1B-01

周期変動や急激な変化に重畳するインパルス状外れ値の除去アルゴリズムの開発

中原　崇†

株式会社日立製作所　研究開発グループ†

1. はじめに

自然現象を計測し制御や過去の類似現象表示に役立てるためにセンサが使用される。例えば気温センサは一日単位で周期的に変動する気温を計測して空調制御の指標に活用され、流量センサは豪雨などにより河川の流量が急激に増加し、やがて晴天時の定常状態に戻る現象を捉え、安全管理に活用される。

これらの計測において、校正や電気系の不良などによりセンサ値を突発的に変化させるインパルス状の外れ値が発生することがある。具体的には数十以上のサンプル単位で時間差分値の符号が反転するセンサにおいて、インパルス状の外れ値により数サンプル単位で時間差分値の符号が反転する。外れ値を除去するにあたり、時間差分値の閾値判定や周波数操作では、周期的または急激に変動するセンサ値上の外れ値の除去と真値の両立が困難である。学習するにも教師データが不足し、作成工数が発生する。

従来の閾値や学習による除去手法では生じやすい誤検知を防ぐため、外れ値の時間差分値がクラスタリングにより少数のクラスタまたは中心から外れたクラスタに属すると仮定して統計的に除去する、教師なし外れ値除去アルゴリズムを開発し、模擬センサ値を用いて評価した。

2. 外れ値除去アルゴリズム

周期的または急激に変動するセンサ信号において、数サンプル単位で時間差分値の符号が反転しインパルス状に重畳する外れ値を除去することが課題である。数サンプル単位で符号が反転する外れ値の時間差分値も、十倍以上のサンプル数の時間差分値の大きさとしては例外であることに着目する。

教師なしアルゴリズムであるクラスタリングを用いて例外を検知し除去する外れ値除去アルゴリズム(以下提案手法と称す)を考案した。

外れ値除去アルゴリズムの動作原理について Fig. 1 に示す。センサの時間差分値を計算し、最適クラスタ数を自動推定可能な X-means[1]法により一定期間の時間差分値にクラスタリングを施す。クラスタ内の時間差分値が少数の場合と、それ以外のクラスタにおいてクラスタ中心からの距離が閾値よりも大きい場合の時間差分値を外れ値と判定する。

その結果、元データ上昇時の場合、元データ一定時に生成したクラスタ1に加えて、新たにクラスタ2が生成される。外れ値の属するクラスタは少数の時間差分が属するクラスタとして削除される為、残った全てのクラスタとも距離が離れている点を外れ値と判定される。

なお、時間差分値が外れ値と判定された時刻では前後の外れ値ではない時間差分値で補間して外れ値を除去する。閾値については一定期間における時間差分値の標準偏差の一次式で表す。

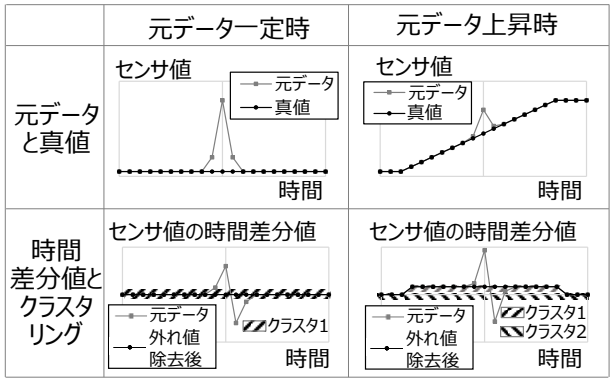

Fig. 1　外れ値除去アルゴリズムの動作原理

3. 評価方法と結果

3.1 評価方法

外れ値除去アルゴリズムの評価にあたり、自然環境で観察される周期変動や急激な変化を抽象化した以下2点の模擬データ値を作成した。

(A)周期変動データ：一日単位で周期的に変動する気温を式(1)により抽象化した。

Development of an algorithm to remove impulsive outliers that superimposed on periodic fluctuations and abrupt changes† Takashi Nakahara, Hitachi, Ltd., R&D Group.

$$y(t) = (T_{max} - T_{min})\sin\frac{t}{t_d} + T_{min} \cdots (1)$$

ここで、T_{max}は気温の最大値、T_{min}は気温の最小値、t_dは気温の周期を示す。サンプリング周期ΔT_1を$t_d/288$とする。

(B) 急激変動データ：豪雨などにより貯水池の水の流量が急激に増大し定常状態に戻る現象を、式(2)により抽象化した。

$$y(t) = \begin{cases} (A_1 - A_0)e^{-\frac{(t-t_s)}{t_x}}\left(1 - e^{-\frac{(t-t_s)}{t_y}}\right) + A_0 & (t \geq t_s) \\ 0 & (t < t_s) \end{cases} \cdots (2)$$

ここで、A_0は流量の最小値、A_1は流量の極大値、t_xは減衰の時定数、t_yは 1 次遅れの時定数、t_sは立ち上がり時刻を示す。サンプリング周期ΔT_2を$t_x/10$とする。

以上、式(1)と式(2)のデータに重畳した外れ値の仕様を Table 1 に示す。

Table 1 外れ値の仕様

変動種類	外れ値 No.	模擬対象となる現象	重畳時刻	重畳期間	外れ値 最大/最小値
(A)周期変動	#1	真値増大時における、校正による外れ値の発生	$\frac{1}{12}t_d$	$5\Delta T_1$	$2(T_{max} - T_{min})$
	#2	真値減少時におけるセンサの信号線の断線	$\frac{5}{12}t_d$	$2\Delta T_1$	0
(B)急激変動	#1	流量立ち上がり時における、センサの信号線の接触不良	t_s, $t_s + \Delta T_2$	ΔT_2 $(2Times)$	$(A_1 - A_0)$, $1.2\ (A_1 - A_0)$
	#2	真値減少時における、校正による外れ値の発生	$4.5t_s$	$5\Delta T_2$	$0.5\ (A_1 - A_0)$

各模擬データと真値のトレンドグラフを Fig. 2 に示す。各模擬データに先述の閾値判定手法と提案手法を施し、誤差率を比較することで提案手法を評価した。各手法の評価条件表を Table 2 に示す。

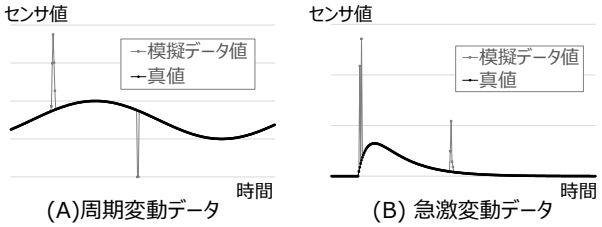

Fig. 2 模擬センサ値のトレンドグラフ

Table 2 評価条件表

ケース番号		1	2	3
手法		閾値判定手法		提案手法
閾値	(A)周期変動	$\frac{2(T_{max} - T_{min})\Delta T_1}{t_d}$	$\frac{8(T_{max} - T_{min})\Delta T_1}{t_d}$	0.5σ+0.5 (σ:一定区間の時間差分における標準偏差)
	(B)急激変動	$\frac{10(A_1 - A_0)\Delta T_2}{t_x}$	$\frac{40(A_1 - A_0)\Delta T_2}{t_x}$	

ここで、誤差率e (%)を式(3)で示す。$x(t)$はセンサ値、$x_c(t)$は真値、x_{max}はセンサ値の最大値、$\Delta T(= \Delta T_1\ or\ \Delta T_2)$はサンプリング周期、$N$は総サンプル数を示す。

$$e = \frac{100}{N}\sum_{i=1}^{N}\frac{|x(i\Delta T) - x_c(i\Delta T)|}{x_{max}} \cdots (3)$$

3.2 適用結果

閾値判定手法と提案手法を模擬データ値に適用したときの外れ値除去結果を Fig. 3 に示す。

(A)周期変動データに関して、ケース 1 では閾値が小さく、ケース 2 では閾値が大きいことから外れ値を除去できない部分がある。ケース 3 では真値が増大・減少する場合でも外れ値を除去できることを確認した。一定期間における時間差分値のうち、時間差分値が多数含まれるクラスタとして真値を抽出し、時間差分値が少数属するクラスタを削除して残った全クラスタとの距離で閾値判定しており、既存のセンサ値だけで外れ値を適切に検知しているためと考える。

(B)急激変動データに関しても、ケース 3 で外れ値の除去を確認した。急激な真値の変動における時間差分値を多数含むクラスタとして抽出でき、誤検知を防いでいるものと考える。

また、周期変動データでは誤差率が元データの 1.43%、ケース 1 の 3.06%、ケース 2 の 0.09%に比べてケース 3 では 6.38×10^{-4}%に低減した。一方、急激変動データでは、誤差率が元データの 3.12%、ケース 1 の 12.8%、ケース 2 の 0.64%に比べてケース 3 では 0.17%に低減した。よって、ケース 3 の提案手法が有効であることを確認した。

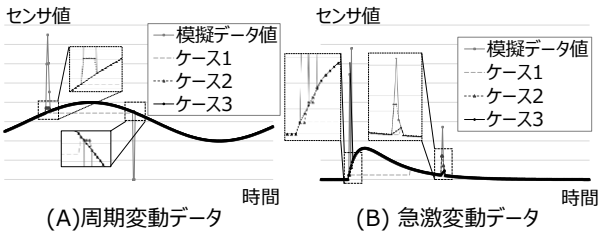

Fig. 3 外れ値除去結果

4. まとめ

自然環境でよく観察される周期変動や急激な変化に重畳するインパルス状の外れ値に対して誤検知を防ぐ為、X-means 法による教師なし外れ値除去アルゴリズムを開発した。外れ値を重畳したデータに本手法を適用した結果、閾値判定手法では困難な外れ値の除去と真値の維持を両立できることを確認した。

参考文献

[1] Pelleg, D., et. al. "X-means: Extending K-means with efficient estimation of the number of clusters", ICML-2000