TREE-BASED NONPARAMETRIC PREDICTION OF NORMAL SENSOR MEASUREMENT RANGE USING TEMPORAL INFORMATION

*Kosuke Akimoto [1], Naoya Takeishi [1], Takehisa Yairi [1], Koichi Hori [1], Naoki Nishimura [2], Noboru Takata [2]

[1] *School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan, E-mail: akimoto@ailab.t.u-tokyo.ac.jp*

[2] *Japan Aerospace Exploration Agency, 7-44-1 Jindaiji Higashi-machi, Chofu-shi, Tokyo, Japan*

ABSTRACT

Currently, limit-checking on telemetry sensor data of a spacecraft is widely used to detect its faults and anomalous behavior. Since classical limit-checking usually considers only *a priori* fixed pair of upper and lower bounds for each sensor variable, it sometimes fails to detect phenomena that are anomalous only in certain operating modes. To handle this problem, we present a method to predict normal ranges of sensor measurements adaptively based on status variables of telemetry data and temporal information. In the proposed method, a regression tree is learned using status variables, and each data point is labeled according to the terminal node of the tree it reached. Three new temporal features are generated from the sequence of the label, and a quantile regression forest is learned using both status variables and the generated features. Normal ranges are calculated from approximate distribution predicted using the quantile regression forest. We apply this method to actual telemetry data with simulated anomalies, and confirmed that the proposed method can detect temporal anomalies with a lower false alarm rate than the previous method.

1 INTRODUCTION

Classical limit-checking usually considers only one fixed pair of upper and lower bounds for each sensor variable, and this simplicity is one of the reasons for its popularity in actual operation of a spacecraft as an anomaly detection method. However deciding normal ranges of sensor measurements is usually a laborious task, since spacecraft is a highly complex system that consists of tens of thousands of components. In this paper, we propose a method, in which normal ranges of sensor measurements are predicted without prior knowledge using data-mining technique.

Chandola et al. [1] classified anomalies into following three categories from the perspective of anomaly detection, and examples of them are shown in Figure 1.

- **Point Anomaly**

  "If an individual data instance can be considered as anomalous with respect to the rest of data, then the instance is termed as a point anomaly" [1] .

- **Contextual Anomaly**

  "If a data instance is anomalous in a specific context (but not otherwise), then it is termed as a contextual anomaly" [1].

- **Collective Anomaly**

  "If a collection of related data instances is anomalous with respect to the entire data setm it is termed as a collective anomaly. The individual data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together as a collection is anomalous" [1].

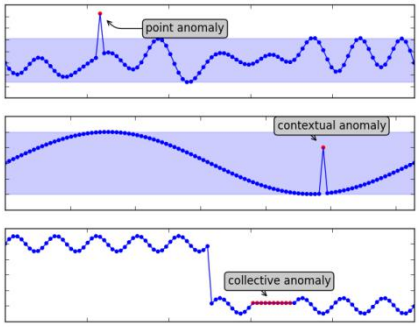

Figure 1: Examples of the three types of anomaly.

Among these three categories of anomaly, classical limit-checking can detect only point anomalies since it applies one fixed normal range for all data instances.

To solve this problem, Yairi et al. proposed Adaptive Limit Checking [5]. In this method, a regression tree [1] is learned using status variables in spacecraft's telemetry data as indicator variables and

target sensor measurements as a target variable. Normal ranges are predicted for each node of the tree based on the values of the target sensor measurements that reached to each node, and a normal range for a new data instance is decided adaptively based on the node of the tree it reached. Since status variables of telemetry data contain important ones such as operating modes of each subsystem and shine/shading, the learned tree has rich information about context of spacecraft operation, and it makes this method effective in detecting contextual anomalies.

Even in Adaptive Limit Checking, a single normal range is applied to data instances that reached same node of the tree, and temporal information of the telemetry data is not considered, which makes its performance limited for collective anomalies. Besides, this method outputs only upper and lower bounds of predicted nominal range, which is not sufficient to describe non-Gaussian behaviors of the sensor values.

The proposed method in this paper is based on Adaptive Limit Checking and has two new features. One is generation of temporal variables, and the other is calculation of approximate distribution using both status variables and generated temporal variables. Normal ranges are determined based on the approximate distribution, and they are effective against temporal anomalies since it is calculated using not only status variables but also temporal variables.

The rest of this paper is organized as follows. Section 2 describes a detail of the proposed method. In Section 3, we apply the method to the actual telemetry data and evaluate the results. Section 4 concludes the paper.

2 PROPOSED METHOD

In this section, we describe a detail of the framework of the proposed method. The following is the definition of terms used in this paper. *Status variables* are discrete variables in telemetry data of a spacecraft. *Target sensor measurements* are continuous values measured by a specific sensor, which is equipped on the spacecraft and monitored for anomalies. *Normal range* is a range within which sensor measurements are supposed to be when a spacecraft is normally operating, and it is estimated and used to detect anomalies in this method.

2.1. Overview of the Proposed Method

The proposed method consists of two phases as shown in Figure 2. Suppose we have telemetry data from spacecraft, which consist of status variables and target sensor measurements .

In the first phase, spacecraft's operating modes are estimated for each data instances, and temporal variables are generated from the sequence of the estimated operating modes.

In the second phase, normal ranges are predicted based on both the status variables and the generated temporal variables.

Detailed methods for the two phases are described in the following subsections.

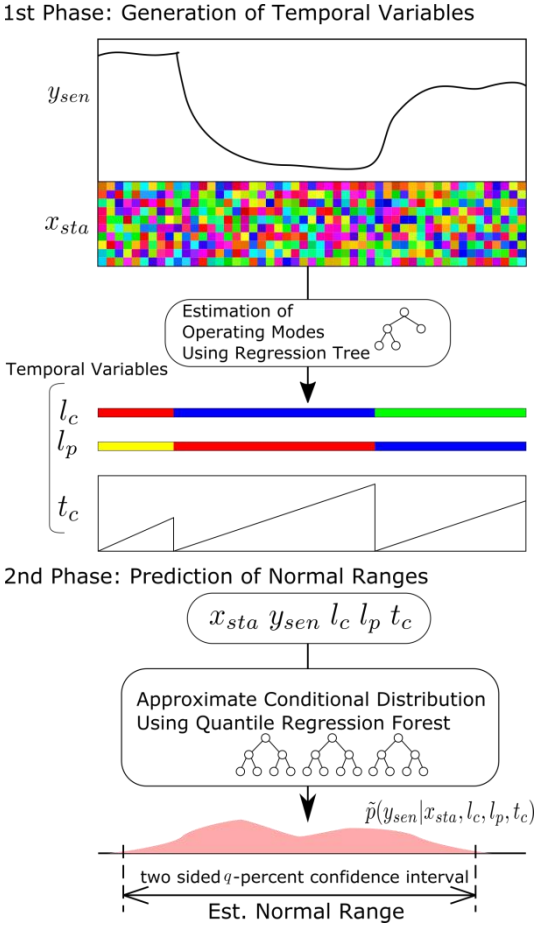

Figure 2: Overview of the Proposed Method.

2.2. Generation of Temporal Variables

As in Adaptive Limit Checking, a regression tree is learned to estimate spacecraft's operating modes.

Training data for the regression tree is the telemetry data from a spacecraft that consist of status variables and target sensor measurements . The status variables are used as indicator variables and the sensor measurements are used as a target variable. As for training method, variance is used as impurity function, and cost-complexity pruning and 10-fold cross-validation are used for pruning. The reader is referred to [1] for a detail description of the learning method of a regression tree.

After training, each data instance is labeled according to the terminal node of the tree it reached[1]. From these labels, which consists of the following quantities: current labels ; previous labels ; and elapsed time since the label changed .

2.3. Calculation of Approximate Distribution and Normal Ranges

In the proposed method, Quantile Regression Forest [4] is used to calculate the approximate distribution of the conditional distribution . Learning method of the quantile regression forest is almost the same as that of a random forest [2], and their sole difference is that not average but order statistics of the training data is saved at each node when training a quantile regression forest. A quantile regression forest is used to predict arbitrary quantiles of the target variable, and the approximate distribution is calculated using predicted quantiles.

Normal ranges can be determined based on calculated approximate distribution. In the proposed method, normal ranges are determined as two-sided -percent confidence interval of the distribution (is a constant value).

3 EXPERIMENTS AND RESULTS

In this section, the proposed method is applied to the actual telemetry data. For all the following experiments, the number of trees in a quantile regression forest is set to 30, and training data for each tree is generated by bootstrapping. In training trees in the forest, 30 percent of input variables are randomly chosen and evaluated at each split. Stopping rule is reaching maximum depth of 8, or reaching minimum number of instances in a node that is set to 5.

3.1. Applied Data

The proposed method is applied to the actual telemetry data of small demonstration satellite SDS-4 [6] which is operated by JAXA. Only *visible data* – data obtained when the satellite is visible – are used in the experiments since other data are not frequent enough to investigate temporal behavior of the spacecraft. We tested the method on temperature sensor measurements since their patterns of temporal behavior obviously differ depending on the operating mode of the spacecraft. Some of the status variables that contain no valid value or only one value for all data instances are ignored, and missing values in other status variables are filled by zero-order interpolation.

[1] If a data instance does not reach any terminal node due to missing values, it is labeled by the latest label. Its effect is negligible when the frequency of missing value is lower than that of change of spacecraft's operating mode.

3.2. Approximate Distribution

In this subsection, we compare the proposed method to Adaptive Limit Checking qualitatively. The approximate distribution of sensor measurements is calculated using the two methods[2], and the results are shown in Figure 3 and Figure 4.

In both figures, green plots are actual sensor values, and the brightness shows quantiles of the approximate distribution. It can be seen that the approximate distribution of the proposed method better predicts the declining behavior of the actual sensor values.

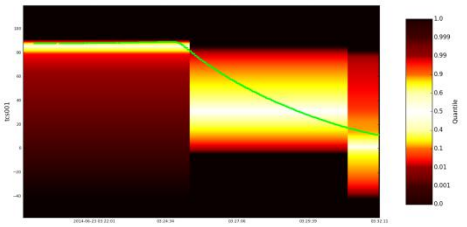

Figure 3: Approximate Distribution (Adaptive Limit Checking).

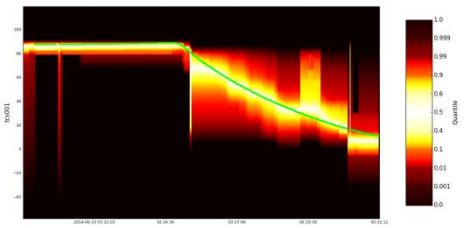

Figure 4: Approximate Distribution (the Proposed Method).

3.3. Performance against Artificial Anomaly

In this subsection, we compare the two methods' ability to detect temporal anomalies. We add simulated anomalies to the actual telemetry data, and an example of these data is shown in Figure 5. The red section in the graph represents simulated anomalies, and these artificial anomalies simulate behavior of sensor measurements with temporary bias.

In this experiment, we created 120 artificial anomalous data set from 12 actual visible data set, and applied both methods to these data. Normal ranges are defined as percent region of the approximate distribution. By setting q closer to zero, true positive rate and false positive rate increases at

[2] Adaptive Limit Checking does not predict approximate distribution. In this experiment, distribution of training data that reached each node of a regression tree is substituted for the approximate distribution.

the same time, and vice versa. We compare the two methods quantitatively by measuring with how small false alarm rate the proposed method can achieve at a specific true positive rate compared with Adaptive Limit Checking.

The result is shown in Figure 6. Left and right graphs show the result for big and small temporal changes respectively. The horizontal axes show true positive rate, and the vertical axes show the difference between false alarm rates of both methods (negative value means that the proposed method can achieve specific true positive rate with smaller false positive rate than the previous method). A black line represents the average of the results, and edges between colored regions represent each quartile of the results.

From Figure 5, it can be seen that the proposed method achieves same true positive rate with smaller false positive rate than the previous method.

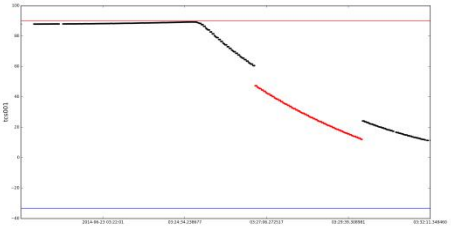

Figure 5: An Example of Simulated Artificial Anomalies

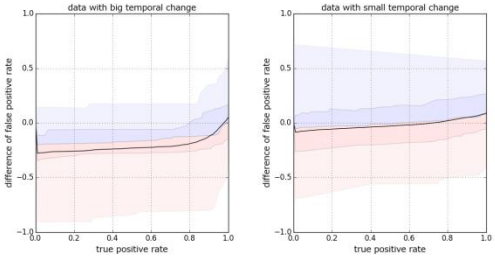

Figure 6: Difference between False Alarm Rate of the Both Methods.

4 CONCLUSION

In this paper, we proposed a method, in which normal ranges of sensor measurements are predicted based on both status variables of telemetry data and temporal information generated from telemetry data. We show that the temporal information as simple as elapsed time since spacecraft's operating mode changed improves performance of prediction of normal ranges, and it may be because of the high regularity of spacecraft operation.

Utilizing more sophisticated methods such as hidden Markov models for operating mode estimation or using results from regression or forecasting method as temporal information may further improve precision of the method.

Acknowledgement

The actual telemetry data used in this paper is offered by JAXA. This study was partially supported by JSPS KAKENHI Grant Number 26289320.

References

[1] Breiman, Leo, et al. Classification and regression trees. CRC press, 1984.
[2] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
[3] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." ACM computing surveys (CSUR) 41.3 (2009): 15.
[4] Meinshausen, Nicolai. "Quantile regression forests." The Journal of Machine Learning Research 7 (2006): 983-999.
[5] 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.
[6] Yosuke, Nakamura, et al. "Small Demonstration Satellite-4 (SDS-4): Development, Flight Results, and Lessons Learned in JAXA's Microsatellite Project." (2013)