2B-01

稀なイベントの検出とデータ拡張による予測

小島 亮一 †
株式会社 KDDI 総合研究所 †

南川 敦宣 †
株式会社 KDDI 総合研究所 †

1 はじめに

IoT 機器と回線の普及でリアルタイムに多種多様な時系列データ収集ができるようになった一方、大量の学習データを前提とした機械学習モデルは稀なイベントが起きるとその予測精度が低下してしまうことが知られている。本稿では稀なイベント発生時にもその少量データを時系列的な複数解像度による特徴を考慮して拡張することで迅速に再学習しリアルタイム予測可能な機械学習ワークフローを提案する。そして稀なイベントの一例として大雪時の 5 分後断面交通量予測実験を行い、提案ワークフローが平時交通量との違いを速やかに検出した上で予測精度も維持できることを確認した。

2 データセットと実験設定

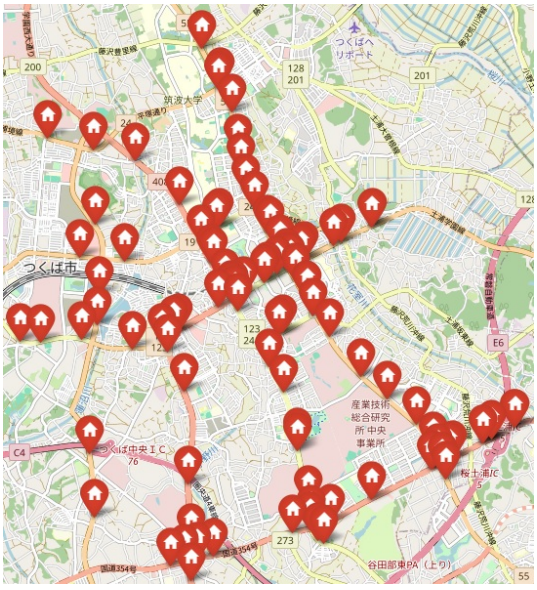

図 1 つくば市周辺の断面交通量計測地点

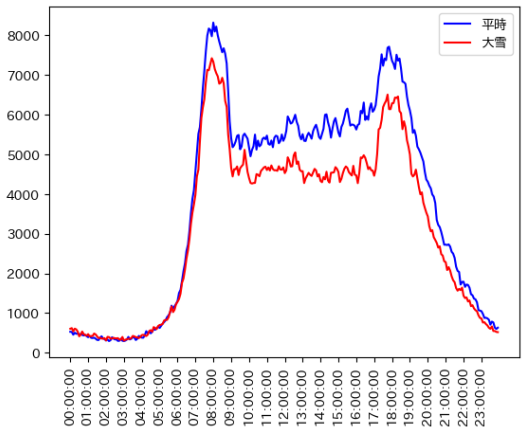

図 2 平時と大雪時の断面交通量

提案ワークフローの適用対象をつくば市周辺 (図 1) の 5 分後断面交通量合算値予測とする。断面交通量は道路上の車両感知器で 5 分おきに計測されたカウントデータであり、日本道路交通情報センター (JARTIC) により公開されている [1]。本稿では 97 箇所ある車両感知器ごとではなくそれらの断面交通量の合算値を用いる。実験対象期間は 2022/1/1〜2/10 で、2/10 の大雪を稀なイベントとする。平時と大雪時 (2/10) の 1 時間ごと断面交通量は図 2 のようであり、大雪時には日中の断面交通量が大きく減少していることが見て取れる。本稿で提案するワークフローの特徴である稀なイベント検出とその稀なイベントのデータ拡張による迅速な再学習の効果を以下の実験設計で評価する。

1. 平時の期間 2022/1/1〜2/9 で 5 分後断面交通量予測モデルを学習する。
2. 2/10 の大雪を稀なイベントとし、5 分後断面交通量予測を 1. で学習しただけの場合と提案ワークフローで再学習した場合とで精度評価する。
3. 提案ワークフローが稀なイベントを検出するまでのラグを評価する。
4. データ拡張と再学習にかかる時間を評価する。

Detection of rare events and prediction with data augmentation
†Ryoichi Kojima, Atsunori Minamikawa, KDDI Research, Inc.

3 提案ワークフロー

提案ワークフローは、(A)(B)(C)(D) の 4 つのモジュールよりなる。以下で各モジュールについて詳説する。

(A) 5 分後断面交通量予測モジュール
U-Net(1D CNN)[2][3] により時系列的複数解像度 (週 (2016 コマ)、日 (288 コマ)、時 (12 コマ) ごと) で畳み込んで 5 分後断面交通量予測モデルを学習する。ここで U-Net は画像セグメンテーション向けに考案された深層学習モデルであるが、入力と出力を同じとするため AutoEncoder などと同様に潜在表現抽出に有益であり時系列データを 1D として複数解像度別で畳み込む実装 [4] を使う。

(B) 平時に学習した潜在表現保存モジュール
(A) で学習した時系列的複数解像度別の平時潜在表現を保存する。

(C) 稀なイベント検出モジュール
運用時の時系列的複数解像度別のリアルタイム潜在表現が、(B) で保存した平時の時系列的複数解像度別の潜在表現と標準偏差 1 つ分以上乖離したら稀なイベントが起きたとしてその時系列的解像度 (週 or 日 or 時) を検出する。

(D) データ拡張モジュール
(C) で検出した時系列的解像度箇所のみデータ拡張し、(A) で高速に再学習する。

4 実験結果

平時データで学習しただけの U-Net(1D CNN) と提案ワークフローで再学習した U-Net(1D CNN) の 5 分後断面交通量の予測誤差 MAE を 7 時〜9 時、9 時〜17 時、17 時〜24 時の各レンジ内で平均した結果を表 1 に示す。ここで提案ワークフローにおける再学習によって 7 時〜9 時においては約 3 倍、9 時〜17 時においては約 4 倍高精度な予測ができていることが確認できた。そして前章の (C) が標準偏差 1 つ分以上の乖離を検出したのは 7 時 55 分であり、図 2 によれば断面交通量に差が出始めるのは 7 時頃なので稀なイベント検出までのラグは 55 分ほどであるといえる。さらに、(D) によるデータ拡張にかかる時間と (A) による再学習時間の合計は約 12 分であった。つまり、提案ワークフローは大雪によって 7 時頃に断面交通量に差が出始めてもその約 55 分後に (C) が稀なイベントと検出し、そこから約 12 分後に (D) によるデータ拡張と (A) による再学習を完了して予測、とほぼリアルタイムに追随した対応ができたことを確認した。

なお、本実験を行ったマシンのスペックは CPU:Intel Core i7-9800X 3.80GHz, Memory:16GB, GPU:NVIDIA TITAN V (12GB) である。

	7 時〜9 時	9 時〜17 時	17 時〜24 時
U-Net	678.9	702.7	79.7
提案	531.0	211.5	71.2

表 1　5 分後断面交通量予測誤差 MAE の平均

5 おわりに

一般的な機械学習モデルは学習データとテストデータとで分布が変わってしまうとモデルの予測精度が落ちてしまう。本稿では稀なイベントが起きて未知のデータ分布になってもロバストに対応可能な機械学習ワークフローを提案した。大雪を稀なイベントの一例としての断面交通量予測実験を行い、稀なイベント検出までの早さに優れ、検出した時系列的解像度箇所のみデータ拡張して高速に再学習することでほぼリアルタイムに高精度予測できることを確認した。今後は本提案ワークフローを事故や大雪以外の自然災害（地震、台風など）といった突発的で稀なイベントでも検証し、そして他の異常検知手法とも比較する予定である。

謝辞

本研究開発の一部は、NICT 高度通信・放送研究開発委託研究（持続性の高い行動支援のための次世代 IoT データ利活用技術の研究開発（課題 227））において実施されたものである。

参考文献

[1] https://www.tmt.or.jp/research/index9.html
[2] Olaf Ronneberger, et al. U-Net: Convolutional Networks for Biomedical Image Segmentation. https://arxiv.org/abs/1505.04597 (2015)
[3] https://huggingface.co/docs/diffusers/api/models/unet
[4] https://www.kaggle.com/code/super13579/u-net-1d-cnn-with-pytorch/