Applying Machine Learning-Based Diagnostic Functions to Rotorcraft Safety

Daniel R. Wade, and Andrew W. Wilson

Aeromechanics Division, Aviation Engineering Directorate,
Bldg. 4488, Martin Road, Redstone Arsenal, Alabama, 35898, United States of America

Abstract

The United States Army has recently funded future vertical lift aircraft manufacturers to investigate and integrate machine learning based technologies into Health and Usage Monitoring Systems and Flight Data Recorders. Furthermore, the Army has started a project to employ these methods to improve diagnostic classifiers fielded to Soldiers as part of the Condition Based Maintenance Program. Since the 1990s, the Army has accumulated a large and diverse dataset of rotorcraft flight characteristics, operational behaviour, and component and system failure data that could be used in machine learning based applications. As the airworthiness authority of the Army rotorcraft fleet, the Aviation Engineering Directorate has embarked on a project to define the necessary qualification and substantiation documentation required to use machine learning based diagnostics to monitor aircraft safety. This paper presents a qualification sub-process that uses machine learning classifiers to develop safety based diagnostics for rotorcraft resulting in actionable information for pilots, maintainers, engineers, logisticians and project managers.

Keywords: Airworthiness, Machine Learning, Aviation Data Science

Introduction

The United States Army Aviation Engineering Directorate (AED) is the Army's airworthiness certification authority; it also provides engineering services to the Army's Program Executive Office for Aviation (PEO-AVN). One portion of this service is data analysis and management of Health and Usage Monitoring System (HUMS) data collected on all Army rotorcraft. Under funding from PEO-AVN and internal research resources, AED started a machine learning program to improve diagnostic classifiers fielded to Soldiers as part of the Army's Condition Based Maintenance (CBM) Program. The AED is defining the necessary qualification and substantiation documentation required to use machine learning based diagnostics to replace existing maintenance practices. This paper presents a portion of a process for employing machine learning classifiers at the enterprise and field levels.

The data used to demonstrate the proposed process in this paper comes from aircraft Nose Gearboxes (NGBs) operated on a fleet of Army rotorcraft for over a decade. HUMS data has been collected on this NGB since the inception of the CBM program via the Modernized Signal Processing Unit manufactured by Honeywell Aerospace. Data shown in this paper ranges across combat and garrison flight activity in a broad range of environmental and topographical conditions.

Airworthiness qualification criteria for implementing machine learning based diagnostics require a modification to the current qualification methods outlined in Aeronautical Design Standard 79 (ADS-79) [1]. Appendices D and I of ADS-79 govern the qualification of classifiers, known as Condition Indicators (CIs). The main method for qualifying the classifiers outlined in Appendix I is through demonstration of an acceptable Bernoulli trial from field or laboratory data. ADS-79 makes fundamental assumptions regarding the classifiers, namely, they are **physics-based algorithms** developed during the engineering design process, and qualified through aircraft flight testing, or failure testing in laboratories. Engineers that have developed the diagnostics for the last few decades have approached problems from a physics-based perspective. The Army airworthiness authority has generally only accepted diagnostics associated with absolute threshold values, e.g. 5.0 inches per second max periodic velocity of a shaft. Unfortunately, it has been demonstrated that the unique aspects of each aircraft make it such that the absolute limits often do not perform with acceptable classification error [2].

Machine learning models are powerful, and are able to fit data **without the constraints** of physics-based models. This is why they could potentially be used to improve the performance of the HUMS, e.g. acting as a fusion algorithm for existing CIs or as a new way to generate features. AED demonstrated the power of a relatively simple model, principal component analysis (PCA), to engineer features associated with bearing failure in a rotorcraft oil cooler [3]. Prior to this work, *simple* models that used linear bearing energy methods (e.g. Root Sum Square Energy over a frequency band, and demodulation techniques) were unable to separate healthy and faulted bearings with *airworthiness-acceptable* classification accuracy. With the proper controls, airworthiness substantiation can be demonstrated by a system built through a machine learning process, particularly if that system is built upon a physical understanding of the sensor choices, placements, and sampling methods.

This paper has one goal, to define the process of selecting a single, airworthy machine learning-based diagnostic classifier that replaces a suite of fielded CIs. This paper focuses on developing aerospace specific metrics for model selection. The high level process for development of machine learning based diagnostics was presented previously by AED [4], it focused on the use of segregated test vaults that were not accessible to model developers during training. It showed that a *pre-testing* vault can be used to mitigate program risk due to stringent airworthiness requirements for test activities. The critical airworthiness requirement for using machine learning techniques is that multiple models cannot be used during the test activity.

Model Selection for Airworthiness Applications

For applications where machine learning is applied to airworthiness, process control during model training and testing is required. A high level flow diagram is shown in Figure 1. The steps in this flow comprise the following: problem definition, model acceptance criteria based on airworthiness credits, model space and search method(s), data curation, data vaulting, model training, model decision based on acceptance criteria, final model testing against acceptance criteria, and delivery of the model in a format usable in the enterprise or field environment.

This process is important to the reduction of bias, which is a significant aspect of training machine learning models. AED has carefully curated a demonstration dataset from the NGB data collected from the fielded HUMS that has been divided into three data vaults: training, pre-test, and testing. The make-up of these vaults comprises over 600 examples of ground truth assets of which there are over 40 examples of faults. The three vaults are sampled across those examples in a 60/20/20% split respectively. At time of authorship, the pre-test vault has been used for preliminary validation activities as described later in this paper. The test vault has not been opened. Relative to Figure 1, the project is currently in the *Train Models* block.

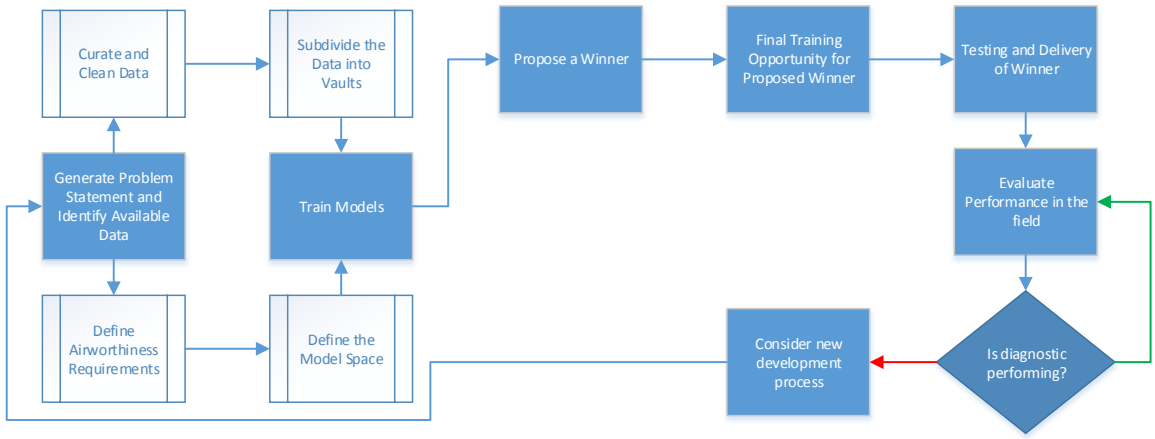

Fig. 1. High level process flow for machine learning airworthiness substantiation.

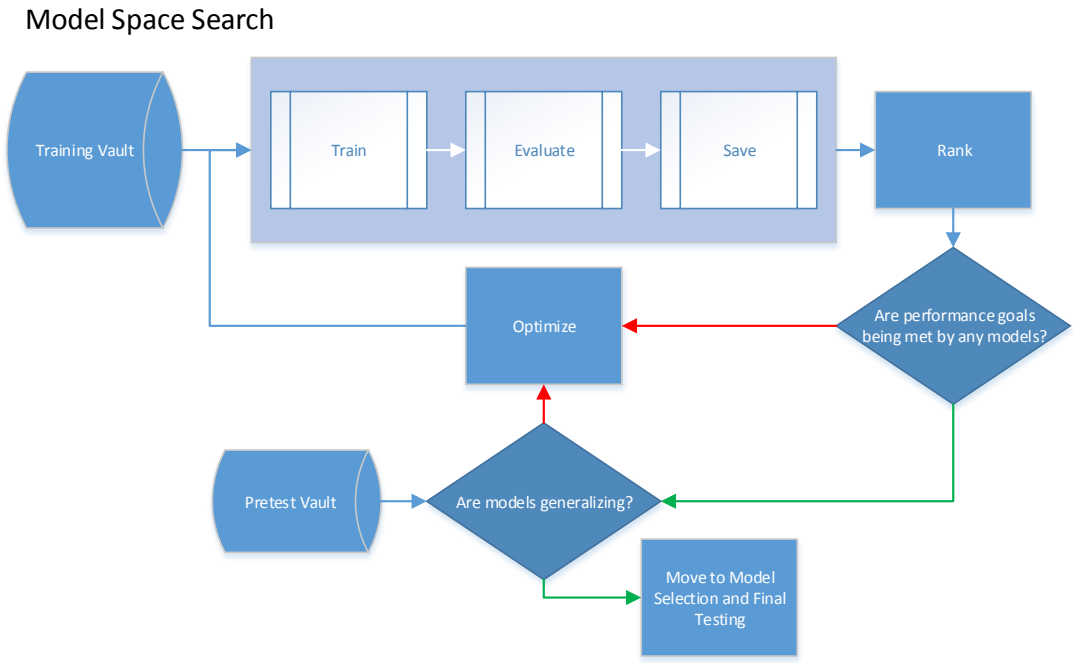

Fig. 2. Illustration of the model space search recursion and optimization.

A detailed sub-process of the *Train Models* block is shown in Figure 2. A winning model will be selected from the thousands of trained models prior to moving into final testing. The AED is currently using several metrics to support model selection decisions. The remainder of this paper reviews the most promising metrics. Model selection for testing is a problem that involves balancing model performance against the requirements, such as minimum True Positive and maximum False Positive Rates (TPR and FPR), or their counterparts False Negative and True Negative Rates (FNR and TNR), respectively.

Bookmaker's Informedness

The first criteria used by AED for model ranking and selection is Bookmaker's Informedness [5], or simply *informedness* (Eqn 1). Informedness is particularly interesting because of its resistance to skewed class problems. Typical aviation component health problems will present as significantly skewed class problems where healthy components comprise more than 99% of the population. Traditional metrics, such as accuracy, are unable to represent model performance in skewed class problems. A classifier that makes no errors has an informedness of 100%. A classifier that is doing no better than random guessing scores 0%, and one that is actually misclassifying scores between 0 and -100%. ADS-79E infers that a model must have an informedness above 95% to be considered for airworthiness. Informedness is not itself a metric, it must be measured during training (In-Sample, or cross-validation), pre-testing (validation), or testing (Out-of-Sample performance estimation).

$$Bookmakers\_Informedness = TPR - FPR \quad (1)$$

Historical and Asset Based TNR

The classification error for all data points in the ground truth that are associated with known healthy components is the Historical TNR, TNR_{hist}. This is an optimistic estimate of the out-of-sample TNR (and FPR). It can be visualized as a time history of classification or presented as a proportion of correct classifications to total points. A trained logistic regression model using the NGB data is shown in Figure 3.

Fig. 3. Example visualization of Historical and Asset Based TNR ratio calculation for a logistic regression model learned from the NGB training vault data.

The classification error for the data in the ground truth associated with known healthy components expressed as the proportion of assets correctly identified as healthy is the Asset

TNR, TNR_{assets}. For example, take the history of all healthy gearboxes, as evaluated by a candidate model, classify a gearbox as a TN if it is identified as healthy by the candidate model **for all data in the history,** or classify it as a FP if it is identified by **one or more** data points in the history. This metric is a pessimistic estimate of the out-of-sample FPR.

Together, TNR_{hist} and TNR_{assets} help the analyst bound the expected out-of-sample error for healthy components. These two bounds are critical for understanding the business and logistics consequences for fielding the model. ADS-79 recommends maintaining a high TNR, greater than 90%, but preferably higher. These two bounds are easy to find during training because of the ease of finding guaranteed healthy assets.

Informedness Based Metrics for Faulted and Healthy Assets

The classification error expressed as informedness at the end of life of all components in the training vault is In-Sample Informedness, I_{IS}. This is the most optimistic estimate of TPR and FPR. The classification error expressed as mean informedness at the end of life of all components in the training vault in cross-validation (mean across folds by model) is cross-validation informedness, I_{CV}. This is an estimate of out-of-sample TPR and FPR. It is less optimistic than the In-Sample estimate but it is not a lower bound. Additional measures of generalization are necessary to inform the analyst of out-of-sample performance.

Estimating Generalization

The first estimate of generalization is a ranking of candidate models by the absolute difference between I_{CV} and I_{IS}, $|I_{CV} - I_{IS}|$. A near zero value of this metric is highly desirable and a value greater than 10 is undesirable.

The second metric of generalization is called *position shuffle*, P_s. Position shuffle uses the training data and a validation dataset not used during training. For the purposes of this paper, the contents of the pre-test vault are the validation set. P_s is computed using one of the aforementioned metrics such as I_{CV}, I_{IS}, TNR_{hist}, and TNR_{assets}. The chosen metric is computed on the training data and then computed on the pre-test data from a sub-set of analyst selected models. The output of the metric is a score representing how models change ranking positions from training to pre-testing. After all scores are computed, the minimum shift is enforced so that anonymity can be maintained among the output values. For the NGB case, AED has used two as the minimum Ps for any model. This ensures that analysts have difficulty identifying which models are performing the best; thus P_s is actually a metric **about the data-model combinations** rather than individual model performance. The RMS of the shuffle is computed across the results to give a final, single metric. This can be compared against a Monte Carlo simulation to determine if the data-model combinations are behaving in a similar manner both in-sample and out-of-sample.

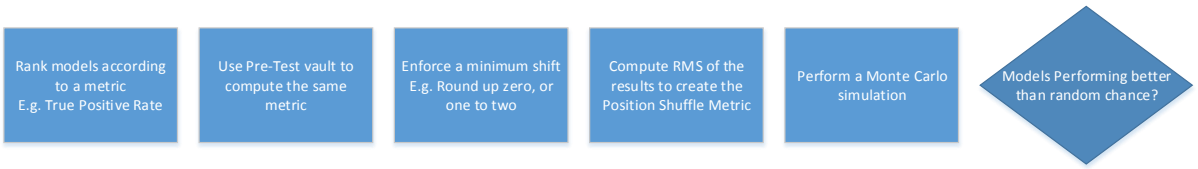

Fig. 3. Steps for computing the position shuffle of a group of models over a chosen metric.

Position shuffle is best used over a large set of models that are diverse in the way they select from the available training data, feature extraction methods, regularization methods, and other meta-tuning parameters [6]. As an example, after running 6,000 models per the procedures explained in Ref. 5, AED used this metric to demonstrate over-fitting without actually receiving validation results from a pre-test dataset. The value of P_s found in this situation was 9.2, which was higher than the mean (~8) indicated by the Monte Carlo simulation. This means that the models are doing worse than a random shuffle of the top models and are likely over-fitting. To combat against the over-fitting, the authors have re-curated the dataset to utilize additional faulted assets, thus increasing the number of examples from which to learn. Results from this re-curation will be published at a later date.

Conclusion

This paper has presented the reader with a suite of metrics and a process for model selection that can be used to select a single machine learning based diagnostic classifier for fielding directly to Army maintainers. By following this process, the single model is certified for use in directing maintenance activity and replaces a suite of physics based CIs with a single output that is actionable by the field maintenance personnel. Future work should include the results of the test activities, and a full demonstration of how the model is fielded Soldiers.

References

1. *ADS-79E-HDBK Condition Based Maintenance for US Army Aircraft Systems*, AMCOM Standardization Technical Data Management Division, Redstone Arsenal, 2016.

2. Krick, S., Wade, D., and Pipe, K. "Evaluation of a Novel Adaptive Threshold and Trend Alert Generation Technology on a HUMS Equipped Fleet", *Proceedings of the 68th Annual Forum of the American Helicopter Society*, Ft Worth, Texas, United States, May 2012.

3. Szelistowski, M., "Determining Test Stand Usability through Frequency Response and Principal Component Analysis", *Proceedings of the American Helicopter Society CBM Specialists Meeting*, Huntsville, AL, United States, Feb 2015.

4. Wade, D., and Vongpaseuth, T., et. al., "Machine Learning Algorithms for HUMS Improvement on Rotorcraft Components", *Proceedings of the 71st Annual Forum of the American Helicopter Society*, Virginia Beach, VA, United States, May 2015.

5. Powers, D., "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation", *Journal of Machine Learning Research*, 2:37-63, 2011.

6. Wilson, A., Wade, D., Albarado, K., Partain, J., and Statham, M., "A Classifier Development Process for Mechanical Health Diagnostics on US Army Rotorcraft", *Proceedings of the 1st ML and PHM Workshop SIGKDD 2016,* San Francisco, CA, United States, Aug 2015.