An Investigation of the Expectation Maximization Technique for an Unsupervised Machine Learning Application of Rotorcraft HUMS Data*

Lance Antolick
Aerospace Engineer
RMCI, Inc
U.S. Army Aviation Engineering Directorate
Redstone Arsenal, AL

Abstract
The Army is currently conducting research on the use of machine learning algorithms for refining diagnostics as part of the Condition-Based Maintenance (CBM) Program for Army rotorcraft. A major objective of this research is the ability to fuse the outputs of various data sources in order to create a health algorithm that recommends specific maintenance actions and estimates remaining useful life of components. Machine learning offers a variety of potential techniques and methods, such as the expectation maximization (EM) technique, to accomplish this objective. EM is an iterative process that maximizes the likelihood that a particular data point belongs to a unique Gaussian distribution within the dataset. This paper will explore the process and results of applying EM on a subset of AH-64 nose gearbox ground truth database that is representative of the faulted/non-faulted distribution of the fleet. The results of the algorithm will be analyzed and a determination will be made on how to integrate the EM algorithm into the machine learning process.

Introduction
The United States Army has implemented a Condition-Based Maintenance (CBM) program for its rotorcraft fleet. The CBM program relies heavily upon digital source collectors, commonly known as Health and Usage Monitoring Systems (HUMS), which the Army has installed on over 3,000 helicopters. The purpose of the HUMS is to perform rotor track and balance functions; monitor the health of the helicopter’s drivetrain components, including shafts, gears, and bearings; and record aircraft parametric data. The drivetrain components are monitored using a suite of accelerometers from which the HUMS collects vibration data. The vibration data is then processed to extract specific vibration fault features. These vibration fault features are processed into a single value referred to as a condition indicator (CI). CIs are designed to detect specific failure modes of a shaft, gear, or bearing. CIs can be used individually or as part of a calculation with other CIs to produce a health indicator (HI), which displays the remaining useful life of a specific component or line replaceable unit (LRU) to the end user.

A governance document for the CBM program has been created and is titled the Aeronautical Design Standard (ADS) 79 Handbook for Condition Based Maintenance System for US Army Aircraft. ADS-79 prescribes that each HI have a true positive rate of 90% with minimal false positive and false negative rates (Ref. 1). The early HUMS configurations produced large quantities of false positives, and undesirable true positive and false negative rates. Therefore, an effort was launched to improve the HUMS’ accuracy. Components suspected of having internal faults were sent to undergo a teardown analysis (TDA) after being removed from the aircraft. The results of the TDAs are documented and an estimated remaining useful life is associated with each component. These TDAs were used to create a ground truth database. The ground truth database is used to grade the effectiveness of CIs, adjust CI thresholds, create new CIs, and evaluate new techniques being considered for HUMS implementation.

The TDA program has vastly improved the accuracy of the HUMS; however, the goals set forth in the
ADS-79 are yet to be obtained for most HUMS-monitored components. The reasons the ADS-79 objectives have not been met can be narrowed down to two main issues. These issues are the use of a universal static threshold across the fleet of aircraft and a low fault signal-to-noise ratio, which can be attributed to either excessive drivetrain noise or a poor vibration transfer path.

Static Thresholds

Static thresholds tend to work well with shafts and with some other dynamic components, but do not work well with all components. There are two primary reasons for this. First, due to differences in manufacturing, installation, and fault progression, the same type of fault may not produce equivalent CI magnitudes. As can be seen in Figure 1, a handful of CI values exceed the yellow threshold for the bearing fault, while the majority of points are considered nominal in the time period leading up to gearbox replacement (November – January 8).

Figure 1: Tail Rotor Gearbox Bearing Fault and Associated CI

Meanwhile in Figure 2, the majority of the CI values are above the yellow threshold while two CI values surpass the exceedance threshold. The same CI magnitudes are not achieved for what appears to be equivalent bearing damage when compared to Figure 1.

Figure 2: Tail Rotor Gearbox Bearing Fault and Associated CI

The most straightforward solution to this discrepancy would be to make a threshold adjustment; however, the variances in vibration signatures throughout the fleet may not allow for the threshold to be lowered without an increase in the false positive rate. Another potential solution would be to implement an aircraft-unique threshold for each component based upon the subject component’s nominal values. This approach presents its own difficulties, such as infant mortality and configuration changes. If this approach did prove to be viable, it would not be practical to implement in the Army’s current software environment.

The second problem with the static threshold is CI variability. As can be seen in the two aforementioned examples, the CI oscillates between nominal and faulted. Depending on when and how the user examines his or her data, they may never know that there is a faulted component operating on their aircraft. An intuitive solution to this problem is a rule-based ‘latching’ alert, where the component remains faulted in the HUMS software until the user or maintenance software acknowledges that maintenance has been performed. Again, this approach would require a significant alteration to the current software environment.

Low Fault Signal-to-Noise Ratio

The second issue with the current HUMS configuration is a low fault signal-to-noise ratio. In
some cases, the fault signature is masked by the normal operation of a large magnitude and/or quantity of the dynamic components’ vibration (this does not include components in which the fault signature can be extracted via advanced signal processing techniques). This is particularly an issue with larger gearboxes such as the Apache main transmission or the Black Hawk main module. The low fault signal-to-noise ratio problem can also be attributed to the lack of a sufficient vibration transfer path from the fault initiation site to the monitoring sensor. One proposed solution is the installation of a waveguide sensor. The waveguide serves as a pseudo conduit between the component of interest and the sensor (Ref. 2). Another future solution may include embedded sensors, where a sufficient vibration transfer path exists from every dynamic component to its associated sensor. Incorporation of this technique may utilize waveguide technology.

HUMS Refinement Efforts

Several efforts have been launched and demonstrated over the past few years to improve HUMS performance. Although the implementation of the TDA program has greatly enhanced the HUMS, the goal of a 90% true positive rate with minimal false negatives and false positives is yet to be obtained. The TDA program and the ground truth database have allowed for the evaluation of other HUMS improvement programs. Three of those programs are discussed further below.

Bearing Resonance

The Bearing Resonance Project was undertaken to characterize the frequency response functions from the bearing’s fault initiation site to the HUMS sensor for all the monitored drivetrain components on all Army rotorcraft (Ref. 3). New bearing diagnostics were then produced, based on the features analyzed within the frequency response functions. It is estimated that the diagnostic accuracy for components without a ground truth database increased by approximately 20% (Ref. 4). The Bearing Resonance Project also revealed that some components had a poor vibration transfer path from the fault site to the sensor. In these cases, the magnitudes of the fault signatures for these components may not exceed the vibration noise floor of the component. Therefore, a useful vibration diagnostic may not exist for these components in their current configuration.

CI Magnitude Change and Trend Detection

One technique developed to overcome the shortcomings of static thresholds and CI variability is CI trend detection. Instead of utilizing a static threshold, these methods generate an alert based upon step changes in the data or changes in trend rates or data scatter. Application of these techniques resulted in a 40% increase in diagnostic accuracy in some cases (Ref. 5). An evaluation of the algorithms did show that they are sensitive to noise and normal fluctuations of gearbox and drivetrain usage that are not necessarily faults. Further refinement of these algorithms would be required prior to fielding. These algorithms would also be difficult to implement into the current Army HUMS software structure.

Machine Learning

The Army’s current efforts to improve HUMS performance have centered around machine learning techniques and algorithms. Perhaps the most enviable capability of machine learning for HUMS applications is hidden-patterned data discovery. The human mind is limited by the dimensionality of the data that it can process. A computer is much less limited. Machine learning algorithms may be capable of finding attributes of the vibration and parametric data that are specific to fault types and create satisfactory separation from the healthy population in cases where a HUMS engineer cannot. In other words, machine learning may be able to turn what were previously classified false negatives into true positives. It may also be capable of improving the overall diagnostic accuracy of the HUMS.

Machine learning algorithms are split up into two general types: supervised and unsupervised learning. Supervised learning uses a training dataset and a test dataset. The test dataset is typically much smaller than the training dataset and is used to evaluate the supervised learning algorithm. The training dataset includes both the input data and the responses, which are used to mature the supervised learning model. Once the model is considered mature, it is evaluated using the input data from the test dataset and then scored based on the number of correct predictions of the response.

Unsupervised learning uses the entire input dataset (no responses) to find hidden patterns in the data or to group the data into clusters. Unsupervised learning is often used prior to using a supervised learning technique to reduce the dimensionality of the dataset. This paper explores the unsupervised learning technique expectation maximization (EM).

A paper has been published that outlines the overall process for implementing machine learning for an airworthiness application. This paper is titled: Using Machine Learning Algorithms to Improve HUMS Performance (Ref. 7). This paper outlines the rigorous process that the Army is implementing to build, test, evaluate and deploy a machine learning
algorithm for a CBM application that meets airworthiness standards.

**Expectation Maximization**

The EM technique is a method to find the maximum likelihood estimation of a subject parameter \( \theta \) of a probability distribution given observed data \( y \). This work specifically investigates EM clustering, also commonly known as fitting a Gaussian mixture model. Simply put, EM assumes that the data originated from a particular distribution, and the EM technique estimates that distribution for each data point. This technique is of interest to HUMS applications, since ideally EM will group the data into healthy and faulted distributions. The resulting model can then be used to classify new input data. The process of applying this technique is discussed in the next section.

**Theory**

The EM technique has two steps: the expectation step and the maximization step. The expectation step estimates the probability of each data point generated by a particular Gaussian. The maximization step modifies the statistical parameters of the model to maximize the likelihood of the data. The EM algorithm iterates between the expectation step and the maximization step until convergence occurs. The result of the technique is an estimate of the data boundaries for each distribution. These steps are further described as follows:

**Step 1: Initialize**

Choose initial guesses for: the probabilities \( \omega_j^{(0)} \), means \( \mu_j^{(0)} \), and variances \( \Sigma_j^{(0)} \), where \( j = 1, \ldots, k \) distributions within the dataset

Calculate the initial log-likelihood over \( n \) data points where \( \phi \) is denoted as the normal density:

\[
L^{(0)} = \frac{1}{n} \sum_{i=1}^{n} \log \left( \sum_{j=1}^{k} \omega_j^{(0)}\phi(y_i|\mu_j^{(0)}, \Sigma_j^{(0)}) \right) \quad (1)
\]

**Step 2: Expectation step computation**

Calculate the responsibilities \( \gamma \) and sum \( n_j^{(m)} \):

\[
\gamma_{ij}^{(m)} = \frac{\omega_j^{(m)}\phi(y_i|\mu_j^{(m)}, \Sigma_j^{(m)})}{\sum_{j=1}^{k} \omega_j^{(m)}\phi(y_i|\mu_j^{(m)}, \Sigma_j^{(m)})} \quad (2)
\]

for \( i=1,\ldots,n, j=1,\ldots,k \)

\[
n_j^{(m)} = \sum_{i=1}^{n} \gamma_{ij}^{(m)} \quad (3)
\]

**Step 3: Maximization step computation**

Calculate the mixing probabilities, and weighted means and variances:

\[
\omega_j^{(m+1)} = \frac{n_j^{(m)}}{n}, j = 1, \ldots, k, \quad (4)
\]

\[
\mu_j^{(m+1)} = \frac{1}{n_j^{(m)}} \sum_{i=1}^{n} \gamma_{ij}^{(m)} y_i, j = 1, \ldots, k, \quad (5)
\]

\[
\Sigma_j^{(m+1)} = \frac{1}{n_j^{(m)}} \sum_{i=1}^{n} \gamma_{ij}^{(m)} (y_i - \mu_j^{(m+1)}) (y_i - \mu_j^{(m+1)})^T, j = 1, \ldots, k \quad (6)
\]

**Step 4: Check for convergence**

Compute the new log likelihood:

\[
L^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} \log \left( \sum_{j=1}^{k} \omega_j^{(m+1)}\phi(y_i|\mu_j^{(m+1)}, \Sigma_j^{(m+1)}) \right) \quad (7)
\]

If \( |L^{(m+1)} - L^{(m)}| < \) a predetermined threshold, end the algorithm, otherwise return to the expectation step (Ref. 8) and (Ref. 9).

**Application**

The EM technique using a Gaussian Mixture Model was applied to the Apache Nose Gearbox (NGB) ground truth database. The Apache NGB database was selected because it has a large quantity and variety of faults. In order to properly deploy a machine learning algorithm, the training datasets must be proportional to the number of healthy and faulted components of the deployed aircraft fleet at any given time. The ground truth dataset had a large number of faults but few known healthy components. To address this issue, the training dataset was completed with historic data from assumed healthy components based on operation time and the definition of remaining useful life. For these purposes, a component was considered healthy if it had more than 200 flight hours of remaining useful life remaining. If a component was installed on aircraft for 1500 hours, it is assumed that through 1300 flight hours the component was healthy. The additional 50 flight hours at the beginning of operation were given to account for run-in time.

For this work, the training dataset was populated with 9 CIs selected by subject matter experts. These 9 CIs have been found to be the most effective in the detection of Apache NGB dynamic component faults. A fleet-proportional training dataset was then completed using these 9 CIs.
2 Clusters

The EM technique requires the user to input the number of distributions or data clusters that are present in the dataset. There are techniques that can automatically detect the number of clusters within a dataset, but those are outside the scope of this work. The most basic application of the EM technique is to input 2 clusters with the intention that the EM algorithm will separate healthy from faulted data. Figures 3 and 4 display the results in 2 dimensional space for 4 CIs from the application of the EM technique for 2 clusters.

Figure 3: EM GMM for 2 Clusters AFD CIs

Figure 4: EM GMM for 2 Clusters STACIs

The EM algorithm was effective in separating the higher magnitude CI values from the lower magnitude CI values. Histograms of the clusters for each CI were calculated (Figures 5 and 6), confirming that the higher CI values were grouped into one cluster.

Figure 5: Histogram of CI for 2 Clusters

Figure 6: Histogram of CI for 2 Clusters

The EM algorithm determined the clusters to have a distribution of 82% for the blue distribution (assumed healthy), and 18% for the red distribution (assumed faulted). The actual fleet distribution has a much higher healthy proportion; therefore, the rates for the accuracy, true positive, false negative, and false positive were calculated to determine the effectiveness of the EM technique:

Accuracy = 83.4%
True Positive Rate = 76.6%
False Negative Rate = 23.5%
False Positive Rate = 16%

The EM technique utilizes a posterior probability metric to assign each data point to a cluster. A probability of 1 suggests that it is very likely that the subject data point belongs to its assigned cluster, whereas a probability near 0.5 has a low confidence that the subject data point belongs to its assigned cluster. The probabilities for the false positive and false negative data points are shown in Figures 7 and 8. It was found that the vast majority of the data points have a high probability that they belong to their assigned distributions.
The training dataset included several faulted and healthy types. It is helpful to know the type of fault associated with a particular assumed faulted data point. The type of fault has severity and prognostic capability implications. The data was labeled either healthy or faulted, and if faulted, included the type of fault. Possible labels included healthy, output duplex bearing fault, gear fault, input roller bearing fault, input triplex bearing fault, shaft fault, and combination fault. The clusters were then evaluated to determine which types of fault belonged to each cluster. As can be seen in Figure 9, the ‘healthy’ cluster contains over 80% of the known healthy data, all of the input triplex bearing faults, a small portion of the output duplex bearing faults, and the majority of the shaft faults. The ‘faulted’ cluster contains nearly 80% of the output duplex bearing faults, over 90% of the gear faults, all of the input roller bearing faults, and the majority of the combination faults.

Several other cluster inputs were attempted to improve performance with little success. In an effort to discriminate each fault type, 6 distributions were input into the algorithm. The EM technique did successfully separate the bearing and gear faults (Figure 10), but it also distributed the healthy data among the clusters, which is not a desirable outcome. Another attempt was made to run the EM algorithm with 2 clusters, then re-run the algorithm using the ‘faulted’ data in order to separate the faults, but this too was unsuccessful.

A fault case was selected to determine how a particular fault would be classified by the EM algorithm as the fault progressed. The simple 2 distribution input EM algorithm was selected for this case. A faulted input roller bearing was examined (Figure 11). Two CIs reacted to the fault. High Frequency Energy (Figure 12) trended past the caution threshold prior to the gearbox being removed.
from service on July 28. Mid Frequency Energy (Figure 13) trended past its exceedance threshold giving several red indications.

Figure 11: Input Roller Bearing Fault

Figure 12: Roller Bearing Fault CI vs. Time

The results of the test are summarized in Figure 14. The blue dots represent the normalized Mid Frequency Energy CI classified by the EM algorithm as part of the healthy cluster. The red dots represent the normalized Mid Frequency Energy CI classified by the EM algorithm as part of the faulted cluster. The probability that the subject data point belongs to its assigned cluster is represented by the green line. The majority of the data points are determined to be healthy by the EM algorithm prior to the CI beginning to trend in June. The probability, however, is highly variable until the CI begins to trend, then the confidence is high that the component belongs to the faulted dataset. Once the gearbox is replaced at the end of July, the gearbox is considered to be in the healthy cluster by the EM algorithm. The variance of the probability and the intermittent cluster selection suggest that the EM technique may be able to detect a fault prior to the CI trending upward.

Figure 13: Roller Bearing Fault CI vs. Time

Figure 14: Cluster and Probability Change of Roller Bearing Fault vs. Time
Conclusions

Efforts to improve diagnostic techniques for Army rotorcraft HUMS have been discussed and evaluated. The work presented herein shows that machine learning may be used to further enhance the diagnostic and prognostic capability of the HUMS. The expectation maximization technique in particular has been reviewed, and it has been determined that it may be a useful technique as part of the machine learning process. It was discovered that EM performed the best when 2 distributions were specified. This result may change when more parameters are introduced into the input matrix and greater distinction can be made between the fault modes. In its current configuration, EM cannot be deployed as a standalone function due to its inability to meet ADS-79 criteria. It may, however be a valuable function for dimension reduction, or just an important input into the supervised machine learning dataset. This will only be known after the supervised learning models are evaluated.

Future Work

The exploration of the EM technique is in its infancy. Future work will include introducing the following features into the training dataset: additional CIs, the results of principal component analysis, trend detection features, and 120KTA data. Processes for the auto-detection of the number of clusters within a dataset will also be researched.

References


