D-044

機械学習によるセンサー異常値検出
Outlier Detection in Sensor Data Using Supervised Learning

栗原 慶典† 根山 亮† 三宮 千尋† 那和 一成†
Keisuke Kurihara Ryo Neyama Chihiro Sannomiya Kazunari Nawa

1. はじめに

2015 年は IoT 元年と言われ、自動車や携帯端末、さらには家電など、あらゆるものから大量のストリームデータが収集できる環境が整いつつある。これらの大量のストリームデータと Web 上のデータなどと合わせてデータセンターに集積し、自動車のドライバや携帯端末利用者へ種々の支援を行ったり、HEMS などのエネルギー管理システムを提供したりする Cyber Physical System (CPS)を構築できる。これにより、従来よりも利用者と協調できる情報提供サービスの実現が期待できる[1] (図 1)。また、昨今では自動車の自動運転技術の開発や運転支援サービスの提供が活発化しており、これらのための自動車センサーデータの解析が進められている。

しかしながら自動車のセンサーデータには時系列的な物理値をサンプリングした数値データだけでなく、スイッチの ON/OFF など限られた数値しかとらないフラグデータも混在しており、多種多様な特性を持っていることが知られている。そのためデータ解析にあたってはデータの特性に合わせた適切な解析を行うことも重要である[2]。

このようなデータ解析において、信頼性の高いデータを使用することは重要である。しかし、自動車のセンサーデータにはセンサーや ECU の異常による異常値が含まれており、データから異常値を事前に取り除くクレンジング処理が必要である。

クレンジング処理を人手で行うには多くの手間とデータに対する深い理解が必要である。そのためデータ量が増え続けるとクレンジングの処理時間も増え続けてしまう。前述のように自動車のセンサーデータが多種多様な特性を持っていることもその要因として大きい。

そこで、本研究では、既知の異常値を学習した教師あり機械学習によって新しいデータの異常値検出を行う手法を評価した。この手法により、既に判明している異常値をあらかじめ検出でき、クレンジング処理の負荷を軽減できると考えた。また、機械学習を行うにあたりデータの特徴量を算出するが、平均値や標準偏差など、データの基本的な特性を表す基本統計量を使用することで新しい種類のデータが登場しても容易に適用できると考えた。

本論文では教師あり機械学習によって異常値を検出し、その検出精度を評価した結果を報告する。2 章では、本研究で対象とする異常値として自動車のセンサーデータで発生しやすい異常値を紹介する。3 章では異常値を教師あり機械学習で検出するための手法を提案する。4 章では提案手法を用いて実験を行い、検出精度を評価する。

†(株)トヨタ IT 開発センター TOYOTA InfoTechnology Center Co., Ltd.

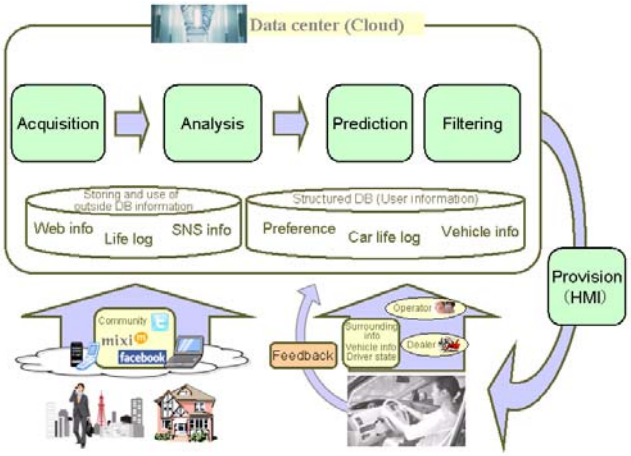

図 1 データ提供サービスのための Cyber-Physical System

2. 本研究で扱う異常値について

本研究は、自動車のセンサーデータにおける代表的な異常値を対象とする。対象とする 4 種類の異常値を以下に示す。これらはバイナリデータの加工を誤ったり、センサーやデータロガーに異常が発生したりする場合に現れる異常値であり、実際に自動車のセンサーデータに現れる場合がある。

1) ax+b 型異常値（以下、ax+b）
単位違いや基準値違いによって全体的に正常な値の定数 a 倍＋バイアス項 b で表現される値となる異常値である。図 2.1 に ax+b の波形例を示す。

2) 突発値型異常値（以下、突発値）
ノイズなどにより突然大きな値が発生している異常値である。図 2.2 に突発値が発生した場合の波形例を示す。

3) 欠損値型異常値（以下、欠損値）
センサーがデータを取得できない場合に、その部分が 0 や取得可能な値の最大値で置き換えられている異常値である。図 2.3 に欠損値が発生した場合の波形例を示す。

4) サンプル漏れ型異常値（以下、サンプル漏れ）
データロガーの異常などにより一時的にデータが取得できない場合に値が記録されず、それまでのデータにデータロガーが復旧してからのデータが連続して記録されている異常値である。図 2.4 にサンプル漏れが発生した場合の波形例を示す。

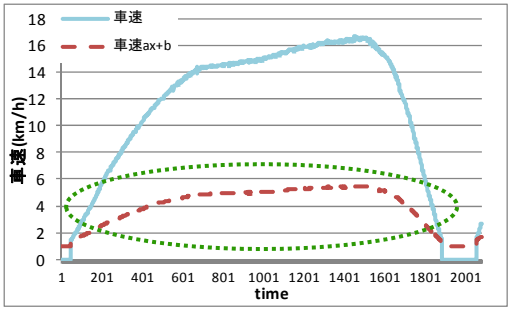

図 2.1　ax+b

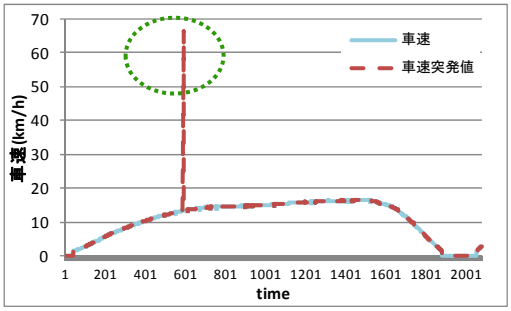

図 2.2　突発値

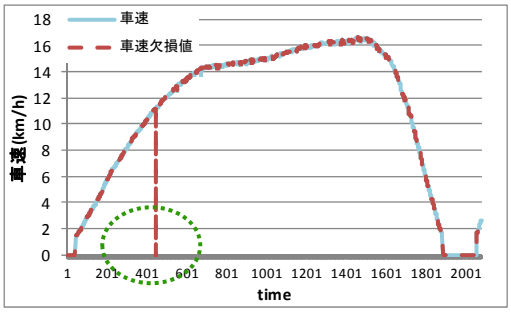

図 2.3　欠損値

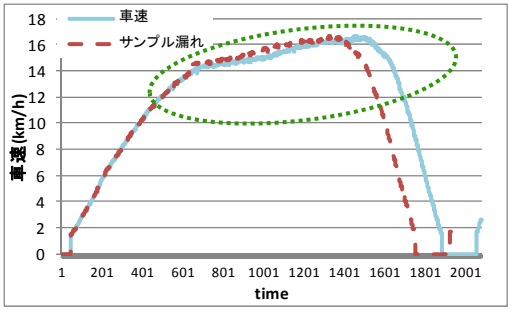

図 2.4　サンプル漏れ

3. 異常値検出手法の検討

2 章で述べた 4 種類の異常値を精度よく検出するため、アルゴリズムに入力する特徴量と候補となるアルゴリズムを検討した。

3.1 特徴量の検討

本研究は、データの特性に依存せずに異常値検出を行うことを目的としている。そのため、特徴量は特定の領域知識に依存しない普遍的な値であることが望ましい。そこで使用する特徴量には対象データの「平均値」、「標準偏差」、「値の種類数の平方根」、「最大値」、「最小値」、「中央値」の 6 種類の値、さらにデータのフレーム間の差分値を算出し、同様に「差分値の平均値」、「差分値の標準偏差」、「差分値の値の種類数の平方根」、「差分値の最大値」、「差分値の最小値」、「差分値の中央値」の 6 種類の値を算出した計 12 種類の値を使用した。

特徴量を算出するにあたり、時系列データから 1s の時間窓を、0.1s シフトで切り出してから特徴量を算出した。時間窓を切り出すことで、膨大な時系列データの中での異常値が存在する時間帯を知ることができる。また想定している異常値のうち、突発値と欠損値は局所的に異常が発生するため窓単位での特徴量算出が適していると考えた。

3.2 検出アルゴリズム

本研究では分類アルゴリズムに教師あり学習の Random Forest を使用して異常値検出を行うことにした。

Random Forest はあらかじめ判明している正常値と異常値を学習することで、新規に分析する値が過去に学習した正常値と異常値のどちらに分類されるかを判断することが出来るアルゴリズムである。他のアルゴリズムとして SVM も考えられるが SVM はパラメータチューニングを慎重に行う必要がある。Random Forest はパラメータチューニングが SVM より簡易でありながら精度が高く使いやすいと判断し、これを採用した[3]。

4. 異常値検出の実験

3 章で提案した手法で異常値検出の精度を求めた。

実験に使用したデータは東京都港区赤坂周辺を周回した車両走行データ 90 トリップで、今回は、その中から車速のデータを選択した（表 4.1 参照）。4.1 節で詳しく述べるが、90 トリップのデータに 4 種類の異常値をパラメータを変えながら適用し、異常度合の異なる複数パターンの異常値を生成した。すべての実験データで窓毎に特徴量を算出し、少なくとも 1 つの値が異常値である窓を「異常」、異常値を 1 つも含まない窓を「正常」とラベル付けした。

4 種類の異常値のパターンごとに教師データを 90 トリップ中 50 トリップ、評価データには残り 40 トリップを選択し、教師データを学習して評価データの分類を行い「異常」の検出結果を得た。

「異常」の検出における適合率、再現率が共に 80%以上となることを目標とし、検出結果の評価を行った。

表 4.1　実験データ諸元

データの種類	車両走行データ
場所	赤坂周辺
コース	3 コース（A,B,C）
件数（トリップ数）	90
サンプリングレート	約 91Hz
使用するデータ属性	車速

4.1 異常値の生成

異常値には車速のデータに人工的に 4 種類の異常値を発生させたものを使用した。また、異常度合を調整するパラメータを変えながら複数パターンの異常値を用意した。

4.1.1 ax+b の生成

ax+b は実際の値を x として ax+b に変換した。異常度合を示すパラメータ a、b を a=[0.25,0.5,0.57,1.25,1.5]　b=[0.25σ,0.5σ]（σは 1 トリップ全体の標準偏差）とし、網羅的に組み合わせて複数パターン生成した。

4.1.2 突発値の生成

突発値は 1 トリップのデータから無作為に 10 か所を選択し、突発値(s)と置き換えた。異常度合を示すパラメータ s には s=[-2σ,-σ,-0.5σ,0.5σ,σ,2σ]（σは 1 トリップ全体の標準偏差）を適用し複数パターンを生成した。

4.1.3 欠損値の生成

欠損値は 1 トリップのデータから無作為に k か所を選択し、値を 0 もしくは 16 ビットの符号なし整数における最大値である2^{16}-1 で置き換えた。異常度合を示すパラメータ k には k=[1,10,50]を当てはめ複数パターンを生成した。

4.1.4 サンプル漏れの生成

サンプル漏れは 1 トリップのデータから無作為に 10 か所を選択し、n フレーム分削除して前後のフレームを連結した。異常度合を示すパラメータ n は n=[100,200,500]とし、複数パターン生成した。

4.2 教師あり学習による異常検出実験結果

検出結果を表.4.2-1～4 に示す。表中の値は評価データ 40 トリップにおける「異常」の検出結果から適合率と再現率の平均値を示したものである。検出結果から以下のことが判明した。

1) ax+b （表 4.2-1）

a=1.25 のみ適合率と再現率が 90%を下回るがそれ以外は 90%以上であった。b の影響は大きく表れなかった。

2) 突発値（表 4.2-2）

0.5σだけ適合率が低く、誤検出が多く発生していることがわかる。他は 90%近い精度で検出ができた。突発値の値が＋方向、または－方向に大きくなるほど精度が高くなる傾向が現れた。

表 4.2-1　ax+b の検出結果（%）

a	b=0.25σ		b=0.5σ	
	適合率	再現率	適合率	再現率
0.25	100	100	99	99
0.5	97	97	97	97
0.75	94	93	94	93
1.25	86	82	88	85
1.5	95	94	95	95

表 4.2-2　突発値の検出結果（%）

s	適合率	再現率
−2σ	100	100
−σ	88	100
−0.5σ	88	100
0.5σ	**69**	100
σ	85	100
2σ	93	100

表 4.2-3　欠損値の検出結果（%）

k	0 埋め		2^{16}-1	
	適合率	再現率	適合率	再現率
1	96	100	100	100
10	99	100	100	100
50	100	100	100	100

表 4.2-4　サンプル漏れの検出結果（%）

n	適合率	再現率
100	**24**	85
200	**51**	87
500	**58**	92

3) 欠損値（表 4.2-3）

全体的に高い精度で検出できた。

4) サンプル漏れ（表 4.2-4）

すべてのパラメータで適合率が 80%に満たず、誤検出が多いことがわかる。

4.3 「異常」を検出できない要因の考察

4 種類の検出結果の内、全てのパラメータで適合率が 80%を下回ったサンプル漏れについて、検出精度の低かった原因を推測する。

サンプル漏れで検出精度が低かったのは n=100 の時で適合率は平均 24%、再現率は平均 85%あった。中でも特に検出精度の低かった評価データ D1 では適合率 22%、再現率 68%であった。

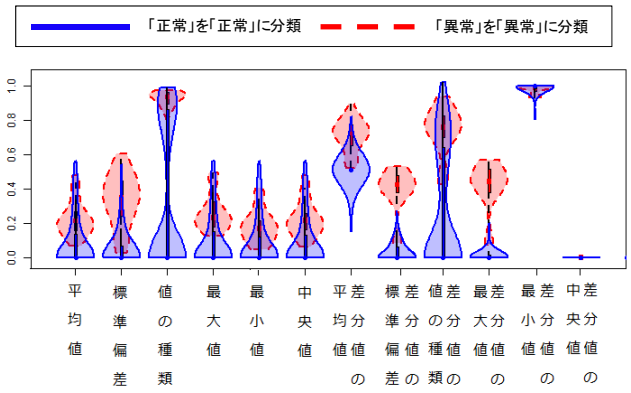

図 4.3-1　正しく検出した評価データの特徴量

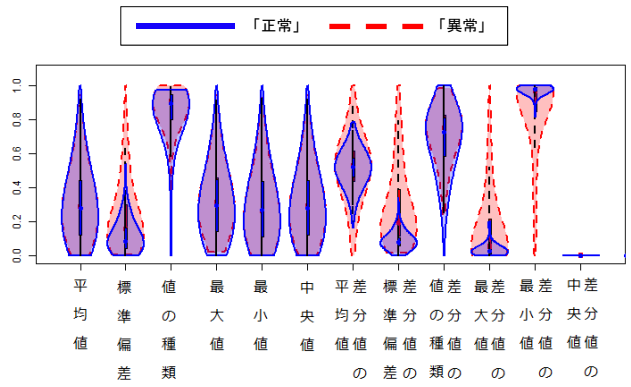

図 4.3-2　教師データの特徴量

評価データ D1 について正しく検出した窓の特徴量の分布をバイオリンプロットで表したものを図 4.3-1 に示す。図形の横に太い部分は特徴量がその値をとる窓が多いことを表している。正しく検出した「正常」の分布と「異常」の分布を比較すると互いの最も太い位置が離れおり、明らかな「異常」しか正しく検出できなかったと言える。この異常検出に使用した教師データをバイオリンプロットで表したものを図 4.3-2 に示す。教師データの「正常」の分布と「異常」の分布を比較すると「正常」の分布を覆うように「異常」が分布しており、重なっている部分が大きかった。「正常」と「異常」の分布が重なっていると、その部分に評価データが分布しても正しく検出するのは難しい。そのため重なりが大きければ大きいほど検出精度が低下すると考えられる。

サンプル漏れの「正常」と「異常」の特徴量に差異が現れにくい原因として、緩やかに減速している場合や、定速走行状態にサンプル漏れが発生した場合、異常値が発生する前後の値の差分が小さいためだと考えられる。強い加減速が発生している場合にサンプル漏れが発生すると、特徴量にも差異が現れると考えられる。

4.4 検出結果のまとめ

ax+b、突発値、欠損値において、適合率、再現率が多くの場合で 90%近い高い検出精度を得られ、異常度合が高いほど検出精度が高いことが分かった。

一方でサンプル漏れに関しては、異常前後の値の差分が小さい場合があり、そのデータが学習データに含まれている場合は検出が難しい。しかし、サンプル漏れが発生している場合は車速以外のデータにおいても同様の異常値が発生していると考えられるため、他のデータと合わせて検出するなど手法の改良をすることで検出が可能になると考えられる。

5. まとめ

本研究ではセンサーデータにおいて発生し得ると考える 4 種類の異常値を挙げ、教師あり機械学習で異常値を検出可能か実験した。特徴量は基本統計量を使用し、検出アルゴリズムに Random Forest を使用することで、新たに入手したデータにも容易に適用できる手法を採用した。

実験結果から、ax+b、突発値、欠損値の 3 種類については高い精度で検出ができ、提案手法が有効であることが分かった。しかし、サンプル漏れの場合は誤検出が多く、提案手法の改良が課題として残った。

実験は車速データのみを用い、教師データと同一条件の評価データを分類する限定的なケースのみであったが、データ解析で大きく影響を及ぼすと考える異常度合が高い異常値を、高精度で検出することが可能であると判明した。これは使用している特徴量が異常値の特徴を捉えているからだと言える。そのため異なるデータ、異なる条件においても異常度合が高い異常値については高精度に検出が可能と考える。

これらの結果から既存の異常値を学習し、新しいデータの異常値を検出できると考えられる。今後の課題として、実際のデータに含まれる異常値が検出可能か検証していく必要がある。本研究では人工的な異常値を対象としたが、実際の異常値には検出できない異常値が存在する可能性がある。

また、本手法は学習データが存在しない状態では使用できないコールドスタート問題を抱える。そのような場合に異常値を検出するには教師なしの機械学習による異常値検出手法も検討が必要となる。

さらに、検出した異常値を学習し次回以降の異常値検出に適用するサイクルを異常値検出フレームワークとして構築することも重要となる。

参考文献.

[1] K. Nawa et al., "Cyber Physical System for Vehicle Application," Proceedings of IEEE CYBER 2012.
[2] 疋田ら,"車両センサーデータを蓄積・活用するためのデータベースシステムの提案", 情報処理学会 DBS 研究会,第 158 回,No.30,2013
[3] Rich Caruana et al, " An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics", ICML'06