Data Prognostics Using Symbolic Regression

V. Hunter Adams[*]
Cornell University
317 Upson Hall Ithaca, NY 14853
vha3@cornell.edu

Mason Peck[†]
Cornell University
249 Upson Hall Ithaca, NY 14853
mp336@cornell.edu

ABSTRACT

This paper describes a general technique for data prognostics using symbolic regression. This analysis treats the characterization of turbofan engine degradation as a particular case for the general technique. The proposed genetic program characterizes engine degradation, and then uses that characterization to both detect engine faults and predict the remaining lifetimes of engines after a fault. The genetic program takes advantage of the fact that engine degradation manifests itself as changing correlations between sensor outputs, and searches for those correlations. NASA Prognostics Data Repository provides a training set in which 100 simulated engines are run to failure, and a test set in which a separate set of 100 simulated engines are shut off before they fail [5]. This analysis trains the genetic program on the training set and verifies functionality on the test set. The genetic program successfully detects the moment that the fault occurs for every engine in the test fleet and accurately predicts the remaining lifetime of the engines after the fault.

Categories and Subject Descriptors

I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search

General Terms

Prognostics, Algorithms

Keywords

Symbolic regression, Genetic Programming (GP), Engines, Machine Learning

1. INTRODUCTION

In complex engineering projects such as turbine development, there is a huge demand for low-cost health monitoring systems. Failure of machines like these leads to extremely expensive or dangerous downtime, and data prognostics methods like the one presented here offer a means for mitigating risk by detecting degradation before failure occurs.[1]

[*]Graduate Research Assistant
[†]Associate Professor

Prognostic health monitoring is an area of very active research and has been approached from a number of different perspectives. Moghaddass and Zuo use a machine learning approach. They characterize engine degradation as a multistage process and treat the prognostics process as a classification problem that places each engine in a particular phase of degradation [1]. Others have approached the problem using belief functions, neural network based particle filtering, and hidden Markov models. [5] [4].

Using symbolic regression to characterize physical systems is not without precedent. Dr. Hod Lipson uses this technique for determining system equations of motion using only data [3], and Byington et. al. of Impact Technologies use symbolic regression on spectroscopy data to evaluate the health of diesel engine lubricating oil [4]. Eureqa by Nutonian, software developed under the advisory of Dr. Hod Lipson, uses symbolic regression to find relationships in complex data sets [6]. Symbolic regression has not been applied to this problem specifically, but its similarity to those mentioned above suggest that it is a promising technique.

Symbolic regression is particularly well suited to data prognostics. The genetic program searches for an integer corresponding to the number of cycles remaining in the engine. The solution is determined via a function of sensor output, which is a structure that symbolic regression easily accommodates.

2. PROBLEM DEFINITION

The input to the system is sensor data from 100 engines. Each engine has 26 sensors that, for the training data set, record information until the engine fails. The output of the system is a single number for each engine, which corresponds to that engine's remaining lifetime. The program evolves function of the form:

$$L_{remaining} = f(sensor_i, sensor_j, sensor_k, \dots) \qquad (1)$$

Where $L_{remaining}$ is remaining engine lifetime, which may be a function of any combination of some or all of the sensor outputs. The program is entirely system agnostic. It treats all sensors purely as sources of data and has no information about what the data represents physically.

3. METHOD

The genetic program operates on a population of solutions. For each generation, the population goes through five distinct steps [2]. These steps include:

1. Ranking the population according to the established fitness criteria
2. Selecting a subset of the population that will survive/breed into the next generation.
3. Performing crossover among the surviving parent population to create a population of children.
4. Performing mutation on the child population.
5. Adding the child population to the parent population to replenish the original population size.
6. Returning to step 1, and repeating for many generations.

Each of the above steps is described in sections 3.1-3.5.

3.1 Rank the Population

Elitist multi objective optimization on a Pareto front is used to maintain genetic diversity in the population[2]. Solutions are optimized along four dimensions:

- Age: Solutions that have been in the population for a shorter amount of time are more fit (in the age dimension) than solutions that have been evolving for a long time. This helps maintain diversity in the population by giving solutions with more potential to evolve an advantage over those that have become stagnant. During crossover, the child adopts the age of its oldest parent. [2]
- Mean Prediction Error: Solutions that have a lower average prediction error (measured across all engines in the fleet) are more fit than solutions that have higher average error.
- Uniqueness: Solutions that have better predictability on particular engines for which other members of the population are unable to predict are more fit in the uniqueness dimension. This helps maintain diversity.
- Worst Prediction: Solutions with lower error on their worst prediction are more fit than solutions with higher error on their worst prediction (even if they have a lower mean error over all engines). This prevents the algorithm from getting stuck at the mean remaining lifetime of all engines in the fleet.

The solutions that compose the population are placed on a series of Pareto fronts according to the NSGA non-dominated sort described by Seshadri in [7]. In brief, the population is sorted in the following way:

1. Initialize the number of individuals that dominate each member of the population to 0 and the members of the population that each member of the population dominates to an empty list.
2. For each member of the population p, loop through every other member of the population q. If p dominates q, then add q to the list of solutions that p dominates. If q dominates p, then increment the domination counter by 1.
3. If the domination counter equals 0 for a particular solution, then add that solution to the leading Pareto front.
4. Initialize a second Pareto front as an empty list.
5. Decrement the domination counter for every solution by 1. If the domination counter for any of the solutions becomes 0, add it to the second Pareto front (because this indicates that it was only dominated by one of the individuals in the first Pareto front).
6. Return to step 3 and continue until all solutions have been placed in a Pareto front

Domination is defined as being the equally or more fit along each of the four dimensions of fitness (age, mean prediction error, uniqueness, and worst prediction), and more fit along any one of the dimensions. Once each solution is in a front, the algorithm moves on to the selection process [8].

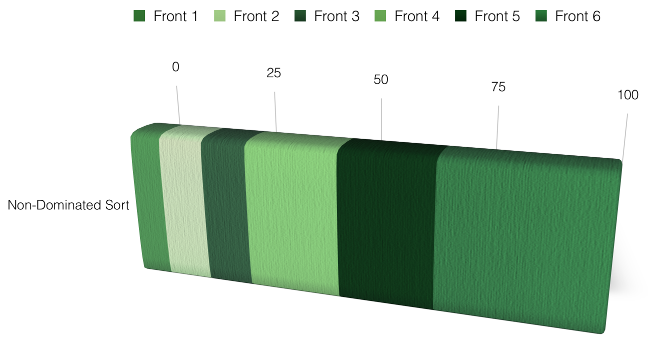

Figure 1: A visualization of how the population is divided into fronts. The relative sizes of the fronts change from generation to generation.

3.2 Selection

The algorithm uses elitism in that any solution that lives in the first Pareto front is guaranteed to survive into the next generation. All members of the population have some probability of surviving, but the solutions that occupy the more fit Pareto fronts have a higher probability of making it to the next generation.

The algorithm works with a selection pressure of 0.4. This is an empiracally determined parameter that can be tuned for different applications. Tighter selection pressures led to homogeneity in the population. After the members of the leading Pareto front are added to the surviving population, 80 percent of the remaining survivors are picked from the top 60 percent of the old population, and 20 percent of the remaining population are randomly generated *new* solutions. It is rare that any of these solutions have better predictive abilities than the older solutions that have been evolving for longer, but because they are younger than the rest of the solutions a few are able to survive to the next generation [2]. This helps maintain diversity in the population.

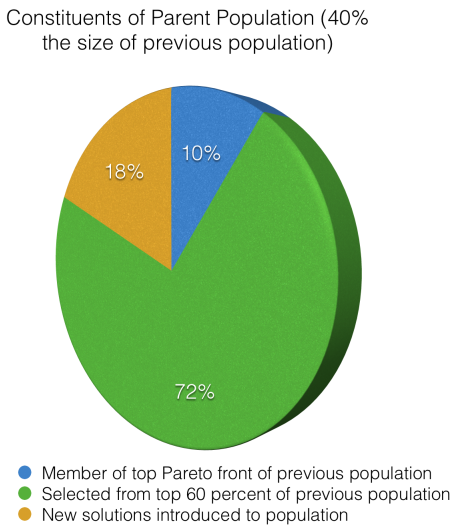

Figure 2: Constituent members of parent population after selection. The relative sizes change for each generation, depending on the size of the leading Pareto front.

3.3 Crossover

In order to replenish the population to its original size, crossover is performed to produce child solutions from the surviving parent solutions. The mother is preferentially chosen to be among the elite members of the population (ranking somewhere in the top 20), but the father is randomly selected from the parent population. This practice leads to useful diversity in the children. When both parents are randomly selected from the surviving population, they produced very diverse children, none of whom are particularly fit.

Because the members of the population are functions that are represented as a tree structure, crossover amounts to swapping branches between parents to produce a new child that has traits of both mother and father [3]. The depth of crossover is randomly determined for each parent every time crossover occurs.

3.4 Mutation

After crossover creates a new child, it is mutated before being placed into the population. While crossover allows the population to strategically explore new parts of the optimization landscape, mutations are small variations that allow solutions to climb to the nearest peak. The mutations are constructed such that they are not disruptive to the good genes in the genome.

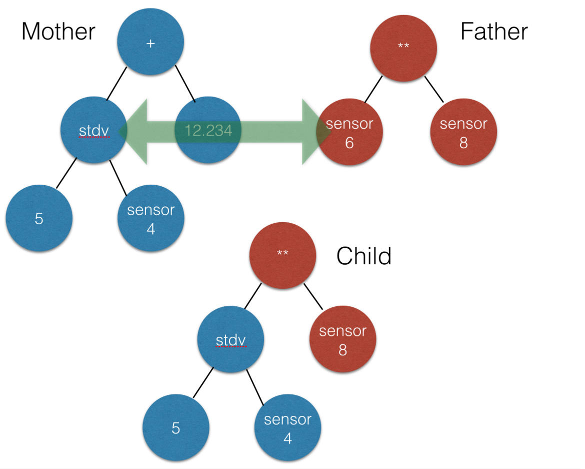

Figure 3: Two parent solutions producing a child solution by swapping branches.

The mutation rate cools as the algorithm runs, until it reaches the 100th generation. At this point the mutation rate heats back up before cooling off again over the course of the next 100 generations. The algorithm has the ability to mutate the value of a constant, change a constant to the output of sensor, or change a sensor variable to a constant. The algorithm may not mutate operators, because this sort of mutation is often extremely disruptive to the existing genes.

3.5 Replacement into Population

In order to maintain diversity, the children are replaced into the population using a form of deterministic crowding. After a child is created, the program checks whether it dominates either of its parents. If the child dominates a parent, then it replaces the parent.

Because the algorithm uses a cooling (and periodically heated) mutation rate, there are a few generations for every 100 for which mutations are extremely rare. If crossover occurs between two solutions that are not very deep (contain very few branches), it is possible for the child to be identical to one of its parents. In order to prevent duplicate solutions in the population, every child is compared with every member of the population before being injected into it. If a child does not replace a parent, and it is not identical to a solution that already lives in the population, then it enters the population.

4. RELATED WORK

Data prognostics is an area of very active research that has been approached from a number of different directions. Moghaddass and Zuo applied a different sort of algorithm to this particular data set, while Dr. Lipson and the team at Nutonian tackle the general problem of data prognostics using a technique similar to that which I have implemented. [1] [6]

Moghaddass and Zuo make the point that a multistage degradation can either be considered a continuous or a discrete process. By modeling degradation as a continuous process, one defines a threshold beyond which the machine is considered failed. In a discrete consideration of the same problem, the degradation process is quantized into discrete levels ranging from perfect operation to total failure. Traditionally, a discrete representation has been applied more to data diagnostics than data prognostics (that is, it answers the question ”what is the current state of the machine?” rather than answering the question ”how much longer until this machine fails?”). Moghaddass and Zuo are unique in that they do not assume a constant transition rate between states in order to estimate remaining engine lifetime, but they instead used non homogenous continuous-time semi-Markov processing for estimating the rate of transition between states. This allowed them to use traditional classification methods to place the engine in a particular state of degradation, and then estimate remaining lifetime based purely on its current state by estimating transition rates. Each state had an estimated probability of transitioning to another state in a given amount of time. This technique showed convergence on true remaining life for nearly all engines. [1] This analysis approaches the same problem using symbolic regression rather than the method explained above.

Eureqa by Nutonian is a software package that uses symbolic regression to find trends and relationships among in general data sets. Eureqa offers the user the ability to ”set a target,” where he/she specifies the particular variable that should be modeled by the other variables in the data set. This is similar to the problem which is approached in this paper. In this case, the genetic program is asked to provide remaining engine lifetime as a function of sensor outputs [6].

5. SYSTEM ARCHITECTURE

5.1 Structure

The problem is framed in a machine learning context. The genetic program evolves tree data structures, where each node is an operator and each leaf is either a constant or an array of sensor output. The operators are arranged in a dictionary, and they include:

- Arithmetic: Left leaf and right leaves are combined according to the arithmetic operator (+, -, *, /). These are four separate operators.
- Trigonometry: Left leaf is multiplied by the result of acting one of the three main trigonometric functions on the right leaf. These are three separate operators.
- Exponentiation: Left leaf is multiplied by the exponential of the right leaf.
- Logarithm: Left leaf is multiplied by the natural logarithm of the right leaf.
- Noise: Left leaf is multiplied by a Gaussian smear of the right leaf.
- Standard Deviation: Left leaf is multiplied by standard deviation of right leaf.
- Gradient: Left leaf is multiplied by gradient of right leaf.
- Second Gradient: Left leaf is multiplied by second gradient of right leaf.
- Window Operators: Left leaf is multiplied by one of the above 3 operators (standard deviation, gradient, or second gradient) acting only on the most recent 10 data points from the right leaf. These are three separate operators.

Sensor output is arranged in a separate dictionary, with each key corresponding to a separate engine and the value of each key being a list of 26 lists, each list corresponding to a different sensor. The GP evolves functions with the variables represented as keys of these dictionaries. When a function is evaluated for fitness, it is evaluated on every engine in the training fleet.

At a high level, the program works with objects of two classes. The ”person” class is a tree structure. An object of the person class has the ability to generate predictions for a particular engine, to evaluate its own depth, to perform crossover with another object of the person class, to mutate itself, to determine if it is dominant to another object of the person class, and to evaluate the accuracy of its predictions are on a particular engine. There are other methods associated with this class, but they are all helpers to these main methods.

The ”population” class is, fundamentally, a list of objects of the person class. The population has the ability to perform operations on the population as a whole. It may add a person to the population, gather the traits of each member of the population (uniqueness, age, predictability, etc.), rank the population according to their traits, select a parent population from the entire population, and breed the members of the parent population to replenish its original size. These functions are combined into broad methods that attempt to maximize the fitness of the entire population.

As input, these broad optimization methods take mutation rate, population size, selection pressure, elitism pressure, and maximum allowable depth for the tree structures that compose the population.

5.2 Data

The data used for both training and testing comes from NASA’s Prognostic Data Repository. Each dataset consists of multivariate time series, each variable corresponding to a different sensor or operational setting. The data was produced by C-MAPSS simulation software, which is the industry standard for simulating transient effects in turbofan engine degradation. For each engine in the training dataset, the engine starts under normal operation, develops a fault, and runs to failure [5] [1]. In the test dataset, the engines all start under normal operation, develop a fault, but are *not* run to failure. This analysis uses dataset FD001, which is composed of engines of all the same type. Minimal preprocessing is required for the genetic program to begin exploring the data. Each dataset is loaded into the python module and parsed into a dictionary.

6. EXPERIMENTAL EVALUATION

6.1 Methodology

The fitness of each potential solution is judged in the training dataset alone, according to the criteria discussed in section 3. For sake of making accurate predictions, however, the important variable is prediction error. The other dimensions of the Pareto front exist to provide diversity and to help make improvements along this one dimension of actual concern. The constituent members of each population compete in the training set, and the performance of the GP is baselined against that of the hill climber and random search in the training data set. Every variable that composes a dimension of the Pareto front is a dependent variable, and the sensor outputs form the independent variables. The performance data, however, is *mean prediction error in the training data set.* The solutions that better characterize the training dataset are considered better solutions than those that cannot make as accurate predictions. Comparisons are made between populations through performance curves that show predicability plotted against the number of evaluations.

Performance on the test dataset is evaluated by applying the function generated on the training dataset and comparing the predictions against the true values. These predictions can be judged on an engine-by-engine basis by plotting the true remaining lifetime of the engine (for every moment in time) along with the prediction generated by the function (for every moment in time). These are the sorts of plots that Moghaddass and Zuo used to make empirical judgements of the performance of their algorithm [1].

6.2 Results

6.2.1 *The Toy Problem*

In the real problem , the GP is asked to return some combination of sensors that outputs the remaining lifetime of the engine. A priori, however, there is no guarantee that such a solution even exists. For that reason, it is important to create a toy problem that verifies the GP will be able to find such a solution, if it does exist. This toy problem should be one where the solution exists and is known. In other words, there exists a combination of sensors in the toy dataset that will return remaining engine lifetime with absolute precision.

This can be accomplished by giving the GP, hill climber, and random search access to a clock on each engine. The clock is represented as another sensor, the output for which starts at 0 and linearly increases until the engine fails. The remaining lifetime of the engine, therefore, is just a scaling of this sensor's output. The optimal solution should include just the output of the clock and a scaling. Fig. 4 shows that the GP finds the optimal solution in the toy dataset, and it does so faster than either the random search or the hill climber. This suggests that, if such a solution exists in the real dataset, the GP will be able to find it.

6.2.2 *The Real Problem*

By removing the clock sensor from each engine, all optimization techniques are forced to search for a function that yields remaining engine lifetime based *solely* on sensor output. Figure 5 shows that the GP lost much of its advantage over the hill climber in this situation. The GP still finds a

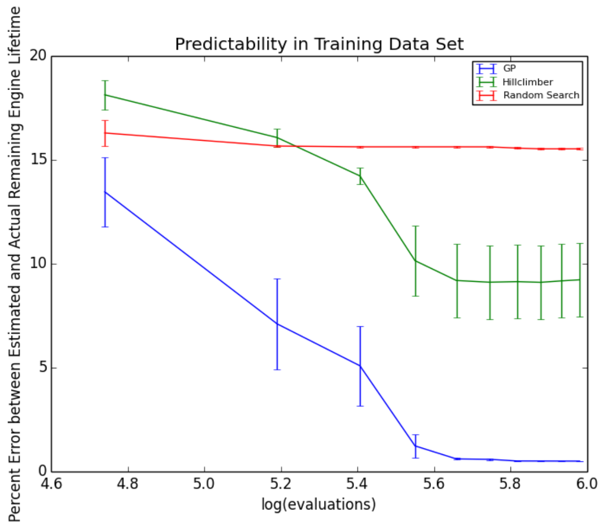

Figure 4: *Toy Problem*: **Percent error between estimated and true remaining engine lifetimes in the training data set. Shows GP finding the optimal solution constructed for the toy problem.**

better absolute solution than the hill climber, but the error bars overlap (Fig. 5). Among the best solutions that the GP found are:

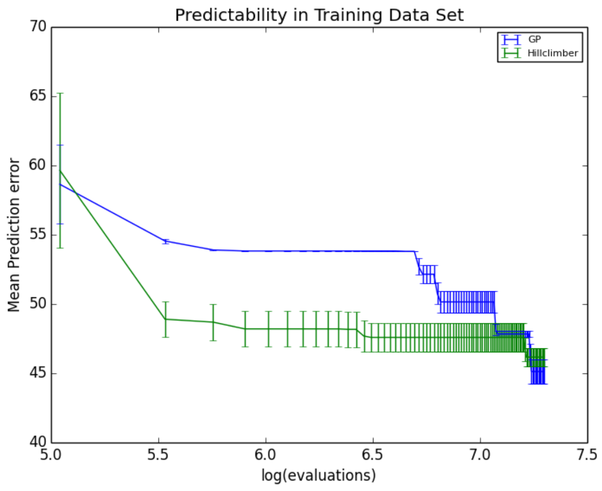

Figure 5: *Real Problem*: **Percent error between estimated and true remaining engine lifetimes in the training data set. GP loses some advantage over hill climber. Y axis is average difference between true and estimated number of remaining engine life cycles.**

$$Lifetime = -119.95\sin(sensor23 - sensor6) \tag{2}$$
$$Lifetime = -126.7854\cos(sensor15) \tag{3}$$

Fig. 6 shows the estimated vs. actual remaining engine lifetime of the first 20 engines in the test data set. The

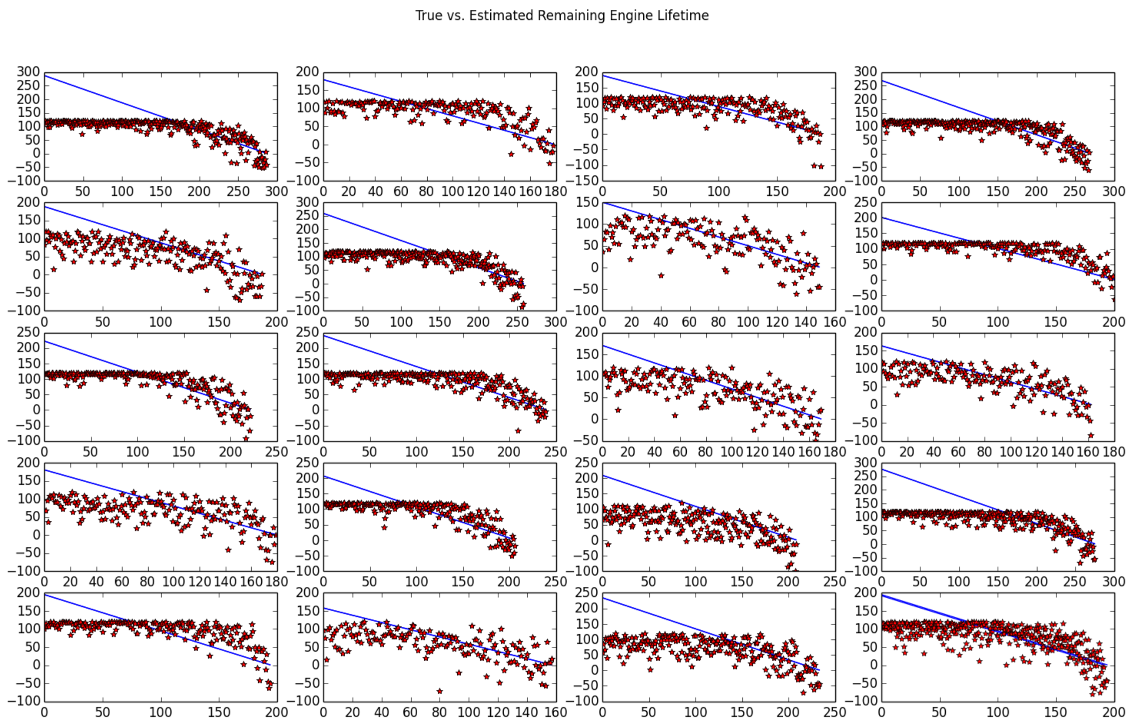

Figure 6: ***Real Problem*****: Estimated vs. actual remaining engine lifetime. GP successfully captures the moment of each engine's failure.**

GP not only captures the moment that the fault occurs in each engine (represented by the elbows in each curve), but also creates an empirically close estimate for each engines remaining lifetime after the fault.

6.3 Discussion

The genetic program described in this paper is able to determine the moment of failure for each engine tested. Although the GP's advantage over the hill climber lessens in the real problem (as opposed to the toy problem), the GP does a better job characterizing engine failure in both situations. The correlation that the GP isolates as being indicative of engine failure is the difference between the outputs of sensors 23 and 6. It is able to use this information to determine when each engine fails, and to accurately predict remaining lifetime after failure.

7. FUTURE WORK

Currently, fitness for each function in the population is evaluated on every engine in the fleet. By instead representing the fleet of engines as a population, and co-evolving a subset of engines for testing the fitness of each function, the GP will likely converge to a solution with less prediction error more quickly. A fit engine in the engines population would create disagreement among the functions in the solution population. It is also possible that the GP will do a good job answering a question that is different from the one asked in this analysis. Instead of asking for a number corresponding to the remaining engine lifetime, it may be possible to ask for a boolean that answers the question "will this engine fail in the next x cycles?"

8. CONCLUSION

This paper presents a general technique for performing data prognostics using genetic programming and symbolic regression. The validity of the technique is proven with the particular case of turbofan engine degradation. The genetic program successfully identifies the moment that each engine in the test fleet experiences a fault, and accurately predicts the remaining lifetime of each of those engines after the fault.

9. ACKNOWLEDGEMENTS

Thank you to Dr. Hod Lipson for his help.

10. REFERENCES

[1] Moghaddass, R., & Zuo, M. J. (2014). An integrated framework for online diagnostic and prognostic health monitoring using a multistate deterioration process. Reliability Engineering & System Safety, 124, 92-104.

[2] Lipson, H. (Director) (2014, September 1). Evolutionary Computation. CS 5724. Lecture conducted from Cornell University, Ithaca.

[3] Schmidt, M., & Lipson, H. (2010). Symbolic regression of implicit equations. In Genetic Programming Theory and Practice VII (pp. 73-85). Springer US.

[4] Byington, C. S., Mackos, N. A., Argenna, G., Palladino, A., Reimann, J., & Schmitigal, J. (2012). Application of symbolic regression to electrochemical

impedance spectroscopy data for lubricating oil health evaluation. ARMY TANK AUTOMOTIVE RESEARCH DEVELOPMENT AND ENGINEERING CENTER WARREN MI.

[5] Overview. (2008, January 1). Retrieved November 1, 2014, from http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/

[6] Nutonian, Inc. (2011, January 1). Retrieved December 14, 2014, from http://www.nutonian.com/products/eureqa/

[7] Deb, K., Agrawal, S., Pratap, A., & Meyarivan, T. (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. Lecture notes in computer science, 1917, 849-858.

[8] Horn, J., Nafpliotis, N., & Goldberg, D. E. (1994, June). A niched Pareto genetic algorithm for multiobjective optimization. In Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on (pp. 82-87). Ieee.