Examining the Role of Genetic Programming in the Analysis of Crowdsourced Data

Student Author: Robert Swain
Advisors: Josh Bongard and Paul Hines

Missing data is a problem faced by many researchers, yet is often left unexplored due to traditional data sets typically having few missing values. With the rise of the internet and cheap data storage, large data sets with high percentages of missing values are becoming increasingly common. Specifically, we look at crowdsourced data sets that not only contain a high percentage of missing data but also exhibit non uniform sparsity: most users participate very little, while a core group of users participate a great deal. It has been shown that as the percentage of missing data in a data set increases, the choice of how to deal with that missing data becomes increasingly important, yet many researchers continue to use antiquated techniques that result in biases in subsequent analyses. Because of this, here we investigate how linear regression and genetic programming perform as sparsity increases and the distribution of missing values becomes increasingly non uniform. We find that regression models produce lower forecasting errors than GP, but that the regression models also become much more complex.