

# F-032 天文測光データにおける突発検知問題を対象とした ウェーブレット変換と DeepAnT の可用性調査

西川 侑志, Thanapol Phungtua-eng, 山本 泰生

静岡大学 総合科学技術研究科

## 1 研究概要

近年、情報化社会の発展により様々な分野で大量のデータが日々生成されている。本研究では、タイムドメイン天文学において観測される測光ビッグデータを扱う。このデータは高時間分解能の天文観測から得られるものだが、データ量が膨大で、専門家の手による既存の解析パラダイムでは処理が追いつかないという問題点がある。例えば、木曾観測所では、シュミット望遠鏡に搭載された CMOS カメラ Tomo-e Gozen [1] を用いて、毎晩 30TB もの測光データを取得しており、これをリアルタイム解析することが求められている。

このような背景の中、機械学習やデータマイニングの諸技術を用いることで、大量データの中に潜む価値のある天体現象を発見する取り組みが進められている。例えば、先行研究 [2] では、動的ビンニング処理に基づくオンライン異常検知法を提案しており、測光データ中の突発的な変動現象を検出する問題に適用している。

本研究では、時系列データに対する深層学習モデルベースの異常検知法として知られる DeepAnT [3] を用いて、突発現象をどの程度検出できるか実証的に評価する。さらに、汎用の時系列解析技術であるウェーブレット変換 [4] を用いたノイズ除去の効果を DeepAnT に適用し検証する。本稿では、はじめに DeepAnT とウェーブレット変換を概説し、次に、生データとノイズ除去後のデータそれぞれにおいて生成した DeepAnT モデルの性能評価を示す。

## 2 背景

### 2.1 可用性調査の狙い

先行研究 [2] では、PAA, SAX, MatrixProfile 等のデータ要約法に基づく突発検知の性能を実証的に評価している。一方、時系列データを対象とする異常検知法には問題依存性があることが指摘されている [6]。本研究では深層学習モデルに基づく代表的な異常検知法である DeepAnT に焦点を当て、測光データの突発検知問題への応用可用性を明らかにする。

測光データは、大気ゆらぎ、気象条件、測定誤差等の多くのノイズを含む。ノイズ成分から突発的なバースト信号を検出するため、本研究ではウェーブレット変換を用いたノイズ除去を試みる。このように測光データに適したデータ前処理及び異常検知法を探索することが本研究の目標である。

### 2.2 DeepAnT

DeepAnT は、深層学習モデルを用いた時系列データの異常検出法 [3] である。回帰を行う時系列予測器 (前段) と回帰予測をもとに異常判定する異常検知器 (後段) の 2 つのモジュールから構成される。前段では、スライディングウィンドウをもとに過去の時系列データ (サイズ  $w$  のウィンドウ) から次のタイムスタンプの値 ( $p_w$ ) を予測する。後段では、真値と予測値をもとに正常か異常かを判定する。

時系列予測器のアーキテクチャを説明する。入力層は、ウィンドウ内の  $w$  個の時系列データをベクトルに変換するため、 $w$  個の入力ノードを持つ。2 つの畳み込み層とそれに続く最大プーリング層が使用されている。各畳み込み層は 32 個のフィルター (Conv1D) とそれに続く

要素ごとの活性化関数 (ReLU 関数) から形成されている。アーキテクチャの最後は FC 層 (完全結合層) であり、各ニューロンは前の層の全てのニューロンと結合されている。出力層で使用されるノード (ノード数 1) は  $p_w$  を与える。

損失関数として、平均絶対誤差 (MAE) (式 (1)) が採用されている。真値と予測値の誤差を小さくすることによって、ネットワークは時系列の正常な挙動を予測するように学習することが可能になる。 $y_j$  は真値で、 $\hat{y}_j$  は予測値である。 $n$  は訓練データのサンプル数から過去の時系列データ ( $w$ ) を引いた値である。

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j| \quad (1)$$

異常検知器としてのモジュールにおいて、時系列予測器での予測値が本モジュールに渡され、真値と予測値の差が計算され、不一致の尺度 (異常スコア) として、式 (2) で与えられるユークリッド距離が使用される。 $y_t$  は実測値で、 $\hat{y}_t$  は予測値である。

$$abnormal(t) = \sqrt{(y_t - \hat{y}_t)^2} \quad (2)$$

この手法は回帰モデルに基づく教師なし異常検知法であるので、正解のわからない未知の突発現象の検知にも適用できると考えられる。

### 2.3 ウェーブレット変換

ウェーブレット変換は、タイムスタンプごとに周波数特性を抽出する手法である。この手法では、時系列データを窓関数と信号との相互相関関数として表現する。周波数に応じて窓幅を変化させるウェーブレット変換には、連続ウェーブレット変換と離散ウェーブレット変換の 2 種類が存在する。

ウェーブレット変換では、まずウェーブレット関数 ( $\psi_{a,b}(t)$ ) のスケール係数 ( $a$ ) を変化させ、時間軸方向の変数 ( $b$ ) を平行移動した関数 (式 (3)) を考える。ただし、 $R$  は実数全体を、 $L^2$  は二乗可積分関数全体を表している。

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right), \quad (3)$$

$$\psi(t) \in L^2(R), a > 0, b \in R$$

このとき、ウェーブレット変換は、元信号 ( $x(t)$ ) とウェーブレット関数 ( $\psi_{a,b}(t)$ ) の内積を用いて、式 (4) のように定義される。 $\overline{\psi_{a,b}(t)}$  は  $\psi_{a,b}(t)$  の複素共役である。

$$W(a,b) = \int_{-\infty}^{\infty} x(t) \overline{\psi_{a,b}(t)} dt \quad (4)$$

本研究では、スケールパラメータが離散値を取るため、離散ウェーブレット変換を用いる。また、突発信号の検出課題を想定し、本稿では、ウェーブレット関数として Haar (式 (5)) を用いる。

$$\psi(t) = \begin{cases} 1 & (0 \leq t < \frac{1}{2}) \\ -1 & (\frac{1}{2} \leq t < 1) \\ 0 & (\text{elsewhere}) \end{cases} \quad (5)$$

ウェーブレット変換 (式 (4)) を行うことで、元信号から係数  $W(a,b)$  を取得する。この  $W(a,b)$  に対してス

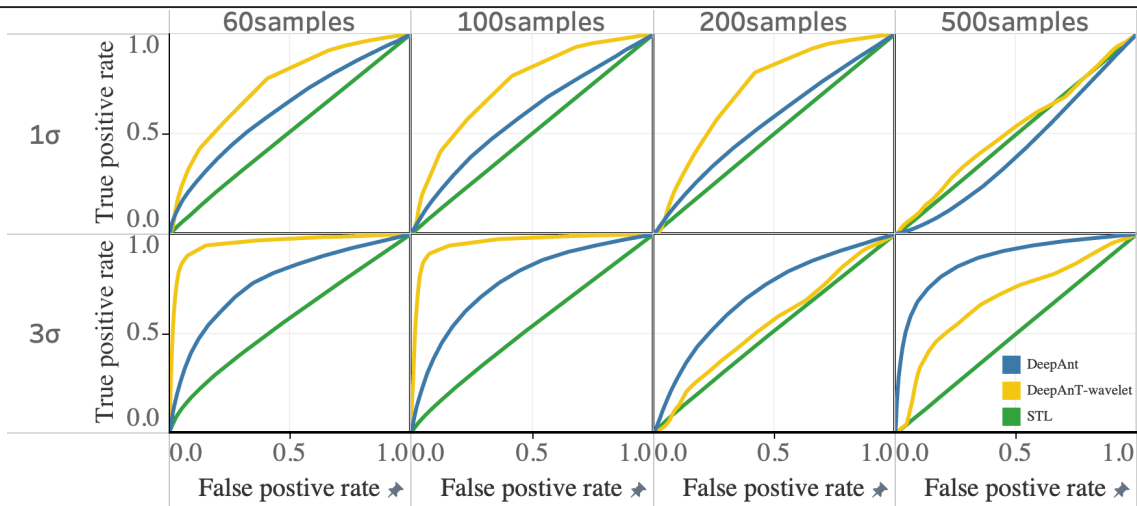


図1: 各パターンごとのROCカーブ

ケーリング係数 (レベル) 毎の閾値処理 [4] を適用し、逆ウェーブレット変換を通してノイズ成分を除去した信号成分を再構成する。

逆ウェーブレット変換を式 (6) に示す。

$$\hat{x}(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(a,b)\psi_{a,b}(t)dadb \quad (6)$$

そして、逆ウェーブレット変換 (式 (4)) を行うことによって、元信号の近似が可能となる。したがって、生データを用いた DeepAnT の場合より、高精度に異常が発見できるものと考ええる。

### 3 実験設定

実験で用いるデータは、Tomo-e Gozen で得られた星の明るさを表す測光データである。生データとして、108 個のデータファイル (各ファイルは、4,260 タイムスタンプを含む) があり、このデータには 1 種類の人工的な突発信号がランダムに挿入されている。挿入する突発信号の振幅は、各データの標準偏差を  $\sigma$  とすると、 $1\sigma$  と  $3\sigma$  の 2 種類である。挿入する異常値としての標準偏差とは、各星データの全サンプルにおける標準偏差である。さらに、信号の区間 (サンプル数) として 60, 100, 200, 500 の 4 種類を用意する。前述の 8 種類の挿入パターンがあるので、合計で 864 件のデータがベンチマークとなる。

しかし、DeepAnT ではウィンドウサイズを  $w$  とすると、 $1$  から  $w$  までの予測ができないので、その間の異常は検出することができない。その上で、本実験では  $w$  を 500 とし、異常値が 500 以降であるデータファイルを選定した。選定したベンチマーク数は、表 1 より合計で 773 件となった。今回の実験では、訓練データとしてベンチマークの元データを用意した。一方、テストデータは人工的に突発信号を付加している測光データである。ウェーブレット変換によるノイズ除去を実施する場合、この訓練データとテストデータに対してウェーブレット変換及びベイズによるノイズ除去、閾値ルールを実施する。そして、そのデータを DeepAnT に適用する (DeepAnT-wavelet)。

表 1: 本実験におけるベンチマーク数

標準偏差 ( $\sigma$ )	1	1	1	1	3	3	3	3
区間	60	100	200	500	60	100	200	500
ベンチマーク数	101	96	97	94	93	93	98	101

### 4 実験結果と考察

DeepAnT, DeepAnT-wavelet, および Thanapol による STL [2] の研究成果を、ROC カーブを用いて比較する。ROC カーブは、真 (正常値) と偽 (異常値) に分類

する際の閾値を変更した時の真陽性率と偽陽性率を表現した曲線である。図 1 に、突発信号の挿入パターンごとに表した ROC カーブを記載する。図 1 から、STL より DeepAnT が多くのパターンにおいて精度が向上する結果となった。さらに、サンプル数が 60, 100 に関して、生データを用いた DeepAnT よりノイズ除去後のデータを用いた DeepAnT-wavelet が高精度になった。これにより、短期間における突発信号に対してデータ前処理であるノイズ除去を実施することで、より効果的な異常検知が可能になるとわかる。一方、( $1\sigma$ , 500samples) では、どの手法においても同様な結果となった。さらに、( $3\sigma$ , 500samples) では、DeepAnT において生データを用いた方が精度が向上する結果となった。したがって、本実験においてサンプル数が課題となることがわかり、この点を改善するために、ウェーブレット関数の選択や DeepAnT のパラメータ調整が必要になると考える。

### 5 まとめ

本稿では、DeepAnT での異常検知法を実施した。さらに、ウェーブレット変換を用いたノイズ除去のデータ処理を実施した。そして、生データとノイズ除去後のデータそれぞれにおいて生成した DeepAnT モデルの性能評価を実施した。実験結果より、時系列データに対して深層学習モデルベースのアプローチとノイズ除去による有用性を検証することができた。今後の展望として先進的なオンライン異常検知技法をプースティング等により効果的に連結することで、より頑健な検知手法を開発し、この研究課題へと適用したいと考える。

### 参考文献

- [1] S. Sako et al., The Tomo-e Gozen wide field CMOS camera for the Kiso Schmidt telescope, *Ground-based and Airborne Instrumentation for Astronomy VII*, volume 10702 of Proc. SPIE 2018.
- [2] R.B. Cleveland et al., STL: A Seasonal-Trend Decomposition Procedure Based on Loess, *Journal of Office Statistics*, 1990.
- [3] M. Munir et al., DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series, *IEEE Access*, December 2018.
- [4] Denoising, MATLAB Documentation <https://jp.mathworks.com/help/wavelet/denoising.html>
- [5] T. Phungtua-eng et al., Elastic Data Binning: Time-series Sketching for Time-domain Astrophysics Analysis, *SAC'23*, March 2023.
- [6] S. Schmidl et al., Anomaly Detection in Time Series: A Comprehensive Evaluation, *Proceedings of the VLDB Endowment*, 2022.