日本オペレーションズ・リサーチ学会
2021年 秋季研究発表会

システムにおける異常値などの検知について

01991726　久留米大学　*譚康融

1. はじめに

高度情報化社会になった今日では、システムの異常値や、その他のレア・イベント等の検知・予測、そしてその対応が重要な課題になっている。

コンピュータシステムにおける不正侵入や、ネット攻撃を受けたとき、いち早くそれらを検出し対処するのは急務である。また経済システムの分野においては、種々の経済指標が乖離している傾向を見せているとき、その背後、システムの構造変化 (Structural Change) や、恒常的だと思われたパラメーターの変化 (Change Point) が考えられる。

本研究はこれらの異常値の検出について、いくつかのアプローチを用いて、数値実験の結果により、ハイブリッドな検出方法を提案する。ここで主に従来の判別分析法、SVM(Support Vector Machine)、One-Class SVM、および Bayesian 法との比較を行い、それぞれの方法による異常値などの検出のメリットとデメリットを検討する [1][2][3]。

2. SVM および One-class SVM について

この節では、SVM および One-Class SVM 方法の概要をまとめておく。簡単に

1)C-support vector classification(C-SVC)

2)ν-support vector classification(ν-SVC)

3)Distribution estimation (One-class SVM)

の順番に説明していく。本研究は主に 2 種類の状態しかない問題を考慮する。さらにカーネル手法 (Kernel Trick) が計算処理に適用させた。

一般的にはカーネル手法によって、特徴空間に超平面を生成し異なるデータグループの分離ができるようになる。カーネル関数は $k(x,y) = (\phi(x) \cdot \phi(y))$ と定義されている。色々なカーネル関数が利用されているが、主にガウス、多項式、RBF がよく用いられる。例えば、ガウスカーネルは $k(x,y) = e^{-||x-y||/c}$ となる。ここで c はカーネル関数のパラメータであり、$||x-y||$ は所謂不一致性測度 (dissimilarity measure) である。

1) と 2) は教師信号ありの学習であり、それらの結論を簡単にまとめた。

1) においては、すべての $x_i (i = 1, \ldots, m)$, に対して、$y_i = -1, 1$ とし、その初期問題と双対問題によって以下の決定関数が得られる。

$$f(x) = sgn(\sum_{i=1}^{m} y_i \alpha_i K(x_i, x) + b). \qquad (1)$$

2) においても以下の結論がある。ただし、$\nu \in (0, 1]$ であり、エラーの割合を調整するパラメーターである。決定関数は以下になる。

$$f(x) = sgn(\sum_{i=1}^{m} y_i \alpha_i (K(x_i, x) + b)). \qquad (2)$$

3)One-class SVM

One-class SVM は Schölkopf によって最初に提案された。教師信号なしのデータ $x_i (i = 1, \ldots, m)$ に対して、その初問題の設定は

$$\begin{aligned} \min_{w,\xi,\rho} \quad & \frac{1}{2} w^T w + \frac{1}{\nu m} \sum_{i=1}^{m} \xi_i - \rho \qquad (3) \\ & w^T \phi(x_i) \geq \rho - \xi_i, \\ & \xi_i \geq 0, i = 1, \ldots, m. \end{aligned}$$

となる。またその双対問題は

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \alpha^T Q \alpha \\ & 0 \leq \alpha_i \leq 1/(\nu m), i = 1, \ldots, m, \quad (4) \\ & e^T \alpha = 1, \end{aligned}$$

となる。ここで、$Q_{ij} = K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$. 一方、次のように書き換えることができる。即ち、

$$\begin{aligned} \min \quad & \frac{1}{2} \alpha^T Q \alpha \\ & 0 \leq \alpha_i \leq 1, \qquad i = 1, \ldots, m, \\ & e^T \alpha = \nu m. \end{aligned}$$

同様に Lagrange 法を用いて、決定関数が得られる。

$$f(x) = sgn(\sum_{i=1}^{m} \alpha_i K(x_i, x) - \rho). \qquad (5)$$

3. 数値実験の結果

この節では、二つの実験結果を示しておく。1）正弦関数にノイズを加えた例、2）ネットワークトラフィックの異常値の例。

[数値実験 1: 正弦関数にノイズを加えた例]

図 1 に示されている様に、正弦関数にノイズを加えたが、One-class SVM によって、異常値の検出を試みたが、大きな異常値はほぼ検出されて、黒い三角マークで表されている。

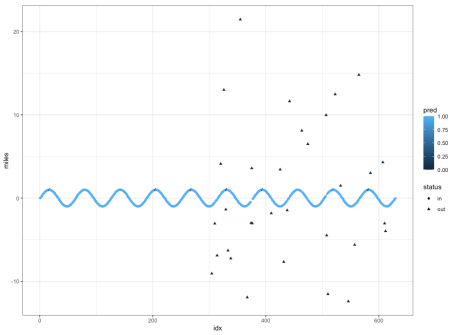

図 1: Sine wave with outliers

繰り返しシミュレーションの結果を踏まえて、平均正解率は約 70-80%となる.

[数値実験２:ネットワークトラフィックの異常値]

実験の設定：通常トラフィック量は N(100,10) に従うとし、通常は 100 前後となるが、時にはネット攻撃を受け、例えば、DoS 攻撃を受け、 700 になったりする。システム管理者はいち早くその事象をキャッチし、解析してアクションを起こさなければいけない。図２ではトラフィックの変化を示している。設定した 12 点の異常値は７００前後になっていて、One-class SVM を用いた解析した結果は、12 点全部を異常値として検出し認識できた一方、稀に通常期の高いスポット値も異常値として挙げられていることが指摘できる。

4. デスカッション

SVM などの教師信号のある分類法は、一般的には精度がやや高いと思われる。しかし現実の環境では必ずしも最初から教師信号が得られる訳ではない。そのため、教師信号なしの One-Class SVM がより精度の高い検知ができる。一方、比較的に長いスパーンで変動している場合は、MCMC を利用した Bayesian 推論の方が効率よく、パラメーターの変換点を推定できるが [3]、他方、1 番例の様に異常値がバラバラ撒いている場合は検出はやや難しい。よって、これら検出方法の特性を活かして複数の方法の併用が可能なハイブリッド検出法が望ましく、よりよい検知の結果を得られると期待できよう。

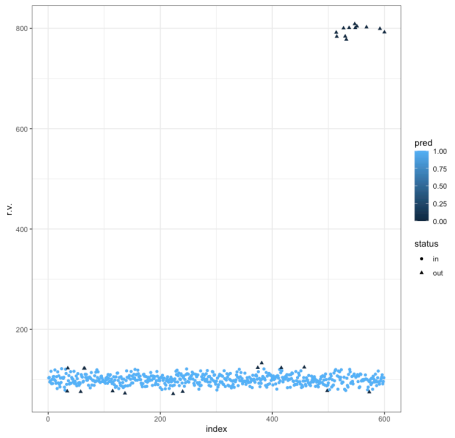

図 2: predict values

5. おわりに

本研究はシステムに発生し得る異常値などの検出について、従来の諸方法 (SVM、One-Class SVM, Bayesian 推論) についてのメリットとデメリットについて検討した。従来、ケースバイケースで特定の方法を用いることは多いが、しかしながら、これらの方法を同時に実施し、それぞれのアプローチによって、得られた結果を総合的に吟味することにより、より良い精度の高い結果が得られるではないかと思われる。本研究の一部は、JSPS 科研費の助成 (基盤研究 (c)18K04626,) によるものであり、ここで JSPS に感謝致します。

参考文献

[1] V. Vapnik, Statistical learning theory, Wiley, New York, 1998.

[2] B. Schölkopf, J. Platt et al., Estimating the support of a high-dimensional distribution, *Neural Computation*, 13, 1443-1471,2001.

[3] K.R. Tan, Detecting structural changes in stochastic differential equation system based upon a Bayesian approach, *Journal of Institute of Industrial Economics Research*, 58(1-2), 51-67, 2018.