プローブデータを用いた地点別速度の表現学習

松本　拓樹[1]・塚井　誠人[2]

[1]学生会員　広島大学大学院　先進理工系科学研究科（〒739-0047 広島県東広島市鏡山 1-4-1 A2-543）
E-mail:m200795@hiroshima-u.ac.jp

[2]正会員　広島大学准教授　先進理工系科学研究科（〒739-0047 広島県東広島市鏡山 1-4-1 A2-543）
E-mail:mtukai@ hiroshima-u.ac.jp

自然言語処理や画像処理などの様々な分野で従来手法より高い成果を上げる深層学習は，交通分野でも適用が進みつつある．深層学習が高い性能を上げるのは，ニューラルネットワークがビッグデータから効果的な特徴量を得る能力を持つためである．この能力は「表現学習」と呼ばれており，予測や分類タスクだけでなく，データ分析にも活用が期待されている．本研究では，表現学習に基づくデータマイニング手法に着目し，プローブデータへの適用を試みた．具体的には，表現学習を利用した特徴抽出アルゴリズムである Word2vec を利用して，幹線道路上の車両の走行軌跡を捉えたベクトルを得る．ベクトルの類似度を用いて，対象区間における走行軌跡を進行方向別・時間帯別に可視化した．また，進行方向別・時間帯別の結果から，交差点の特性や信号制御の影響を考察した．

Key Words: *big-data, skip-gram, word embedding*

1. はじめに

深層学習は，多層のニューラルネットワークを用いた機械学習手法で，近年，音声や画像，自然言語などの分類・予測タスクで他の方法を圧倒する高い性能を示している[1)]．近年は交通分野でも，時系列予測や画像認識のタスクにおいて深層学習の利用が進みつつある．

永廣ら[2)]は，高速道路の所要時間の予測精度改善を目的に，ニューラルネットワークを用いた所要時間予測モデルを作成した．その結果，事故等の特異事象の発生時を除いて，現在の方式よりも大きく予測精度が改善することを示した．小川ら[3)]は，深層学習を利用した時系列データ予測手法である Long-Short Term Memory (LSTM) を用いた交通量予測システムを構築した．予測精度向上のため，イベントの有無や降水量を変数として用いたほか，誤差関数に平均二乗対数誤差を用いることで，過小な予測の削減を狙っている．分析の結果，誤差関数に一般的な最小二乗誤差を用いた場合と比べて平均予測精度はやや悪化したが，過小予測の削減につながった．佐藤ら[4)]は，信号制御に深層学習と強化学習を組合わせた Deep Q-Network による，動的な信号制御手法を提案した．また，交差点内の状況の認識に，画像認識手法である畳み込みニューラルネットワークを用いた点も特徴である．この研究では交通流シミュレーター上に十字交差点を再現し，シミュレーターの画像から畳み込みニューラルネットワークによって状況判断を行うとともに，Q-Network によって待ち台数や待ち時間を減らす信号制御を行うように学習を行った．この研究は右左折が考慮されていないことや他の交差点の信号制御を考慮していないことなど，実環境の反映には課題が残るものの，静的な制御と比較して渋滞の発生が抑制された．

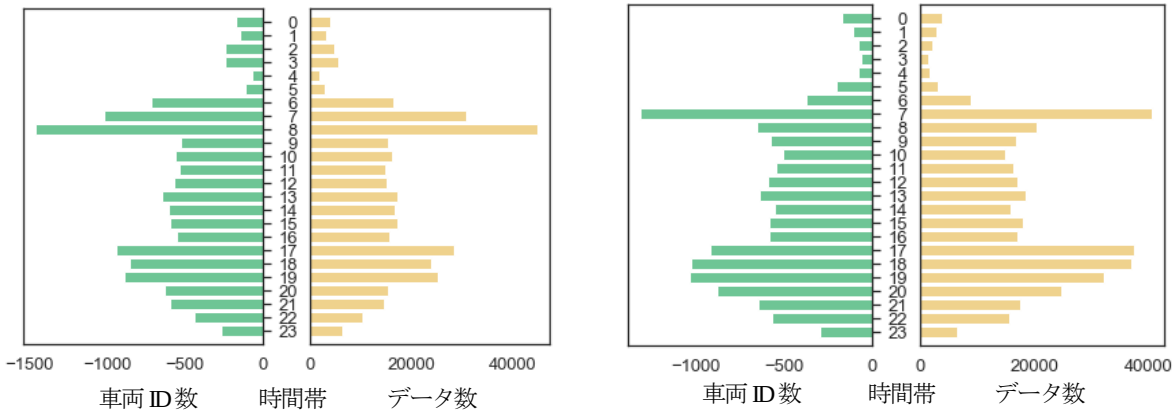
これらの深層学習が高い性能を上げるのは，ニューラルネットワークがデータから効果的に特徴量を得る能力を持つためである．この能力は「表現学習」と呼ばれ，ビッグデータの利活用の際に大きな効果を発揮する．表現学習は，予測や分類だけでなく，動的制御や動的学習の改善をもたらす基礎分析として活用が期待される．

そこで筆者らは，表現学習に基づくデータ分析手法として，Word2vec に着目した．Word2vec は，自然言語処理の分野において，語彙の意味をベクトル化する手法として開発された．Word2vec により獲得できる意味ベクトルは学習に用いた自然言語データの特徴を反映する．その特性を用いて，意味ベクトルから文書中に含まれる特徴の分析が行われている．例えば，Nikhil Garg ら[5)]は，Word2vec および Word2vec と同様の意味ベクトル獲得手法である GloVe を用い，1900 年代以降の新聞記事から語彙の意味ベクトルを獲得した．それらを用いて，新聞記事に含まれる性別や民族に関する偏見を時系列で分析し，特定の期間に偏見の増減が見られることなどを示した．土田ら[6)]は Twitter に投稿されたテキストに Word2vec を適用して，語彙の意味を表すベクトルを獲得した．ベクトルの演算によって，「東京の東京タワーと大阪の通天

図-1　対象区間と交差点番号

表-1　使用するプローブデータの概要

収集方式	パイオニア社のカーナビで収集
記録間隔	速度に関わらず3秒ごとに記録
データ項目	・車両ID　・走行距離　・緯度，経度 ・速度　・記録日時　・進行方位　など
収集期間	2015年4月1日~2015年12月31日の9か月間



(a) 東向き　　(b) 西向き

図-2　時間ごとの車両 ID 数とデータ数

閣の意味ベクトルは同等」のように，特定の地方のランドマークやイベントと対応する，ほかの地方のランドマーク・イベントが適切に特定できることを示した．さらにこの研究では Word2vec のパラメーター決定についても考察している．

また，Word2vec の特徴抽出特性を，自然言語以外のデータに活用した事例も見られる．Chung Park ら [7] は Word2vec を拡張したモデルである Road2vec をタクシーの移動経路データに適用して，道路リンクのベクトルを得た．獲得したベクトルのクラスタリングおよび演算によって，タクシー交通を代替するバス路線の最適化を行っている．名渡山ら[8] は，Word2vec を商品の購買履歴に適用して，各商品のベクトルを得た．商品の類似性を元にした独自の評価法を定めて，商品ベクトルを算出する際のパラメーター決定手法を定量的に示した．

椿ら [9] は，Word2vec と GloVe を用いて，タンパク質を構成するアミノ酸の配列よりタンパク質のベクトルを獲得した．その上で SVM を用いて，タンパク質のベクトルよりその構造を予測するタスクを行い，既存手法に比べて高精度な予測が可能なことを示した．

Word2vec を用いてデータ分析を行う研究は様々な分野でなされてきた．本研究ではプローブデータへの Word2vec の適用を行う．対象区間内の代表的な走行軌跡（以下走行軌跡）を特徴ベクトルとして抽出して，一般道路における走行軌跡の可視化を目指す．獲得した分散表現の考察により，対象とした道路における車両の走行パターンを明らかにするとともに，進行方向別・時間帯別に比較して，交差点の特性や信号が交通に与える影響を考察する．これらを踏まえて信号制御の評価や予測タスクへの導入可能性を検討する．

2.　対象区間と使用データ

(1)　分析対象区間

分析対象区間は，**図-1** に示す国道 2 号上に位置する広島市中区から南区にわたる 6 交差点とした．いずれも片側 3 車線以上で比較的交通量の多い信号交差点である．特に交差点 2，4，5，6 は従道路の交通量も非常に多い．

(2)　使用データ

本研究では，パイオニア社の提供する民間プローブデータを用いる．本データの概要を**表-1** に示す．本データの特徴として，GPS 衛星から受信する電波のドップラー効果により，瞬間速度を収集していること，3 秒ごとのデータ観測により，低速時のデータの収集性に優れることがあげられる．特に，本研究のように平均速度の低い一般道を対象とした分析は，ETC2.0 のように一定距離

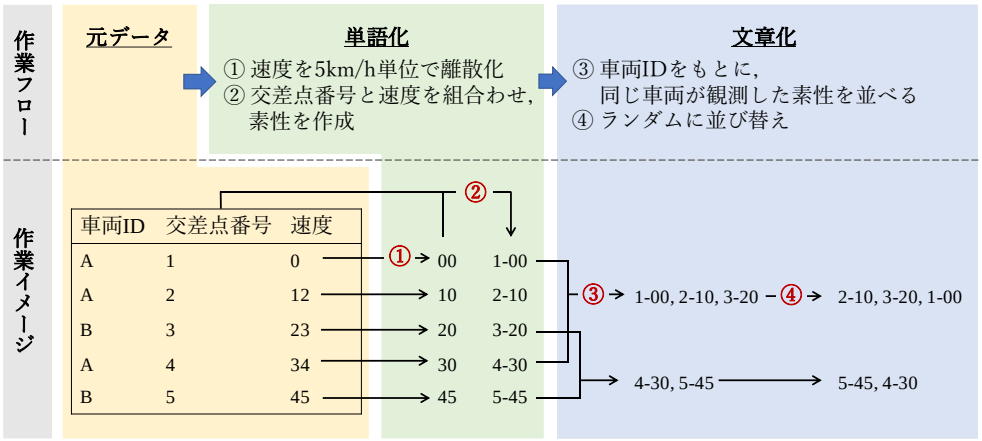
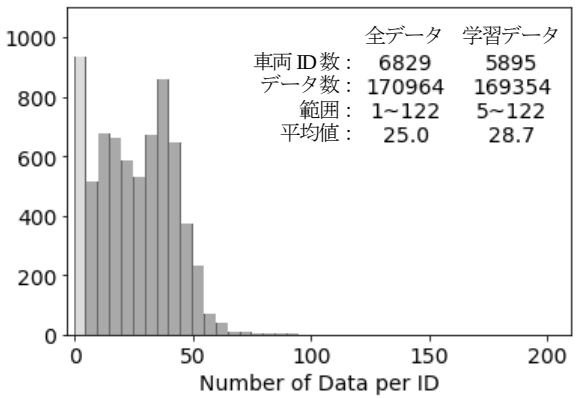
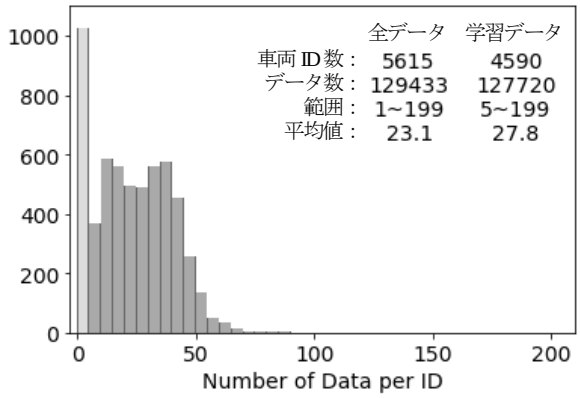


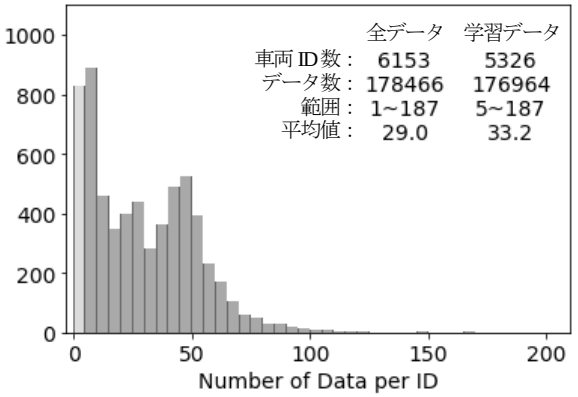
図-3　データセットの作成プロセス



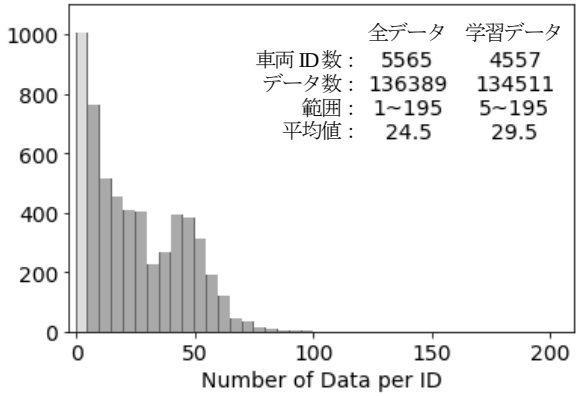
(a) 東向き・混雑時



(b) 東向き・非混雑時



(c) 西向き・混雑時



(d) 西向き・非混雑時

図-4　各データセットの ID あたりデータ数

間隔でデータを記録するとデータ密度が低くなりやすい．そのため，一定時間間隔でデータを収集する本データを用いた分析が有効といえる．

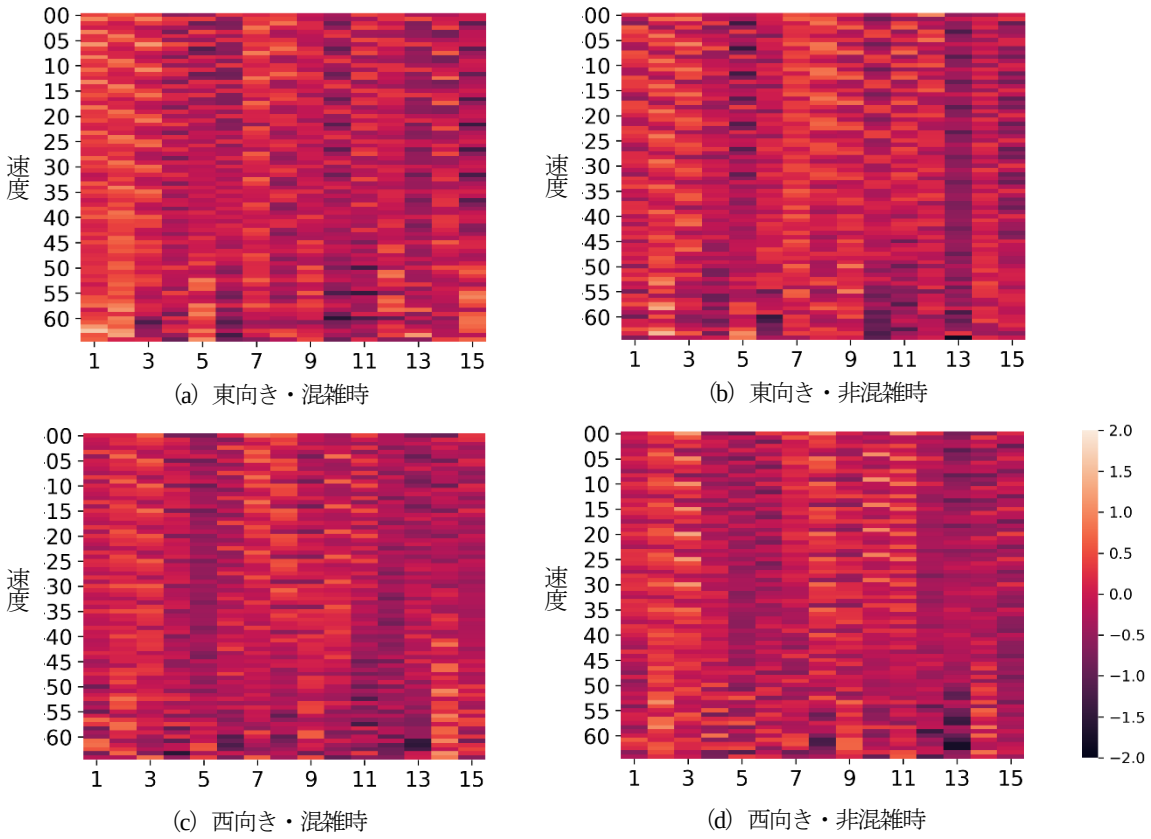
3. 方法

(1) Word2vec の概要

Word2vec は，Mikolov らによって提案された，ニューラルネットワークを利用した特徴抽出手法である．詳細は既往の文献 [10] に譲り，以下では Word2vec の概要を記す．

自然言語処理において，語彙の意味をとらえたベクトル（意味ベクトル）を獲得する手法は主要な課題である．最も単純な手法は，1 つの要素に 1 つの語彙をあてる方法で，one-hot ベクトルと呼ばれる．しかしこの方法では全語彙数に等しい要素数が必要である．つまり語彙が異なれば別の要素（次元）が必要となるため，意味の近さを表せない．意味ベクトルを獲得する手がかりとして，「語彙の意味はその周辺に出現した語彙によって決まる」という分布仮説が提案されている．

Word2vec は分布仮説に基づいた意味ベクトルの獲得手



(a) 東向き・混雑時　(b) 東向き・非混雑時

(c) 西向き・混雑時　(d) 西向き・非混雑時

図-5　獲得した分散表現

法である．すなわち，Word2vec とは，特定の単語w_iの周辺に現れる単語$w_{i\pm c}$を学習し，ベクトル化する手法である．このように学習した意味ベクトルは分散表現と呼ばれ，学習元語彙数よりも低次元（少ない要素数）で，各要素は実数で記述される．なおcはウィンドウサイズと呼ばれ，いくつ離れた単語まで学習するかを示すパラメーターである．

Word2vec には，CBoW と Skip-gram と呼ばれる 2 つのアルゴリズムがあり，一般的に精度が良いとされる Skip-gram を用いた．本研究では，Python のライブラリである gensim を実装して，分散表現の学習を行う．

(2)　データセットの作成とその概要

パイオニアプローブデータから，国道 2 号上で観測されており，かつ交差点中心から 100m 以内の位置情報を抽出する．平日と休日・祝日で走行軌跡が異なると考えられるため，抽出したデータから休日および祝日の観測分は削除して，初期データセットを作成した．初期データセットの時間帯別の ID 数およびデータ数を**図-2** に示す．以下特に混乱が無い限り，観測データ数を交通量と呼ぶ．6～8 時台および 16～19 時台の通勤・通学の時間帯は特に交通量が多い一方で，深夜の交通量は日中に比べて極端に小さい．

前述のとおり Word2vec は自然言語に適用される．そのため，抽出したデータを整形して自然言語を模した入力セットを作成する．そのプロセスを**図-3** にまとめる．単語化は，交差点番号と離散化した速度を，“交差点番号-離散化した速度”のように組み合わせることである．例えば“1-00”は，交差点 1 で観測された時速 0~4km のプローブデータを示す．以上の手順で単語化したプローブデータを，以下素性と呼ぶ．

文章化とは，作成した素性を並べることでそれらの共起セットを作成することである．走行軌跡をとらえた分散表現を得るための工夫として，対象区間を通過する 1 車両を表す素性を並べて 1 つの文章とみなした．なお以下では車両を ID と呼ぶ．分散表現は，共起しやすい素性ほど類似するように学習される．ここで共起とは，2 素性以上の素性を同一 ID が観測することをいう．また素性を時系列に並べた場合は，ID が観測した順に素性が並ぶため，位置が離れた交差点ほど共起をとらえにくくなる．本分析では ID 単位の草稿履歴に基づく面的な素性（交差点別・地点別）の特徴を捉えることが目的のため，ID 別の素性の並びはランダムとして，離れた地点の素性の共起が捉えられるように工夫した．

なお，進行方向によって走行特性は異なると予想されるため，進行方向（東向き，西向き）別にデータセットを作成したほか，時間帯によって異なる素性を獲得するため，時間帯でもデータセットを分割した．時間帯は，**図-2** を考慮して混雑時（6～8，17～19 時）および非混雑時（9～16 時）とし，渋滞がほとんど発生せず，交通制

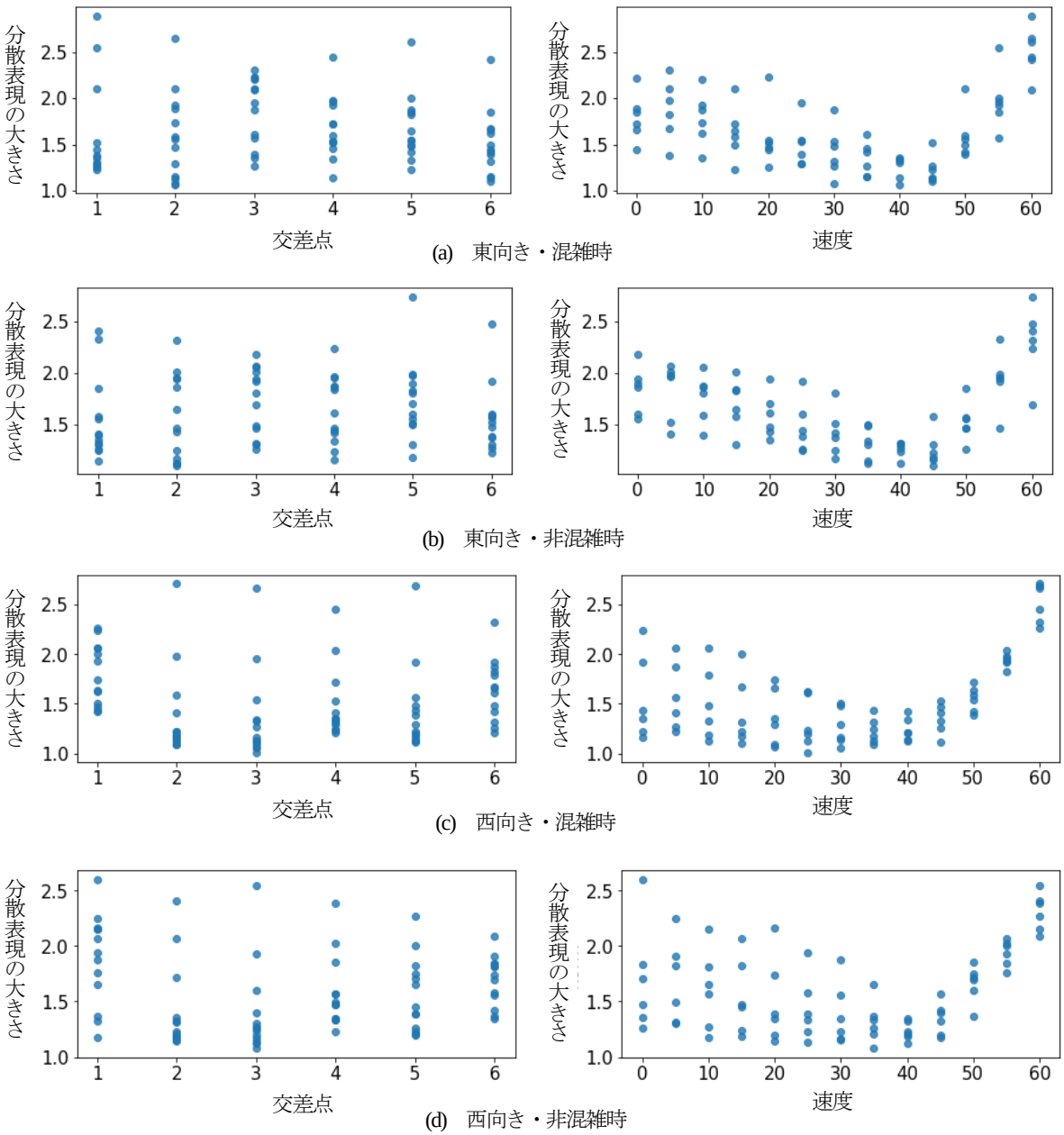


図-6 分散表現の大きさと，交差点および速度の散布図

御の必要性が低い夜間は，学習対象から除いた．

以上の手順により進行方向・時間帯の組み合わせにより 4 つのデータセットを作成した．作成したデータセットの，IDあたりデータ数を**図-4**に示す．同図より，観測データ数が 4 以下の ID が最も多い．しかし，Word2vec の特性上，ID あたりデータ数が少ない場合，うまく分散表現を学習できない可能性がある．そこで，観測数 4 以下の ID は学習対象から除外した．

(3) Word2vec の適用による分散表現の獲得

Word2vec をデータセットに適用するにあたり，パラメーターの検討を行う．Word2vec の主要なパラメーターには，アルゴリズム，ウィンドウサイズおよび分散表現の次元数の 3 つがある．データセット間の比較を可能とするため，これらは全データセットで同一とする．アルゴリズムは前述の通り Skip-gram を用いて，ウィンドウサイズは 10 と決定した．なお本研究ではデータセットを作成する際に素性の並びをランダムにしているため，ウィンドウサイズは重要ではない．また，分散表現の次元数は，学習時の損失値と学習の安定性を考慮して 15 とした．

分散表現の直接観察による考察が困難な場合は，類似度が用いられることが多い．本研究でも素性w_i，w_j間の類似度として，式(1)に示すコサイン類似度を用いる．

$$cos(w_i, w_j) = \frac{w_i \cdot w_j}{|w_i||w_j|} \quad (1)$$

コサイン類似度の特性を示す．コサイン類似度はベクト

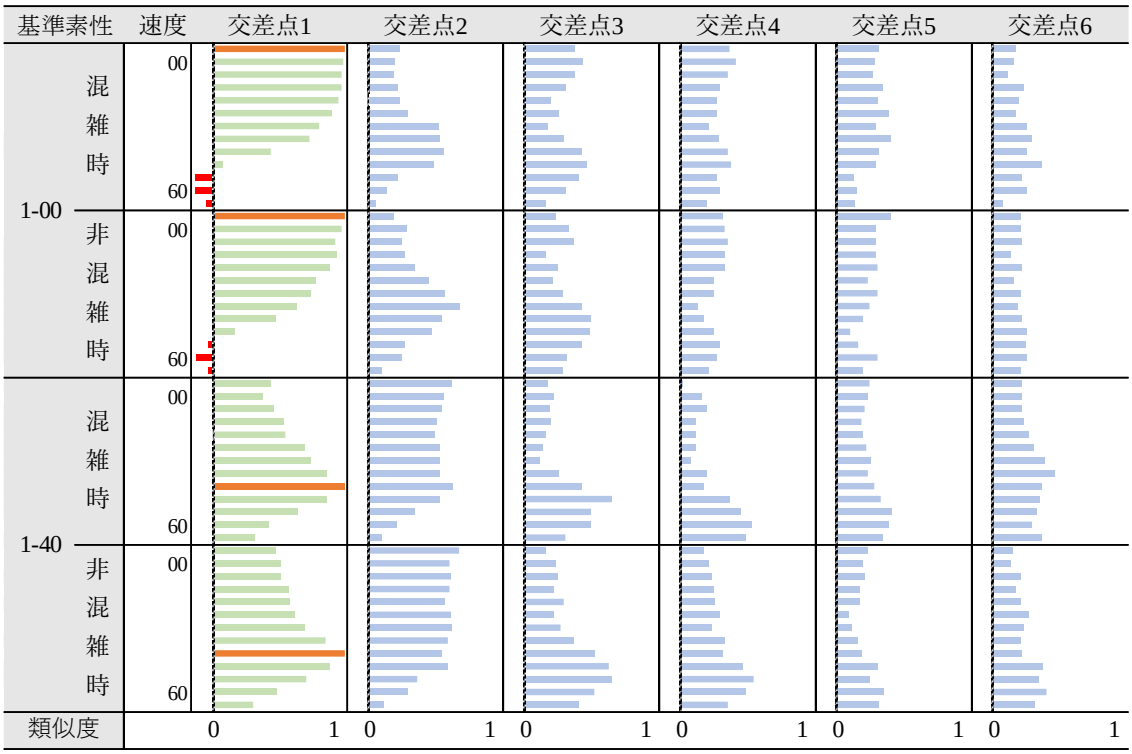


図-7　東向きの類似度分布図

ルw_iとw_jの余弦で，-1 から 1 の値域をとる．コサイン類似度が大きいとき，2 ベクトル間の角度は小さく，類似している．

一方でベクトルの大きさは，一般にベクトルの L2 ノルムで計測される．分散表現の大きさに関連する文献は少ないが， Kim ら [11] は，分散表現の加法構成性を基に，分散表現のノルムは語意の多様性を示すとした．以下では，L2 ノルムと素性間類似度に着目して，分散表現が表す内容について考察する．

4.　結果と考察

(1)　獲得した分散表現の可視化

以上の手順でプローブデータを Word2vec に適用し，分散表現を得た．その結果を**図-5** に示す．**図-5** の横軸は分散表現の次元を示し，縦軸は素性の速度を示している．全素性を速度順に，同じ速度の素性は交差点番号順に並べた．図中の色は凡例に示した各次元の値をあらわす．東向き・混雑時の第 15 次元のように，一部に速度との関係性が予想される次元がある．しかし，**図-5** に基づいて，分散表現の特徴を直接考察するのは難しい．そこで以下では，（1）分散表現の大きさ（L2 ノルム），および（2）素性間の類似度の 2 指標に基づいて，獲得した分散表現を考察する．

(2)　分散表現の大きさ

図-6 は，縦軸に分散表現の大きさ，横軸を交差点番号および速度とした散布図である．進行方向や時間帯にかかわらず，交差点と分散表現の大きさの関係はほとんど見られない．一方で速度との関係性は強く，どの図でも時速 40km で分散表現の大きさが最も小さくなる．また，時速 40km を基準に，速度が大きい側ではより急激に分散表現が大きくなる．

Kim らの解釈に基づくと，プローブデータの交差点別速度カテゴリを素性とするデータセットに Word2vec を適用して得られる分散表現は，共起する素性の多様性を表わすと考えられる．対象区間の制限時速は 50km だが，比較的交通量の多い区間なので自由走行が難しく，実際には 40km 程度が巡航速度になっていると考えられる．また，面的に系統制御されている信号群のスルーバンド設定速度も，時速 40km 程度と思われる．図-6 より，対象区間を 40km よりも高速，または低速で通過する車両の走行軌跡には多くのパターンが現れるため，40km 以外の素性は，40km よりも走行軌跡が複雑になると思われる．なお以上の傾向は，素性においても何らかの加法構成性が成立する可能性を示唆しているがその検証は困難なため，以下では行わない．

(3)　素性間の類似度

本研究では同じ ID が観測した素性を並べて文章化することで分散表現を獲得した．これにより共起しやすい

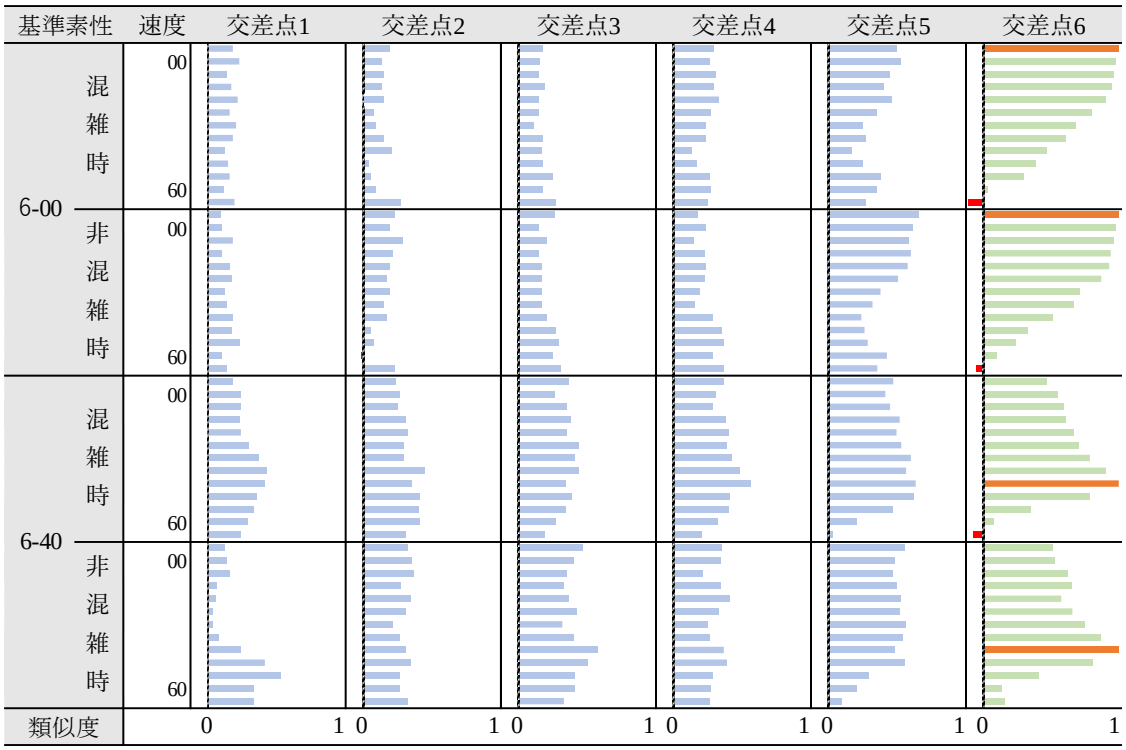


図-8　西向きの類似度分布図

素性の組み合わせは，データセット内で頻出する車両の走行軌跡を表すと考えられる．本節ではこの特徴を踏まえて，走行軌跡の可視化を試みる．

獲得される走行軌跡の特徴は，例えば，交差点 1 の信号で停車した車両は，他の交差点では停車するか，無停車なら速度はどの程度か，などである．この走行軌跡は，交差点間における素性の共起のしやすさによってとらえられる．そこで，素性の共起のしやすさを表す類似度を用いて，以下の手順で走行軌跡を把握する．進行方向別に最初に通過する交差点において，停車（時速 00km）もしくは巡航（時速 40km）を示す素性に着目する．次に着目した素性との類似度を，全素性で算出する．それらを交差点別・速度順に並べた，交差点別類似度分布図を作成する． ここで巡航を 40km としたのは，前節より，分散表現の大きさより時速 40km 付近が巡航速度と示唆されたためである．

図-7，**図-8** の縦軸は速度を，横軸は類似度を表している．また，オレンジで示す交差点・速度の組み合わせは基準素性をあらわしており，定義より基準素性自身との類似度は常に 1 になる．緑は基準素性と同じ交差点の類似度，青は異なる交差点の類似度の大きさを表す．各交差点の素性は，速度の小さい素性から順に昇順で並べた．同図において，青で示す基準素性とは異なる交差点の類似度が，速度によって異なるとき（つまり類似度分布に山が見られる場合）は，その交差点の通過速度は，基準素性の交差点での速度と共起する，つまり典型的な走行軌跡を表すと考えられる．一方で類似度分布が平坦で速度による差が見られない場合は，基準素性とその交差点での通過速度の関係がみられないことをあらわす．以下簡単のため，類似度分布の形状に着目する．類似度に閾値を設定して考察する手法も考えられるが，離れた交差点ほど自然に走行軌跡が多様になると予想されるため，この点を踏まえた走行軌跡抽出は，今後の課題とする．

図-7 は東向きの結果を示す．停車時（1-00）を基準素性とした場合，混雑時・非混雑時ともに交差点 2，3 では中速域で類似度が大きい．つまり，交差点 1 で停車する車両は，交差点 2，3 を無停車で通過することが多い．また，交差点 3 の低速域の類似度が大きく，時間帯別に比較すると混雑時の方がわずかに大きい．これは交差点 1 で停車した車両は交差点 3 で速度を落とすことを示している．この要因には，信号待ちや右左折による速度低下などが考えられ，その影響は混雑時の方が大きい．そのほかの時間帯による差はあまり見られない．次に巡航時（1-40）を基準素性とした場合は，交差点 2 の類似度は低～中速域では平坦で明確な傾向が見られない．しかし，交差点 3，4 では高速域の素性と類似度が高くなっている．つまり，交差点 1 を無停車で通行した場合，交差点 2 では速度を落とすか，そのままの速度で通過して，交差点 3，4 では速度を少し上げたまま無停車で通過することが多いといえる．

以上より，東向きに進行する車両において，混雑時と非混雑時の走行軌跡の差はあまり見られなかった．特に

巡航速度を基準素性とした場合に，交差点 3 から 4 にかけて高速域との類似度が高い結果となった．信号の系統制御により，交差点 3 から 4 は停車せずに通過できることが多いと考えられる．

図-8 は西向きの結果である．交差点 6 の停車時を基準素性としたとき，交差点 5 の低速域の素性も類似度が高いことから，交差点 6 で停車する車両は交差点 5 でも停車しやすい．また，時間帯別で比較すると非混雑時の類似度が少し大きい．交差点 4 以降の交差点では特定の傾向は見られず，交差点 4 以降の走行軌跡は交差点 6 での速度とのかかわりが小さい．巡航時を基準素性とすると，交差点 5 で低～中速域の素性の類似度が高い．交差点 4 以降の交差点では，明確な傾向が見られない．交差点 6 を時速 40km で通過すると，交差点 5 では速度を落とすかそのままの速度で通過することがわかる．また，非混雑時には交差点 1 で高速域との共起が見られるが，この要因の考察は同図からは難しい．

以上より，西向きに進行する車両も，混雑時と非混雑時の走行軌跡に大きな差は見られなかった．交差点 5 における共起は見られるが，それ以外の交差点との共起傾向はほとんど見られない．東向きと比べて走行軌跡は明確ではなく，信号制御により東向きに進む車両を優先的に通過させている可能性がある．

5. おわりに

本研究では，プローブデータを Word2vec に適用し，対象区間における走行軌跡を分散表現として学習した．学習にあたって，同一車両の観測した素性を並べて 1 文章とする工夫を加えた．獲得した分散表現の大きさと速度の間には関連性が見られ，プローブデータの分散表現に何らかの加法構成性の存在が示唆される．また，対象区間における走行軌跡を可視化したところ，走行軌跡の獲得ができていたことがわかった．進行方向別に走行軌跡を分析した結果，各進行方向の特徴を捉えることができた．しかし時間帯別による差はほとんど見られなかった．交通量が増加した際も，効果的な信号制御によって，スムーズに走行ができることが要因として考えられる．

一方で課題として，交差点における車両の右左折を考慮できていない点があげられる．対象区間を直進する車両の挙動を考察したが，作成したデータセットは対象区間内の交差点で右左折する車両も多く含む．今回用いたプローブデータでは直進車のデータ数が少ないため分散表現の学習が難しいものの，直進車のみを抽出してデータセットを作成することで，より明確に直進車の走行軌跡を把握できると思われる．

本研究で示した手法は，交差点番号ではなくリンク番号を位置情報として用いるなどの拡張ができる．交通量予測などの深層学習を用いたタスクの特徴量として，走行軌跡をとらえた分散表現を取り入れることで，更なる予測精度の向上が期待できる．

参考文献

1) 岡谷 貴之：機械学習プロフェッショナルシリーズ 深層学習，pp.v-vi，講談社，2015.
2) 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.
3) 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.
4) 佐藤季久恵，高屋英知，小川亮，芦原佑太，栗原聡：Deep Q-Network を用いた交通信号制御システムの提案，第 31 回人工知能学会全国大会，2017.
5) Nikhil Garg，Londa Schiebinger，Dan Jurafsky，James Zou：Word embeddings quantify 100 years of gender and ethnic stereotypes，Proceeding of the National Academy of Sciences，vol.115，No.16，E3635-3644，2018.
6) 土田 崇仁，遠藤 雅樹，加藤 大受，江原 遥，廣田 雅春，横山 昌平，石川 博：Word2Vec を用いた地域やランドマークの意味演算，DEIM Forum，2016.
7) Chung Park, Jungpyo Lee, So Young Sohn : Recommendation of feeder bus routes using neural network embedding-based optimization, Transportation Research Part A, 126, pp.329-341, 2019.
8) 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.
9) 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.
10) Tomas Mikolov，Kai Chen，Greg Corrado，Jeffrey Dean：Efficient estimation of word representations in vector space，Proceedings of Workshop at International Conference on Learning Representations，2013.
11) Geewood Kim，横井 洋，下平 英寿：単語埋め込みの二種類の加法構成性，言語処理学会 第 26 回年次大会 発表論文集，pp.724-727，2020.

(????. ?. ? 受付)