1C-06

携帯電話ユーザの位置分布情報を用いたリアルタイムイベント検出方式

西山　智　　木村　健斗、中島　純、加島　伸二

KDDI

1．はじめに

近年、都市計画や防災、商圏分析など幅広い分野での応用を目指して、携帯電話ユーザから得られる位置情報を利用した人口動態分布を求める研究が行われはじめている。例えば数時間後の人口分布が推定できれば、災害発生時の帰宅困難者数などが予測でき減災に有効である。しかしながら従来研究では人の日常行動がモデル化されており、花火やコンサートなどのイベントが表現されていない。このため、災害時にイベントが行われていた場合、数時間後の分布予測を誤る可能性がある。そこで本稿では、携帯電話ユーザの位置情報の統計値(以下では推定ユーザ数と呼ぶ)からリアルタイムにイベントの発生を検出する手法について提案する。

2．背景

2.1　既存研究

推定ユーザ数からイベントの検出を行う課題は、センサー値から異常値検出を行う問題の一種である。正解データ無しの手法として、データの平均値や分散などを閾値とする統計的手法、データ間の距離に基づく手法、密度に基づく手法など多数が提案されている（詳細は、[1]参照）。しかし、ユーザ推定数の統計的特性が場所によって異なるため、統計的手法は適用が難しい。また他の手法は計算量が多く、リアルタイムでの検出に適していない。

通信ログ(CDR)を用いたイベント検出事例として、LPE(Localized P-value Estimator)[2]がある。LPE は、イベント無しの日を正解データとする教師あり学習によるもので、1 時間ごとのユーザ推定数を一日分(24 点)入力値とし、検出対象日と学習データ(イベント無しの日)との距離（例えばユークリッド距離）の k-近傍値(k-Nearest Neighbor)を、正解データ間の k-近傍値と比較し、指定日との k-近傍値が指定する比率(α)以上上回っていた場合にイベントと判定する手法である(図 1、図 2)。[2]ではバルセロナ市の CDR データを用いて LPE が教師無し学習方式である OC-SVM より検出精度が良いと報告している。しかし多数の場所で LPE が必要とする正解データを用意することは困難である。また、一日分のユーザ推定数を特徴量として使用するため、そのままではリアルタイム検出に適用できない。

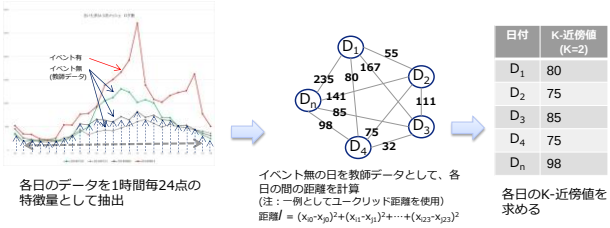

図 1 LPE 法(教師データによるモデル生成)

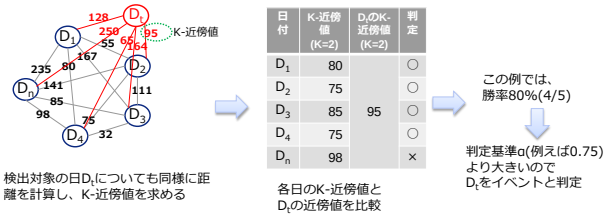

図 2 LPE 法(イベントの判定方式)

2.2　使用するデータ

本稿では、個別に許諾を取得した au ユーザについて、匿名化処理を行った CDR から位置を推定し、1/2 基準メッシュ(メッシュサイズ約 500m)ごとに 5 分ごとのユーザ数を集計して使用した。分析に当たっては常にメッシュ集計値を利用し、個別ユーザの位置情報は一切用いていない。なお CDR からの推定位置は数百メートルの推定誤差があり、メッシュ集計値もその影響を受けている。

3．提案手法

3.1 LPE をベースとするリアルタイム検出

LPE の判別コストは、教師データとの k-近傍値の計算(O(n)、n は教師データ数)+教師データ間の k-近傍値との勝ち負け判定(O(logn))であり、使用する特徴量もユークリッド距離などを用いれば時間軸上で差分計算が可能である。学習コストも $O(n^2)$ であり多数のメッシュに個別に判別器を作成できる。このため LPE の判定区間を短縮(例：4 時間)し、一定間隔の差分計算によりリアルタイムでイベントを検出する(図 3)。

LPE の課題は正解データを必要とすること、及

Real-time Event Detection Method using Location Information from Mobile Phone Handsets by Satoshi NISHIYAMA, Kento KIMURA, Jun NAKAJIMA and Shinji KASHIMA (KDDI)

び判定区間を短時間とした場合、常に少しだけ上回っている（あるいは下回っている）データを イベントとして誤検出することが多くなることである。これらの課題に対して以下の手法により課題の軽減を試みた。

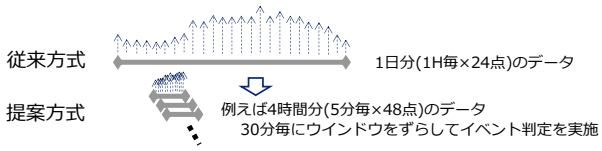

図 3 LPE によるリアルタイムイベント検出

3.2 正解データの自動生成

機械的なクラスタリングなどの手法では 100%イベント無しの日を分類することは難しいが、イベント有無が偏るように分類することは可能である。一方 LPE は k-近傍値を距離指標としており、正解データ(イベント無し)に少し例外(イベント有り)を含んでも判定可能と考えられる。そこで、機械的なクラスタリング(例えば k-means)によりすべてが正解ではないがより正解データが多いと予想されるクラスタを生成し、その要素を正解データとすることとした。

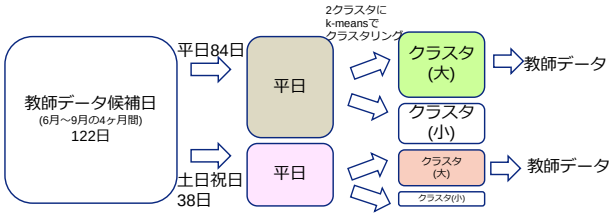

図 3 クラスタリングによる正解データの自動生成

3.3 誤検出対策

今回は 1 万人程度の大規模なイベントを検出することを想定し、距離の k-近傍値が正解データ間の k-近傍値の平均値と比較して大幅に大きい場合（ここでは平均値の 2 倍）のみをイベント検出とする。

このほか、ユーザ数の絶対数が少ない場合(1/2 基準メッシュあたり実人口で約 2,000 人未満の場合)、および教師データの平均より下回っている場合（例えばお盆の都心）も検出対象外とした。

４．実装と評価

4.1 実装

提案方式を Hadoop 上に実装し、ユーザ推定数のメッシュ集計値を用いてエミュレーションを行った。関東圏を包含する 12 個の 1 次メッシュに含まれる全ての 1/2 基準メッシュ(30.72 万箇所)について、2014 年 6 月～9 月の 4 ヶ月間のデータを基に、平日と土日休日の 2 つに分けて、クラスタリングにより正解データを生成し、ユークリッド距離を使用して LPE の判別器を生成した。評価対象として同年 7 月および 8 月に大規模(1 万人以上を目安)と想定されるイベントから 50 件を選定し、その開催地と目される 33 箇所の 1/2 基準メッシュを評価対象とした。イベント判定は 30 分毎に行い、イベント開始後 30 分(2 回)以上イベント検出が遅れた場合にイベント検出失敗、イベント検出を 2 回連続で通知したにもかかわらず、インターネット上の調査などでイベントの存在が検索できなかった場合に誤検出と定義した。

なお、LPE の k-近傍値の k は正解データの 1-2 割となるよう、平日休日それぞれ k=7、k=3 を使用した。また、判定基準は α>0.9 とした。

4.2 評価結果

50 イベントに対して検出成功 31 件、検出失敗 19 件の結果となり、再現率は 31/50=0.62 となった。また 7 月 8 月の 2 ヶ月間に提案方式がイベントと判定した 261 件のうち、イベントが見つからなかったものは 52 件であった。従って適合率は (261-52)/261=0.80 となった。これから F 値は 0.70 となった。なお、ユーザ推定数の定数倍を閾値とする統計的手法では、平均値の 1.8 倍～2.0 倍付近で F 値の極大値 0.62 となった。

５．考察

再現率が低い理由は、評価対象に選定した 50 件のイベントのうち、イベントで集まった人数が実際には少なく目視でもデータ上差が見られない場合(15 件)が含まれていたためである。これらを除くと再現率は 0.89、F 値は 0.85 となる。残りの失敗事例は、イベントを実施している日のほうが多くそちらのクラスタを正解データとしてしまった場合(4 件)であった。

６．おわりに

本稿では、LPE をベースに携帯電話のユーザ推定数からリアルタイムにイベントの発生を検出する手法を提案した。実データにより評価を行った結果、統計的手法と比較して高い精度でイベント検出できた。今後、より小型のイベントが検出できるよう手法を改良する予定である。

本研究は総務省直轄研究「G 空間プラットフォームにおけるリアルタイム情報の利活用技術に関する研究開発」の成果である。

参考文献

[1]Kriegel, H-P., Kroger, P. and Zimek, A., Outlier Detection Techiniques, Tutorial Notes, The 2010 SIAM International Conference on Data Mining, 2010.

[2] Neumann, J., Zao, M., Karatzoglou, A. and Oliver, N., Event Detection in Communication and Transportation Data, in Proceedings of IbPRIA 2013, 828-838, 2013.