべてのカウンタは同一マシン上に置かなければならない。今後は、分散型メモリにも応用できるように、スキャンメカニズムと同様の効果を持つ、新しいメカニズムを開発する予定である。新しいメカニズムでは、各カウンタに「寿命」という概念の導入を考えている。寿命メカニズムは、各マシンが自身の持っているカウンタの寿命に基づいてカウンタの削除を管理して、非同期で動作できるようにしている。現在、Storm というストリーム処理フレームワークで実装中である。

移動軌跡データからの意味情報推定については、まだ精度が十分ではないため、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。

５．主な発表論文等

〔雑誌論文〕（計　１件）

[1] Kazuhiro Seki, Ryota Jinno, and Kuniaki Uehara: Parallel Distributed Trajectory Pattern Mining Using Hierarchical Grid with MapReduce, International Journal of Grid and High Performance Computing, 査読あり, Vol.5, No.4, pp.79-96 (2013).

〔学会発表〕（計　３件）

[1] 田中優子, 上原邦昭: 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, 第29回人工知能学会全国大会, (2015年5月30日)（北海道・函館市）.

[2] 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, 第 28 回人工知能学会全国大会, (2014年5月12日)（愛媛県・松山市）.

[3] 王一驄, 司南, 関和広, 上原邦昭: データストリーム手法による行動軌跡パターン検出と時空間情報の可視化, 第 28 回人工知能学会全国大会, (2014年5月12日)（愛媛県・松山市）.

〔図書〕（計　１件）

[1] Yicong Wang, Kazuhiro Seki and Kuniaki Uehara: Detection of Trajectory Patterns and Visualization of Spatio-Temporal Information Based on Data Stream Approaches, F. Bian and Y. Xie (Eds.) Geo-Informatics in Resource Management and Sustainable Ecosystem, pp.204-214, Springer (2015).

６．研究組織

(1)研究代表者

上原　邦昭（UEHARA, Kuniaki）

神戸大学・システム情報学研究科・教授

研究者番号：60160206

(2)研究分担者

関　和広（SEKI Kazuhiro）

甲南大学・知能情報学部・准教授

研究者番号：　30444566