Genetic Algorithm-R Package Final Report

David Chen, Qi Chen, Emily Suter and Xinyi(Cindy) Zhang

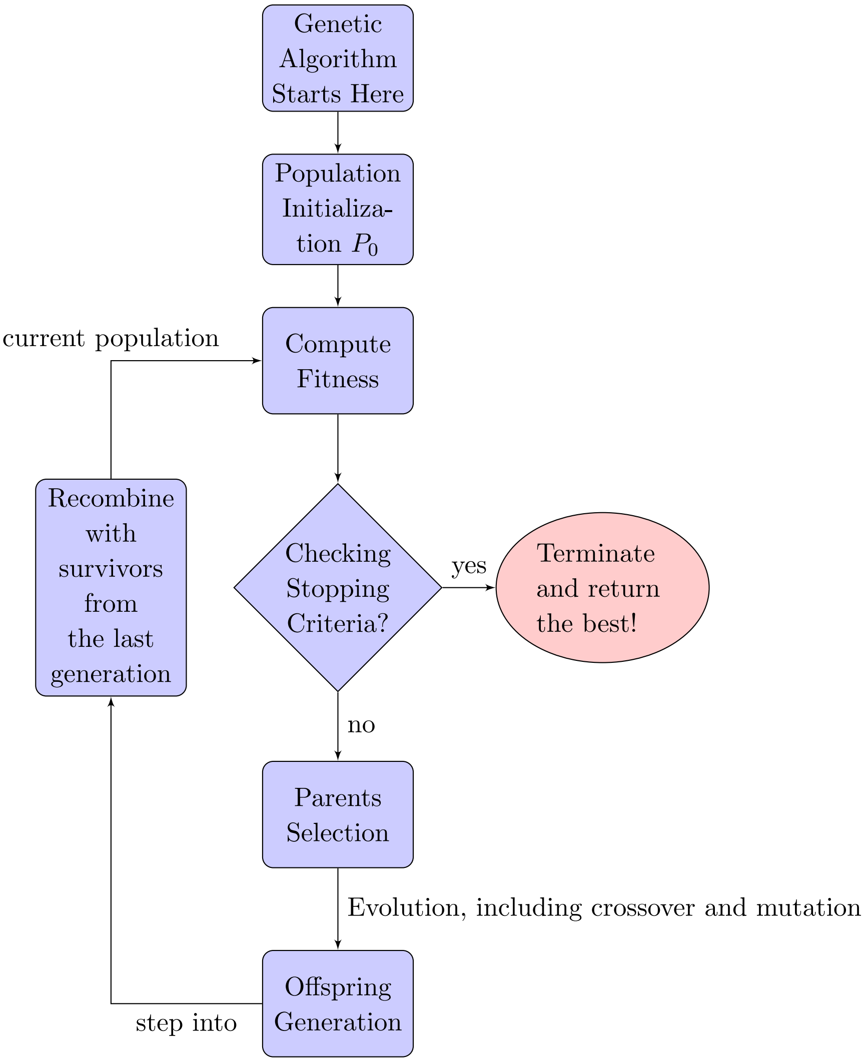
December 10, 2017

Abstract

Genetic Algorithm (GA) is a search-based optimization technique based on the principles of Genetics and Natural Selection. It is frequently used to find optimal or near-optimal solutions to difficult problems which otherwise would take a lifetime to solve. In this report, we will first introduce how we set up the genetic algorithm and the main steps. We then describe the testing procedure carried out. In section 3, we include the results from the example we have taken to apply our GA algorithm. Contributions of each team member is collected in the last part, section 4.

1 Introduction to Our Genetic Algorithm

To better introduce our genetic algorithm Figure 1, a flowchart is displayed as follows


Genetic Algorithm Starts Here
Population Initialization P_0
current population
Compute Fitness
Recombine with survivors from the last generation
Checking Stopping Criteria?
yes
Terminate and return the best!
no
Parents Selection
Evolution, including crossover and mutation
Offspring Generation
step into

2 Testing

3 Application

In this section, we considerde two applications of our genetic algorithm, one on the dataset generated from a linear regression model, which aims to evaluate whether the designed algorithm can successfully select those important features, given known relevant variables combined with some noise terms. Another application is based on the baseball data collected from the textbook "Computational Statistics" Givens and Hoeting (2013). Details will be illustrated as follows.

3.1 Application on Data Generated from Linear Regression Model

In this section, we will first introduce how we generate the covariates and responses for evaluating our genetic algorithm on feature selection. We first generate covariates from the multivariate normal distribution $N_p(\mathbf{0}, \Sigma_x)$, where the (j, j') Σ_x satisfies

$$\Sigma_{x(j,j')} = 0.5^{|j-j'|}, \text{ for } 1 \leq j, j' \leq p.$$

Setting sample size $n = 300$ and feature dimension $p = 20$, we then fit the linear regression model

$$Y = X\beta,$$

where $\beta = (\beta_1, \cdots, \beta_6)^{\mathrm{T}}$ is generated from univariate normal distribution $N(3, 25)$. Moreover, we add some noise $\epsilon = (\epsilon_1, \cdots, \epsilon_n)$ to the covariates generated as above. ϵ_i, for $1 \leq i \leq n$, is generated from $N_{10}(\mathbf{0}, \Sigma_x)$, which has the same parameter setting as X except the dimension. We then combine X and the noise terms together. The resulting covariates matrix is given by

$$\tilde{X} = (X|\epsilon).$$

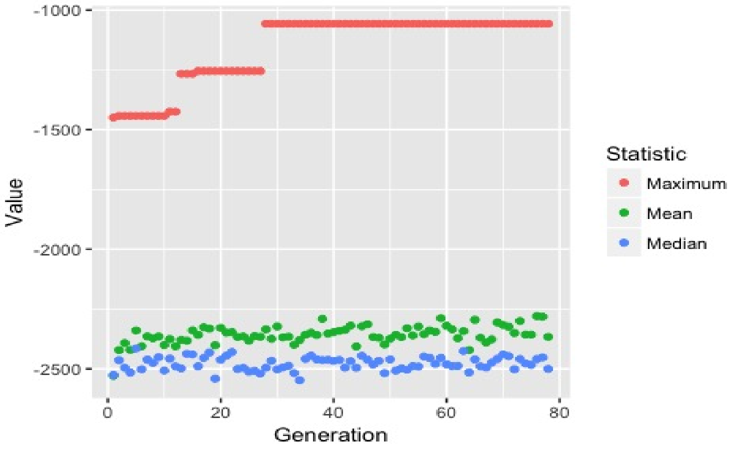
Apply our genetic algorithm to fit a linear model regressing $Y \in \mathbb{R}^n$ on $\tilde{X} \in \mathbb{R}^{n \times 30}$. Given 20 relevant covariates and 10 noise terms, the designed GA algorithm gives the following results for feature selection

$$\text{genotype} = (0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1).$$

One can see that our GA algorithm can select most of the relevant variables, but will also include some irrelevant noise terms denoted in red.

A plot monitoring the optmization process is also included:

Figure 1: How it evolves



3.2 Application on Baseball Data

4 Group Member Contributions

David Chen:

Qi Chen:

Emily Suter:

Xinyi(Cindy) Zhang:

References

Anupriya Shukla, H. M. P. and Mehrotra, D. (2015). Comparative review of selection techniques in genetic algorithm .

Givens, G. H. and Hoeting, J. A. (2013). *Computational Statistics*. Wiley.