Final Project Report

Data Mining & Optimization Tools
For
Developing Engine Parameters Tools

Sponsor: NASA

Duration: 6/1/98-8/30/98

Submitted December 21, 1998

Atam P. Dhawan, Ph.D., Principal Investigator

Final Project Report

Data Mining & Optimization Tools
For
Developing Engine Parameters Tools

Sponsor: NASA

Duration: 6/1/98-8/30/98

This project was awarded for understanding the problem and developing a plan for Data Mining tools for use in designing and implementing an Engine Condition Monitoring System. From the total budget of $5,000, Tricia Erhardt, a graduate student in the College of Engineering was hired during this time. The remaining money was spent on travel to NASA Lewis Research Center, Cleveland to discuss the problem and issues related to Engine Condition Monitoring System.

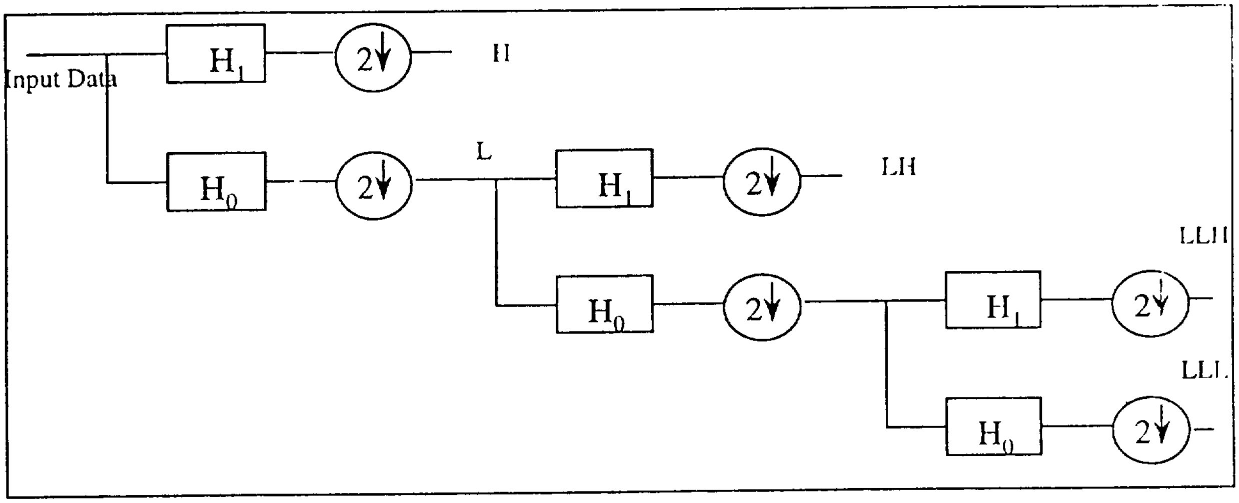
Tricia and I studied the problem domain for developing an Engine Condition Monitoring system using the sparse and non-standardized datasets to be available through a consortium at NASA Lewis Research Center. We visited NASA three times to discuss additional issues related to dataset which was not made available to us. We discussed and developed a general framework of data mining and optimization tools to extract useful information from sparse and non-standard datasets. These discussions lead to the training of Tricia Erhardt to develop Genetic Algorithm based search programs which were written in C++ and used to demonstrate the capability of GA algorithm in searching an optimal solution in noisy datasets. From the study and discussion with NASA LeRC personnel, we then prepared a proposal, which is being submitted to NASA for future work for the development of data mining algorithms for engine conditional monitoring. The proposed set of algorithm uses wavelet processing for creating multi-resolution pyramid of the data for GA based multi-resolution optimal search.

Wavelet processing is proposed to create a coarse resolution representation of data providing two advantages in GA based search:

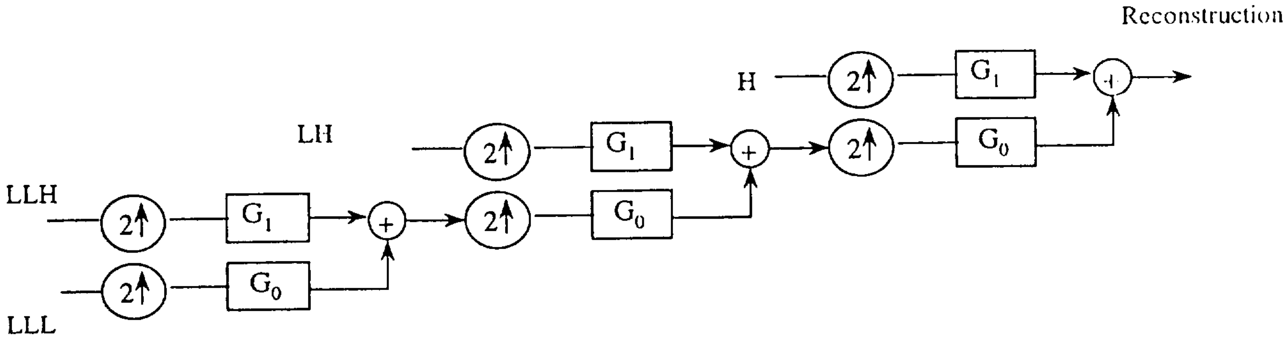
1. We will have less data to begin with to make search sub-spaces.
2. It will have robustness against the noise because at every level of wavelet based decomposition, we will be decomposing the signal into low pass and high pass filters. With multi-level decomposition, not only the data will be reduced through LLL component (see below), it will create several band-pass

filter characteristics through H, LH, and LLH components. The band-pass filter components can be used to define specific events for signal characterization such a transient of a specific frequency with a temporal localization. This should be noted that even if the actual dataset is not sparse, the information in the band-pass filter components will be quite sparse. This means that the data will not be distributed uniformly over the entire temporal space.

The three levels of decomposition would H, LH, LLH, and LLL versions of the signal. (L means the signal passed through low-pass filter, H means the signal passed through high-pass filter; H_1 is high pass filter and H_0 is low pass filter).



Using the selective thresholding schemes, one can remove the noise from the band-pass filter components and use reconstruction filters (G_0 and G_1) to bring data back to the original resolution. The advantage is that the reconstructed data will have less or no noise compared to the original data.



Now the question is how are we going to process the data. Each engine sensor data is 1-D temporal sequence which will be processed through wavelet decomposition to create different versions of the signal, e.g., LLL (number of samples are one-eighth). At the third level of decomposition, we also have LLH as band-pass component.

We now create the multi-dimensional search space with the data obtained at the third level of decomposition (just for the sake of example, more decomposition levels could be used for more data reduction). If we are searching using the GA. We need to start a seed solution of n out of m dimensions, where n is the selected number of variables by the GA or seed solution, and m is the total number of variables present in the database. For discovering a correlation/event in the selected search space, we need to cluster the data using fuzzy clustering algorithm and label it in the context of the event. To do data mining, one has to have an evaluation function to discover correlated events or patterns in the subspace provided by the GA search algorithm. Thus, we can clusters in the data space according to the classes or events. These clusters are used to compute correlation for the evaluation function. We propose to use fuzzy radial basis function network to compute the fitness evaluation function in GA algorithm. We can do the entire search in the original data at the acquired resolution but it will be inefficient and noise sensitive. Once we start searching on LLL component, we can apply fuzzy clustering approach to see which clusters are representing a homogeneous class. This part of the sub-space does not need to be resolved any further. This means that only clusters with heterogeneity (having more than one class) needs to be resolved further on the finer resolution data obtained through reconstruction filters. In other words, we can easily map the subspace in the finer resolution version of the data (obtained through reconstruction filters) which needs to be searched further.

In the implementation, the seed solution and the initial clusters for the potential solution is provided by the GA initiator search module. That means, the clusters will be resolved until a specified decomposition level. This is necessary to make the search efficient by not reaching to the original resolution level until the entire search space is processed. For example, the above-described method is used on the datset provided by LLL, LLH and LL, and LH. Once the solution is found at the LL level, the solution is projected over the L, H levels and then the original resolution by the GA search verifier module, which verifies that the solution is still valid. This check of validity and verification depends on the type of problem and its specifications. The specific details will be designed and implemented following the discussion and understanding of the statistical nature of the datasets.