(10) 機械学習による複数種類の正常値・異常値を含む複雑データに対する同時分類技術の開発

木村　延明[1]・馬場　大地[2]

[1]正会員　農研機構　農村工学研究部門（〒305-8609 茨城県つくば市観音台 2-1-6）
E-mail: kimuran590@affrc.go.jp

[2]非会員　㈱アーク情報システム アドバンストテクノロジー部（〒102-0076 東京都千代田区五番町 4-2）

本研究では，計測機器の不具合などで発生する複数種類の異常値や異なるパターンの正常値を有する時系列データの品質保証を行うために，機械学習を用いた複数種類の異常値・正常値を同時に分類する技術開発を行った．農業水利施設等で観測される水位データは，主にスパイクノイズやスライドずれの異常値や正常値でも，常時と洪水時の異なる水位変化が見られる．本研究では，これらの 4 項目を同時に分類可能な自己組織化マップ（SOM）を導入して分類を行い，クラスタリング手法（K-means 法，Ward 法，多数決法）を用いて，4 値分類における分類精度の評価指標（f1 値）を計算し，分類状態を 2 次元マップ上に可視化した．f1 値では多数決法の値が相対的に良好で，且つ，可視化でも真値が多数決法でクラスタリングされた領域に概ねプロットされることを示した．

Key Words: *self-organizing maps* (*SOM*), *water leveal data, anomaly detection, multiple classification*

1. 背景と目的

近年，情報通信技術（ICT）の発展で大量のデータ収集が可能になった．しかし，計測センサーやデータ通信機器の不具合等でランダムに発生する異常値がデータに含まれることが多く，異常値をマンパワーで取り除くためには膨大なコストが必要である．機械的に異常値の除去を行うために，統計手法を含む様々な手法が開発されてきた．例えば，正常値 or 異常値（つまり，Yes or No の二値）で判断する検知技術（以下，「二値分類」という）が開発されている．しかし，複数のタイプの異常値や異なるパターンの正常値を有するデータに対して，二値分類の適用は難しい．従って，二値分類のみでは各々の種類を分類できない課題がある．この課題解決のために本研究で用いる手法について，統計手法と比べて，時系列パターンの特徴を捉え，長期・短期のトレンドや季節変動を学習することができる機械学習にフォーカスする．

機械学習を利用し，時系列データの異常検知技術に関連した国内外の他研究との対比を検討する．例えば，脈拍のような規則性のある身体計測信号について，画像分類型の深層学習を用いて正常信号の波形図を学習し，それから外れた信号波形図を異常値と見なす手法[1)]や時系列予測に特化した深層学習を用いて，正常値のみを学習し，予測値が観測値から乖離した場合に異常値と判断できる推定器を開発した事例[2)]等がある．しかし，異常値と正常値，さらにそれぞれの複数種類を同時に行う分類（以下，「多値分類」という）手法を実装した異常検知技術は現時点で見られない．

複数種類の正常値と異常値を有する時系列データに対して，各々の種類を同時分類するために教師なし機械学習アルゴリズムの自己組織化マップ（Self-organizing maps, SOM）[3)]の導入を試みる．近年，SOM は様々な気象データをデータの特徴に基づいて，台風等の異常気象や梅雨期の豪雨関連の気象を同時に分類する試みがなされている[4)]．他方，農業水利施設等の現地観測で収集されるような水位データでは，異常値は必ずしも一種類ではない（例：スパイクノイズやスライドずれ等）．加えて，正常値であっても降雨による出水時の水位変化と無降雨状態の水位変化では時系列パターンが異なる．出水時の水位の値の中には，スパイクノイズの値と区別ができない場合もある．このような複雑な特徴を持つデータに対して，異常値のみを検出するのではなく，正常値の種類も含めて各々の種類を同時分類できるものが SOM と仮定する．

本研究では，時系列データに対して，類似性のある複数のタイプの正常値・異常値を同時に多値分類できるSOM 型の機械学習を導入した異常検知技術（**図-1**）を開発することが目的である．

2. 方法

(1) 利用データ

低平地で利用される排水機場や用排水路等の農業水利施設で現地観測されるような水位データを対象にする．データの特性は**図-2** に示すような複数の異常値（スパイクノイズ，スライドずれ）と正常値でも異なるパターンのデータ（通常時の水位，洪水イベントの水位）を有するものである．精密センサー等で実際に観測されるデータでは，異常値を含むことが少ないので，**図-2** の特徴を一定の割合で含む仮想的な時系列データ（2 万点）を人工的に生成した（**図-3**）．このデータでは，現地観測データにおいて 1 時間おきに計測された時間観測に相当する．スパイクノイズとスライドずれはランダムに発生させ，それらの個数の全区間に占める割合は，7～8%であり，洪水イベントも同様の割合（22 個の洪水イベントに相当）である．なお，**図-2** の正常値 A は，低平地の排水機場や用排水路で規則的に水位調整を行う際に観測されるような水位の微小変化である．本研究では，この仮想データについて，異なるタイプの異常値・正常値の分類を行う．

(2) 機械学習モデル

異常検知技術の機械学習アルゴリズムにデータのパターンから類似分類を行う自己組織化特徴マップ（SOM）を用いる．SOM は，教師なしニューラルネットワークで，入力層と競合層からなる 2 層構造を有する．SOM の機能的な特徴について説明する．**図-4**（左図）に示すように，一般的に任意の数のノードが 2 次元マップのように配置され，水位や雨量等の変数の組み合わせで，ベクトル化された入力データが，その特徴量に基づき，類似性のあるノードの近傍に配置される．一方，対象の入力データが非類似的であれば，距離が離れた位置のノードに配置される．SOM の分類機能の特性から**図-4**（右図）のように計 4 項目の異常値と正常値を含む観測データについて，それぞれの項目に分類（4 値分類）できる可能性がある．

一般的に SOM を使った分類では，入力データの設定が良好な分類のカギになる．時系列データは過去から連続的に変化するものなので，過去情報を生かすために，判別を行う対象値の現時点から，過去に 10 ステップ遡ったデータを一括りとして，11 個の値で 1 つのベクトル化された入力データを生成する（**図-2**）．これを，時

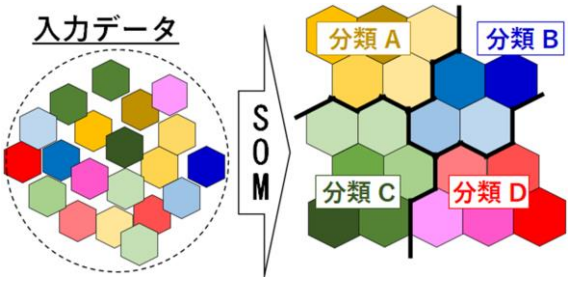

図-1 提案する分類手法の概念と処理フロー

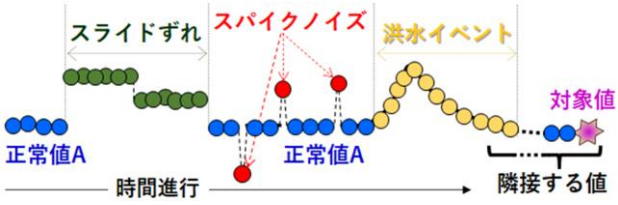

図-2 複数種類を含む典型的な水位の時系列データ

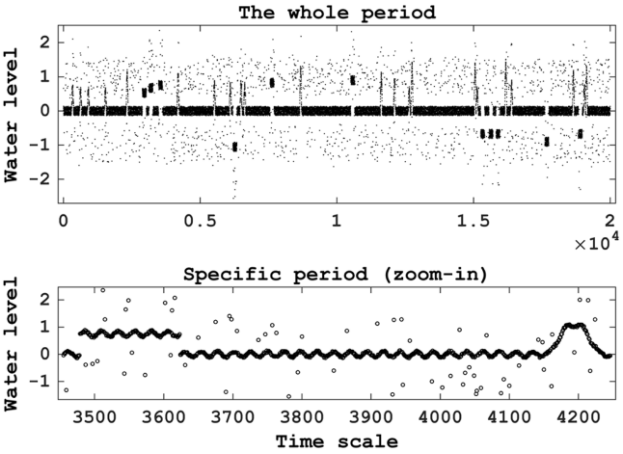

図-3 本研究で用いた仮想時系列データ（上：全区間，下：スパイクノイズ，スライドずれ，洪水を含む区間）

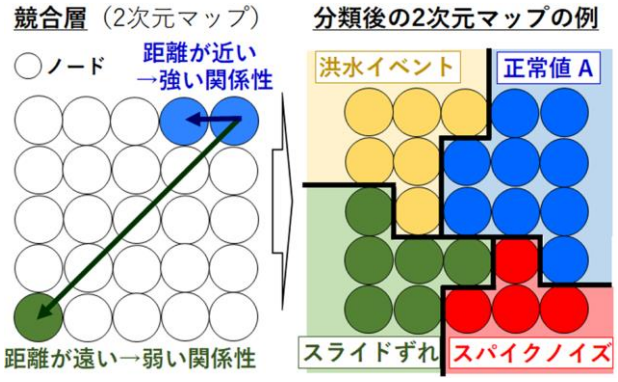

図-4 SOM の分類方法．左図：類似性が強い場合に近傍に，弱い場合に遠方に配置されるイメージ図，右図：理想的に分類された場合のイメージ図．

間ステップ毎に行う．なお，この過去に遡るステップ数は試行錯誤を行って決定した．また，SOM の学習回数は 10 回とした．

(3) クラスター手法

SOM によって判別された各時系列パターン（判別対象の値と隣接する 10 個の値を含む入力データ）をノード数が 50×50 の 2 次元マップ上に振り分け，近接したノードが互いに類似した特徴を持つように配置される．近接の複数のノード群は，ある特徴を持ったクラスターを形成していると考えられる．なお，50×50 の 2 次元マップは，本研究で用いるデータ数から適切な値を決定した．SOM で得られたマップを可視化するためにクラスター化を行い，K-means 法 [5]，Ward 法 [6]の各手法を用いる．加えて，本研究では，精度向上のために多数決法を導入した（**図-5**）．ここで用いた多数決法とは，1 つのノードに複数の項目が重なった場合に，最大数の項目を選択し，ノードの代表的な項目とするものである．

(4) 評価指標

正常値と異常値で二値分類する場合，異常値の検知精度を定量評価するために，f1 score（f1）を導入する．f1 は，異常と予測したデータの内，実際に異常である割合（precision）と実際に異常であるデータの内，異常と予測できた割合（recall）で定義される．これらの関係式は，**表-1** の正常値・異常値のポジティブ（p）/ネガティブ（n）の定義に基づき，調和平均として次式のように示される．

$$f1 = \frac{2 \times precision \times recall}{precision + recall} \quad (1)$$

ここで，precision＝tp/(tp+fp)，recall＝tp/(tp+fn)である．しかし，本研究では，４項目（正常値 A，洪水イベント，スパイクノイズ，スライドずれ）について，４値分類を行うので，統計的分類における混同行列は**図-6** に示されるように，4×4 の行列となる．各項目について，2×2 の行列に帰結して（**図-6**），各々の f1 を求める．さらに，個数による重み付きを考慮すれば，次式のように macro f1 が定義される．

$$f1_{macro} = f1_{nA} w_{nA} + f1_{nF} w_{nF} + f1_{aS} w_{aS} + f1_{aZ} w_{aZ} \quad (2)$$

ここで，w＝各項目の重み付き係数，f1＝各項目の調和平均を示し，添え字 nA=正常値 A，nF=洪水イベント，aS=スパイクノイズ，aZ=スライドずれをそれぞれ示す．

３．結果

４値分類における分類精度を定量的に示す $f1_{macro}$ を計算し，実際に入力データとしての時系列パターンが，3 つのクラスタリングを用いて，2 次元マップ上にどのように配置されるかを確認した（**図-7**）．４項目の正誤個数を示した図表では，K-means 法，Ward 法ともに真値の正常値Ａについて，多くがスパイクノイズと誤って判断

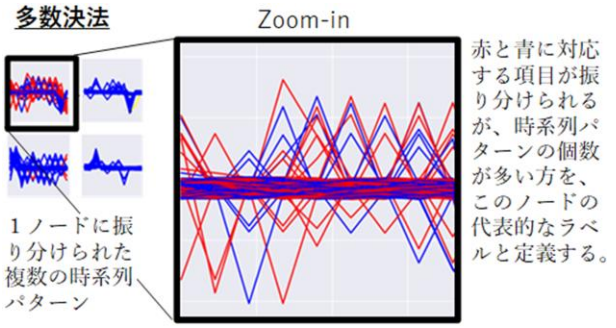

図-5　多数決法のイメージ図

表-1　二値分類の混同行列

		予測値	
		異常（＋）	正常（－）
真値	異常（＋）	tp 異常を正しく異常と判定したデータ数	fn 異常を誤って正常と判定したデータ数
	正常（－）	fp 正常を誤って異常と判定したデータ数	tn 正常を正しく正常と判定したデータ数

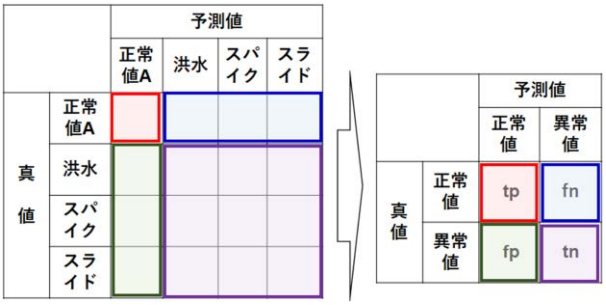

図-6　４値分類の混同行列と二値分類の混同行列の関連性
正常値Ａの $f1_{nA}$ の導出の事例

されている．洪水の真値は，概ね予測値の洪水と一致している．スパイクノイズの真値は，半数以上が正常値Ａと誤って判断されている．スライドずれの真値は，約半数が洪水と誤って予測されている．一方，多数決法は，各項目の真値は，概ね正しく予測されている．これらの正誤個数から計算された定量的な評価指標 $f1_{macro}$ は，それぞれ K-means 法＝0.56，Ward 法＝0.58，多数決法 ＝0.90 であった．さらに，2 次元マップ上に可視化される分類結果について，K-means 法，Ward 法ともにクラスタリングされた領域（４項目毎に色分けされた領域）に真値である各項目（色付きのドッド）の配置が約半数の割合で一致しなかった．特に，真値の正常値Ａは，スパイクノイズの領域に半数以上が配置された．一方，多数決法では，各項目のクラスタリングされた領域と真値の配置が概ね一致した．これらの結果から，クラスタリングで多数決法を用いることで，相対的に良好な真値と予測値の一致が見られた．

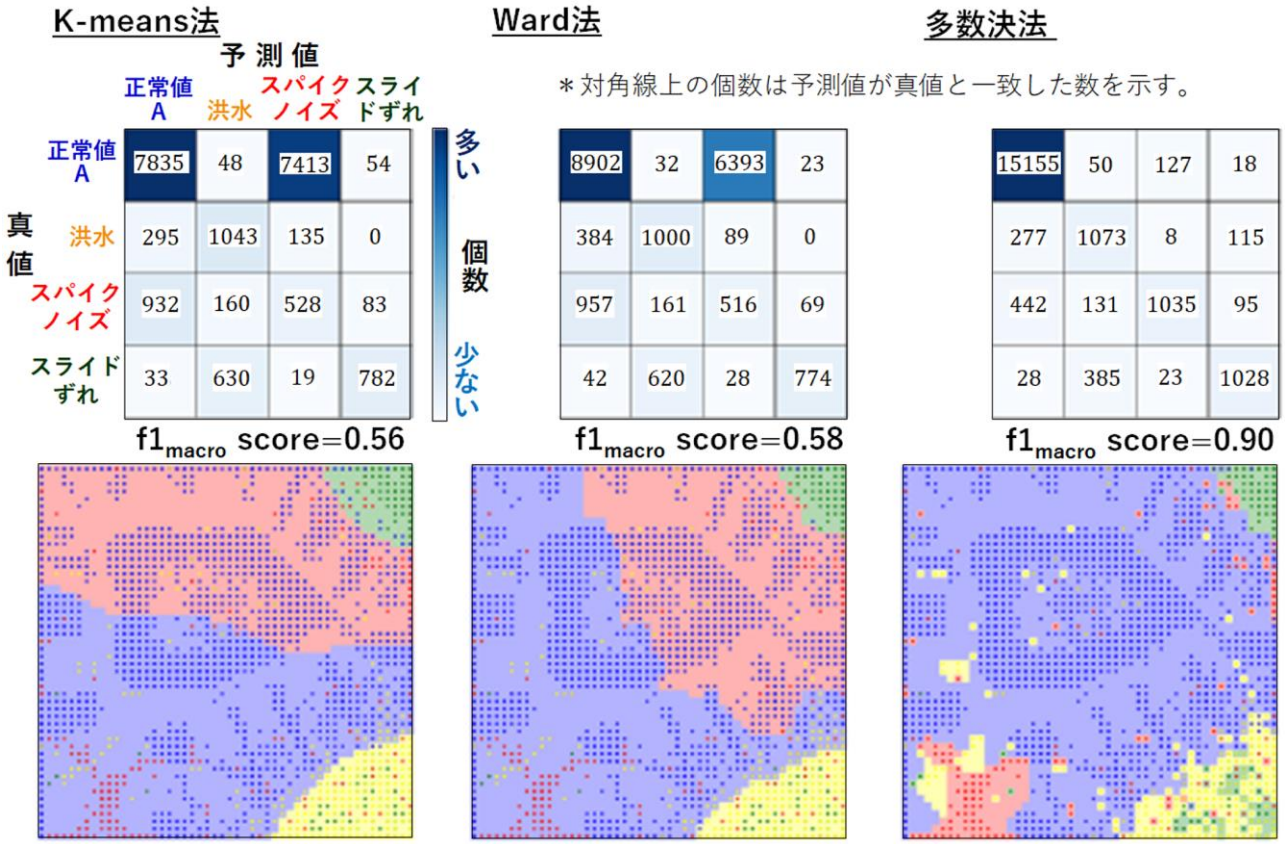

図-7　SOM とクラスタリング手法によって分類された正誤個数の表（上段）と 2 次元マップ（下段）：ドット（点）の色＝真値；背景の色＝クラスタリングの結果．青＝正常値 A，黄色＝洪水イベント，赤＝スパイクノイズ，緑＝スライドずれ．

4．まとめ

本研究では，時系列データに対して，複数種類の正常値と異常値を同時に 4 つの項目（正常値 A，洪水イベント，スパイクノイズ，スライドずれ）に分類可能な SOM を実装した異常検知技術の開発を行った．時系列データの判別対象の値を含む 10 個のデータを入力データ（時系列パターン）とし，3 つのクラスタリング（K-means 法，Ward 法，多数決法）を用いた時系列パターンの類似性を 50×50 のマップ上で可視化した．成果を以下にまとめる．

・4 値分類を二値分類に帰結して計算される，macro f1 score は多数決法が高いスコアーを示した．

・K-means 法と Ward 法ともにクラスタリングされた領域に真値である各項目が配置されなかったものの，多数決法は概ねクラスタリング領域と真値の配置が一致した．

今後の課題として，実用に耐え得るような予測精度（f1≈1.0）に向けて，隣接する値の個数やノード数の最適な個数の探索，さらに，時系列データの特徴をより詳細に抽出できるような機能（例：Encoder-Decoder ネットワーク）を SOM の前処理機能として導入するなどして，分類の予測精度の向上を計る必要がある．

謝辞：本研究は，農林水産省委託プロジェクト研究「AI 等の活用による利水と治水に対応した農業水利施設の遠隔監視・自動制御システムの開発」JPJ009837 の支援を受けて実施した．ここに記して深謝の意を表す．

参考文献

1) Tang, Z., Chen, Z., Bao, Y. and Li, H. : Convolutional neural network‐based data anomaly detection method using multiple information for structural health monitoring, *Structural Control and Health Monitoring*, Vol.26, No.1, 2019.
2) 一言正之，川越 典子，橋田創，清雄一，房前和朋：水位推定誤差の確率分布に基づく河川水位観測データのリアルタイム異常検知，土木学会論文集 B1（水工学），Vol.75，No.2，pp.I_193-I_198，2019.
3) Kohonen, T. : Self-Organized Formation of Topologically Correct Feature Maps, *Biological Cybernetics*, Vol.43, No.1, pp.59–69, 1982.
4) 西山浩司，白水元，朝位孝二：自己組織化マップに基づく九州地方における豪雨の発生時間帯の傾向に関する分析，土木学会論文集 B1（水工学），Vol.77，No.2，pp.I_1135-I_1140，2021.
5) Vesanto, J. and Alhoniemi, E. : Clustering of Self-Organizing Map, *IEEE Transactions on Neural Networks*, Vol.11, No.3, pp.586-600, 2000.
6) Ward, J. H., Jr. : Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, Vol.58, pp.236-244, 1963.