ARGEN + AREPO: Mixing the Artificial Genetic Engineering and Artificial Evolution of Populations to Improve the Search Process

Agustín León-Barranco[1]
agustinleonb@inaoep.mx

Sandra E. Barajas[1]
sandybarajas@inaoep.mx

Carlos A. Reyes[1]
kargaxxi@inaoep.mx

[1]Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE).
Departamento de Ciencias Computacionales
Luis Enrique Erro #1, Tonantzintla, Puebla, 72840, México
Telephone Number: 01(222) 266 31 00. Exts. 8302, 8308
Fax Number: 01(222) 266 31 52

ABSTRACT

In this paper we analyze the performance of several evolutionary algorithms in the feature and instance selection problem. It is also introduced the ARGEN + AREPO search algorithm which has been tested in the same problem. There is no need to adapt parameters in this genetic algorithm, except the population size. The reported preliminary results show that using this technique in a wrapper model to search data subsets, we can obtain accuracy similar to the obtained with some of the genetic algorithms models here presented, but with less data.

Categories & Subjects

I. Computing Methodologies, I.5 PATTERN RECOGNITION, I.5.2 Design Methodology, Feature evaluation and selection.

General Terms

Algorithms, Design, Experimentation, Performance.

Keywords

Hill-Climbers, genetic algorithms, data selection, classification, nearest neighbors, wrapper model, artificial genetic engineering, artificial evolution of populations, infant cry.

1. INTRODUCTION

Two of the most used search strategies in the search of feature and instance subsets are Hill-Climbers (HCs) and Genetic Algorithms (GAs). The kind of GA presented in this paper keeps the advantages of both searching techniques and, besides, it is reinforced by means of the ARGEN concept here described. The result is an *"artificial algorithm with a more natural behavior"*.

2. ARGEN + AREPO

ARGEN (*ARtificial Genetic ENgineering*) + AREPO (*ARtificial Evolution of POpulations*) is an improved version of AREPO, an adaptive evolutionary algorithm. AREPO is a population level based evolutionary algorithm that moves from Classical Genetic Algorithms towards *Artificial Evolution of Populations*. It is composed by individuals able to meet and interact. AREPO is a technique for setting the genetic parameters during a run by adapting the population size and the operator rates on the basis of the environmental constrain of a maximum population size. In addition, genetic operators are seen as alternative reproduction strategies, fighting among individuals is also applied [1].

ARGEN is the reinforced part of AREPO. GEN (*Genetic ENgineering*) biologically is a set of techniques that allow us to modify the organism's characteristics in a predetermined sense, by means of the alteration of its genetic material. ARGEN (*ARtificial GEN*) is a term that is introduced in this paper to include all kinds of alterations made to an individual (in GAs) to improve its fitness. Three levels of alteration have been distinguished: "*bit-level*", where the changes are directly made to the chromosome's bits (mutation), "*individual-level*", where the alteration is given by the way in which new information is introduced to one chromosome (crossover) and finally, the "*population-level*" which is the way the chromosomes are selected to mate and reproduce (*artificial & natural selection*). The difference with the traditional way is that this alterations poof-based let us adapt the individual to a certain environment, i.e., adapt better the individuals to a certain problem.

The general ARGEN + AREPO algorithm is the following:

```
1. Initialize population & Evaluate population
2. Alter best individual (bit-level)
3. While not termination condition
   3.1. Get two individuals (natural & artificial selection)
   3.2. If "meeting"
        3.2.1. If "reproduction"
               3.2.1.1. Crossover (individual-level) & Evaluate
        3.2.2. Else
               3.2.2.1. Fight
   3.3. Else
        3.3.1. Simple Mutation & Evaluate
4. Alter best individual (bit-level)
```

2.1 Adaptation Rules

One of the main features of ARGEN+AREPO algorithm is the adaptability of the parameters as presented in [1]. Population size is constrained by the environmental limit and its dynamics are determined by the meeting probability, reproduction and competition rules among individuals. These three adaptive rates are defined as:

$$Pm = \frac{Cp}{Mp} \qquad \Pr = 1 - Pm \qquad Pc = 1 - \Pr$$

GECCO'05, June 25-29, 2005, Washington, DC, USA.
ACM 1-59593-010-8/05/0006.

Where *Pm* is the meeting probability (the population density), *Pr* is the reproduction rate, *Pc* the competition rate, *Cp* the current population size and *M*p the maximum population size.

2.2 Meeting and Competition

When two individuals meet they can interact in two ways: by reproducing (crossover) or by fighting for natural resources (the stronger kills the weaker), otherwise the current individual can generate a new individual by a simple mutation.

2.3 Initialization and Finalization

In ARGEN + AREPO the initial size of the population is a random number, limited by the maximum size of individuals. All individuals are unique, ranked in the population by their fitness. In the initial population we alter the best chromosome by means of a hill-climber algorithm with the intention of improving its fitness. The ending of the algorithm is given in the same way. Like in Genitor, ARGEN + AREPO algorithm ends when a maximum number of iterations is reached.

2.4 Natural and Artificial Selection

Artificial selection is the modification of the hereditary constitution of the offspring by controlling the crossover between the parents. The idea presented in ARGEN + AREPO combines the *Natural and Artificial Selection* because the first is based upon the notion that all individuals in nature have a chance of mating. So, at each iteration, we pick the i^{th} individual of the population, for *i* from 1 to current population size, then, applying artificial selection, the second individual is selected by ranking.

2.5 Simple Mutation and Alterations at the Individual-Level

In ARGEN + AREPO mutation is performed according to the adaptive reproduction rate *Pr* and when it occurs, in the bit-string (representation of the chromosome), *k* bits are randomly flipped. The mutated individual does not replace the original one; simply it is ranked in the population, so the population size increases in one unit. When the population size reaches its maximum value, the offspring is ranked in the population and the least fit individual is removed. Crossover is performed according to the adaptive rate *Pr*, the alterations made between individuals is directed by the 2-point "reduced surrogate" crossover operator, and if it occurs the resulting offspring would be simply ranked in the population. In this way, population increases by two new elements.

3. PRELIMINARY RESULTS

For the data selection problem, six GAs were selected for evaluation and to compare their performance with the novel algorithm presented in this paper. The chosen algorithms are the following: SGA (*Simple Genetic Algorithm)* [4], TDA (*Traditional Genetic Algorithm)* which differs from the SGA in that one or more individuals are passed to the next generation by elitism, CHC [2], GENITOR [5], AREPO [1] and CFRSC [3]. The seven GAs were evaluated in a wrapper model where the classifier used was Distance-Weighted K-Nearest Neighbors, the results are presented in Table 1; the column "method" shows the searching algorithm used in the wrapper model. The stop criterion was to reach a maximum of 8000 evaluations of the fitness function, for all the searching techniques. In SGA and TGA a population of 50 individuals was used with a crossover rate of 0.8 and a mutation rate of 0.01. CHC uses the same number of individuals and a rate of 0.35 for the cataclysmic mutation. The population size in CF/RSC is 10, and, the individuals generated by each of them are 10. In GENITOR, AREPO and ARGEN + AREPO the population used was 1000 individuals.

3.1 Automatic Infant Cry Recognition

It was chosen an Infant Cry database to show the performance of ARGEN + AREPO. The Automatic Infant Cry Recognition Process detects pathologies of recently born babies by means of their cry, it is very similar to Automatic Speech Recognition. The process is divided in two parts, the first one corresponds to signal processing, or acoustic feature extraction, and the second part is the pattern classification. The patterns are represented by vectors of Mel Frequency Cepstral Coefficients (MFCC). The used dataset has 1376 instances of 305 features each. There are three kinds of infant cry: normal, hypo-acoustics and asphyxia. Ten experiments were performed. At each experiment, 60% of data for training and 20% for evaluation were randomly selected from the database; both directed to the wrapper model. The remainder 20% was left to evaluate the accuracy of the feature and instance subset given by the wrapper model. The "Average" and "Best" columns from "Method Accuracy" show the average accuracy of the 10 data subsets given by the wrapper model over their respective evaluation subsets, and, the best accuracy obtained from the 10 experiments, respectively. In "Real Accuracy" are shown the average and best accuracy of the 10 data subsets returned by the wrapper model, but, evaluated over their respective test subsets. Finally the "Storage" column shows the percentage of the original training set that is necessary to store to classify new instances, these results are also an average of the 10 experiments.

Table1. Results at the classification of normal and pathological baby cry, this last divided in two classes: asphyxia, and hypoacustics.

Method	**Real Acc. (%)**		**Method Acc. (%)**		**Storage**
	Average	Best	Average	Best	
SGA	98.88	99.64	99.64	100	24.68
TGA	98.84	100	100	100	12.94
CHC	99.31	100	100	100	18.43
CFRSC	97.97	98.91	99.93	100	1.92
GENITOR	99.24	100	100	100	18.07
AREPO	99.16	100	99.96	100	20.57
ARGEN + AREPO	98.19	99.64	100	100	2.94

4. CONCLUSIONS

We obtain high accuracy while the wrapper model is applied, but, this accuracy falls down when new unseen instances are being classified. In fact, CHC is the less affected because, as can be seen, its accuracy is the best of all. Nonetheless, ARGEN + AREPO obtains high accuracy needing a lower amount of training data.

5. REFERENCES

[1] Annunziato, Mauro., and Pizzuti, Stefano. *"Adaptive Parameterization of Evolutionary Algorithms Driven by Reproduction and Competition"*. ESIT 2000, Aachen, Germany.

[2] Eshelman, Larry J. *"The CHC Adaptive Search Algorithm: How to Have Safe Search When Engaging in Nontraditional Genetic Recombination"*. In G. Rawlins, editor, Foundations of Genetic Algorithms. p. 265-283. Morgan Kaufmann, 1991.

[3] Guerra, Cesar., Chen, Stephen., Whitley, Darrell., and Smith, Stephen. *"Fast and Accurate Feature Selection Using Hybrid Genetic Strategies"*. CEC99: Proceedings of the Congress on Evolutionary Computation. 1999.

[4] Kuri, Ángel., and Galaviz, José. Algoritmos Genéticos. IPN, UNAM & FCE. p. 13-30, 53-112. 2002.

[5] Whitley, Darrell. *"The GENITOR Algorithm and Selective Pressure: Why Rank-Based Allocation of Reproductive Trials is Best"*. Technical Report CS-89-105. 1989.