Data Prognostics Using Symbolic Regression

V. Hunter Adams

***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 application for the general technique. The proposed genetic program (GP) 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 exploits the fact that engine degradation manifests itself as changing correlations between sensor outputs. The 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. The GP uses the training fleet of engines to identify the sensor relationships that indicate engine fault and predict remaining lifetime, and then observes the learned sensor relationships in the test fleet. 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.**

Keywords*—*artificial intelligence, estimation, genetic algorithms, jet engines, learning systems.

I. Introduction

Data prognostics is an area of active research that has been approached from a number of different directions. 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]. To solve the related problem of characterizing a physical system using data, Dr. Hod Lipson uses symbolic regression to arrive at system equations of motion from dynamics data alone. [1] [3]

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. This analysis approaches the same problem using symbolic regression rather than the method explained above. In doing so, the degradation process is treated as a continuous one. [1]

Schmidt and Lipson use symbolic regression to identify physically relevant conserved quantities in experimentally gathered dynamics data. In doing so, they can experimentally arrive at Hamiltonians, Lagrangians, and other laws of conservation. As in this analysis, Schmidt and Lipson evolve functions of operators and constants that take the form of tree structures, and they perform mutation, crossover, and selection by similar means to that presented here. Depending on the building blocks provided to the GP, Schmidt and Lipson could discover different sorts of equations that describe the system (manifold equations, energy equations, equations of motion, etc.). The GP described in this paper is similar in structure to that of Schmidt and Lipson, but it departs from their research in that it is constructed to be prognostic rather than diagnostic. Furthermore, the GP described in this paper is designed such that it targets a particular relationship among sensors, as opposed to the GP of Schmidt and Lipson that has no target, but instead arrives organically at a number of equations describing aspects of the system in question. [3]

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

II. 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 functions of the form:

$$L_{remaining} = f(sensor_i, sensor_j, sensor_k, \ldots) \quad (1)$$

Where $L_{remaining}$ is remaining engine lifetime, which may be a function of any combination of some or all of the sensor

V. Hunter Adams is with the Department of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY, 14853 USA e-mail: vha3@cornell.edu

outputs. The program is entirely system agnostic. Each sensor is treated purely as a source of data without any connection to the physical world.

III. 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 population to its original size.
6) Returning to step 1, and repeating for many generations.

Each of the above steps is described in subsections IIIA-IIIE.

A. 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 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].

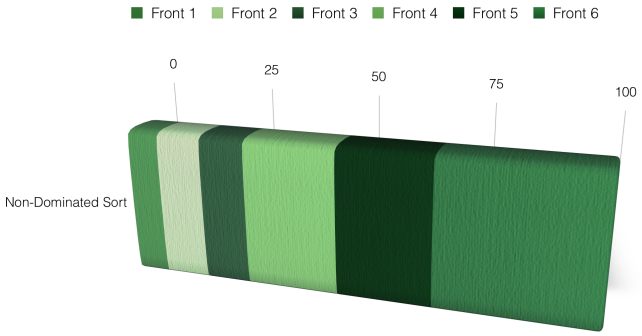

Fig. 1. Visualization of NSGA non-dominated sort. The relative sizes of the Pareto fronts change from generation to generation.

B. 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.

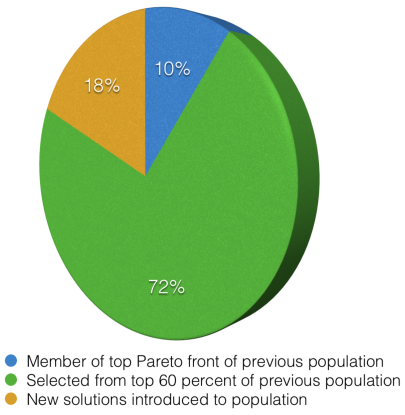

Fig. 2. Constituent members of parent population after selection. The relative sizes change for each generation, depending on the size of the leading Pareto front.

C. 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 solutions), 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. When both parents are selected from the elite members of the surviving population, the population loses diversity.

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.

D. 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.

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

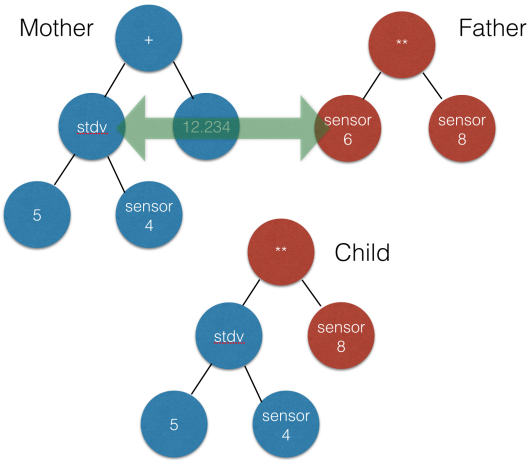

Fig. 3. Two parent solutions producing a child solution by swapping branches.

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.

E. 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.

IV. System Architecture

A. 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 leaf 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 the right leaf, which has been modified by additive Gaussian noise.
- 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.

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.

B. Data

The data used for both training and testing comes from NASA's Prognostic Data Repository. Each dataset consists of multivariate time series, with each time series corresponding to a different sensor or operational setting. The data was produced by C-MAPSS simulation software, 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.

V. Experimental Evaluation

A. 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].

B. Results

1) The Proof of Concept: In the engine prognostics problem, the GP is asked to return a function of sensor output that returns the remaining lifetime of the engine. *A priori*, however, there is no guarantee that such a solution 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.

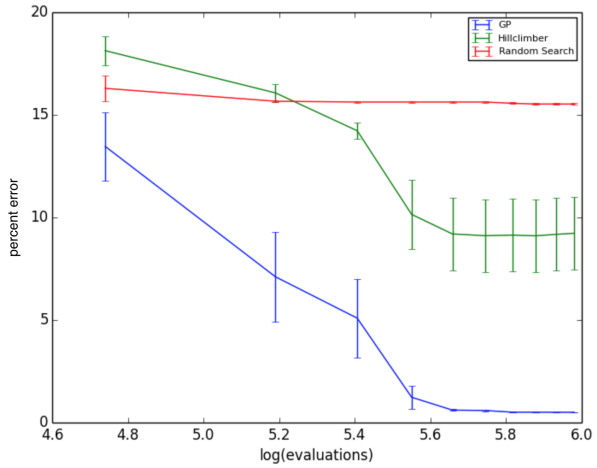

Fig. 4. *Proof of Concept*: Percent error between estimated and true remaining engine lifetimes in the training data set. Shows GP finding the optimal solution constructed for the toy problem.

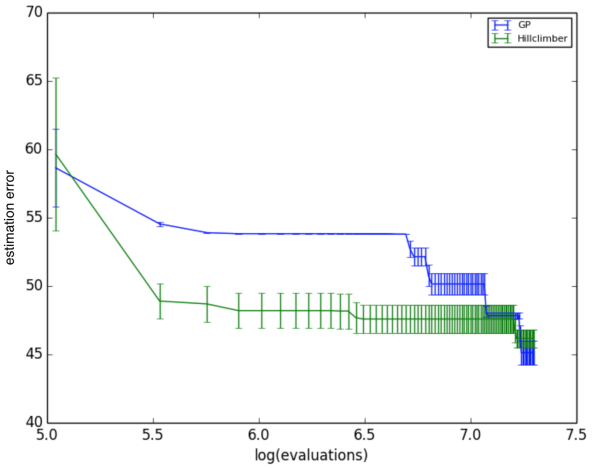

Fig. 5. *Engine Prognostics*: Difference between estimated and true remaining engine lifetimes in the training data set. GP loses some advantage over hill climber. Y axis is average difference between true and estimated number of remaining engine life cycles.

2) Engine Prognostics: 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 some of its advantage over the hill climber in this situation. The GP generally finds a better absolute solution than the hill climber, though not always (as indicated by the overlapping error bars). Among the best solutions that the GP found are:

$$Lifetime = -119.95 \sin\left(sensor23 - sensor6\right) \quad (2)$$

$$Lifetime = -126.7854 \cos\left(sensor15\right) \quad (3)$$

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

VI. Conclusion

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 engine prognostics problem (as opposed to the toy problem), the GP generally 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.

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?"

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.

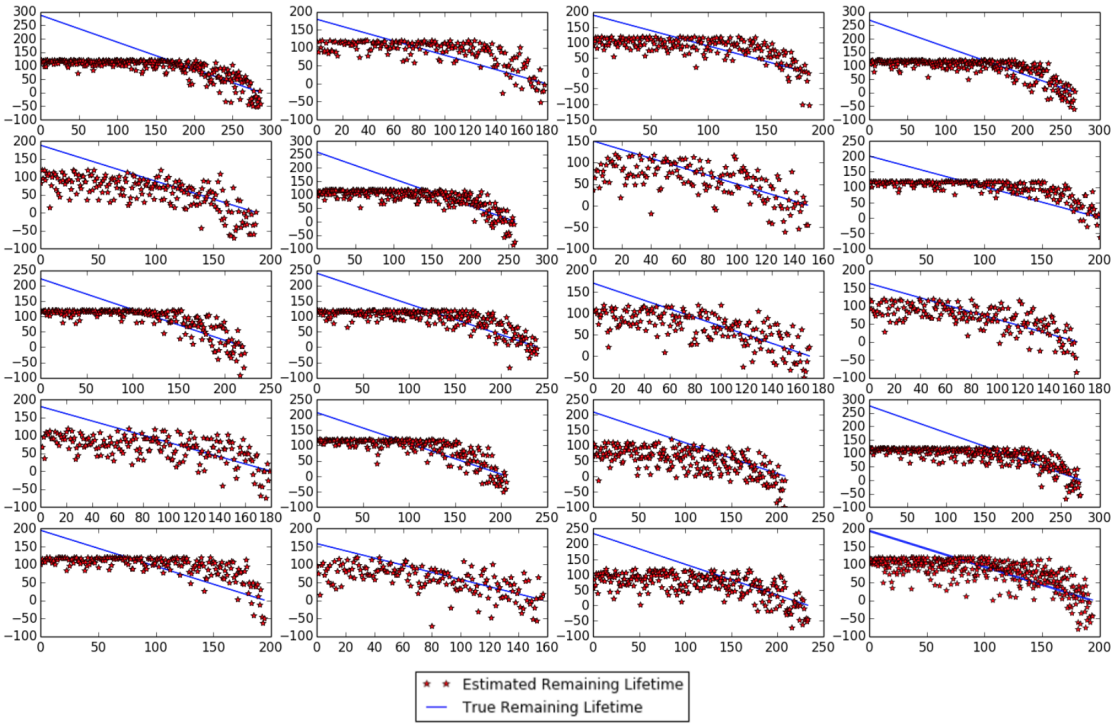

Fig. 6. *Engine Prognostics*: Estimated vs. actual remaining engine lifetime. GP successfully captures the moment of each engine's failure and estimates remaining lifetime after failure.

ACKNOWLEDGEMENTS

Thank you to Dr. Hod Lipson for his advice.

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, Michael, and Hod Lipson. "Distilling free-form natural laws from experimental data." science 324.5923 (2009): 81-85.

[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.