

緯度経度情報追加による行動履歴分散表現の高精度化

Accuracy Improvement of User Move Embedding with Latitude and Longitude Information

佐賀健志 *1*2
Takeshi Saga

田中宏季 *1*2
Hiroki Tanaka

中村哲 *1*2
Satoshi Nakamura

*1 理化学研究所 革新知能統合研究センター 観光情報解析チーム
Riken, Center for Advanced Intelligence Project, Tourism Information Analytics Team

*2 奈良先端科学技術大学院大学
Nara Institute of Science and Technology

Recently, with the rise of the popularity of wearable devices like smartphones, it is becoming possible to use user's moving records as a part of big data. In previous research, user movements were modeled by bi-directional LSTM trained with time series of Mesh-ID. However, since Mesh-ID was assigned artificially, its physical distance and relative position were not considered properly. Therefore, it might be difficult to train the model effectively by using only Mesh-ID. In this research, to solve this problem, we used additional latitude and longitude information in the new model. As a result, we confirmed the accuracy improvement for Mesh-ID prediction and the difference between the output of clustering with user embeddings for each model.

1. はじめに

スマートフォン等の携帯端末は今や1人に1台が当たり前の時代になっている。そのような携帯端末はユーザの行動記録装置として健康維持から行動パターン解析まで幅広く応用されている。特に私たちの研究チームでは観光情報解析を目的として、これまでにLSTMや双方向LSTM(Bi-LSTM)による行動履歴埋め込み生成手法を提案してきた[久保19, 田中20]。しかし、これまでの手法では機械的に割り振られたMesh-IDのみを入力としているため、その離散性から各ランドマーク間の位置的關係や距離等をうまく学習できていない可能性があった。

そこで、本研究ではネットワークへの入力として、Mesh-IDだけではなく、連続的な緯度と経度の値を追加することで位置的關係性も学習させ、Mesh-ID推定性能の向上を試みた。本稿ではその際に使用した学習手法の詳細や、学習後の埋め込みベクトルを用いたクラスタリング結果について述べる。

2. 関連研究

Ermagunらは旅行中の行動傾向予測に対して解釈性の高いNested Logitと呼ばれる経済学的手法と汎化性能が高いと言われる機械学習手法であるRandom Forestをそれぞれ適用し、Random ForestがNested Logitよりも2倍程度の予測能力があることを示した[Ermagun15]。

Hirotaらは地図をメッシュ状に区画を分けてIDを付与し、ユーザの移動経路をID配列で表現することで、これまで自然言語処理で使われていたWord2Vecをユーザの経路予測に応用した[Hirota19]。しかし、Word2VecはID予測に予測対象の前後のIDのみしか利用していないため、系列情報を最大限生かし切れていない可能性があった。

Liuらは行動系列予測において各系列点間で時間的・位置的な線形性が考慮されていない問題を指摘した。この問題に対応

するために、RNNの入力特徴量として位置情報と時間情報を加えることで予測モデルを構築した[Liu16]。

CrivellariらはWord2Vecを用いて獲得された各ランドマークに対応する分散表現に対して、コサイン類似度やt-SNEを適用することで各ランドマークや各移動経路の意味的類似度について分析を行った[Crivellari19]。これによりWord2Vecによる分散表現の行動分析への応用可能性を示した。

また、その後の研究でCrivellariらはLSTMを用いて入力行動系列から次の場所を予測する研究も行っている[Crivellari20]。LSTMは入力系列の全ての情報を隠れ状態として伝播させるため、Word2Vecよりも情報を効率的に利用していると考えられる。

久保らはBi-LSTMによって分散表現を獲得し、それらを用いて経路の階層的クラスタリングを行った[久保19, 田中20]。LSTMは順方向にしか情報を伝播しない一方で、Bi-LSTMは逆方向にも情報を伝播するため経路の順番に依存しない意味的類似性を考慮した分析において有効な手法であると考えられる。しかし、Mesh-IDでは相対的位置が明示的に示されていないため、位置的距離や相対位置などをうまく学習できていない可能性があった。そこでそのモデルをベースに、本研究では緯度経度の値を学習に利用することとした。

3. 提案手法

3.1 モデル構造

モデル構造の概略図を図1に示す。先行研究で使用されていたモデルをベースにしているが、入力ベクトルとして緯度経度を追加している[田中20]。図1における各Predictionは全結合層によって構成されている。緯度経度予測の出力は各1次元のスカラー値、Mesh-ID予測の出力は学習データのクラス数(学習データに含まれるMesh-IDの種類数)を次元とするone-hotベクトルである。

3.2 損失関数

Mesh-IDに緯度経度を加えたモデルでは離散値であるMesh-IDと連続値である緯度経度を同時に学習させる必要がある。しかし、離散値に対する損失関数と連続値に対する損失関数は

連絡先: 佐賀健志 : takeshi.saga@riken.jp

連絡先: 田中宏季 : hiroki.tanaka.ks0@riken.jp

連絡先: 中村哲 : satoshi.nakamura.vr0@riken.jp

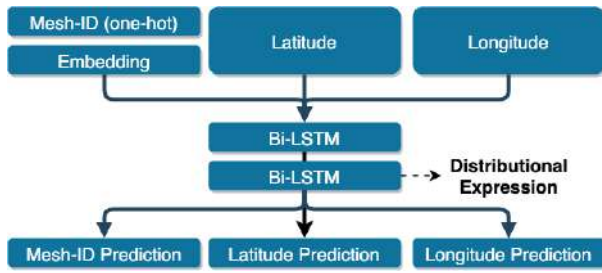


図 1: Mesh-ID と緯度経度を入力とする Bi-LSTM

異なるため、今回は以下のような数式によって統合損失 L_{total} を算出して学習全体の損失関数とした。

$$L_{total} = \alpha * L_{mesh} + \beta * (L_{lat} + L_{long}) \quad (1)$$

ここで L_{mesh} は Mesh-ID に対する交差エントロピー損失、 L_{lat} と L_{long} はそれぞれ緯度と経度に対する MSE、 α と β は損失値のバランスを調整するハイパーパラメータである。

4. Mesh-ID 推定

4.1 学習データ

学習には 2020 年 7 月から 10 月までの 4ヶ月間のスマートフォンユーザの日次時系列行動履歴を使用した*1。

今回は正確な分析を行うために元データに対して位置誤差 100m 以下で 30 系列以上存在するユーザのデータをフィルタリングしている。さらに、計算量を削減するためにその中から出発 (Departure) と到着 (Arrival) のタグがついているデータのみを解析対象としている。以上のフィルタリングの結果、学習解析対象データの総系列長が 8,768,203 となった。

また、これらのデータはそれぞれ学習用・検証用・テスト用に 8:1:1 の比率で分割されている。

4.2 詳細設定

最適化関数には学習率 0.001 の Adam を使用した。Bi-LSTM の隠れ層の次元数は各層 300 次元 (順方向と逆方向でそれぞれ 300 次元であるため、各 Bi-LSTM 層全体で連結すると 600 次元になる)、drop out は 0.5 に設定している。 α と β は事前実験においてテストデータに対する Accuracy が最も高かった 1 と 100 の組み合わせを用いている。

モデルへの入力系列長はデータの最大系列長に合わせて 260 とし、それより短い系列に対しては系列の末尾に 0 埋めを行っている。緯度経度は変化幅が非常に小さいため、平均値が 0、分散が 1 になるように標準化してから入力している。

4.3 学習結果

Mesh-ID 予測における Mesh-ID のみのモデルと Mesh-ID に緯度経度を加えたモデルのテストデータに対する Accuracy 比較を表 1 に示す。Mesh-ID だけで学習した場合よりも緯度経度情報を追加した場合のほうが高い Accuracy となった。

4.4 考察

緯度経度を加えたことによる影響を調べるために学習曲線の分析を行った。緯度に対する MSE と Mesh-ID に対する交差エントロピー損失の学習推移をそれぞれ図 2 と図 3 に示す (経度に対する MSE 推移は類似していたため省略する)。Mesh-ID に対する損失値は徐々に低下している一方で、緯度に対する損

表 1: Mesh-ID 予測における Accuracy 比較

モデル	Accuracy
Mesh-ID のみのモデル	0.32
緯度経度を加えたモデル	0.34

失値は序盤に大きく減少した後は改善が見られなかった。この結果より、学習初期は緯度経度情報が利用され、学習が進むにつれて Mesh-ID 情報が学習に利用されている可能性が示された。

さらに、Mesh-ID に緯度経度を加えたモデル (図 3) と Mesh-ID のみのモデル (図 4) の Mesh-ID に対する交差エントロピー損失の推移を比較すると、Mesh-ID に緯度経度を加えたモデルのほうが序盤の損失の減少が早いことが確認できる。このことから、Mesh-ID 間の相対的位置関係を明示的に示す緯度経度情報を加えることで、学習序盤における損失値の減少を早めることにつながったと考えられる。念のため、別々に 3 回学習を行ったが、いずれの試行においても同様の傾向が確認された。

5. 階層的クラスタリング

5.1 詳細設定

先行研究に倣い、Mesh-ID のみのモデルと Mesh-ID に緯度経度を加えたモデルによって獲得した分散表現に対してそれぞれ階層的クラスタリングを行った。

入力系列に対する分散表現には最終層の Bi-LSTM の順方向と逆方向の隠れ状態を結合して用いた。この際、Mesh-ID 推定に使用したデータでは多すぎて結果の解釈が難しいため、系列長 200 以上のデータを解析対象に変更している。クラスタリング手法にはコサイン類似度を用いた Ward 法による階層的クラスタリングを使用している。最終的なクラスタ作成時にはクラスタ間距離が 1.2 になるように設定している。

5.2 学習結果

Mesh-ID のみのモデルと Mesh-ID に緯度経度を加えたモデルによって獲得した分散表現に基づくクラスタリング結果をそれぞれ図 5 と図 6 に示す。各図の上部にコーフェンの相関係数の値を示しているが、それぞれの Bi-LSTM で獲得した分散表現が異なるため単純比較はできない点に注意が必要である。

Mesh-ID のみのモデルと Mesh-ID に緯度経度を加えたモデルに対して、クラスタ間距離 1.2 (図中点線) の位置でクラスタを分けると、それぞれ 2 つと 3 つのクラスタに分割されていることが確認できる。

参考までに Mesh-ID のみの場合は全く異なるクラスタに分類されていたにもかかわらず、緯度経度を追加することで同一クラスタ内の極めて近い距離に分類された例としてユーザ 17 とユーザ 21 をそれぞれ図 7 と図 8 に示す。これらのユーザーはどちらも中央線・総武線の沿線に位置した軌跡であることがわかる。

また、同様に提案手法によって極めて近い類似度を示したユーザ 12 (図 9) と 13 (図 10) は、両者ともにときわ台駅を中心とした軌跡となっていた。

6. まとめ

本研究では Bi-LSTM を用いた先行手法に緯度経度を追加して学習させることで Mesh-ID 予測 Accuracy が向上し、移

*1 データ提供:株式会社 Agoop

動系列情報に対する分散表現の表現力が向上することが確認された。

しかし、現在の手法では地図情報と行動系列を見比べながら発見的に分析する必要があるため、解析者の技量に依存するだけでなく、時間と人手もかかってしまう。この問題を解決するために行動傾向予測に役立つような位置情報付近の人気店などのランドマーク情報や地理的特徴などを組み込んだ解析手法について検討が必要である。

参考文献

[Crivellari 19] Crivellari, A. and Beinat, E.: From Motion Activity to Geo-Embeddings: Generating and Exploring Vector Representations, Traces and Visitors through Large-Scale Mobility Data, in *ISPRS Int. J. Geo-Inf.* (2019)

[Crivellari 20] Crivellari, A. and Beinat, E.: LSTM-Based Deep Learning Model for Predicting Individual Mobility Traces of Short-Term Foreign Tourists, *Sustainability*, Vol. 12, No. 1 (2020)

[Ermagun 15] Ermagun, A., Rashidi, T., and Lari, Z.: Mode Choice for School Trips: Long-Term Planning and Impact of Modal Specification on Policy Assessments, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2513, No. 1, pp. 97–105 (2015)

[Hirota 19] Hirota, M., Oda, T., Endo, M., and Ishikawa, H.: Generating Distributed Representation of User Movement for Extracting Detour Spots, in *MEDES '19: Proceedings of the 11th International Conference on Management of Digital EcoSystems* (2019)

[Liu 16] Liu, Q., Wu, S., Wang, L., and Tan, T.: Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30 (2016)

[久保 19] 久保基, 田中宏季, 中村哲: 観光行動理解のための分散表現に基づくユーザクラスタリング, 人工知能学会全国大会 (2019)

[田中 20] 田中宏季, 久保基, 中村哲: 時系列分散表現に基づく東京都の人流クラスタリング, 観光情報学会研究発表会 (2020)

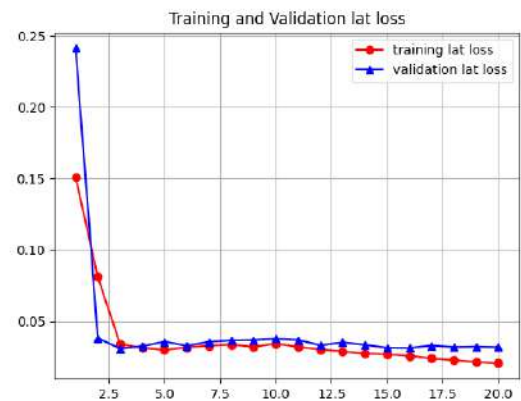


図 2: Mesh-ID に緯度経度を加えたモデルにおける緯度 MSE

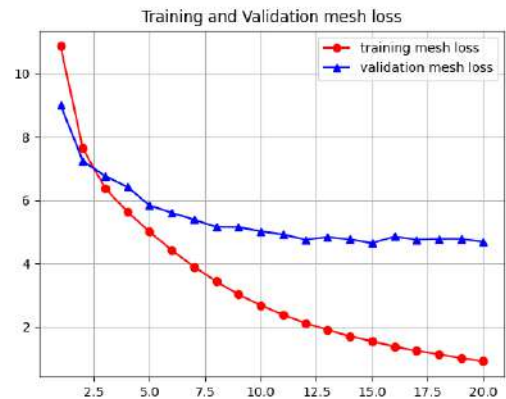


図 3: Mesh-ID に緯度経度を加えたモデルにおける Mesh-ID 交差エントロピー損失

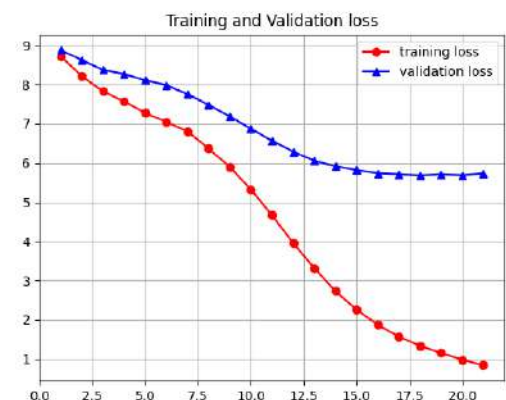


図 4: Mesh-ID のみのモデルにおける Mesh-ID 交差エントロピー損失

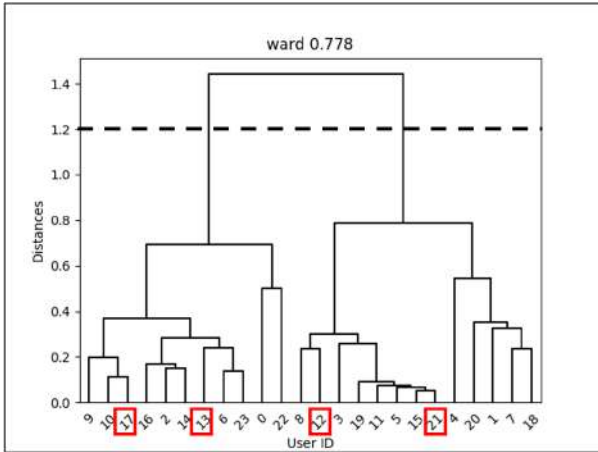


図 5: Mesh-ID のみのモデルを用いたクラスタリング結果

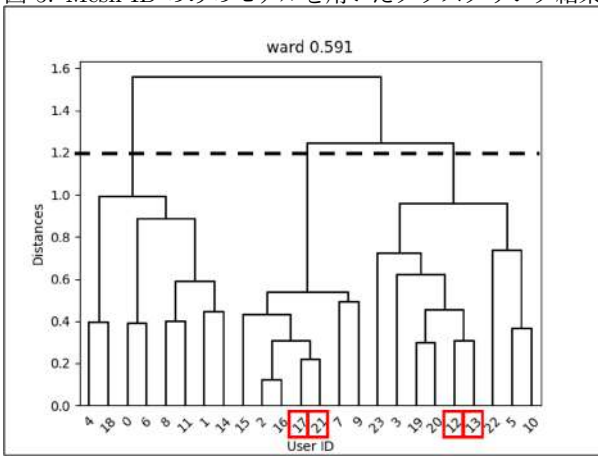


図 6: Mesh-ID に緯度経度を加えたモデルによるクラスタリング結果



図 9: ユーザー 12 の行動軌跡



図 7: ユーザー 17 の行動軌跡



図 8: ユーザー 21 の行動軌跡



図 10: ユーザー 13 の行動軌跡