

プローブデータの地点別速度の機械学習

広島大学 学生会員 ○松本 拓樹
 広島大学 正会員 塚井 誠人

1. はじめに

近年、高精度な予測や分類を実現する機械学習手法として、深層学習が注目を集めている。交通分野でもプローブデータのようなビッグデータに深層学習を適用した研究が見られる。永廣ら¹⁾はプローブデータを教師データに用いて、車両感知器による情報から旅行時間を予測する手法を示した。また、小川ら²⁾はイベントの有無や降水量を考慮した交通量予測手法を提案している。これらの研究では、予測精度の向上を狙って深層学習を適用しており、今後も予測タスクを中心に深層学習の利用が進むと考えられる。

深層学習はニューラルネットワーク(以下 NN)を利用し、データの特徴の効果的な抽出により、予測・分類を行う。この、NN による特徴抽出は表現学習と呼ばれ、深層学習は表現学習を繰り返すことで抽出した特徴により、高精度な予測や分類を可能としている。しかし、表現学習により得られる特徴は機械的な処理に適するが、その内容の直感的な理解は難しい。

表現学習によって抽出される特徴が示す内容が理解可能な知見として蓄積されれば、効果的な分析や予測タスクなどの性能向上に寄与できると考える。しかし、交通分野で表現学習に着目した文献は筆者の知る限り存在しない。そこで本研究では、NN にプローブデータを投入して特徴を抽出する。特に、交差点付近における速度の共起を特徴として抽出して考察し、抽出された特徴と交通工学的知見との対応を明らかにする。

2. NN による特徴抽出手法

NN を用いた特徴抽出手法である Word2vec を用いる。Word2vec (skip-gram モデル) は自然言語処理において単語の意味を表すベクトル(分散表現)を、「周囲の単語を予測する」ことで獲得する手法である。本研究では、隠れ層が1層のみの浅層の NN により表現学習を行い、データの特徴を簡潔に抽出する。Word2vec の自然言語以外への適用例に、名渡山ら³⁾の商品ベクトル獲得や

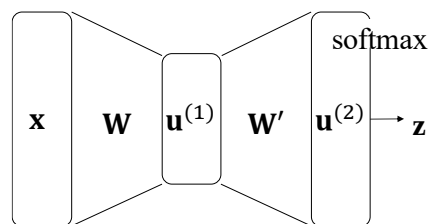


図1 Word2vec で用いる NN

椿ら⁴⁾のたんぱく質ベクトル獲得があげられる。

Word2vec では、図1に示す NN を用いる。この NN では、入力ベクトル $\mathbf{x}$ から出力ベクトル $\mathbf{z}$ が以下のように計算される。

$$\mathbf{W} \cdot \mathbf{x} = \mathbf{u}^{(1)}, \quad \mathbf{W}' \cdot \mathbf{u}^{(1)} = \mathbf{u}^{(2)}, \quad \mathbf{z} = \mathbf{f}(\mathbf{u}^{(2)})$$

なお、

$$\mathbf{u}^{(1)} = \begin{bmatrix} u^{(1)}_1 \\ \vdots \\ u^{(1)}_J \end{bmatrix}, \quad \mathbf{u}^{(2)} = \begin{bmatrix} u^{(2)}_1 \\ \vdots \\ u^{(2)}_K \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_i \end{bmatrix}$$

$$\mathbf{z} = \begin{bmatrix} z_1 \\ \vdots \\ z_K \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} w_{11} & \cdots & w_{1J} \\ \vdots & \ddots & \vdots \\ w_{J1} & \cdots & w_{JJ} \end{bmatrix}$$

$$\mathbf{W}' = \begin{bmatrix} w'_{11} & \cdots & w'_{1J} \\ \vdots & \ddots & \vdots \\ w'_{K1} & \cdots & w'_{KJ} \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} f(u_1) \\ \vdots \\ f(u_K) \end{bmatrix}$$

$$f(u_k) = \frac{\exp(u_k)}{\sum_{k=1}^K \exp(u_k)}$$

である。

具体的には、単語 $n(n=1 \sim N)$ 、語彙 $v(v=1 \sim V)$ より構成される文書をもとに、周囲 1 単語の予測を行うことで語彙 v の分散表現 $\mathbf{w}_v$ を得る手順を以下に示す。

- 1) 全単語を one-hot ベクトル $\mathbf{w}_v^{\text{onehot}} \in \{\mathbf{e}_v\}_{v=1}^V$ に変換
- 2) 単語 n に該当する語彙 v の $\mathbf{w}_v^{\text{onehot}}$ を $\mathbf{x}$ に代入して $\mathbf{z}$ を計算する。周辺単語 $n'(n' = n \pm 1)$ のそれぞれに該当する語彙 v' の $\mathbf{w}_{v'}^{\text{onehot}}$ を教師ベクトル $\mathbf{y}$ として、 $\mathbf{z}$ と $\mathbf{y}$ の誤差 E を減らす方向に $\mathbf{W}$, $\mathbf{W}'$ を調整
- 3) $n=1 \sim N$ に対し 2) を行い、最適な $\mathbf{W}$, $\mathbf{W}'$ を求めたとき、 $\mathbf{W} \cdot \mathbf{w}_v^{\text{onehot}}$ を語彙 v の分散表現 $\mathbf{w}_v$ とする

キーワード プローブデータ、ニューラルネットワーク、機械学習、交通流、信号制御、分散表現

連絡先 〒739-8527 広島県東広島市鏡山 1-4-1 A2-543 広島大学先進理工系科学研究科 地球環境計画学研究室

T E L 082-424-7825

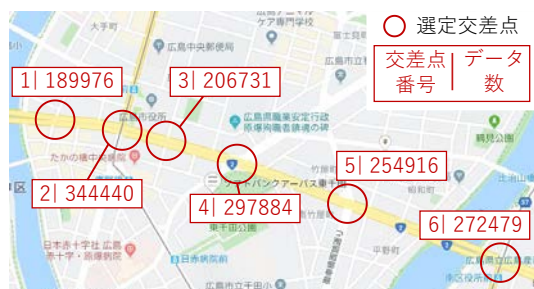


図2 対象交差点とデータ数

ここで、誤差 E は

$$E = - \sum_{k=1}^K y_k \log z_k$$

と定義する。また、one-hot ベクトルは V 次元の基底ベクトルで、1 語彙に 1 要素を対応させ、単語に該当する要素を 1、他を 0 としたベクトルである。

分散表現の次元数は任意である。関数 f はソフトマックス関数と呼ばれ、全要素の総和が 1 となるため、 $\mathbf{z}$ は各語彙の予測出現確率とみなせる。本研究では最も確率が高いと予測した語彙の正解率をもとに次元数を決定した。次元数の増加に伴い正解率は上昇するが、十分に次元数が大きいと正解率の上昇は見られなくなる。そこで、正解率が一定となった時の次元数を採用する。

3. 使用するプローブデータの概要と対象交差点

プローブデータは、GPS を搭載した車両が走行した位置や速度などを、空間的、もしくは時間的に一定間隔で記録したデータである。車載器を搭載した車両しか観測できないが、車両感知器に比べて様々な道路断面のデータを得られる。

本研究ではパイオニア社が提供する、カーナビ搭載車両の位置、日時、速度などを 3 秒ごとに記録したカーナビプローブデータを用いる。本データは速度に関わらず取得されるため、低速時のデータ収集性に優れる。また、衛星電波のドップラー効果により瞬間速度を算出している。データ収集期間は、2015 年 4 月から 12 月の 9 カ月間である。

対象交差点と、各交差点 (中心から 100m 以内と定義) のデータ数を図 2 に示す。広島市中区から南区の国道 2 号上の 6 交差点である。以下では、国道 2 号を主道路、主道路と交差する道路を従道路と表記する。なお、図 2 のデータ数はプローブカーが 3 秒ごとに観測した延べデータ数であり、交差点を通過したプローブカーの台数とは一致しない。時間あたり観測データ数は、最多の交差点 2 で 52 点/h で、最少の交差点 1 では 29 点/h である。

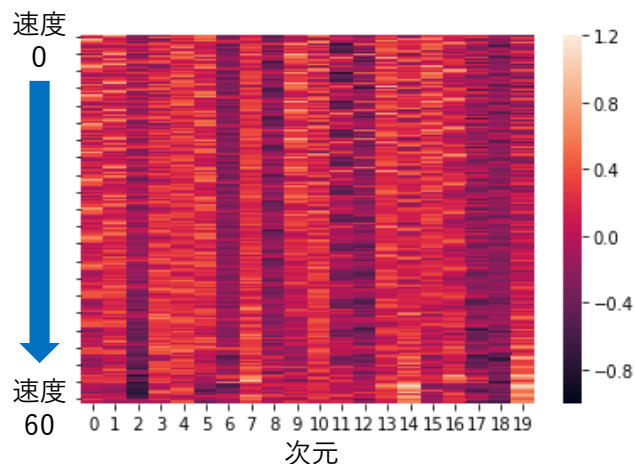


図3 各素性の分散表現

4. 速度データへの Word2vec の適用

Word2vec の適用のため、プローブデータを交差点番号・速度・進行方向で離散化した。速度は 5km/h 単位で離散化し、60km/h 以上のデータは全て 60 とした。進行方向は主道路上の東進を E、西進を W、従道路は NS とする。たとえば、交差点 1、速度 20~24km/h、進行方向が西のデータは、「1-20-E」と表記する。この離散化したデータを「素性」と定義する。得られた 234 素性が自然言語における語彙に相当し、観測データを日時順に並べた文書を作成して、Word2vec を適用することで各素性の分散表現を得る。

Word2vec の適用にあたり、プローブデータの特徴を考慮して周辺単語の定義の工夫を行った。具体的には、周囲の素性を学習するのではなく、同一時間帯の 10 素性をランダムに選択して学習する。一般にプローブデータは同一車両のデータが連続して記録される。そのため周囲の素性から分散表現を学習すると、同一車両が観測した素性を多く学習する。しかし、本研究では交差点間の速度の共起性に注目するため、他車両の素性を学習の方が望ましいためである。また、分散表現の次元数は、次元数の増加に伴う正解率の上昇が見られなくなった 20 次元とした。

獲得した分散表現のヒートマップを図 3 に示す。図 3 の縦軸は速度順の素性を、横軸は各素性の次元を示し、各要素の値を色で示した。獲得した分散表現は、-0.8 から 1.2 の値を要素に持つベクトルとなった。

前述のとおり、分散表現の直接の理解は難しい。そこで、素性間の類似性をもとに素性間の共起を考察する。同一時間帯に観測されやすい素性は互いに似た素性を学習して分散表現を獲得している。つまり、同一時間帯に共起しやすい素性間の分散表現は類似する。

表 1 各クラスターの概要

クラスター	素性数	交差点	主な速度	主な進行方向
1	70	3以外	中	E, NS
2	33	1~4	低~中	E
3	26	3, 4	低	E, NS
4	13	1	低	W, NS
5	62	1~6	中	W, NS
6	21	1~6	高	W
7	9	5以外	高	E, NS

5. 分散表現の類似性に基づく共起性の考察

各素性の分散表現をクラスタリングし、素性間の類似性を判定した。クラスタリングには階層的クラスタリングであるウォード法を用いる。同一クラスターに属する素性の分散表現は類似しているといえる。また、階層的クラスタリングのため、クラスター間距離が近い素性も分散表現の類似性が比較的高く、クラスター間距離が大きいほど類似性が低くなる。クラスター数は、各クラスターの要素数を考慮して7クラスターとした。表1に各クラスターの交差点・速度帯・進行方向を示す。また、図4にクラスター間距離をデンドログラムとして示す。

図4よりクラスター1~5とクラスター6, 7間のクラスター間距離が大きく、低速~中速の素性と高速の素性が獲得した分散表現は大きく異なる。さらに、クラスター1~5は、進行方向がEの素性を主とするクラスター1~3と、Wの素性を主とするクラスター4, 5のクラスター間距離が大きい。進行方向がEとWの素性はそれぞれ異なる分散表現を獲得し、NSの素性はEかWどちらかに似た分散表現となった。以上より、低速~中速の素性と高速の素性の共起性は低く、それぞれが観測されやすい時間帯が異なる。また、低速~中速が主となる時間帯では、EとWの進行方向が観測されやすい時間帯が異なるといえる。前者は時間帯による平均速度の変動を、後者は時間帯による東西方向の交通量変動を示している。

次に、クラスター1を例に同一クラスターに属する素性間の類似性に着目する。クラスター1には交差点5の低速、EとNSの素性が属する。つまり、交差点5では低速のE, NSの共起が起りやすいので、交差点5では進行方向がEとNSの車両が干渉した速度低下（交通干渉）の発生が考えられる。

さらに、クラスター1内に属する素性間の類似性をt-SNEにより可視化する。t-SNEは、高次元空間上で近傍に位置する2点を近傍に、遠い2点は遠く配置した低

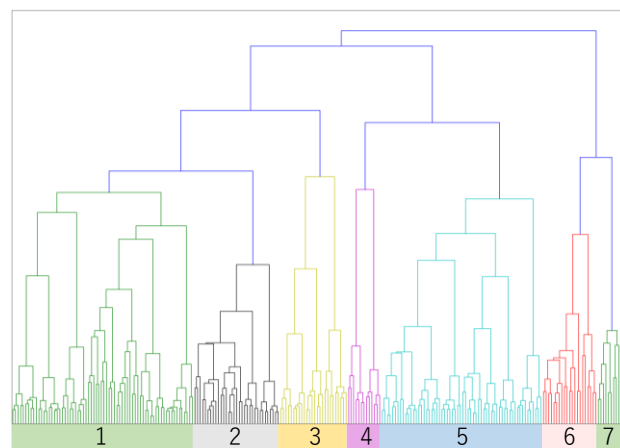


図4 デンドログラム（クラスター間距離は縦軸）

次元空間へと変換し可視化する手法である。つまり、t-SNE図において距離が近い2点は元の空間上でも距離が近く、分散表現が類似する。図5にクラスター1のt-SNE図を示す。図5中央部の交差点5, 6の速度が高い素性間で距離が近く、これらの素性は特に共起しやすい。交差点5, 6間を東進する車両は速度を低下せず通過できることを示し、信号制御時に設定される交差点間のスルーバンドを捉えた可能性がある。

クラスター2~7においても同様に交通干渉および交差点間共起の発生を考察し、図6に整理した。交通干渉（赤矢印）と交差点間共起（緑矢印）を示す。なお、より定量的な判定のため、コサイン類似度を用いる。素性*i*の分散表現 $\mathbf{w}_i$ と素性*j*の分散表現 $\mathbf{w}_j$ 間のコサイン類似度 $\cos(\mathbf{w}_i, \mathbf{w}_j)$ は、

$$\cos(\mathbf{w}_i, \mathbf{w}_j) = \frac{\mathbf{w}_i \cdot \mathbf{w}_j}{\|\mathbf{w}_i\| \|\mathbf{w}_j\|}$$

と定義する。交差点が異なる同一進行方向の素性間の類似度が0.9以上の場合に交差点間共起と判定した。なお、交通干渉の判定はクラスター数に、交差点間共起の判定はコサイン類似度の閾値に依存するが、これらの判定基準は今後の検討課題である。

交差点1, 5, 6では交通干渉が見られる。これらの交差点は川に隣接しており、東西方向の交通量の増加時に、橋を通過する車両が従道路から主道路へ集中し、交通干渉が発生すると考えられる。また、主に交差点1~4間で交差点間共起が見られる。信号制御により、この区間でスルーバンドの設定がされている。交通流上のスルーバンドを素性の共起としてとらえた。

6. クラスターの時間帯別割合

時間帯ごとに観測された素性が属するクラスターを集計し、図7に示す時間帯別のクラスターの割合を得た。クラスター3, 6は1日を通して割合の変動が比較的小さ

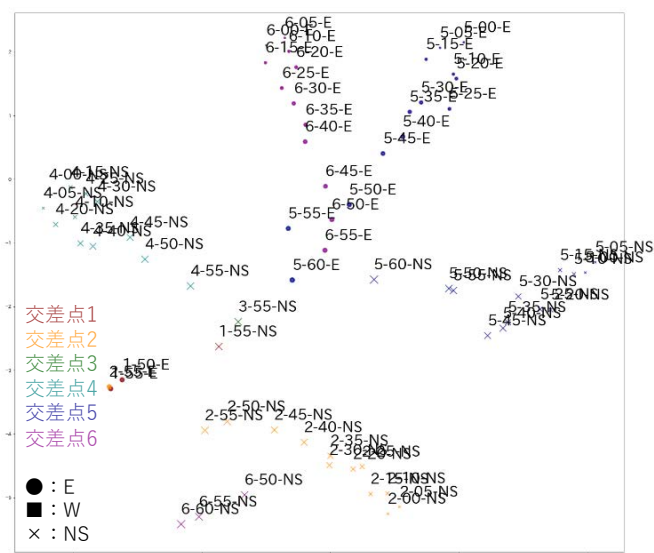


図5 クラスタ1の類似性の可視化

い。それに対し、クラスタ5は日中に、クラスタ7は夜間に増加する。また、クラスタ1は6時台、クラスタ2は8時台、クラスタ4は18時台に増加している。

日中は、クラスタ5で見られる交差点1~4間のスルーバンド（西進）や、交差点5における交通干渉が発生しやすい。また、夜間はクラスタ7で見られる、交差点1~4間の速度帯の高いスルーバンド（東進）が発生しやすい。クラスタ1, 2, 4の示す交通干渉および交差点間共起は、それぞれ6, 8, 18時台に起こりやすい。特に、6時台の交差点5、および18時台の交差点1における交通干渉は、通勤車両による交通量の増加が原因の可能性がある。時間変動の少ないクラスタ3, 6の示す交通干渉およびスルーバンドは、時間帯に左右されず発生するといえる。

しかし、クラスタ6の共起が示す高速のスルーバンドは、交通量の多い日中に発生することは稀と考えられる。本研究では、1素性に対応する1つの分散表現を求めて共起性の考察を行った。そのため、実際は時間帯などの要因により変化する共起の傾向を、時間帯によらず一定と捉えてしまったことにより、獲得した分散表現の表現力が低下してしまった可能性がある。NNの構造を工夫し、時間帯による共起傾向の変化を分散表現の変動としてとらえたい。

7. 結論

本研究では、プローブデータにより作成した素性にWord2vecを適用し、交差点・進行方向別速度の分散表現を求めた。分散表現の類似性に基づき、素性の共起傾向を考察した。その結果、交通干渉スルーバンドを示す交差点間共起を、それぞれの発生交差点や区間とともに

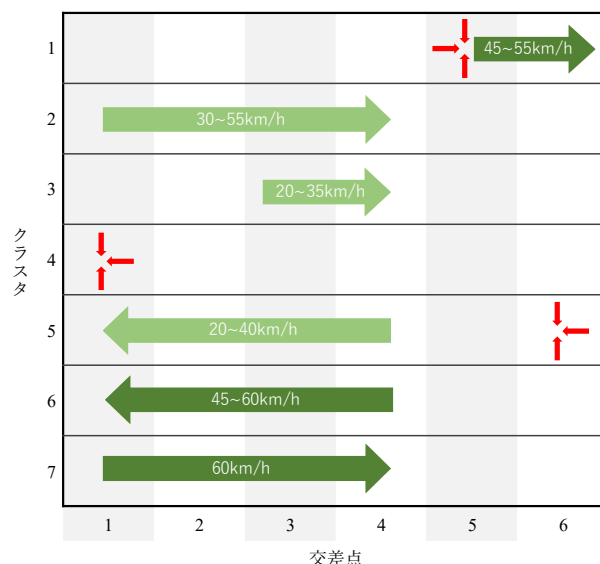


図6 交通干渉と交差点間共起の整理

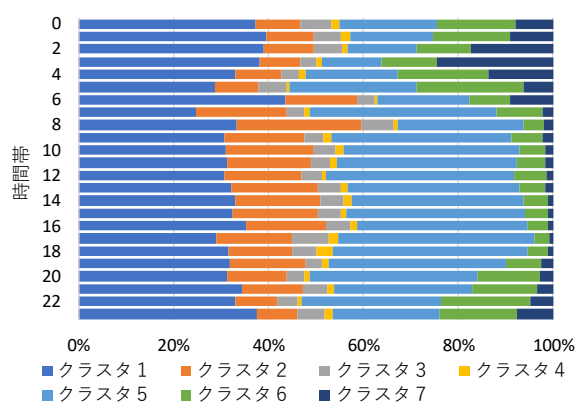


図7 時間帯別のクラスタ割合

捉えた。これらの現象は、交通工学的な知見とも整合している。さらに、時間帯に応じた素性の出現割合から、交通干渉と交差点間共起の時間変動を考察した。

しかし、低速時の交差点間共起はとらえられなかった。また、時間帯により変動する共起傾向を捉える点が課題である。NNの構造の変更などを通して、時間帯別の交通状況の変化を捉える改良を行いたい。

参考文献

- 1) 永廣 悠介, 西岡 悟史, 岡本 博: ETC2.0 データを活用した Neural Network モデルによる所要時間提供情報の精度向上, 交通工学論文集, Vol.5, No.2, pp.B_42-B_48, 2019.
- 2) 小川 晃平, 福田 大輔: LSTM の枠組みによる交通量短期予測の検討: 鎌倉市中心部を事例として, 第 60 回土木計画学研究発表会・講演集 (CD-ROM), No.37-02, 2019.
- 3) 名渡山 夏子, 岡本 一志: Word2vec に基づく購買履歴からのアイテムベクトル学習, 知能と情報 (日本知能情報ファジィ学会誌), Vol.29, No.3, pp.579-585, 2017.
- 4) 椿 真史, 新保 仁, 松本 裕治: タンパク質構造予測のための表現学習, 情報処理学会研究報告, Vol.2015-BIO-44, No.2, 2015