The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

密度球を用いた GraphCNN 深層学習手法による渋滞予測

Congestion prediction using GraphCNN depth learning method using density Sphere

高橋 慧[*1]， 坂本 克好[*1]， 山口 浩一[*1]， 沼尻 匠[*2]， 曽我部 完[*2] 曽我部 東馬[*1*2*3]

[*1] 電気通信大学 大学院 情報理工学研究科 [*2] 株式会社 GRID
[*3] 電気通信大学 i-パワードエネルギー・システム研究センター

In this paper, we study the data clustering in a high dimensional space based on density spheres for traffic data sets with many samples and features, and predict traffic congestion by creating a distance matrix from features with Density Sphere GraphCNN. Density spheres represent the density which serves as a reference for clustering data in a high dimensional space, and it is possible to investigate the relationship of data by considering both data correlation and distance. A mechanism to realize highly accurate congestion prediction will be studied based on the result of predicting the degree of congestion by combining traffic simulation model, which reproduces congestion and compares the prediction accuracy by varying the volume of density balls

1. はじめに

近年、日本の道路交通システムは複雑な道路網や交通需要の増大、都市部への人口の過密化によって、慢性的に渋滞が発生している。そのため最近、交通需要の調整に人工知能を用いて渋滞を解消しようとする試みが行われている。スマートフォンの普及や車両の状態や道路状況などの様々なデータを取得することの出来るコネクテッドカーの登場などにより、車や人間の動きをリアルタイムで捉えられるようになってきた。そうしてリアルタイムの環境や道路のデータを収集することで、機械学習による渋滞予測が可能となった。2017 年には、NEXCO 東日本と NTT ドコモにより東京湾アクアラインにおける AI 渋滞予知が行われ、従来より高い精度の予知が達成されている[*1]。今後、IoT により様々なデバイスから取得したデータが増えれば天候や路面、車両ごとの測定が可能となり、より高精度な渋滞予測が可能になると見込まれる。

そこで本研究では、多くのサンプルと特徴量を持つ交通データセットに対して、高次元空間におけるデータの密度球(Density Sphere)に基づいたクラスタリングを行い、特徴量から距離行列を作成することで畳み込みを行う Density Sphere GraphCNN を用いて渋滞を予測する。密度球とは、高次元空間におけるデータのクラスタリングの基準となる密度を表現したものであり、データの相関と距離を両方考慮してデータの関係性を見ることが出来る。渋滞を再現した交通シミュレーションモデルと GraphCNN を組み合わせることで渋滞の度合いを予測、また密度球の違いによる予測精度の比較し、その結果に基づいて高精度な渋滞予測を実現するメカニズムを検討する。

2. Density Sphere GraphCNN

2.1 GraphCNN

CNN [Lecun 89] は古典的な多層パーセプトロンの延長にあるが、画像の局所的な特徴抽出を行う畳み込み層と、局所ごとに特徴をまとめるプーリング層を繰り返した構造になっている。従来のニューラルネットワークは中間層を増やすことで表現力が増し、表現できる関数や分類できる対象も増えるが、実際には過学習や勾配消失の問題から層を増やすことが容易ではない。一方、CNN は畳み込み層とプーリング層を交互に繰り返すことで層を増やし、ネットワークを深層化することが可能である。そのため CNN はディープラーニングの主力とされ、これまで画像処理、音声認識、コンピュータビジョン、言語処理など多くの分野で成功を収めてきた。しかし、CNN はそうした高い性能を持つ一方で、csv ファイルのような行と列の概念がある非構造化データに対して、適用することは困難とされてきた。

そこで最近、特徴行列の相関行列を用いて非画像データに対して畳み込みを可能とする GraphCNN[Yotam 17] という手法が報告された。GraphCNN は非画像データに対しても、CNN の長所である特徴抽出を発揮できる深層ニューラルネットワークである。先行研究ではデータマイニングコンペティションサイト Kaggle で扱われていた 2153 個の特徴と 6148 個のサンプルデータを持つ“Merck Molecular Activity Challenge” [*2] の分析を行い、DNN と RandomForest を用いた当時の Kaggle コンテストの優勝者のよりも優れた性能を発揮している。

2.2 既存手法の問題点

先行研究の GraphCNN では、特徴量の相関行列を用いて非画像データに対する畳み込みを行っていたが、相関係数にはデータの関係性を表すのに適切ではない場合がある [高橋ら 17]。 図 1 のような 3 つのデータ点がある場合を考える。

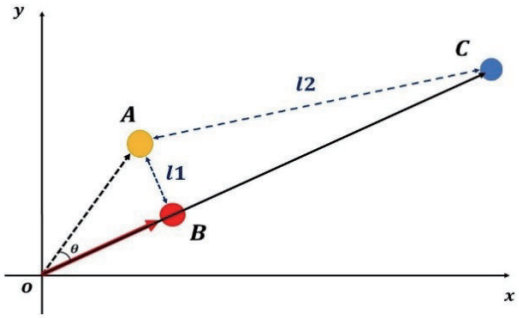

図 1 相関と距離によるクラスタリングの違い

相関係数 r は 2 つのベクトルのなす角度 θ の余弦で表されるため、$\overrightarrow{OA}$ と $\overrightarrow{OB}$ の相関係数は cosθ となり、同様に $\overrightarrow{OA}$ と $\overrightarrow{OC}$ の相

連絡先：曽我部 東馬，電気通信大学 i-パワードエネルギー・システム研究センター， sogabe@uec.ac.jp

*1 https://www.nttdocomo.co.jp/info/news_release/2017/11/30_01.html

*2 https://www.kaggle.com/c/MerckActivity

関係数も cosθ となる。つまり、相関係数を基準に考えたとき、点 A,B,C は同じクラスタに分類される。一方、データ間の距離を基準に考えたとき、点 C は点 A,B から離れているため、一般的に点 A,B と同じクラスタと見なすことは出来ない。

また、図 2 のようなデータ群があったとき、点 A と点 B は同じクラスタであると考えられる。しかし、データ間の距離のみを基準に考えたとき、点 A,B の距離 L1 は、点 A,C の距離 L2 よりも離れているため、点 A と点 B が同じクラスタの場合、点 A と点 C も同じクラスタだと見なされる場合がある。

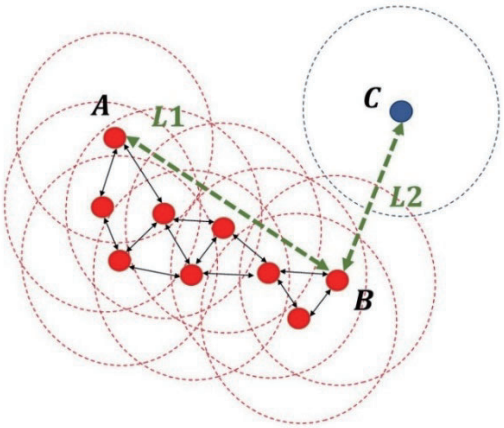

図 2 相関と距離によるクラスタリングの違い

相関係数やデータ間の距離はこうした特性があるため、それぞれを単体で用いると、データの関係性を表するのに最適とは言えない。データの相関と距離の両方を考慮に入れて、関係性を図るのが最善であり、密度球を用いることでそれら両方を考慮することが出来る。

2.3 密度球 (Density Sphere)

高次元の特徴空間において、多くの点が近接しているような領域を高密度領域といい、その領域に属する点は同じクラスタに分類されることが多い。密度球は、その特徴空間におけるデータのクラスタリングの基準となる空間密度を表したものである。図 3 に密度球のイメージモデルを示す。

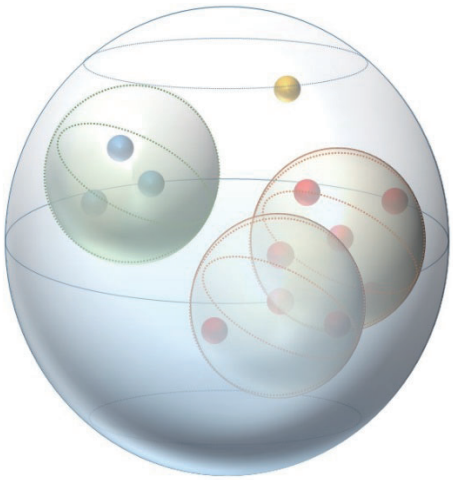

図 3 密度球のイメージモデル

あるデータ点を中心に距離 r 以内の空間に Q 個以上のデータが存在するとき、密度球が定義され、密度球がお互いに隣接するデータを同じクラスタとして扱う。距離 r 以内の空間に Q 個以上のデータが存在しない場合、密度球が定義されず、その点はどのクラスタにも属さない外れ点となる。このように密度球によるクラスタリングを行ったのちに、同クラスタの点同士の距離を求める。同じクラスタに存在するデータとの距離を取ることで、データの相関と距離を両方考慮した距離行列を作成することが出来る。図 2 の場合においても、初めにデータの密度を考慮したクラスタリングを行うことで、点 A,B と点 C が違うクラスタであると認識し、外れ点である青点を除いた点同士の距離を取ることで、データの関係性をより説明することが出来る。

3. データセットの構築と分析

3.1 交通シミュレーションモデル

本研究では、WITNESS™ という汎用シミュレーションソフトウェアを用いて、高速道路で渋滞が発生する要因を再現した簡易的なモデルを作成した。そのモデルを図 4 に示す。

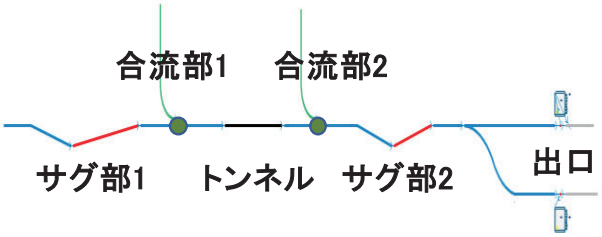

図 4 高速道路のモデル

高速道路で渋滞が発生する要因は主に 4 つあり、サグ部と呼ばれる下り坂から上り坂にさしかかる凸部、トンネル、合流部、そして出口である。渋滞の予測には通常、ある時刻にどれくらいの車が特定のエリア内にいたかという統計データと、そしてどの程度の渋滞が発生したかという道路状況を示すデータが必要になる。今回の実験では、モデルに常時数十台の車を走らせて各エリアごとの車の台数、および車ごとの走行時間を出力し、5510 サンプル 8 つの特徴量を持つデータセットを作成し、学習に用いた。

3.2 データセットの関係性

作成したデータセットの特徴量がどのような関係性を持っているかを調べる。図 5 にデータセットの関係性を表す。左上が特徴量同士の相関行列、右上がガウシアンカーネルにより求められたデータ間の距離行列、そして左下が密度球において Q=2 のときの距離行列、右下が Q=4 をにしたときの距離行列である。ガウシアンカーネルは以下のように表される式であり、データ間の距離が近いほどに値は 1 に近づく。

$$\mathrm{k}(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right)$$

マスの色が白に近いほど特徴量はお互いの変動をよく表している。図 5 を見ると、同じデータから生成したにもかかわらず、どの行列も異なる関係性を表していることが分かる。このように用い

る手法によって、データ間の重要度が変わってくるため、学習の際に真にデータの関係性を表している行列を選択することが重要になる。

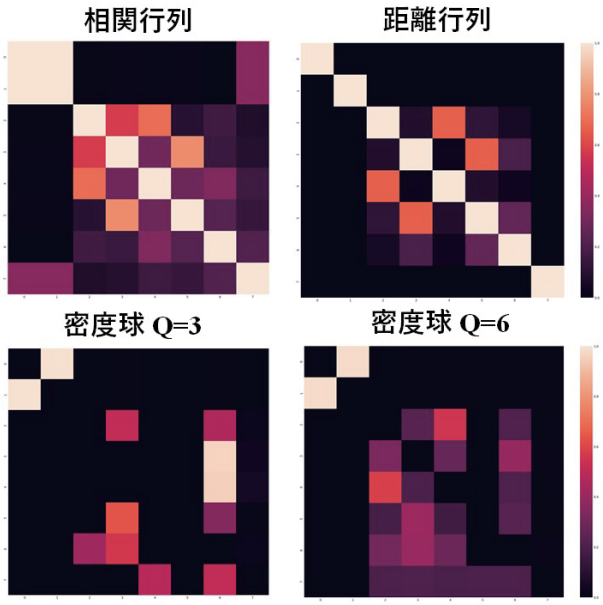

図 5 データセットの関係性

4. 実験および評価

実験には、畳み込みを 2 回繰り返す全 7 層の構造を持つ Density Sphere GraphCNN(D-GCNN)を用いた。また比較のため相関行列を利用した Correlation-GraphCNN(C-GCNN), 距離行列を利用した Kernel-GraphCNN(K-GCNN), 他に回帰予測において一般的に用いられている Neural Network, RandomForest, XGBoost, LightGBM でも実験を行った。

4.1 走行時間の予測

前述の 7 個の手法を用いて、車の走行時間(s)の予測を行った。分析の評価には RMSE を用いた。RMSE は

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=0}^{N}\left(f_i - y_i\right)^2}$$

で定義される関数であり、Y は実際の走行時間、$\hat{Y}$は予測された走行時間を表し、RMSE が 0 に近いほど予測精度が高い。3 つの GraphCNN の構成において、畳み込み 1 層目は 15 枚、2 層目は 20 枚のフィルタ、全結合層のユニット数は 64 個、反復回数は 300 回というパラメータ設定を共通に用いた。

実験結果を図 6 に示す。GraphCNN を用いた 3 つの構成の精度は、従来のニューラルネットワークの精度を大きく上回るということが分かる。GraphCNN は他の手法よりも精度が高く、さらに D-GCNN は最も精度が良いことが分かる。次に、密度球の密度の違いによる予測精度の差を図 7 に示す。図 7 より、Q=3 の時が最も良い精度を持つことが分かった。クラスタリングの際の空間密度は予測精度に影響を与え、密すぎても疎すぎても良くないことが分かる。

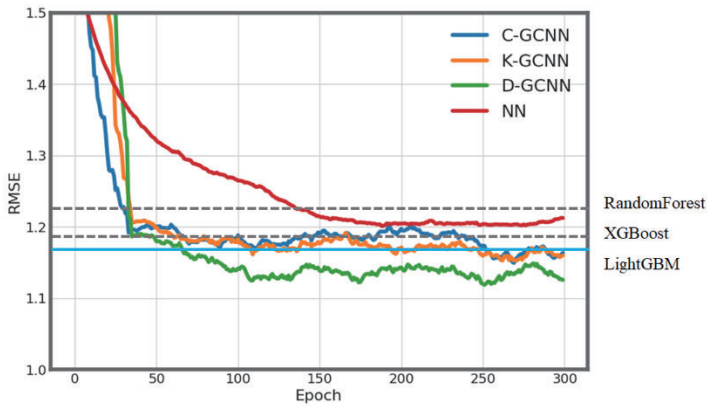

図 6 走行時間の予測における各手法の精度

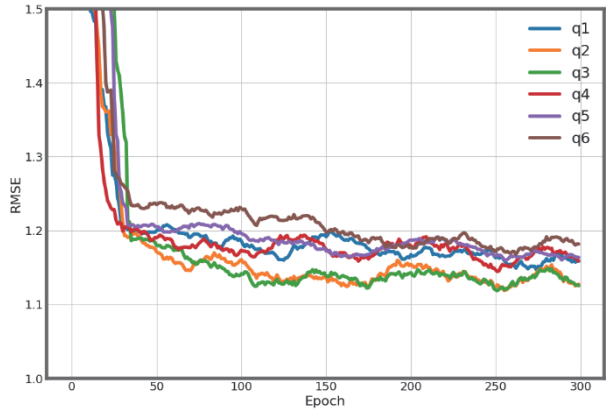

図 7 密度球における Q の値ごとの精度

4.2 渋滞の分類

次に同じデータにおいて、走行時間を「渋滞なし」「やや渋滞」「渋滞」の 3 クラスに分類し、渋滞の度合いを判断する分類問題を行った。その結果を表1に示す。表 1 より D-GCNN がわずかに他の手法の精度を上回っていることが分かった。

表1 各手法における渋滞分類の正答率

Method	正答率 (%)
NN	71.0
RandomForest	68.4
XGBoost	73.0
LightGBM	72.9
Correlation GraphCNN	72.6
Kernel GraphCNN	72.8
Density Sphere GraphCNN	73.3

次に、密度球の密度の違いによる分類精度を表 2 に示す。予測と同様に Q=3 の時が最も精度が良く、Q>5 以上の時に分類精度が低下した。このデータセットにおいて Q>5 の場合、密度球によるクラスタリングが細かく行われすぎたと考えられる。

表 2 密度球における Q の値ごとの正解率

	正答率 (%)
Q = 1	72.7
Q = 2	72.9
Q = 3	73.3
Q = 4	70.7
Q = 5	65.1
Q = 6	65.1

5. まとめ

本研究では、交通データに対する CNN の応用手法として GraphCNN に注目し、その手法の改善案として密度球を用いた Density Sphere GraphCNN を提案した。提案手法の有効性を検証するため、高速道路の渋滞を再現したミュレーションモデルから生成された交通データセットを用いて予測、分類問題の実験を行った。その結果、提案手法である Density Sphere GraphCNN は、他の GraphCNN の予測精度を上回り、密度球の有用性が証明された。また、密度球の密度条件は予測結果に影響を与える、それぞれの課題に合った値を選択する必要があると分かった。今後、データセットの特徴量が増えて関係性が複雑になったとしても、密度球を用いることで予測に必要な特徴量のみを抽出し、より高精度の予測を行えることが期待される。

参考文献

[Lecun 89] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L.D. Jackel, “Backpropagation applied to handwritten zip code recognition”, Neural Computation, vol.1, pp.541-551, 1989.

[Yotam 17] Yotam Hechtlinger, Purvasha Chakravarti, Jining Qin: A Generalization of Convolutional Neural Networks to Graph-Structured Data, arXiv preprint, arXiv:1704.08165 (2017)

[高橋ら 17] 高橋慧、沼尻匠、曽我部完、坂本克好、山口浩一、横川慎二、曽我部東馬、特徴グラフを用いた汎用型 CNN 深層学習手法の開発, 2018 年度人工知能学会全国大会(第 32 回)論文集(2018)