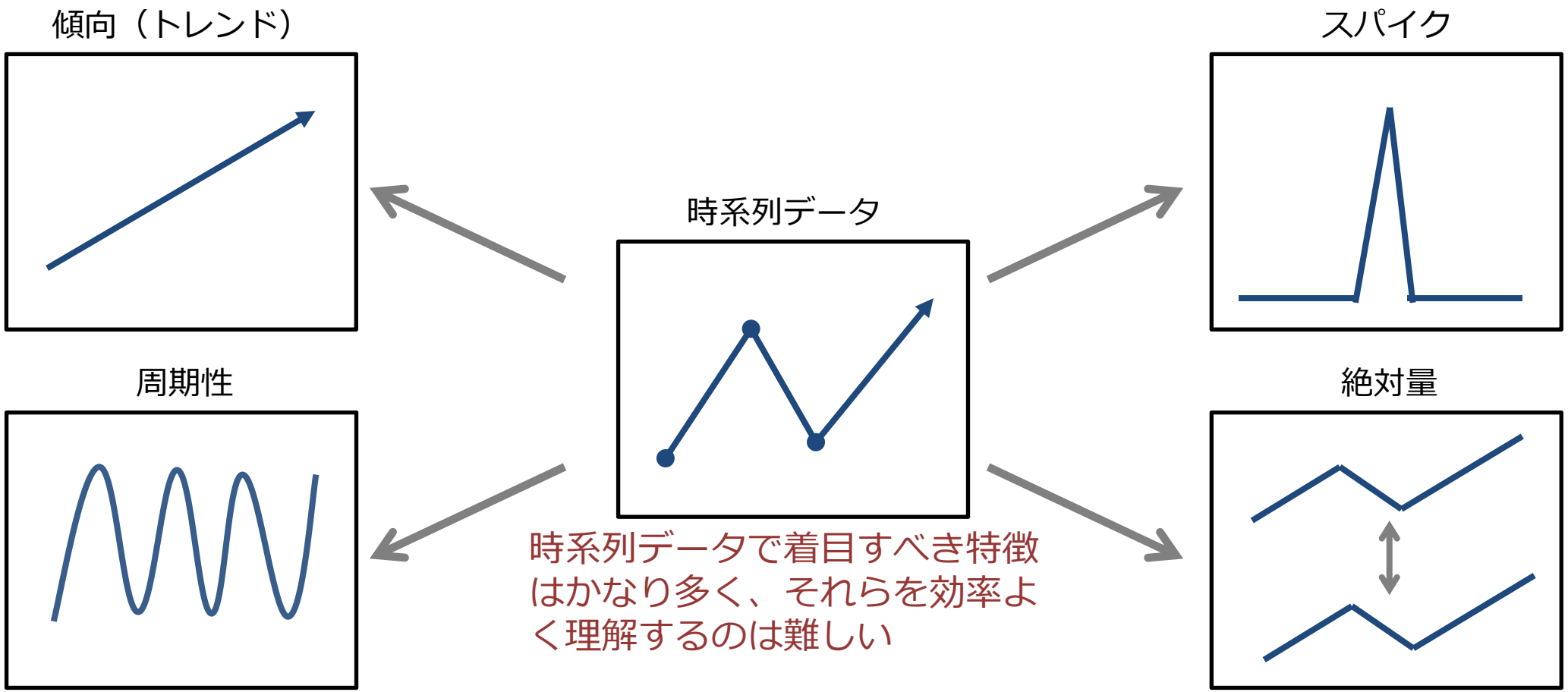
デモ

はじめに

■**時系列データを使ってクラス分類予測を行う**には時系列データの特性を掴み、表現することが重要です。しかし、それは容易ではなく**専門家レベルの知識と高度なプログラミング力を必要**とします



©GRI Inc.

はじめに

■時系列データのクラス分類予測を**自動、高速、かつ高精度**で行う方法をご紹介いたします

自動で特徴量設計、予測

機械学習モデルを構築する際に最も大変な部分である特徴量設計・機械学習モデルの構築を自動で行います。

高速で結果が出る

データがあればすぐに機械学習ができて、予測精度を確認できます。後述する例では3時間ほどで予測モデルが構築できました。

簡単に高精度なモデルを作成

用いている機械学習モデルは高精度ができやすいもので、面倒なパラメータチューニングも不要です。後述する例では一回の試行で元論文の精度を超えました。

時系列データの特性を素早く探索

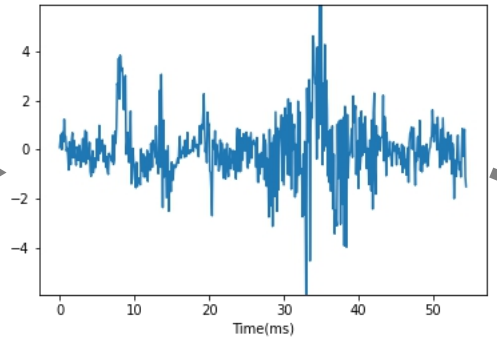
機械学習の結果からどの特徴量が重要かを可視化でき、掴み辛い時系列データの特性を素早く把握することができます。

©GRI Inc.

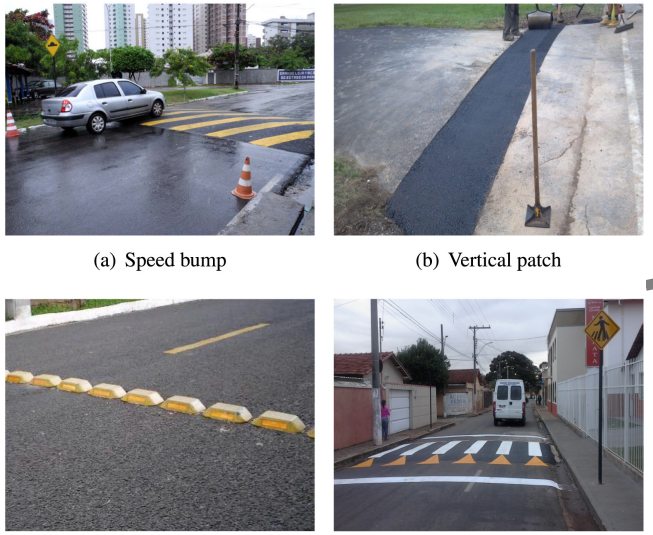
ForecastFlow による時系列データのクラス予測問題に関するサマリー

問題設定

■車に取り付けた加速度センサー(3軸)の速度変化の時系列データを用いて、**どの障害物（横断歩道、ロードマーカー、ハンプ、道路のつぎはぎ）の上を通過したのか**を予測する



障害物



(a) Speed bump (b) Vertical patch

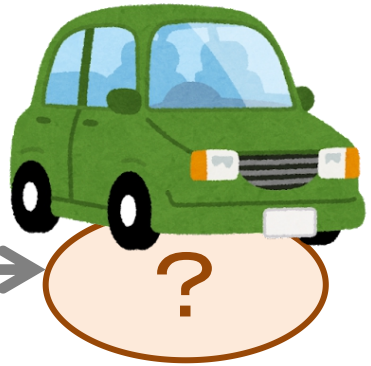
(c) Raised pavement markers (d) Raised crosswalk

pandas tsfresh

学習

ForecastFlow

予測



データソース・論文
Souza V.M.A. Engineering Applications of Artificial Intelligence Volume 74, pp. 198-211.

©GRI Inc.

ForecastFlow による学習の結果概要

■精度指標 F1, Recall, Precision はどの障害物に対しても約80%程度

- ■予測用データに対しての正答率は **89.4%**
- ■**元論文の最大正答率 81.13%を上回っている**
- ■**サンプル数が 400個程度**でもマルチクラス分類が可能

■**特徴量設計、機械学習は自動で設計**

- ■特徴量設計：tsfresh によって Python コード数行で設計
- ■機械学習：ForecastFlow によってコーディングなしで学習（Python API を利用しても数行）

target	Counts	F1	Precision	Recall
raised_crosswalk	80 (20.5%)	93.9%	91.7%	96.3%
raised_markers	94 (24%)	82.2%	86%	78.7%
speed_bump	106 (27.1%)	95.2%	97.1%	93.4%
vertical_patch	111 (28.4%)	86.1%	83.2%	89.2%

©GRI Inc.

ForecastFlow による時系列データの予測モデルの説明

使用データ

■時系列データによるクラス分類問題の事例を集めているデータベースのデータセットを使用

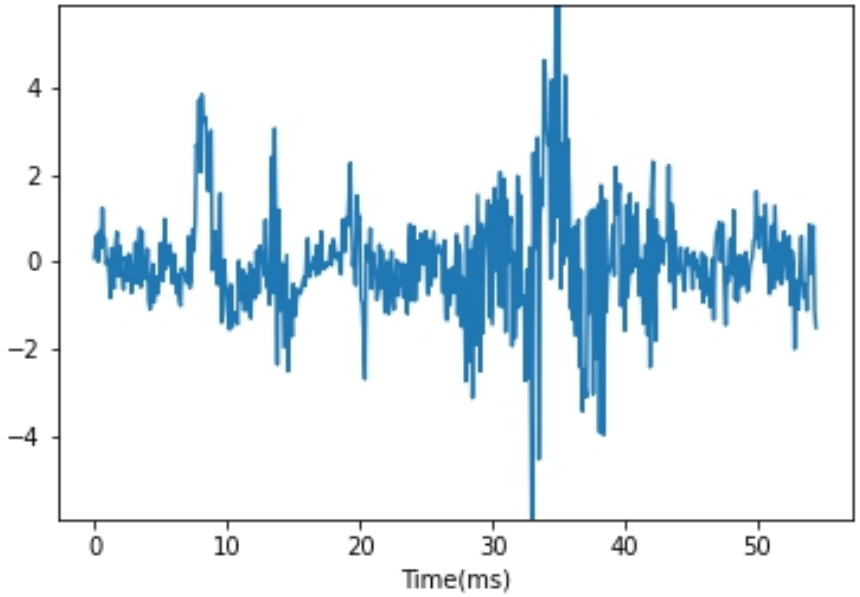
- 車に取り付けた加速度センサー(3軸)の速度変化の時系列データを用いて、**どの障害物（横断歩道、ロードマーカー、ハンプ、道路のつぎはぎ）の上を通過したのか**を予測する
- https://www.timeseriesclassification.com/description.php?Dataset=AsphaltObstacles

■データ数は学習用：390個、予測用：391個

障害物の内訳（サンプル数）

	学習用	予測用
横断歩道	80	80
ロードマーカー	93	94
ハンプ	106	106
道路のつぎはぎ	111	111
計	390	391

センサーデータ（速度変化）



©GRI Inc.

各障害物のセンサーデータの特性

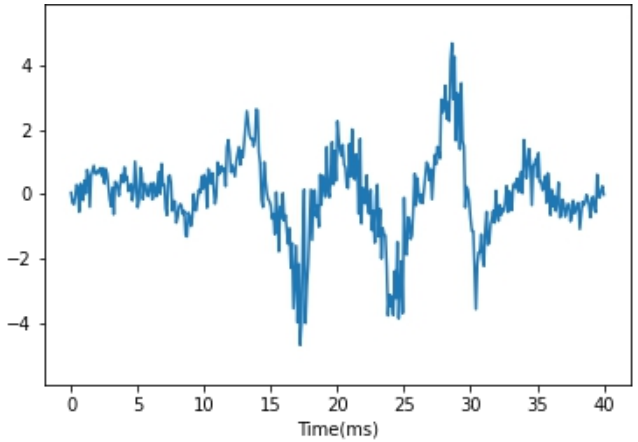
■障害物（横断歩道、ロードマーカー、ハンプ、道路のつぎはぎ）によって、センサーデータの速度変化に以下の大まかな特徴がみられた

- **「ロードマーカー」と「道路のつぎはぎ」の時系列パターンは似通っていて分類が難しいことが予想される**

横断歩道

(rased crosswalk)

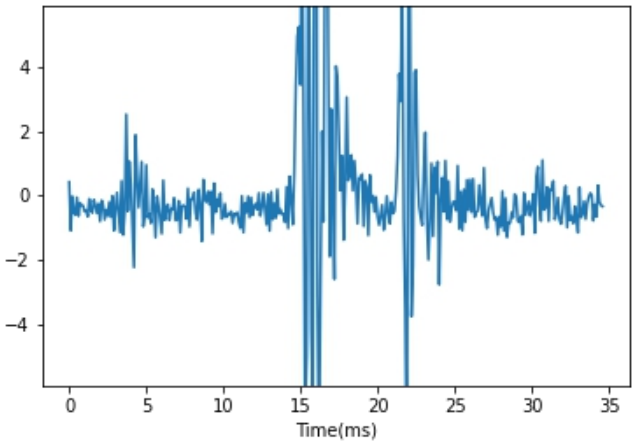
- **三回振動する**



ロードマーカー

(raised markers)

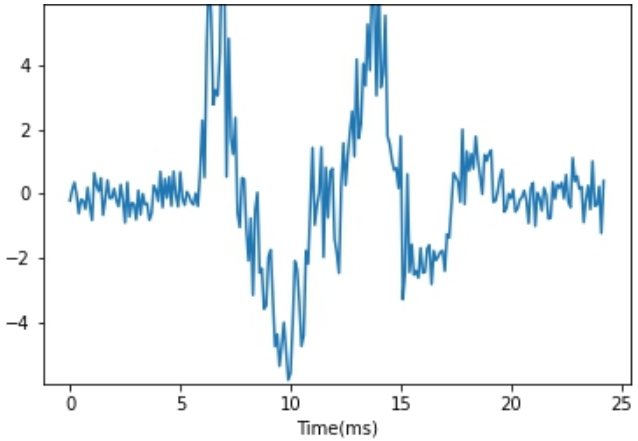
- **二回激しく振動する**



パンプ

(speed bump)

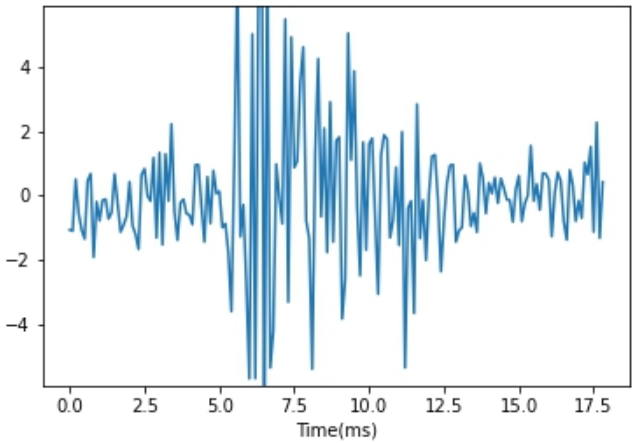
- **二回振動する**



道路のつぎはぎ

(vartical patch)

- **一回長く激しく振動する**

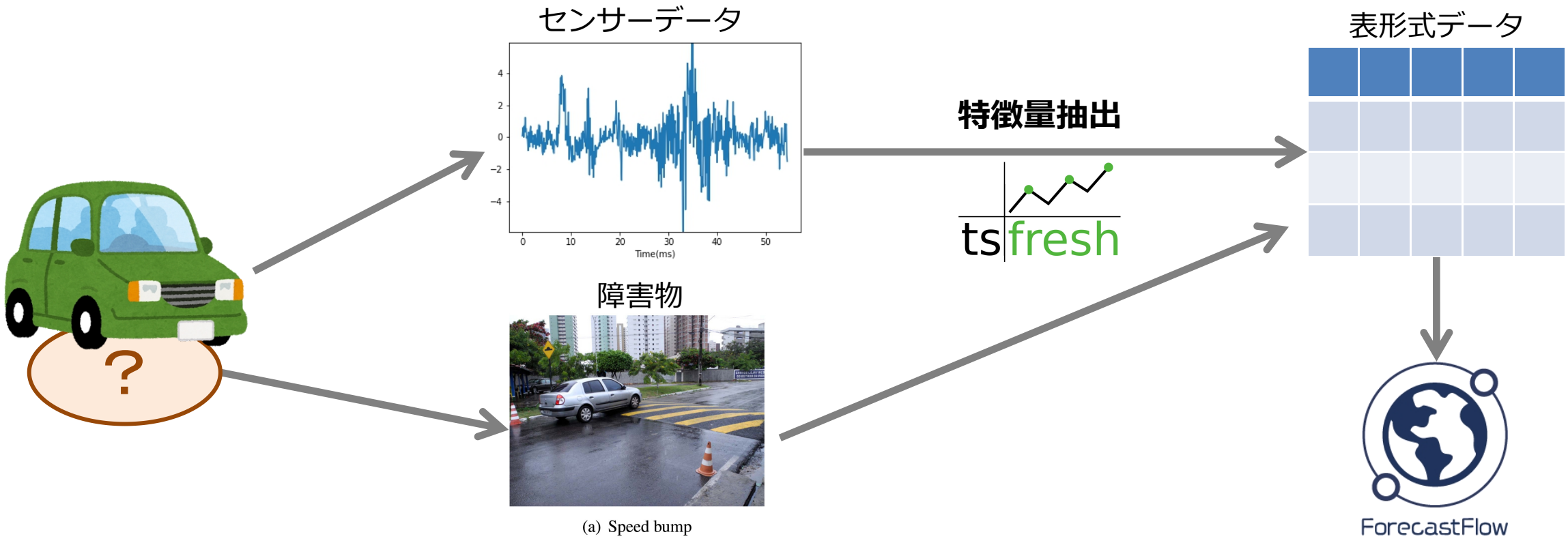


©GRI Inc.

データ整形：フロー

■センサーデータから障害物を予測するために、時系列特有の特徴を抽出した表形式データへと変換、ForecastFlow で学習しスコアを評価

- ■ ForecastFlow には表形式データを入れる必要がある
- ■ 特徴量抽出に Pythonパッケージである tsfreshを使用

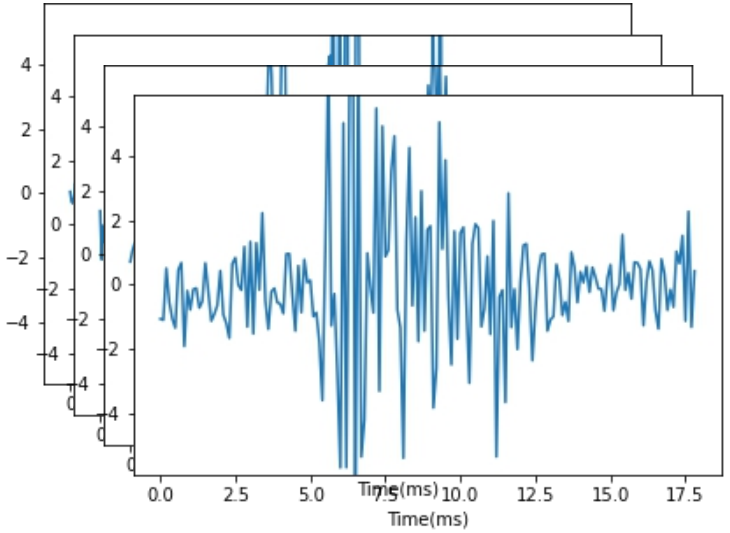


©GRI Inc.

データ整形：特徴量抽出

■tsfreshで**センサーデータから 143個の特徴量を抽出**

- ■tsfresh から生成される特徴量一覧は p26 以降に記載
- ■tsfresh の機能により**障害物と関連性のありそうな特徴量のみを選別**
  - ■欠損値のある特徴量は除外
  - ■障害物を関係性がある特徴量のみを抜き出す



tsfresh

143個

id	特徴量	…	特徴量	障害物
0				横断歩道
1				横断歩道
2				横断歩道
…				…
388				道路のつぎはぎ
399				道路のつぎはぎ

©GRI Inc.

デモ：tsfresh

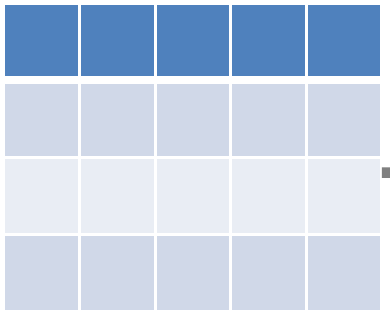
■Jupyter Notebookで tsfreshによる特徴量設計をデモ

©GRI Inc.

機械学習

■学習に **ForecastFlow** を使用

- ■AutoML なので**コーディングが不要**（データがあればすぐに学習ができる、パラメータチューニングも不要）
- ■lightGBM という**決定木モデル**（ルールベースによる切り分け）を改良した機械学習モデル
- ■予測だけではなく、どの特徴量が重要であるかを指標化する**感度分析**も行ってくれる

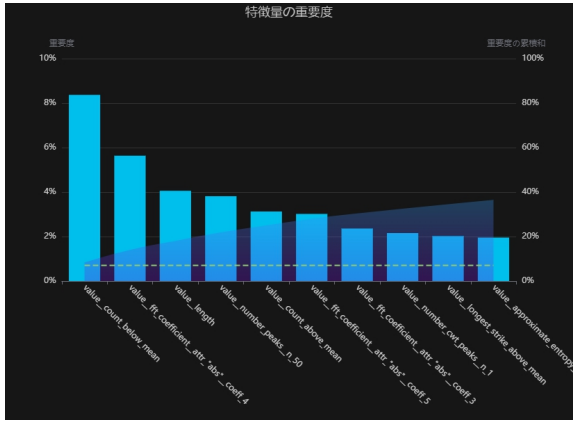


ForecastFlow

学習

学習の評価

感度分析



©GRI Inc.

ForecastFlow による特徴量の選定

■Forecastflow ではワンクリックで相関が強い特徴量を一括で削除可能

- ■相関が強い特徴量を予測モデルの学習で使用すると**過学習を誘発したり、精度や感度分析にも影響**が出る
- ■tsfresh にはこの特徴量同士の強相関を考慮する機能がない → ForecastFlow で補う
- ■今回は精度に影響は無かったが、場合によっては精度が改善する

©GRI Inc.

学習：精度評価

■精度指標 F1, Recall, Precision はどの障害物に対しても約85~90%程度

■予測用データを用いた予測

- ■正答率は **89.4%**
- ■**元論文の最大正答率 81.13%を上回っている**

Confusion matrix		予測			
		横断歩道	マーカー	パンプ	つぎはぎ
正解	横断歩道	79	1	11	0
	マーカー	1	76	0	23
	パンプ	0	1	101	0
	つぎはぎ	0	26	1	102

target	Counts	F1	Precision	Recall
raised_crosswalk	80 (20.5%)	93.9%	91.7%	96.3%
raised_markers	94 (24%)	82.2%	86%	78.7%
speed_bump	106 (27.1%)	95.2%	97.1%	93.4%
vertical_patch	111 (28.4%)	86.1%	83.2%	89.2%

©GRI Inc.

学習：精度評価：強相関特徴量の有無による比較

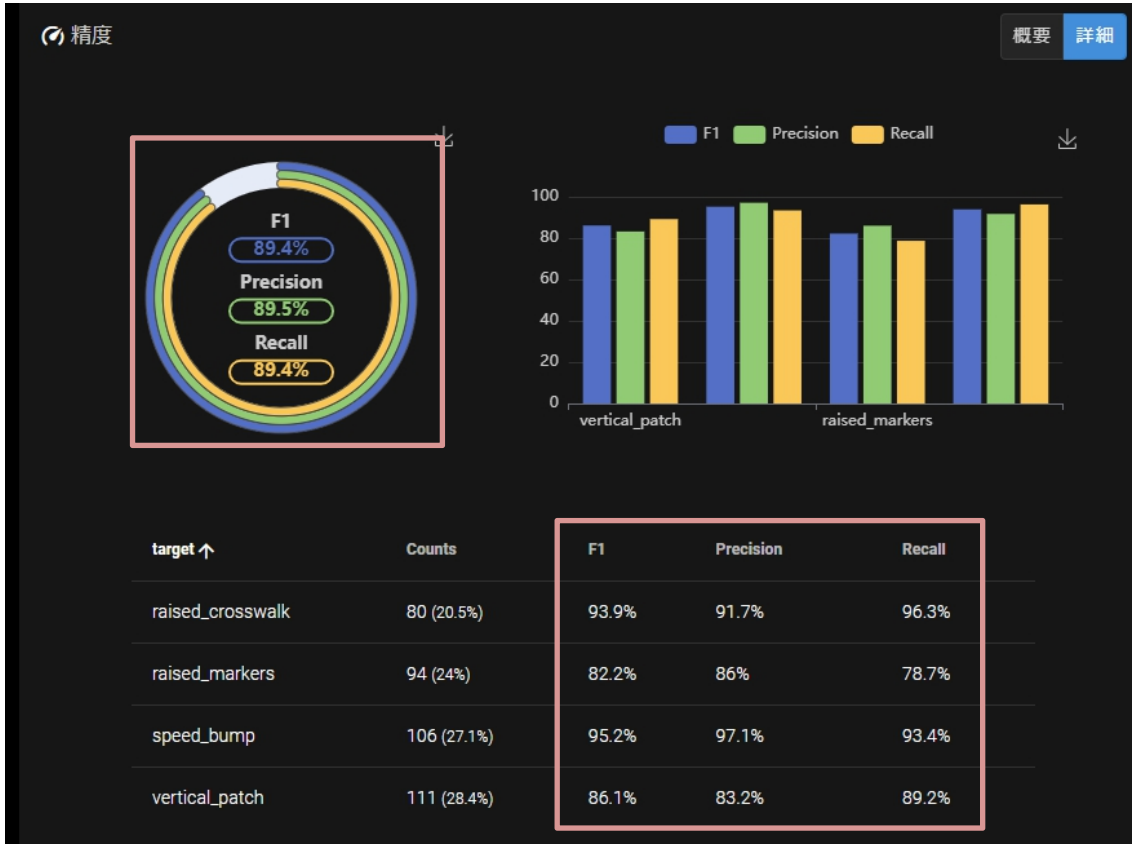
■強相関特徴量を除くか残すかで精度評価スコアが変わるかを比較

- ■両者で精度に差はみられない
- ■強相関特徴量を削除することで精度が改善するケースがあるため、試す価値はある

強相関特徴量を除く

target	Counts	F1	Precision	Recall
raised_crosswalk	80 (20.5%)	93.4%	89.7%	97.5%
raised_markers	94 (24%)	82.2%	86%	78.7%
speed_bump	106 (27.1%)	94.6%	98%	91.5%
vertical_patch	111 (28.4%)	86.1%	83.2%	89.2%

強相関特徴量を残す

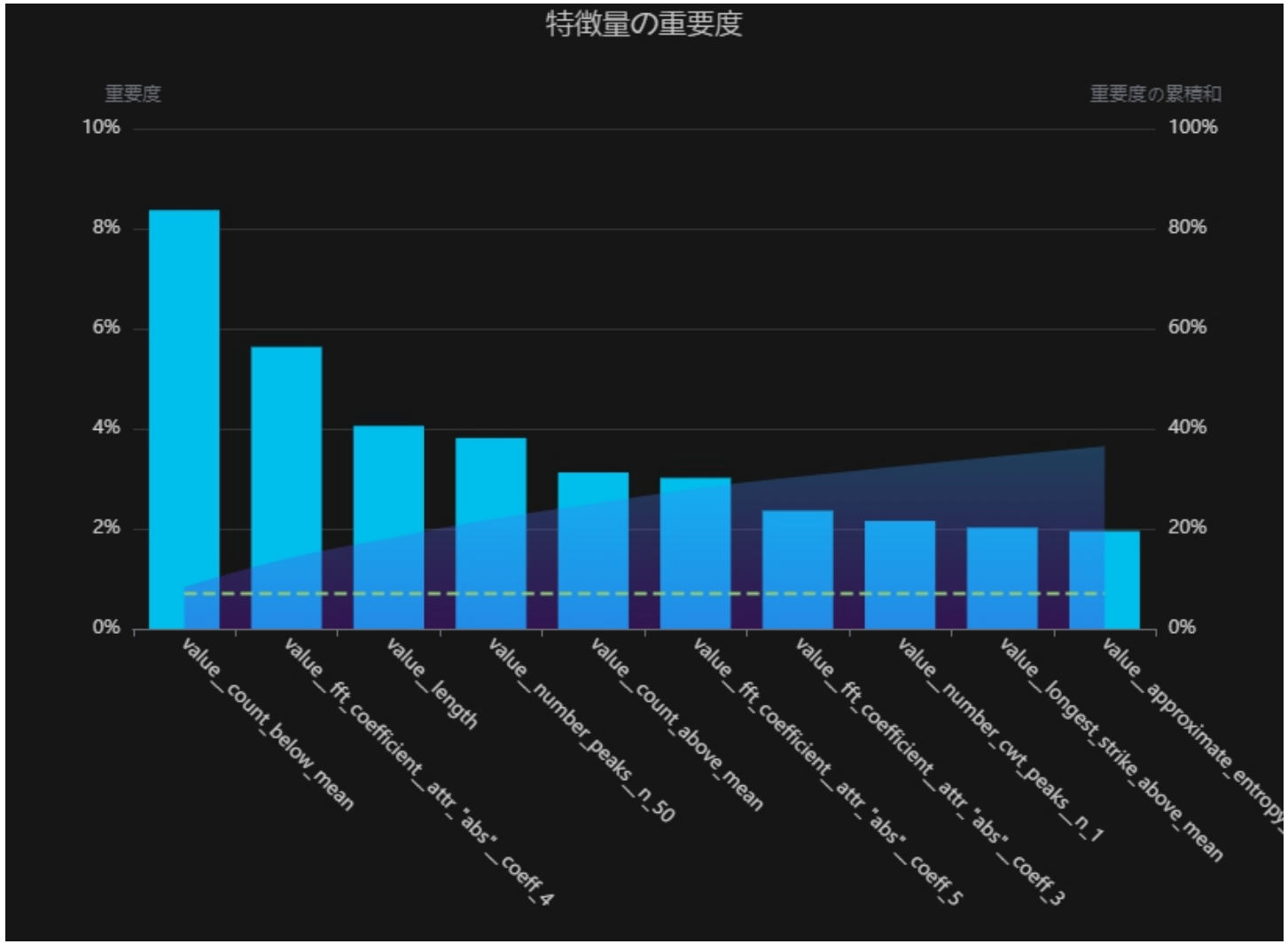


target	Counts	F1	Precision	Recall
raised_crosswalk	80 (20.5%)	93.9%	91.7%	96.3%
raised_markers	94 (24%)	82.2%	86%	78.7%
speed_bump	106 (27.1%)	95.2%	97.1%	93.4%
vertical_patch	111 (28.4%)	86.1%	83.2%	89.2%

©GRI Inc.

感度分析

■予測をする上でよく効いている、重要な特徴量を可視化（上位10個を表示）



©GRI Inc.

感度分析：表

■重要特徴量上位 10個を表示（特徴量については p26 以降を参照）

■絶対量・周期性に関する特徴量が多い

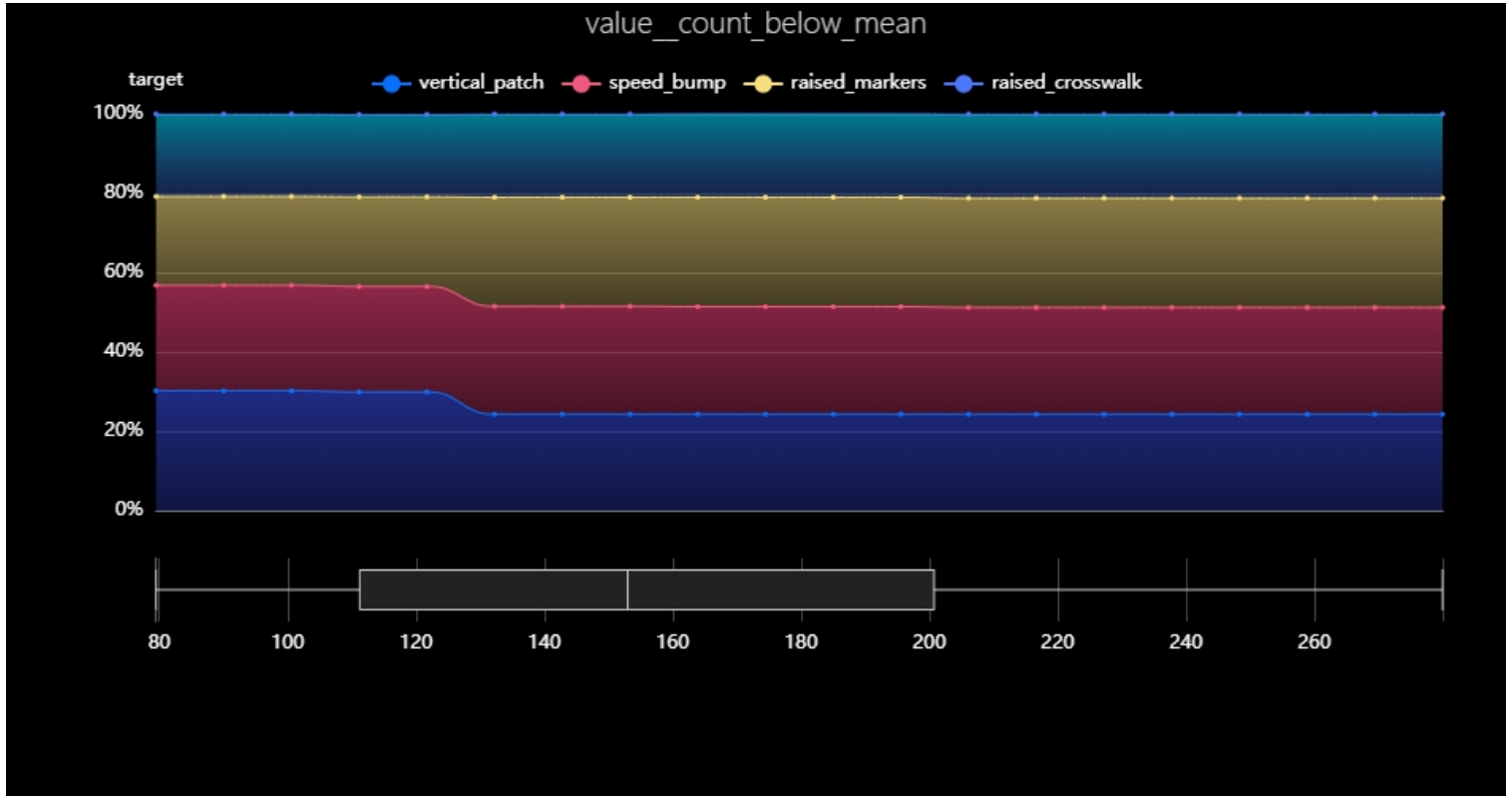
順位	特徴量名	特性
1	count_below_mean	絶対量
2	fft_coefficient__attr_"abs"__coeff_4	周期性
3	length	絶対量
4	number_peaks__n_50	スパイク
5	count_above_mean	絶対量
6	fft_coefficient__attr_"abs"__coeff_5	周期性
7	fft_coefficient__attr_"abs"__coeff_3	周期性
8	number_cwt_peaks__n_1	スパイク
9	longest_strike_above_mean	絶対量
10	approximate_entropy__m_2__r_0.5	周期性 変動

©GRI Inc.

感度分析：部分依存グラフ

■重要特徴量上位 1位の「count_below_mean」の部分依存グラフを表示

- count_below_mean：時系列の平均値より値が小さい観測点の数
- 130 あたりを境に「道路のつぎはぎ」、「ロードマーカー」と予測される確率が変動している
- このように、**学習結果から時系列データの特性を探索することができる**



©GRI Inc.

デモ：ForecastFlow

■ForecastFlow での機械学習モデルの構築をデモ

©GRI Inc.

tsfresh から生成される特徴量について

tsfresh から生成される特徴量一覧1

■tsfresh は時系列の特性（変動・周期性・絶対量など）を抜き出すための計算式とそのパラメータの組み合わせで一つの時系列データから大量の特徴量を生成する（https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html）

特徴量名	特性
abs_energy	絶対量
absolute_maximum	絶対量
absolute_sum_of_changes	変動
agg_autocorrelation	周期性
agg_linear_trend	変動
approximate_entropy	周期性・変動
ar_coefficient	周期性・変動
augmented_dickey_fuller	スパイク
autocorrelation	周期性
benford_correlation	スパイク
binned_entropy	絶対量

特徴量名	特性
c3	変動
change_quantiles	変動
cid_ce	スパイク
count_above	絶対量
count_above_mean	絶対量
count_below	絶対量
count_below_mean	絶対量
cwt_coefficients	周期性
energy_ratio_by_chunks	絶対量
fft_aggregated	周期性
fft_coefficient	周期性

©GRI Inc.

tsfresh から生成される特徴量一覧2

■tsfresh は時系列の特性（変動・周期性・絶対量など）を抜き出すための計算式とそのパラメータの組み合わせで一つの時系列データから大量の特徴量を生成する（https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html）

特徴量名	特性
first_location_of_maximum	絶対量
first_location_of_minimum	絶対量
fourier_entropy	周期性
friedrich_coefficients	変動
has_duplicate	スパイク
has_duplicate_max	スパイク
has_duplicate_min	スパイク
index_mass_quantile	絶対量
kurtosis	変数
large_standard_deviation	変動
last_location_of_maximum	スパイク

特徴量名	特性
last_location_of_minimum	スパイク
lempel_ziv_complexity	変動
length	絶対量
linear_trend	変動
linear_trend_timewise	変動
longest_strike_above_mean	絶対量
longest_strike_below_mean	絶対量
matrix_profile	絶対量・変動
max_langevin_fixed_point	変動
maximum	絶対量
mean	絶対量

©GRI Inc.

tsfresh から生成される特徴量一覧3

■tsfresh は時系列の特性（変動・周期性・絶対量など）を抜き出すための計算式とそのパラメータの組み合わせで一つの時系列データから大量の特徴量を生成する（https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html）

特徴量名	特性
mean_abs_change	変動
mean_change	変動
mean_n_absolute_max	絶対量
mean_second_derivative_central	絶対量
median	絶対量
minimum	絶対量
number_crossing_m	変動
number_cwt_peaks	スパイク
number_peaks	スパイク
partial_autocorrelation	周期性
percentage_of_reoccurring_datapoints_to_all_datapoints	絶対量

特徴量名	特性
percentage_of_reoccurring_values_to_all_values	絶対量
permutation_entropy	変動
quantile	絶対量
query_similarity_count	絶対量
range_count	絶対量
ratio_beyond_r_sigma	スパイク
ratio_value_number_to_time_series_length	変動
root_mean_square	絶対量
sample_entropy	周期性・変動
set_property	不明
skewness	絶対量

©GRI Inc.

tsfresh から生成される特徴量一覧4

■tsfresh は時系列の特性（変動・周期性・絶対量など）を抜き出すための計算式とそのパラメータの組み合わせで一つの時系列データから大量の特徴量を生成する（https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html）

特徴量名	特性
spkt_welch_density	周期性
standard_deviation	変動
sum_of_reoccurring_data_points	絶対量
sum_of_reoccurring_values	絶対量
sum_values	絶対量
symmetry_looking	変動
time_reversal_asymmetry_statistic	周期性・変動
value_count	絶対量
variance	変動
variance_larger_than_standard_deviation	絶対量
variation_coefficient	変動

©GRI Inc.

参考

参考

- ブログ(CO-WRITE)
  - 今回の資料の内容について記載（少し情報が古いので更新しておきます）
  - https://gri.jp/media/entry/5488
  - tsfresh の使い方の記事も纏まり次第公開します
- Jupyter Notebook
  - GitHub で公開予定
  - 5月までには見れるようにします（ForecastFlowの Slackで周知します）

©GRI Inc.