Genetic Algorithm-R Package Final Report

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

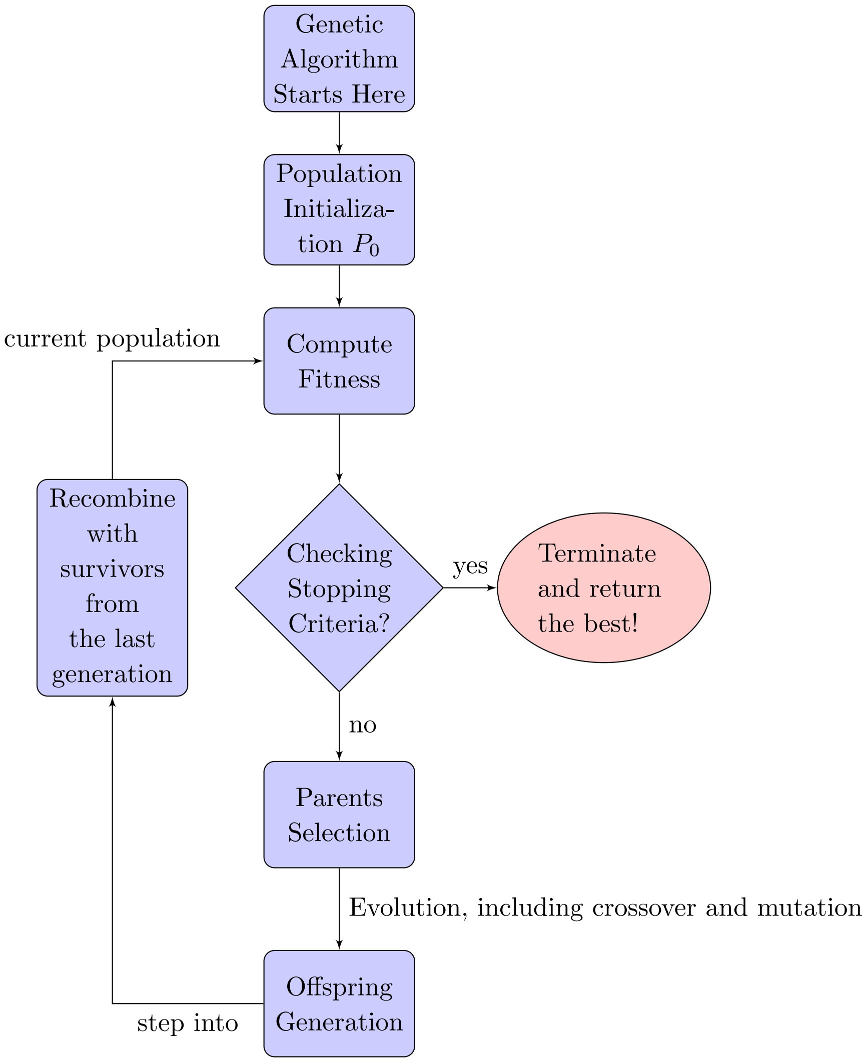
December 13, 2017

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

Ability to find max fitness.
Ability to

3 Application

In this section, we considered 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

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

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.

3.2 Application on Baseball Data

We next test our genetic algorithm on a real data set to demonstrate the ability to select and optimal variable subset. A data set of baseball player statistics and salary numbers was obtained via the website for Computational Statistics, 2nd Edition, by Givens and Hoeting (http://www.stat.colostate.edu/computationalstat

The data set contains 27 different statistics (such as hits and on-base percentage) for 337 players in the 1991 baseball season. Additionally, the data set contains the salaries for the same 337 players in the 1992 season. We used our genetic algorithm to select player statistic(s) that most influence that players salary in the following year.

3.3 Comparison with Global Search for optima

To investigate whether our GA algorithm can attain the same optimum as global search for all possible genotypes, we create another dataset to test this. How data generated is presented as follows:

Consider number of variables $p = 5$, and we generate covariates $X = (X_1, \cdots, X_5)^{\mathrm{T}}$ from different distributions, where $X_1 \sim N(0, 25)$, $X_2 \sim \text{Unif}(0, 1)$, $X_3 \sim \text{Poisson}(1)$, $X_4 \sim \exp(2)$, $X_5 \sim \text{Gamma}(10, 1)$. The responses Y is generated by simply averaging X_1 and X_3, i.e. $Y = \frac{X_1 + X_3}{2}$. Apart from finding optimal genotype via Genetic Algorithm, we also compute all the fitness values for all 32 genotypes to see what the exact global optimum is and thus the "best" genotype. Since the reponse Y is only related to variables X_1 and X_3, we hope that both the global search approach and the GA algorithm can only select these two features but not others. Additionally, we also hope that the results returned by these two methods are consistent. Comparison results are presented below in Table 1.

Table 1: Comparison of optimal fitness values with respect to global search and GA algorithm

Table 2: Global Search

	AIC	BIC
lm	19963.50	19945.41
glm	20175.50	20160.68

Table 3: GA Algorithm

	AIC	BIC
lm	19963.50	19945.41
glm	20175.50	20160.68

One can see from the results in Table 1 that for each combination of fitted model and fitness criteria, the GA algorithm and global search method give exactly the same results. In other words, our GA algorithm does find the global maximum fitness value, i.e global minimum value for AIC or BIC.

Next, we present the genotypes returned by both global search and GA algorithm in Table 4.

Table 4: Comparison of optimal fitness values with respect to global search and GA algorithm

Table 5: Global Search

	AIC	BIC
lm	11100	10100
glm	10100	10100

Table 6: GA Algorithm

	AIC	BIC
lm	11100	10100
glm	10100	10100

In summary, our GA algorithm can find the global optima, results of which are consistent with that from global search.

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.