Aerospace Technologies Advancements: Evaluation of Anomaly Detection Capability for Ground-Based Pre-Launch Shuttle Operations

Rodney A. Martin

April 15, 2009

Abstract

This chapter will provide a thorough end-to-end description of the process for evaluation of three different data-driven algorithms for anomaly detection to select the best candidate for deployment as part of a suite of IVHM (Integrated Vehicle Health Management) technologies. These algorithms were deemed to be sufficiently mature for consideration as viable candidates for demonstration during the launch of Ares I-X. This launch represents the first test flight of Ares I, which will be the successor to the Space Shuttle for NASA's Constellation program. Data-driven algorithms are just one of three different types being deployed [3],[5]. The other two types of algorithms being deployed include a "rule-based" expert system, and a "model-based" system. Within these two categories, the deployable candidates have already been selected based upon non-quantitative factors such as flight heritage and system certifiability. For the rule-based system, SHINE (Spacecraft High-speed Inference Engine) has been selected for deployment, which is a component of BEAM (Beacon-based Exception Analysis for Multimissions) [4], a patented technology developed at NASA's JPL (Jet Propulsion Laboratory) and serves to aid in the management and identification of operational modes. For the "model-based" system, a commercially available package developed by QSI (Qualtech Systems, Inc.), TEAMS (Testability Engineering and Maintenance System) [1] has been selected for deployment to aid in diagnosis. In the context of this particular deployment, distinctions among the use of the terms "data-driven," "rule-based," and "model-based," can be found in [5].

Although there are three different categories of algorithms that have been selected for deployment, our main focus in this chapter will be on the evaluation of three candidates for *data-driven* anomaly detection. These algorithms will be evaluated based upon their capability for robustly detecting incipient faults or failures in the ground-based phase of pre-launch space shuttle operations, rather than based on system certifiability as performed in previous studies [5]. Robust detection will allow for the achievement of pre-specified minimum false alarm and/or missed detection rates in the selection of alert thresholds. All algorithms will also be optimized with respect to an aggregation of these same criteria. Our study relies upon the use of Shuttle data to act as a proxy for and in preparation for application to Ares I-X data, which uses a very similar hardware platform for the subsystems

that are being targeted (TVC - Thrust Vector Control subsystem for the SRB (Solid Rocket Booster)).

The main thrust of the chapter is to provide instructive coverage on the topics of algorithmic optimization and alert threshold selection. All data-driven algorithms under investigation will be optimized and compared using a variety of metrics including the AUC (Area under the ROC (Receiver Operating Characteristic) curve), and an "FPR (False Positive Rate))-limited" variant of the AUC that only considers performance evaluation for algorithmic comparison for low false positive rates. This involves measurement of the partial area under the ROC curve up to a maximum prescribed false alarm rate. We take cues from the fields of radiology and medical diagnostics in regards to the use of this metric (pAUC), whose implementation, development, and evaluation of the advantages and disadvantages of the latter and related methods (e.g. quantification of standard error for the AUC, etc.) are far more mature than their use in the current application [6]-[7]. The literature is sparse on the use of such metrics for aerospace applications, with only a recent study documenting it [2].

Other metrics to be considered use the AUC and pAUC, however with a slightly modified definition of false alarms and missed detections that accounts for predefined latencies and prediction horizons. These definitions are based largely upon pragmatic user requirements in order to prevent the over-penalization of test points that do not match ground truth classification for each individual time point. By providing for a marginal allowance or window of time around the test time point in question, we implicitly allow for a more relaxed definition of false alarms and missed detection rates, thereby counteracting their artificial inflation when adhering to stricter definitions. We can approximate these modified definitions of false alarms and missed detections by "shifting" the ground truth vector by the respective latency and prediction window values, and averaging the resulting AUC and pAUC values across the two extremes.

Alert threshold selection will be the final topic covered in the chapter, which will use the results of AUC/pAUC metric optimization. Due to the fact that these metrics represent overall classification discriminability, optimal algorithmic parameters can be found and used to perform threshold or alert selection based upon the corresponding ROC curve. This allows for alert threshold selection based upon specified minimum false alarm and/or missed detection rates to demonstrate a robust anomaly detection capability. As a practical measure we will also present a performance comparison among the candidate data-driven algorithms by evaluation of their computational complexity as well as the metrics that have been introduced thus far. The final results will also include a formal cross-validation procedure, which will be used to perform optimization and alert threshold selection. Realizations of an unseen hold out test case will finally be used to illustrate the superiority of a particular algorithmic technique.

References

[1] Kevin Cavanaugh. An integrated diagnostics virtual test bench for life cycle support. In *Proceedings of IEEE Aerospace Conference*, volume 7, pages 7–3235–7–3246, Big Sky, Montana, March 2001.

[2] Philip K. Chan and Matthew V. Mahoney. Modeling multiple time series for anomaly detection. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005)*, pages 90–97. IEEE Computer Society, 2005.

[3] David L. Iverson, Rodney Martin, Mark Schwabacher, Lilly Spirkovska, William Taylor, Ryan Mackey, and J. Patrick Castle. General purpose data-driven system monitoring for space operations. In *Proceedings of the AIAA Infotech@Aerospace Conference*, Seattle, Washington, April 2009.

[4] Ryan Mackey, Mark James, Han Park, and Michail Zak. BEAM: Technology for autonomous self-analysis. In *Proceedings of the IEEE Aerospace Conference*, Big Sky, MT, 2001.

[5] Mark Schwabacher and Robert Waterman. Pre-launch diagnostics for launch vehicles. In *Proceedings of the IEEE Aerospace Conference*, Big Sky, MT, March 2008.

[6] James A. Sorenson and X. Wang. ROC methods for evaluation of fMRI techniques. *Magnetic Resonance in Medicine*, 36(5):737–744, 1996.

[7] S. D. Walter. The partial area under the summary ROC curve. *Statistics in medicine*, 24(13):2025–2040, July 2005.