3K3-1

アンサンブル木学習によるノンパラメトリックなセンサー値正常範囲推定

Tree-based Nonparametric Prediction of Normal Sensor Measurement Range Using Temporal Information

秋元　康佑 *1　武石　直也 *1　矢入　健久 *1　堀　浩一 *1　西村　尚樹 *2　高田 昇 *2
Kosuke Akimoto　Naoya Takeishi　Takehisa Yairi　Koichi Hori　Naoki Nishimura　Noboru Takata

*1東京大学大学院工学系研究科
School of Engineering,The University of Tokyo

*2宇宙航空研究開発機構
Japan Aerospace eXploration Agency, JAXA

In usual limit checking on telemetry sensor data of a spacecraft, normal ranges of sensor values are defined by only one pair of upper and lower bounds. It can lead to too optimistic results or a high false alarm rate since it does not consider multimodal behavior and temporal patterns of sensor values. For more precise anomaly detection, normal ranges should be predicted adaptively according to spacecraft's states. The proposed method consists of two phases, one in which regression trees are used to extract temporal information and one in which a quantile regression forest is used to predict target normal range nonparametrically. We apply this method to actual telemetry data with simulated anomalies and confirmed that it can detect temporal anomalies with less false alarm rate than limit checking.

1. 背景

宇宙機は非常に複雑なシステムであるため、送られてくるテレメトリデータを人間が直接監視することはコストが高い。そこで実際の運用現場では監視作業を補助するためにさまざまな自動化手法が用いられている。中でも広く用いられている手法としてリミットチェック (limit checking) がある。リミットチェックはテレメトリデータに含まれるセンサー値をあらかじめ個別のセンサーごとに定めておいた正常範囲に収まっているかどうかによって異常かどうか判断する手法である。単純な手法であり実装も簡単であることが長所であるが、センサー値の正常範囲を求めることは宇宙機システムの複雑さもあり大変な作業である。本稿の目的はリミットチェックに用いられるこの正常範囲を、蓄積された過去のテレメトリデータにデータマイニングの手法を応用することで事前知識なしに求める手法について検討することである。

Chandola らは異常検知の文脈から異常を以下の 3 種類に分類している [Chandola 09]。それぞれの異常の例について図 1 に図示した。

point anomaly
他のどのデータ点と比較しても特異であるとみなせるような異常であり、もっとも単純な種類の異常である。

contextual anomaly
データの背景のある特定の文脈においてのみ特異となるような異常である。

collective anomaly
いくつかのデータ点をまとめて考慮してはじめて特異であるとみなせるような異常である。

前述の通りリミットチェックは正常範囲の設定に労力がかかるという欠点がある。しかし仮にこの設定を適切に行うことができたとしても、単一の正常範囲しか用いていないことから前述した 3 種の異常のうち point anomaly しか検知することができない。

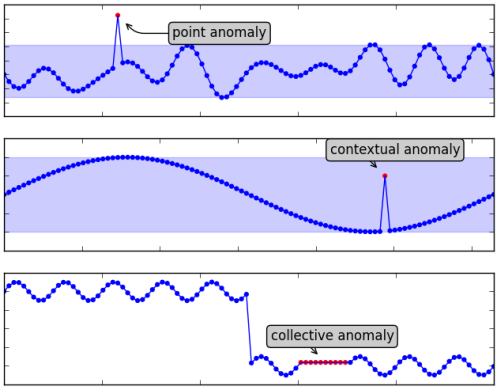

図 1: 異常の分類

こうしたリミットチェックの欠点を改良する手法として Adaptive Limit Checking[矢入 04] がある。この手法では宇宙機のテレメトリデータに含まれる離散値変数であるステータス変数を用いて回帰木 [Brieman 84] を学習し、それぞれのノードにたどりついたデータのみによって正常範囲を推定する。宇宙機のステータス変数には各サブシステムの動作モードや日陰か日照かどうかといった宇宙機の状態に大きな影響を及ぼす状態変数が多く含まれるため、学習して得られた木構造はこれらの変数によって注目しているセンサー値のモードがどのように変化するかを表現していると考えることができる。この手法ではこのような衛星システムのモードに応じた正常範囲が学習されるため、point anomaly だけでなく contextual anomaly の検知に対しても有効な手段となっている。

しかし Adaptive Limit Checking においても、モードが同じ場合には前後のデータ点との関係にかかわらず同じ正常範囲が適用されるため、データ点同士の関係を考慮してはじめて検知できるような collective anomaly を検知することは難しい。本稿では Adaptive Limit Checking の手法をもとに、学習の

連絡先: akimoto@ailab.t.u-tokyo.ac.jp

アンサンブル化や時間的文脈情報の抽出・利用を行うことでこれらの問題を解決できないか検討する。

2. 提案手法

この節では、提案手法の構成と学習法の概要について述べる。

2.1 手法の構成

提案手法は図 2 のような二つの段階に分かれている。最初の段階で評価したい時系列（ステータス変数 x_{sta} とセンサーデータ y_{sen} からなる）の各時刻において動作モードが推定され、この動作モードの変化を時間的な文脈情報として抽出する。続く段階では各データ点に対しステータス変数と最初の段階で抽出された時間的文脈情報の両方を用いて正常範囲を推定する。

それぞれの段階で用いた手法については 2.2 節と 2.3 節において述べる。

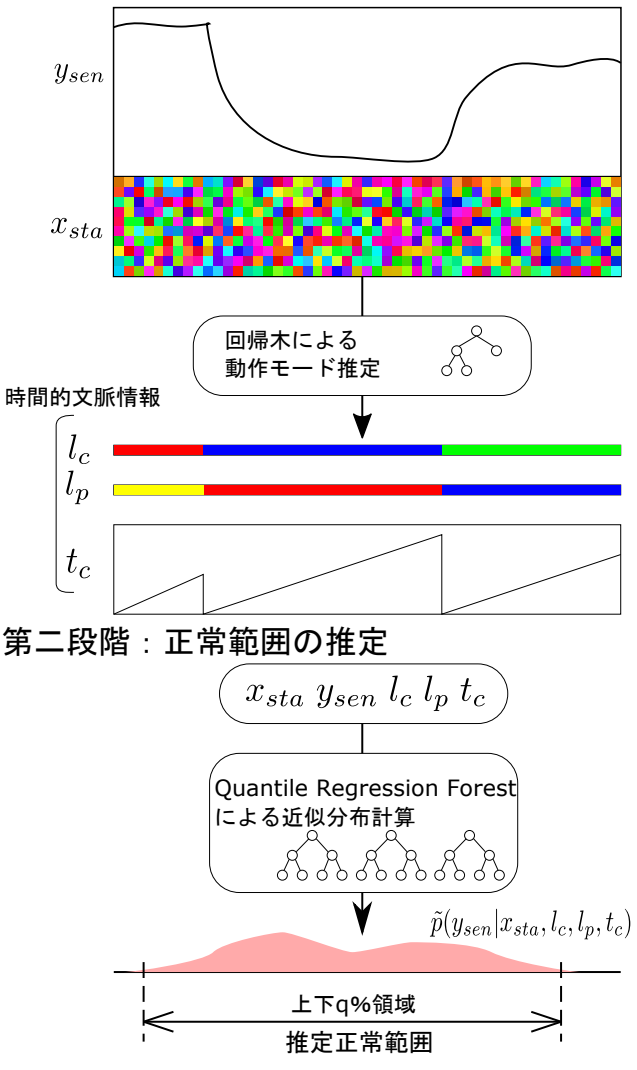

図 2: 手法構成概要

2.2 モード推定と時間的文脈情報の抽出

動作モードの推定は Adaptive Limit Checking[矢入 04] の手法にならい回帰木を用いて行った。回帰木の学習に用いる訓練データは過去のテレメトリデータを用い、説明変数としてステータス変数 x_{sta}、目標変数として注目しているセンサー値 y_{sen} を与えて学習させる。得られた木構造の各末端ノードが注目しているセンサー値の動作モードそれぞれに対応しているとし、各データ点に対しどの末端ノードにたどり着いたかによってラベルを付ける [*1]。回帰木の学習手法については誤差指標として分散を、枝狩りの手法として cost complexity pruning と 10-fold cross-validation を用いた。回帰木の学習手法について詳細は [Brieman 84] を参照されたい。

各データ点に対し動作モードのラベルが付けられた後、このラベルに基づき各データ点ごとに以下のような 3 つの変数を追加した。

- 現在のラベル l_c
- 現在のラベルになる前のラベル l_p
- 現在のラベルに変化してからの経過時間 t_c

2.3 正常範囲の推定

本手法では Quantile Regression Forest[Meinshausen 06] により条件付き分布 $p(y_{sen}|x_{sta}, l_c, l_p, t_c)$ の近似分布 $\tilde{p}(y_{sen}|x_{sta}, l_c, l_p, t_c)$ を求め、ある定数 q に対しこの近似分布の両側 q パーセント領域を推定正常範囲とした。

Quantile Regression Forest の学習は通常の Random Forest[Brieman 01] とほとんど同様であり、学習の際に各ノードで訓練データの平均ではなく訓練データの分布 [*2] を保存する点が異なる。説明変数 x, 目標変数 y を与えて学習した Quantile Regression Forest を用いると条件付き分布 $p(y|x)$ の任意の分位点を推定することができ、これからもとの条件付き分布を近似することができる。今回は説明変数に x_{sta}, l_c, l_p, t_c, 目標変数に y_{sen} を用いた。詳細な学習法は [Meinshausen 06] を参照されたい。

3. 実験

この節では提案手法を実データに対し適用した結果を示す。使用したデータの詳細については 3.1 節に示す。また本節の実験全てにおいて Quantile Regression Forest の木の本数は 30 とし各木の訓練データはもとの訓練データからブートストラップにより生成した。また各ノードでの分割探索時には説明変数の 30 %をランダムに選んで探索した。木の成長の終了条件は木の深さが 8 に達するかノードに達した訓練データが 5 個以下になった場合であるとした。

3.1 使用したデータ

実験に使用したデータは、宇宙航空研究開発機構の運用する小型実証衛星 4 型 (SDS-4)[中村 13] で取得されたテレメトリデータである。本実験ではこのテレメトリデータのうち時系列性を十分に考慮できるだけの頻度でセンサー値が記録されている可視時のデータのみを用いた。また実験対象のセンサーは衛星の動作モードにより時間的な変化のパターンが大きく異なる温度センサーとした。ステータス変数は訓練データ中でまったく値が変化していない変数を除いたすべての変数を用いた。ス

*1 必要なステータス変数が欠損しているなどしてどの末端ノードに到達するか決定できない際は直近のラベルで補完した。値の欠損の頻度よりモード変化の頻度が低い場合には妥当であると考えられる。

*2 [Meinshausen 06] においては訓練データの全てを保存しているが、本手法ではメモリ使用量の削減のためにヒストグラム (bin 数 10) を保存した。

テータス変数についてはデータの欠損が多く認められたので、ある程度時間的に短い欠損に対しては 0 次補間により値を補完し、長期にわたり欠損している場合は欠損したまま扱った。

3.2 既存手法との比較

既存手法である Adaptive Limit Checking との比較を行うため、近似分布の可視化による定性的な比較を行った。図 3 および図 4 はそれぞれ Adaptive Limit Checking および提案手法を実データに適用して得られた結果を可視化したものである [*3]。色の明暗は各時刻におけるセンサー値の近似分布における分位数がどれだけ中心に近いかを示し、緑色のプロットは実際に観測されたセンサー値を示している。実データは途中から徐々に下降するパターンを示しているが、提案手法はこの下降パターンをより良く捉えることができていることがわかる。

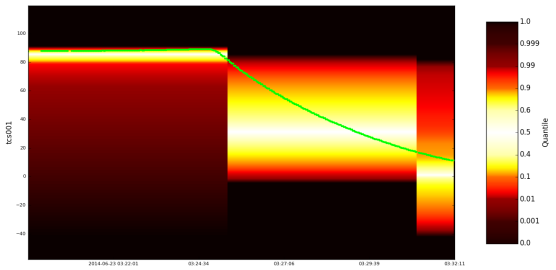

図 3: adaptive limit checking の適用例

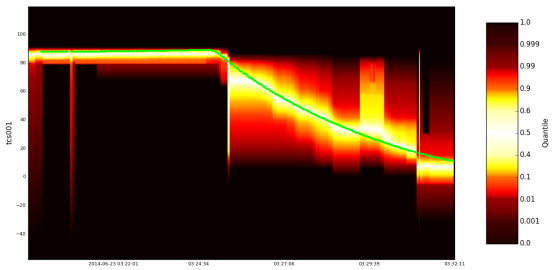

図 4: 提案手法の適用例

本稿では既存手法と提案手法の定量的な比較を行うために実データを加工して実際の異常を模擬した人工データを作成し、この異常の検知を試みる実験も行った。作成した人工的な異常データの一例を図 5 に示す。このデータは例えば温度が徐々に下がる段階で一時的に温度センサーに何らかのバイアスが加わり温度を正確に計測できなくなったような場合を模擬している。今回の実験では 12 個の実データをもとに 120 個の人工的な異常データを作成し、両手法で検知を試みた。この際両手法において正常範囲と判断する両側 q %領域について q を徐々に小さな値に設定すると、正常範囲が狭くなるため多くの異常を検知できる代わりに誤検知率も上昇する。本稿では同じ異常検知率を達成する際に発生する誤検知率を比較することにより定量的な比較を行った。

この実験の結果は図 6 のようになった。左右のグラフはそれぞれ変化の比較的激しいデータと、データ区間でほとんど温度が一定であり変化に乏しいデータに適用した場合の結果を示している。横軸は異常データのうち正しく検知できたデータの割合を表している。一方縦軸は同じ検知率の場合の既存手法に対する提案手法の誤検知率の変化であり、下側になるほど負となるため少ない誤検知率で検知できていることになり望ましい。色のついた領域は作成した 120 個（変化の激しいデータが 30 個、そうでないものが 90 個）の異常データに対する結果における上位 0-25,25-50,50-75,75-100 %の領域を示しており、黒線は平均を表している。グラフから変化が激しいデータ、激しくないデータいずれに対しても、提案手法は既存手法よりも同じ異常検知率を少ない誤検知率で実現できていることが分かる。特に変化の激しいデータについてはこの傾向が顕著であり、時間的な変化のあるデータに対してより有効になっているといえる。

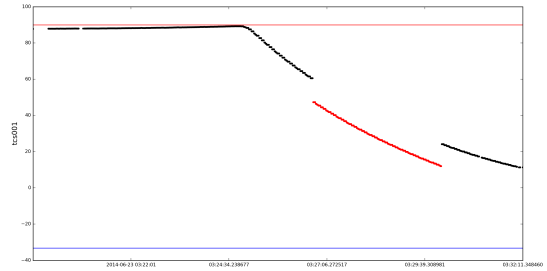

図 5: 人工異常データの一例

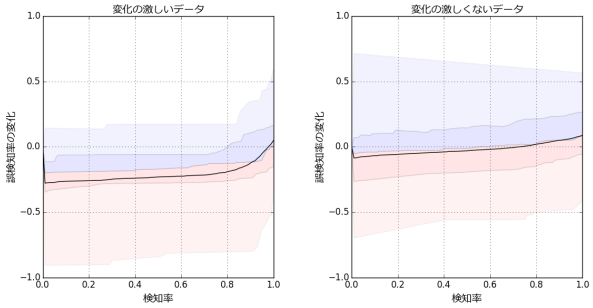

図 6: 人工異常データに対する誤検知率の比較

4. まとめ

本稿では宇宙機から送られてくるテレメトリデータのセンサー値の正常値を推定する際に、各時刻のステータス変数に加えて時間的な文脈情報も用いる手法を提案しその有効性を確かめた。今回加えた時間的文脈情報は動作モードの変化からの経過時間という非常に簡単なものであったが、定常運用中の宇宙機では同じような状態の変化が繰り返されるためにこのような少ない情報を加えただけでも精度を改善することができた。今後の改善点としてはモード推定の際に隠れマルコフモデルなどのより時系列データに関してより洗練された手法を用いてモード推定の精度を改善することや、直近の観測値を用いて動作ごとの変動に対応することなどがあげられる。

*3 Adaptive Limit Checking は近似分布の生成を行わないが、今回は比較のため各ノードに達したデータのヒストグラムを保存し近似分布として代用した。

謝辞

宇宙航空研究開発機構の西村尚樹様、高田昇様には SDS-4 のテレメトリデータを提供いただくとともに、データ解析にあたって貴重な議論や示唆をいただきました。深く感謝いたします。

参考文献

[Brieman 84] Breiman, Leo, et al. Classification and regression trees. CRC press, 1984.

[Brieman 01] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[Chandola 09] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." ACM computing surveys (CSUR) 41.3 (2009): 15.

[Meinshausen 06] Meinshausen, Nicolai. "Quantile regression forests." The Journal of Machine Learning Research 7 (2006): 983-999.

[矢入 04] Yairi, Takehisa, et al. "Adaptive limit checking for spacecraft telemetry data using regression tree learning." Systems, Man and Cybernetics, 2004 IEEE International Conference on. Vol. 6. IEEE, 2004.

[中村 13] Yosuke, Nakamura, et al. "Small Demonstration Satellite-4 (SDS-4): Development, Flight Results, and Lessons Learned in JAXA's Microsatellite Project." (2013).