Feature Selection using Genetic Algorithm: An Analysis of the Bias-Property for One-Point Crossover

Lauro C. M. de Paula
Institute of Informatics
Federal University of Goiás
Goiânia, Goiás, Brazil
laurocassio@inf.ufg.br

Anderson S. Soares
Institute of Informatics
Federal University of Goiás
Goiânia, Goiás, Brazil
anderson@inf.ufg.br

Telma Woerle de Lima
Institute of Informatics
Federal University of Goiás
Goiânia, Goiás, Brazil

Clarimar J. Coelho
School of E. S. and Comp.
Pont. Catholic Univ. of Goiás
Goiânia, Goiás, Brazil

ABSTRACT

Genetic algorithms (GAs) have been used for feature selection with binary representation. Even if binary representation has perfect probability to include or remove a feature in the search process, some works in the field of chemometrics have reported criticism about a high number of features selected by GA implementations. Thus, in this paper, we aim to propose an investigation of the number of features selected on a point of view of the bias-property using implementations from the GA-PLS toolboxes (Genetic Algorithm with Partial Least Square). The study is performed using an one-point crossover operator and a common initialization procedure used in the matlab toolboxes. Results show the existence of such a bias that influences the increase in the number of features over the generations.

Keywords

Genetic Algorithm; Feature Selection; Property Analysis.

1. INTRODUCTION

Genetic algorithms (GAs) have been widely used as a method for feature (or variable) selection. One of the applications of GAs in variable selection is the problem of multivariate calibration in the context of quantitative chemical analysis [3].

Leardi's review [2] report a considerable number of papers in which GAs have been applied for this task. In the case, for instance, a chromosome is made by a very high number of genes (as many as the variables), each of them being just 1 bit long (0 = variable absent, 1 = variable present). On the other hand, some works [1, 5] have criticized the solutions obtained through GAs where they contain a considerably high number of variables when compared to other methods.

The performance of a GA is dependent of many factors, such as, the type of crossover and mutation operator, population size, and the encoding used are just a few examples. The bias is a important property to evaluate the general trend of the encoding and operator in order to ensure the same probability of choice for all solutions [4]. Thus, in this paper we present an analysis of the bias concerned with the number of features selected, using the binary encoding and the one-point crossover operator.

2. PROPERTY ANALYSIS: BIAS

The bias of an encoding in GAs describes whether either the genotype-phenotype mapping or the genetic search operators (such as mutation and crossover) prefers a specific type of solution and leads a population towards this direction. Thus, reproduction operators without bias does not favor the representation of specific solutions. Encodings and biased search operators can be used if there is an initial knowledge that the optimal solution is similar to the set of solutions that are preferred by the bias. On the other hand, representation and non-biased reproduction operators should be used if none specific knowledge about the problem is available. In this paper, our proposal is to analyze the bias of GAs binary encoding and one-point crossover operator for the problem of feature selection towards the number of features selected. Algorithm 1 shows a pseudocode for our proposed GA implementation.

Algorithm 1 Genetic Algorithm for Bias Analysis.

```
Let n be the population size
Initialize the population of individuals using uniformly distributed random numbers
for i = 1 : MaxGenerations
    Evaluate all individuals based on the number of features selected
    Calculate the average number of features selected
    Generate n solutions using one-point crossover
    Replace parents by children
end for i
```


GECCO'16 Companion July 20-24, 2016, Denver, CO, USA

ACM ISBN 978-1-4503-4323-7/16/07.

DOI: http://dx.doi.org/10.1145/2908961.2931636

In our implementation, we do not use selective pressure neither elitism, in which at least one changeless copy of the best solution of current generation is transferred to the new population, so that the best solution can survive to successive generations. Instead, each parent participates in a single crossover and the parents are always replaced by offspring. The application of operators without any selective pressure does not modify the population statistics properties, in this case the average number of features selected. If this number increases, we can claim that the GA is biased to increase the number of features.

3. EXPERIMENTAL

We executed the experiment with two case studies: Let n be the population size, *i*) $n = 50$, $MaxGenerations = 500$; and *ii*) $n = 500$, $MaxGenerations = 5000$. For each combination of n and $MaxGenerations$, 10 trials were performed.

4. RESULTS

Figure 1 presents the histogram of the initial population.

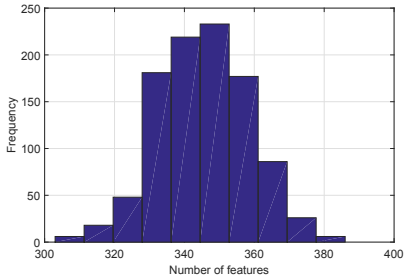

Figure 1: Histogram of the initial population.

Figure 2 presents results for the simulations using $n = 50$; and $MaxGenerations = 500$. We can observe that the average number of features (variables) increases over the generations. Then, there is a bias to increase the number of variables. Furthermore, the minimum number of variables has a tendency to increase, raising the average number of variables of the set of solutions.

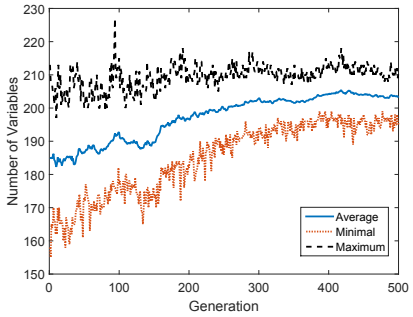

Figure 2: 500 generations to evolve 50 individuals.

In Figure 3, we show the average number of variables selected using $n = 500$; and $MaxGenerations = 5000$. It is possible to notice the same increasing of the average number of variables selected over the generations. In this case, there also exists a bias related to the increasing number of variables.

It is noteworthy that in order to investigate the presence of trend in the crossover operator considered, all simulations were performed without any selective pressure or elitism.

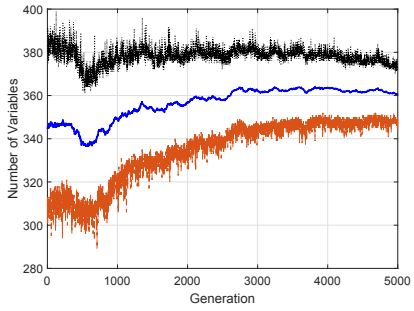

Figure 3: 5000 generations to evolve 500 individuals.

5. CONCLUSIONS

Indeed, GAs have been used as a technique to solve optimization problems such as the feature selection problem. As literature lacks analysis of GAs implementations, this paper aims to propose an analysis of GAs configuration about a bias for the number of features selected in a multivariate calibration problem.

Avoiding the use of selective pressure as well as elitism, it was able to demonstrate that, in fact, there is a bias to increase the number of variables over generations. The main objective of feature selection problem is precisely reduce the number of features. Thus, a bias to increase this number is not desirable. A full research using other operators, encodings and initialization procedure should be done in order to investigate the bias and propose solutions to deal with this problem. Most likely, the problem can be fixed using a proper initialization procedure.

Acknowledgments

Authors thank to the brazilian research agencies CAPES, CNPq and FAPEG for the financial support.

6. REFERENCES

[1] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame, and V. Visani. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, 57(2):65–73, 2001.

[2] R. Leardi. Genetic algorithms in chemometrics and chemistry: a review. *Journal of chemometrics*, 15(7):559–569, 2001.

[3] A. Niazi and R. Leardi. Genetic algorithms in chemometrics. *Journal of Chemometrics*, 26(6):345–351, 2012.

[4] F. Rothlauf. *Representations for Genetic and Evolutionary Algorithms*. Springer-Verlag, 2006.

[5] H. Xu, B. Qi, T. Sun, X. Fu, and Y. Ying. Variable selection in visible and near-infrared spectra: Application to on-line determination of sugar content in pears. *Journal of Food Engineering*, 109(1):142 – 147, 2012.