INTRODUCTION TO EVOLUTIONARY ALGORITHMS FOR DATA MINING

Author:

Patrick Schneider

Course:

MIRI – Algorithms for Data Mining

Date:

20.06.2017

University:

UPC FIB Barcelona, Spain

ABSTRACT

In my past research about data mining algorithms in the subject of algorithms for data mining, I read on several occasions the possibility of improving classification task by applying genetic algorithms. This woke my curiosity about the function of an evolutionary algorithm and the use cases in data mining.

The essay gives an introduction in the field of evolutionary algorithms with the explanation of the main functionalities and the application field.

TABLE OF CONTENT

1 INTRODUCTION

Metaheuristics are problem-specific methods for solving optimization problems by iterative improvement of candidate solutions in relation to a fitness measure. Evolutionary algorithms (EA) are population-based metaheuristics. In those population based metaheuristics a satisfactory solution is evaluated between several solutions and iteratively changed to find a near optimal solution. A EA is based on a population P of potential solutions e_i ∈ P for the optimization problem to be solved or the heuristic. A solution will be developed within the framework of the population based metaheuristics. Evolutionary algorithms are applied in solving combinatorial optimization problems, in machine learning, in engineering sciences, engineering and many more fields.

Biological background

Evolution algorithms are motivated by biological evolution. The driving forces of the biological evolution is variation, selection and gene drift.

A. Variation

A variety is needed for the selection processes of “organism”, which is subject to the evolution. In the biological evolution, the variety is needed for the hereditary variation in the genome (DNA). Variation is in the next generation created by mutations in the DNA and by recombination during reproduction. Both processes are undirected and have no objective. Whether the variations in the gene pool are permanently retained is determined by the selection decided.

B. Selection

Out of the population of different individuals, specific or complex characteristics gets preferred picked for the survival and partner search. This results in selective pressure for these features. A characteristic can be simply represented as a mathematical quantity. For example, we have a starting population with a weak feature size. In the next generation, the gene combination for a strong feature size will be stronger pronounced. The mean value of the strong feature will shift in the subsequent generations to a larger value.

C. **Genetic Drift**

Genetic Drift denotes random influences in the evolution, which lets a population "drift" in a random direction. This development has nothing to do with selection or fitness. The influence of the drift is particularly evident for very small population sizes and are very important for biological evolution.

2 METHODOLOGY

Basically, an evolutionary Algorithm follows the steps:

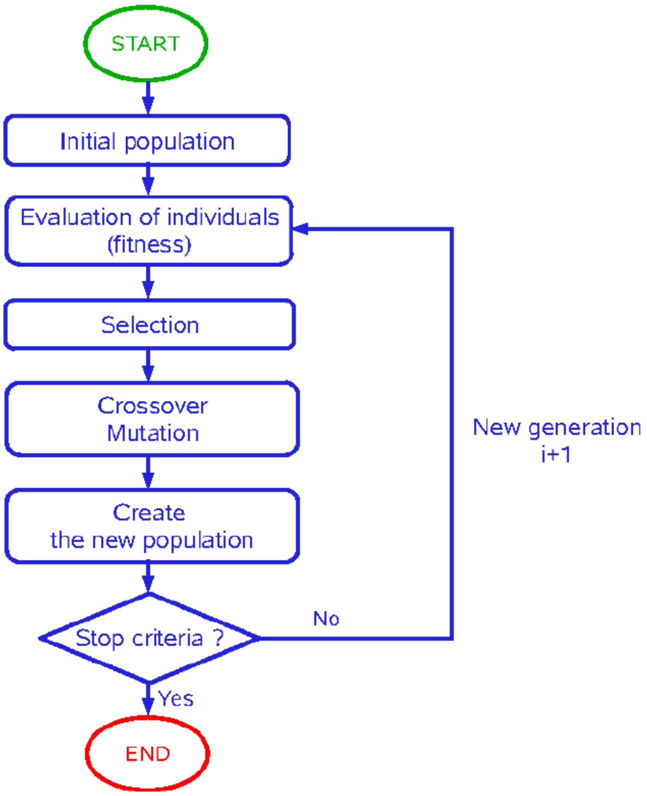


Figure 1: Process of evolutional algorithm source: http://www.cenaero.be/Page.asp?docid=27092&langue=EN

Initialization: a population consisting of random individuals gets initialized.

Evaluation: The fitness f (e_i) is used for all e_i ∈ P is evaluated. The fitness function indicates the value to be optimized of the optimization problem (optimization function).

Selection: Based on the fitness value F (e_i) of the individuals, random individuals are selected who are considered parents for the next generation. Individuals with a high fitness F (e_i) will be preferred for the next generation.

Reproduction: The offspring are generated from the selected parents. This can be obtained by simple copying or also by recombination of 2 or more parents.

Recombination and Mutation: The offspring are recombined and individuals are mutated. This step creates evolutionary variation.

Replacement: The offspring replace parts of the current population or the entire population

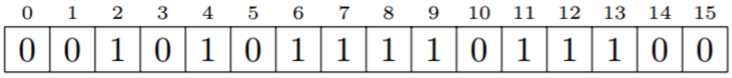
Except the initialization, these steps are repeated until a termination criterion is reached. Such a termination criterion could be, for example, a maximum number of iterations or a maximum fitness achieved.

3 METHODS

3.1 Representation of data

DNA is the molecule in all living cells that carries the hereditary information. The DNA has the form of two strands. The DNA stores the information redundantly. During one of the two Strings are encoded directly, the other is an inverse, complementary copy of the other

The DNA can be modeled in a binary array or tree representation.



0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	1	0	1	0	1	1	1	1	0	1	1	1	0	0

Figure 2: Bit representation

3.2 Selection

In the selection, individuals from the current Population are selected to be used as a parent for the next generation. Stronger individuals become favored. In evolutionary algorithms, this means that individuals with a greater fitness value f (e_i) have a greater chance of being selected.

The following are the most famous selection methods:

1. **Roulette-Selection:** Most used one. Individuals with greater fitness have a better chance of being selected. To illustrate it: You have a roulette and one pin. The individuals with high fitness have a bigger field on the roulette.
2. **Stochastic Universal Sampling:** All individuals will be for the next generation at the same time selected. To illustrate it: Should be k individuals be selected, the idea is that k pins with the same distance to each other determine the k individuals for the next generation.
3. **Tournament selection:** In the tournament selection, t individuals are randomly selected from the Population. The fitness of t the individuals are then compared with each other and the individual with the best fitness selected. If k individuals will be selected, k tournaments are executed.
4. **Rank-based selection:** Instead of choosing the selection probability proportionally to the fitness value f (e_i), will be in the rank-based selection the individuals picked based on their fitness. This is an advantage when in the population few individuals with a very large fitness level are present. In the roulette selection and the stochastic universal sampling are these individuals highly preferred and the population quickly loses its diversity.

3.3 Reproduction

Reproduction is a binary operation that combines two individuals together so that new individuals are generated. Each having characteristics of both parent individuals. [Due to missing time, I will only focus on the linear representation.]

Reproduction in linear representation

A. N-point crossover: Two individual's representations are recombined with each other and thus two new individuals are generated. In the n-point crossover, the linear vectors have n cuts that result in n + 1 pieces and then by changing order reconnected.

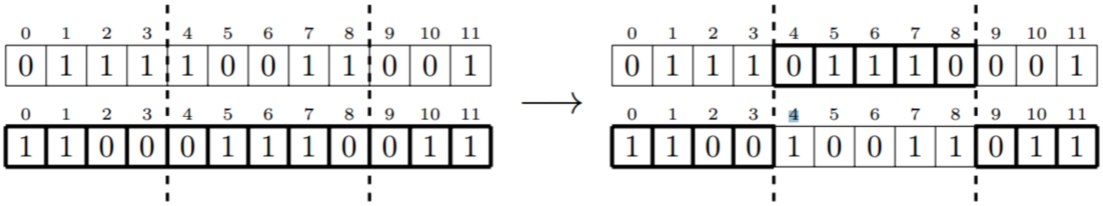


Figure 3: Example of a 2-point crossover of a binary representation

B. Uniform crossover: Each individual symbol is randomly distributed uniformly of one of the two parents with the results that both parents have a more or less even share in each child.

3.4 replacement strategy

Replacement strategies for evolutionary algorithms are used as the current population is replaced by the offspring. There are two main strategies:

A. Generational replacement: In generational replacement, the offspring generation replaces the entire population. In evolutionary algorithms, the population size is generally fixed and a fixed number of offspring must be generated here.

A. Steady-state replacement: With steady-state replacement, only one individual is replaced in every generation. For example, the weakest individual is exchanged.

B. Other strategies: The generational replacement and the steady-state replacement are the case "extreme" approaches. Other approaches follow a middle way. In the case of elitism, the λ weakest individuals are always exchanged. This can result in too fast convergence. Around the search room, a stochastic approach for replacement can be used here.

4 USE CASES IN DATA MINING

Not only in combinatorial problem cases evolutionary algorithms find their use case. Other major fields are in machine and robot learning, where evolutional algorithms are used for classification and prediction, for example, protein structure prediction. In the finance sector, evolutional algorithms are used to develop bidding strategies in emerging markets.

Main application function: Widely applicable for classification for inductive learning, where they provide a practical method for optimization of data preparation and data transformation steps. Here they are an effective tool to use in data mining and pattern recognition.

5 CONCLUSION

Evolutionary algorithms find application in solving many optimization problems, especially those in which no (good) methods are available. The data preparation and data transformation steps of data mining can be improved with this type of algorithm.

Personal conclusion:

I decided to pick this topic because I was reading in the context of sentiment analysis the use of genetic algorithms. To this stage I was not aware about the process of evolutional algorithms. My research began by understanding the basic functions and the idea behind. In the end, I had the feeling that there are not that many use cases in data mining like I had initially expected. In the next weeks, I want to figure out why the scientific papers about specific evolutionary algorithms type are not covered that broad like I expected.

Questions that came up are:

- What are the benefits of using a EV algorithm compared to other algorithms in the field of data mining/classification?
- What are the specific use cases and how do they perform to “state of the art” algorithms in, for example, classification tasks?

To sum it up, it was a great learning experience participating in the subject ADM and the creation of the individual essays and the offered support.

6 REFERENCES

[1] Candida Ferreira, *Gene Expression Programming: A New Adaptive Algorithm for Solving Problems. Complex Systems*, 2001

[2] John Koza und Riccardo Poli, *Genetic Programming*, springer, 2005

[3] Basheer M. Al-Maqaleh, Hamid Shahbazkia, A Genetic Algorithm for Discovering Classification Rules in Data Mining, International Journal of Computer Applications V41, 2012

[4] K.C. Tan, E.J. Teoh, Q. Yu a, K.C. Goh, A hybrid evolutionary algorithm for attribute selection in data mining, Department of Electrical and Computer Engineering, National University of Singapore, 2009