Variable Selection using Genetic Algorithm

Shreyas Battula
Dept. of CSE
PES University
Bengaluru,India
shreyasb2002@gmail.com

Kiran H Kademani
Dept. of CSE
PES University
Bengaluru,India
kademani63@gmail.com

Vinay Kumar
Dept. of CSE
PES University
Bengaluru,India
vinaykumars7244@gmail.com

Abstract*—You always find yourself in ambiguous situation when it comes to choosing features for your analysis. And not including important features would lead to high bias and this would not lead to better prediction .In this article it is studied the application of a genetic algorithm in the problem of variable selection for multiple linear regression, minimizing the least squares criterion. The algorithm is based on a chromosomic representation of variables that are considered in the least squares model. A binary chromosome indicates the presence (1) or absence (0) of a variable in the model. The fitness function is based on the adjusted square R, proportional to the fitness for chromosome selection in a roulette wheel model selection. Usual genetic operators, such as crossover and mutation are implemented.

Keywords*—Genetic Algorithm, crossover, mutation , roulette wheel

I. INTRODUCTION

So variable selection is one of the most difficult yet most important task for data analysis. So genetic algorithm is used for optimization of such NP hard problems.

Genetic Algorithm(Holland, 1975)-

It is an evolutionary algorithm inspired by 'Survival of the Fittest' from Charles Darwin's theory of natural selection . This explains that the fittest individuals are selected to produce offsprings . GA are randomized search algorithms that generate high quality optimization solutions by imitating biologically occuring events like selection ,crossover and mutation.

- Terminologies involved-
  - Population- This contains set of possible solutions for stocastic search process to begin. Genetic Algorithm will iterate over this until the algorithm converges to an optimized solution.
  - Chromosome- This represents each candidate solution in the population. It is also referred to as Genotype.
  - Phenotype – This is the decoded parameter for the genotype that is processed by GA. Mapping is applied to the genotype to convert to a phenotype.
  - Fitness function – This evaluates the individual solution or Phenotypes for every generation for identification of fittest members.
- Genetic operators -
  - Selection – Selection is process of selecting the fittest solution from the population and these solutions act as the parents for next generation. This can be performed from Roulette Wheel.
  - Cross-over – Recombination when genes from two different parents are exchanged to form new genotype. Cross-over can be one point or multiple point based on the parent's segments of genes exchanged.
  - Mutation- Mutation is process to modify a genotype using a random process to generate diversity in the population to find better and optimized solution

II. DATASETS USED

A. Breast Cancer Dataset

We used breast-cancer-wiscousin-data/data.csv dataset from kaggle which had 30 features and 569 records.

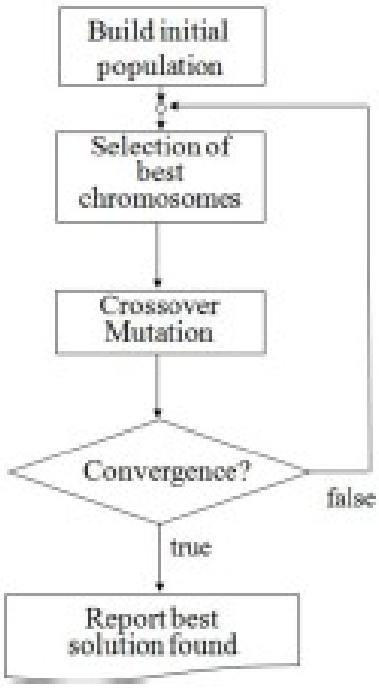
B. Parkinson's disease Dataset

We used Parkinson disease.csv dataset from kaggle which had 22 features and 195 records.

III. METHODOLOGY

The proposed methodology of this paper is as follows:

Initially we use models like LinearSVM , RadialSVM , Logistic Regression,Random Forest,Adaboost,Decision Tree, Gradient Boosting on our datasets we then compared their accuracies.



Later we applied GA on our dataset features with initial population as the total number of features which are initialized with random values.

We run a function to initialize a random population.The randomized population is now run through the fitness function ,which returns the best parents.

Selection from these best parents will occur depending on the n-parent parameter. These selected parents are then sent to cross over and mutation functions.

Crossover is created by combining genes from the two fittest parents by crossing over at random points .

Then mutation is achieved by flipping randomly selected bits for cross over child.

A new generation is created by selecting fittest parents from previous generation and applying cross over and mutation .

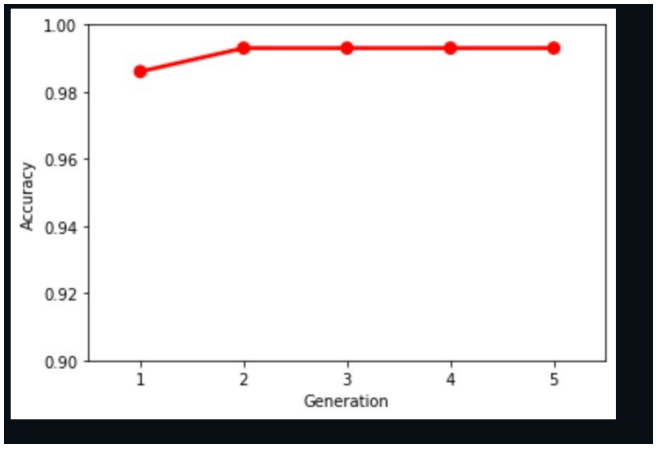
This process is repeated for n number of generations until convergence is observed in the solution.

IV. Conclusion

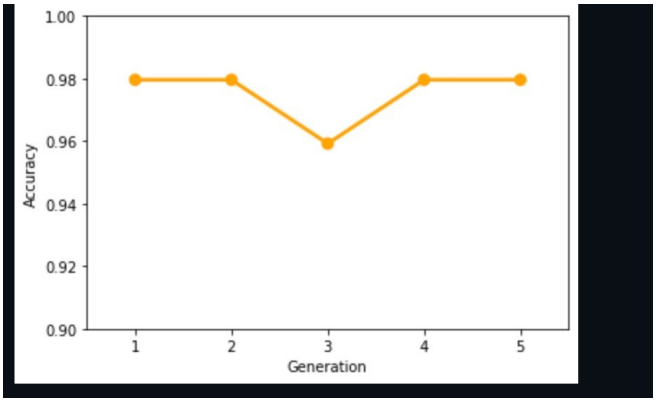
So to finally conclude the genetic algorithm provides best solution amongst set of all possible solution which tells us the maximum number of features that can be chosen for better prediction of the dependent variable without much bias.

Graphs-(Accuracy V/S Generation)

For Breast cancer dataset-



For Parkinson's disease datset-



V. References

Variable Selection in multiple linear Regression using Genetic Algorithm ,
Javier Trejos University of Costa Rica
and Maria Villalobos-Arica University of Costa Rica