時刻表情報を考慮した都営バスのオープンデータによる渋滞検知手法の提案

理学専攻・情報科学コース　畠中　希　(指導教員：小口　正人)

1　はじめに

日本では交通渋滞によって年間 12 兆円の経済的損失，1 人あたり年間 30 時間の時間的損失が生じていると試算されている [1]．このような渋滞によって生じる損失を抑制するため，交通渋滞を検知し回避することは重要である．

渋滞情報の収集源である感知器が存在しない道路では渋滞検知が出来ないという課題を考慮した研究として青柳ら [2] の都営バスのリアルタイム運行データと機械学習を用いた渋滞検知手法がある．この手法では，渋滞を時速 10km 以下で断続的に走行している状態と定義し，連続する 2 つの停留所の実際の発車時刻から算出した走行速度や停留区間におけるバスの移動時間の Z スコア等を特徴量とし，停留所区間ごとに「渋滞」と「非渋滞」という二値分類を行うというものである．しかし，青柳ら [2] の手法では停留所での時間調整や乗客の乗降が考慮されていないため実際の交通速度と異なり渋滞の誤判定が行われる可能性がある．また，渋滞データと非渋滞データのデータ数が不均衡であることにより，accuracy に比べ f1-score が低いという問題点がある．

そこで本研究では f1-score 向上を目的とし，時刻表に関する特徴量を新規に追加する手法と不均衡データの問題解決アプローチを適用した手法の 2 つを提案し評価を行う．

2　提案手法

2.1　提案手法 1

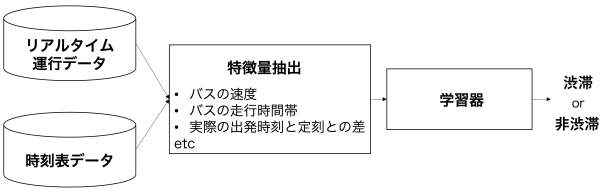

図 1: 提案手法 1 の概要

時刻表を考慮した都営バスのオープンデータを用いた渋滞検知手法の概要を図 1 に示す．まず，バスのリアルタイム運行データ，時刻表データ，渋滞データをサーバに保存する．その後，収集したデータから特徴量として使用するバスの走行速度，バスが走行している時間帯，バスの実際の発車時刻と定刻との差を停留所 2 区間ごとに抽出する．そして，走行区間（停留所 2 区間）ごとに渋滞もしくは非渋滞の二値分類を行う．ただし，渋滞は時速 10km 以下と定義する．

2.1.1　特徴量

特徴量と特徴量の抽出方法について説明する．特徴量を表 1，特徴量抽出に用いる記号と計算方法の説明を図 2 に示す．特徴量は，バス b_i が停留所区間 s_j を走行する際の速度である v_{ij}，1 つ前のバス b_{i-1} が同一停留所区間である s_j を走行する際の速度である $v_{(i-1)j}$，

表 1: 特徴量

特徴量	定義
v_{ij}	停留所区間 s_j 走行時のバス b_i の速度
$v_{(i-1)j}$	同一停留所区間 s_j 走行時の 1 つ前のバス b_{i-1} の速度
c_{ij}	バス b_i の区間 s_j 走行時の時間帯
g_{ij}	バス b_i の停留所 p_j（区間終点）における実際の出発時刻と定刻との差
$g_{i(j-1)}$	バス b_i の停留所 p_{j-1}（区間始点）における実際の出発時刻と定刻との差

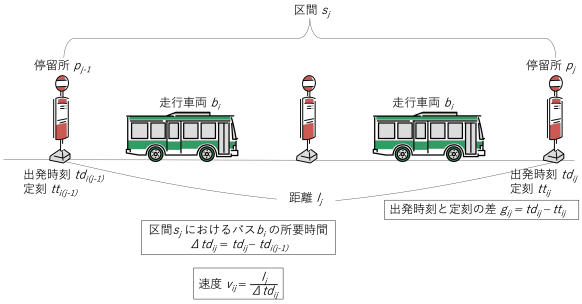

図 2: 提案手法 1 における特徴量の説明

バス b_i が停留所区間 s_j を走行する際の時間帯である c_{ij}，バス b_i の停留所 p_j つまり区間終点停留所における実際の出発時刻 td_{ij} と定刻 tt_{ij} の差である g_{ij}，バス b_i の停留所 p_{j-1} つまり区間始点停留所における実際の出発時刻 $td_{i(j-1)}$ と定刻 $tt_{i(j-1)}$ の差である $g_{i(j-1)}$ の合計 5 つである．ただし，バス b_i は出発順，停留所 p_j は経路順に並んでいるとする．特徴量である速度 v_{ij}，実際の出発時刻と定刻との差 g_{ij} の計算方法は図 2 の通りである．バス b_i の区間 s_j 走行時の時間帯である c_{ij} は 0 時 0 分 0 秒から 0 時 19 分 59 秒を 0 として，以後 20 分おきに 1, 2, 3, …と定義する．ただし，速度 v_{ij}，$v_{(i-1)j}$ はどちらも停留所の出発時刻から算出しているため停留所における時間調整によるノイズが含まれる可能性がある．

2.2　提案手法 2：不均衡データを考慮した渋滞検知手法

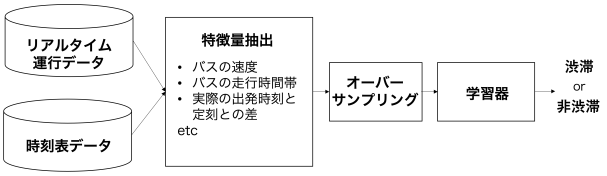

図 3: 提案手法 2 の概要

提案手法 1 である時刻表を考慮した都営バスのオープンデータを用いた渋滞検知手法に対して不均衡アプローチを行った手法の概要を図 3 に示す．

提案手法 1 と特徴量抽出までは同じ流れであるが，不均衡データへの問題解決アプローチによってデータや機械学習アルゴリズムに対して処理を行う．出力は提案手法 1 と同じく連続する 2 停留所区間ごとに渋滞・非渋滞の分類を行う．

2.2.1 不均衡データへのアプローチ

表 2: データへのコスト

少数派クラスへのコスト	$\frac{\text{全てのデータ}}{\text{クラス数}\times\text{少数派クラスのデータ数}}$
多数派クラスへのコスト	$\frac{\text{全てのデータ}}{\text{クラス数}\times\text{多数派クラスのデータ数}}$

不均衡データへの主な問題解決アプローチは，サンプリングアプローチ，コストアプローチ，その 2 つを組み合わせたハイブリッドアプローチの 3 種類に分けられる．

サンプリングアプローチはデータの抽出または合成に対する手法であり，コストアプローチは多数派データよりも少数派データに重点を置くように異なる重みを与える手法であり，ハイブリッドアプローチはサンプリングアプローチとコストアプローチを組み合わせた手法である．

本研究では，サンプリングアプローチとしてはオーバーサンプリングの 1 種であるランダムオーバーサンプリングを行う．ランダムオーバーサンプリングはデータの中からランダムに抽出したデータを複製することでデータを増加を行う手法である．コストアプローチとしては表 2 の計算によって算出されたコストを学習器に与える．ハイブリッドアプローチとしては前処理として前述のランダムオーバーサンプリングを行った後，コストを与えた学習器を用いて分類を行った．ただし，コストはサンプリング前のデータ数のまま表 2 の計算によって算出した．

3 実験

実験対象期間は 2022 年 11 月 30 日〜12 月 13 日，12 月 18 日〜12 月 31 日，実験対象バス系統は都 02，早 77，池 65， 高 71，門 19，平 23 の合計 6 系統である．データ数は全部で 100622 個であり，そのうち渋滞は 3402 個で非渋滞は 97220 個であった．作成したデータセットは 8:2 の割合で訓練データとテストデータに分割し実験を行った．使用したアルゴリズムは提案手法 1 ではランダムフォレスト，AdaBoost，XGBoost，提案手法 2 ではランダムフォレストである．

3.1 提案手法 1 の結果

表 3: 提案手法 1 の実験結果

アルゴリズム	Accuracy	Precision	Recall	F1
ランダムフォレスト	0.971	0.697	0.334	0.451
AdaBoost	0.971	0.686	0.327	0.443
XGBoost	0.970	0.660	0.345	0.453

提案手法 1 の結果を表 3 に表す．1 番 f1-score が高くなったのは XGBoost だがあまり差異はない．

3.2 提案手法 2 の結果

提案手法 2 の結果を表 4 に示す．オーバーサンプリングとコストアプローチを組み合わせたハイブリッドアプローチが f1-score が 1 番高い結果となった． これはサンプルが増えたことによりデータの特徴が掴みやすくなったこととと少数派データをなるべく間違わないようにしたためだと考えられる．

表 4: 提案手法 2 の実験結果

アプローチ	Accuracy	Precision	Recall	F1
サンプリング	0.970	0.633	0.405	0.494
コスト	0.972	0.753	0.314	0.443
ハイブリッド	0.971	0.641	0.409	0.500

3.3 先行研究と提案手法の比較

表 5: 比較結果

手法	アルゴリズム	Accuracy	Precision	Recall	F1
先行研究	XGBoost	0.946	0.602	0.257	0.306
提案手法 1	XGBoost	0.970	0.660	0.345	0.453
提案手法 2	ランダムフォレスト	0.971	0.641	0.409	0.500

表 5 に渋滞検知手法の比較結果を示す．それぞれの手法において 1 番 f1-score が高くなった結果を比較した．ただし，先行研究は異なる正解データを用いているため正確な比較になっていない．比較した結果，提案手法 1，提案手法 2 の両方とも先行研究を上回った．時刻表に関する特徴量を追加したことと不均衡データへの対応手法が予測精度向上につながったと考えられる．また，1 番 f1-score が高くなったのは，提案手法 3 のオーバーサンプリングとコストアプローチを組み合わせたハイブリッドアプローチのものである．これは不均衡データへのアプローチは有効であるためと考えられる．

4 まとめ

本研究では，渋滞検知手法として時刻表情報を考慮した都営バスのオープンデータを用いた渋滞検知手法において，先行研究，時刻表を考慮した特徴量を加え改良した手法，不均衡データを考慮した手法を提案し比較した．先行研究，提案手法 1，提案手法 2 を比較したところ，提案手法 1，2 とも f1-score は先行研究より高くなった．これは時刻表を考慮した特徴量と不均衡データのアプローチを行うことがモデルの精度向上に寄与することを示している．

今後の課題としては，提案手法の更なる精度向上を目指し，新規特徴量の追加やディープラーニングの利用が挙げられる．

参考文献

[1] 国土交通省道路局：平成 18 年度 道路行政の達成度報告書 (2006).

[2] 青柳宏紀, 岡田一洸, 山名早人：都バスのリアルタイム運行データを用いた渋滞検知, DEIM2022 第 14 回データ工学と情報マネジメントに関するフォーラム (2022).