4ZA-05

# 都営バスのオープンデータを用いた渋滞検知に向けた不均衡データ対応手法の適用と評価

畠中 希† 青柳 宏紀‡ 藤田 智也‡ 山名 早人‡ 小口 正人†
お茶の水女子大学† 早稲田大学‡

## 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.

渋滞情報の収集元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.

そこで本稿では f1-score 向上を目的とし，不均衡データの問題解決アプローチであるオーバーサンプリング，コストアプローチ，ハイブリッドアプローチの 3 種類を適用し評価を行う.

## 2 渋滞検知モデル

都営バスのオープンデータを用いた渋滞検知モデルを図 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.

Application and evaluation of imbalance data handling methods for congestion detection using open data of Toei bus
†Nozomi Hatanaka ‡Hiroki Aoyagi ‡Tomoya Fujita
‡Hayato Yamana †Masato Oguchi
† Ochanomizu University
‡ Waseda University

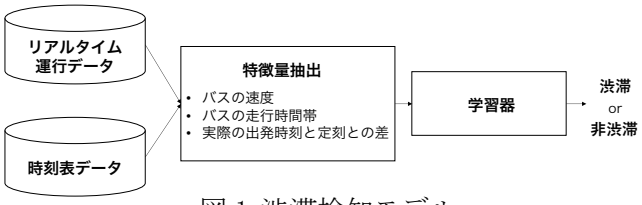


図 1 渋滞検知モデル

### 2.1 特徴量

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$b_i$の停留所区間$s_j$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 |

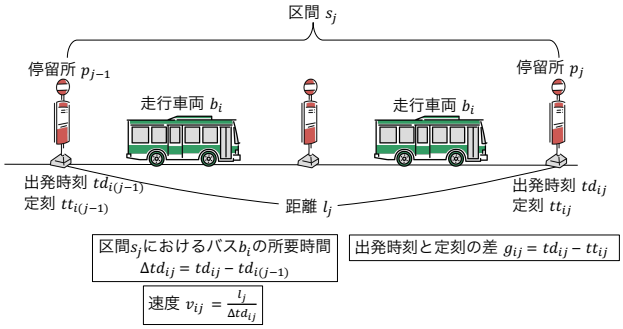


図 2 特徴量抽出に用いる記号と計算方法の説明

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$s_j$を走行する際の速度である$v_{(i-1)j}$，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$c_{ij}$，停留所$p_j$つまり区間終点停留所におけるバス$b_i$の実際の出発時刻$td_{ij}$と定刻$tt_{ij}$の差である$g_{ij}$，停留所$p_{j-1}$つまり区間始点停留所におけるバス$b_i$の実際の出発時刻$td_{i(j-1)}$と

定刻$tt_{i(j-1)}$の差であるである$g_{i(j-1)}$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．

## 3　不均衡データの対応

表 2　コストの計算方法

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

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では少数派データを増やすオーバーサンプリングと多数派データを減らすアンダーサンプリングがある．

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方法である．

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方法である．

本稿では，サンプリングアプローチとして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，コストアプローチとして表 2 の計算によって算出されたコストを学習器に与えた．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．

## 4　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ランダムフォレストである．baseline 手法 は不均衡データへの問題解決アプローチを行わず，ランダムフォレストを使用したものである．

### 4.1 結果

表 3　実験結果

| アプローチ | accuracy | precision | recall | f1-score |
|---|---|---|---|---|
| baseline | 0.971 | 0.696 | 0.325 | 0.433 |
| オーバーサンプリング | 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 |

実験結果を表 3 に示す．全ての不均衡データへのアプローチにおいて baseline 手法より f1-score が高くなった．よって，不均衡データへのアプローチは有効であると考えられる．また，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．

## 5　まとめと今後の課題

本稿では都営バスのオープンデータと機械学習を用いた渋滞検知手法に不均衡データの問題解決アプローチを適用し評価を行った．問題解決アプローチはオーバーサンプリング，コストアプローチ，ハイブリッドアプローチの 3 種類を適用し，その結果，全てアプローチにおいて f1-score が高くなった．特に，ハイブリッドアプローチが 1 番 f1-score が高い結果となり，不均衡データへのアプローチ手法の適用は有効であると考えられる．

今後は他の特徴量追加などにより精度向上を目指す．

## 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. https://www.mlit.go.jp/road/ir/ir-perform/h19/all.pdf

[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, pp. 1-8.