A comparison of bio-inspired metaheuristic approaches in classification tasks

R. L. Oliveira, B. S. L. P. de Lima & N. F. F. Ebecken
Civil Engineering Department,
COPPE/UFRJ - Post-Graduate Institute of the Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Abstract

This paper presents a comparative analysis of three computational tools based in metaheuristics inspired by nature to perform an important data mining task. These tools are employed to generate classification rules from databases. The first one uses the Ant Colony metaphor that is one of the most recent nature-inspired metaheuristics. The second one employs the Artificial Immune System paradigm that is also a relatively new biologically-inspired paradigm. The third one employs a fuzzy genetic approach.

The main motivation for applying those heuristics to data mining is that bio-inspired algorithms have shown to be robust search methods.

In this work, basic concepts of the employed strategies are presented and significant aspects related to each approach are discussed.

Some data sets from the UCI repository were employed to evaluate the performance of the tools. The comparative survey of the classification tasks is performed emphasizing the importance of discovering comprehensible and accurate knowledge.
Keywords: data mining, bio-inspired metaheuristics, Genetic Algorithms, Ant colony optimization and Artificial Immune Systems.

1 Introduction

Researchers of several areas have observed that various principles and theories about nature and the subsequent development of models, based in these systems, have been implemented using computers systems with great potential to solve complex problems. New strategies have been developed, inspired in biological or

WIT Transactions on Information and Communication Technologies, Vol 38, © 2007 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA070031

natural mechanisms, such as Evolutionary Algorithms; Immune Artificial Systems; Artificial Neural Networks, Ant Colony Optimization, etc.

One of the most promising areas has been the development of data mining techniques in large databases, supported by such heuristics that compose computational intelligence. In this work three bio-inspired metaheuristics are employed in the generation of Rule Bases for classification tasks: a fuzzy genetic system, a fuzzy artificial immune system and an ant colony optimizer.

2 Data mining

One of the main purposes of data mining is the extraction of knowledge from available data. It is important to highlight that this activity must evaluate not only the accuracy of the results, but also their comprehensibility [1].

This work presents a comparative analysis between three data mining algorithms using different metaheuristics, for data classification.
In general, classification tasks express the acknowledgement in the form of IF-THEN rules, as:

IF <condition> THEN <class>

The rule antecedent contains a set of conditions, generally interconnected by logic connective. In this work we will refer to each rule condition as a term, so that the antecedent of each rule will consist of a conjunction of logical terms as:

IF $term_1$ AND $term_2$ AND $term_3$ AND....

where each term has a triple parameters < attribute, operator, value >.

The rule consequent refers to the class predicted, which the predictive attributes satisfy all terms specified by the rule antecedent. This knowledge representation presents an advantage of being intuitively understandable if the number of rules, as well the number of terms in each rule are not very large. Therefore in this work, the evaluation parameters will be, the level of accuracy of the data mining algorithm, and the size of the rule base generated (small number of rules and rules with small number of terms).

3 Ant Colony Optimization

In a real ant colony, each ant performs its task in an individual way and independent from the rest of the colony, however those tasks are related with each other, so that the colony as a whole would be capable to solve complex problems through the cooperation between the elements. This interesting ability of ants' society has been studied by scientists. Experiments showed that ants use a chemical substance, named pheromone for exchange information. When an ant moves, it leaves a pheromone trail in its path, signaling the path with this substance. If more ants follow the same path more amount of pheromone will be deposited in it, and it turns to be most attractive for others ants. At the end all ants of the colony follows the same path. This process can be described as a "loop with positive feedback", where the probability of an ant to choose certain path is directly proportional to the number of ants that already had passed there. This fact shows that ants are capable of selecting the shortest path from the nest to the food source, inspiring scientists to develop optimization algorithms.

3.1 Optimization Algorithms using Ant Colony Heuristic

An optimization algorithm inspired in the ant colony behavior includes the mechanisms of cooperation and adaptation. The Ant Colony Optimization (ACO) algorithm [2] is an interesting new approach to optimization that has been applied in several areas.

The first bio-inspired metaheuristics employed in this paper to the generation of rule bases for classification tasks is the Ant-Miner Optimization Algorithm [3]. This algorithm is an Ant Colony-based Data Miner that is able to extract knowledge from data in the form of classification rules.

3.1.1 General description of the algorithm

In an Ant based algorithm each ant can create/modify a solution for a specific problem, incrementally. In classification tasks, the problem consists of discovering classification rules in the form:

IF < $term_1$ AND $term_2$ AND…..> THEN <class>.

where each term has a triple parameter <attribute, operator, value>.

The Ant-Miner algorithm adopts a strategy to discover sequentially a list of classification rules that cover all, or almost all the training cases.

Initially, the list of discovered rules is empty and the training set involves all training cases. At each iteration level, one classification rule is discovered. This rule is added to the list of discovered rules, and the cases correctly covered by the discovered rule are removed from the training set. After that, another iteration is started. This process is repeated to find rules covering most of the cases in the training set.

Each iteration consists of three steps, rule construction; rule pruning and updating the amount of pheromone. In the first step, an ant initiates with an empty rule, incrementally it is added a term to the current partial rule, that corresponds to the partial path followed by this ant. The term that will be added to the current partial rule corresponds to the choice of the direction that depends on both a problem-dependent heuristic function and on the amount of pheromone associated with each term. The ant keeps adding one term at a time to its current partial rule until a stopping criteria is met.

When an iteration is concluded, the best rule produced is selected and added to the set of discovered rules, and a new iteration is processed, restarting all the trails with the same amount of pheromone.

After the rule antecedent is fulfilled, the algorithm selects the best rule consequence that maximizes the quality of the rule, by allocating the majority class to the rule consequent. This procedure is described in details in [3].

4 Artificial Immune Systems

The Artificial Immune Systems (AIS) consist of methods based on the natural immune systems and that are designed to solve problems in the real world [5]. AIS emerged in the 90's as a new computational research area and link several emerging computational fields inspired by nature.

The second bio-inspired approach is the algorithm IFRAIS (Induction of Fuzzy Rules with Artificial Immune Systems) presented in [4]. It proposes a methodology that integrates data mining tasks with AIS and fuzzy logic.

This algorithm is based on the clonal selection principle of biological immune systems. In essence, when an immunological detector system (lymphocytes) recognizes a particular antigen (invading microorganism), it stimulates the proliferation and differentiation of plasmatic cells that produce antibodies.

This process, called clonal expansion produces a great population of antibodies that present high affinity with specific antigens. This clonal expansion, in general, results in the neutralization or destruction of the antigen and, also in the maintenance of some cells in the "immune memory", consequently the immune system can react more quickly when a similar antigen attack occurs. This process is a natural type of selection. The closer is a cell to an antigen, the higher the proliferation rate will be. This is an adaptive process, as the clones suffer an accelerated mutation. This mechanism is called somatic mutation or hyper mutation. In conjunction with the selection process, the somatic mutation improves the ability of clones in to recognize the antigens, producing clones with higher affinity for a particular antigen.

Fuzzy systems use linguistic terms that are represented by membership functions of fuzzy sets. This treatment allows the numerical processing of concepts and is able to incorporate natural vagueness and subjectivity of human reasoning [6,7].

A fuzzy classification rule is a fuzzy if–then rule whose consequent part is a class label. Since the comprehensibility of fuzzy rules by human users is a criterion in designing a fuzzy rule-based system, fuzzy classification rules with linguistic interpretations must be taken into account.

In accordance with [8], in a classification problem with *n* attributes, the fuzzy rules can be written in the form:

$$\text{IF } x_1 \text{ is } A_1^j \text{ AND } \ldots \text{ AND } x_n \text{ is } A_n^j \text{ THAN class } C_j,\ j = 1,\ldots, N,$$

where:

- $x = (x_1, \ldots, x_n)$ is a n-dimensional vector of attributes.
- A_n^j $(i = 1\ldots, n)$ represents a linguistic value of *i*-th, as *small* or *large*
- C is a consequent class.
- N is the number of fuzzy rules.

The antecedent of each fuzzy rule is a combination of linguistic values.

4.1 General description of IFRAIS

The IFRAIS Algorithm [4] discovers fuzzy rules for classification. Basically, it evolves a population of antibodies, where each antibody represents the antecedent part of a fuzzy classification rule. Each antigen represents an example (data instances) of the training database. The rule antecedent is formed by a conjunction of conditions. Each attribute can be continuous or categorical. The categorical attributes are inherently crisp, but the continuous attributes are fuzzyfied by using a set of linguistic terms. In this work the terms are represented by triangular membership function.

Each antibody is codified as a string with n genes, where n represents the number of attributes. Each gene consists of two elements: a value V_{ij}, specifying the value of i-th attribute in j-th condition of the rule and, a boolean value B_i, pointing if the i-th condition occurs or not in the classification rule. Although all antibodies present the same genotype length, different antibodies represent rules with different numbers of condition in their antecedents. The rule antecedent must have at least one condition. The optimum number of conditions in each rule is unknown *a priori*. The rule consequents are not evolved by the system. All antibodies of determined execution are associated with the same class, so the algorithm is executed multiple times to discover rules predicting different classes.

5 Fuzzy genetic approach

Genetic Algorithms (GA) [10] are one of the most popular bio-inspired metaheuristics and have attracted much interest in several research areas. They have shown to be effective in exploring large and complex spaces in an adaptive way, which reflects some of the main aspects of the natural evolution mechanisms such as reproduction, crossover, and mutation. GAs have been widely used as an evolutionary way to discover classification rules from databases [1].

The third bio-inspired algorithm employed in this paper to the generation of rule bases for classification tasks is a Fuzzy Genetic System (FGS) proposed by [11].

Feature selection and classification are performed by a fuzzy genetic system, in which the Takagi-Sugeno-Kang (TSK) fuzzy rules are generated automatically from the datasets and GA is applied to find the shortest and most accurate subset of rules.

6 Comparison of the results

The algorithms studied in this work had been evaluated using six bases of public domain: Wisconsin breast cancer, Credit card approval, Car evaluation, Iris plant, Glass identification and Wine recognition.

These sets of databases are available in the repository of the University of California at Irvine (UCI) in the site http://www.ics.uci.edu/~mlearn/ MLRepository.html. Table 1 describes the characteristics of the used databases

In order to have a better view of the comparison among the bio-inspired metaheuristics, it was also employed two well-known algorithms for discovering classification rules, C4.5 [9] and CN2 [12].

The comparison was carried out using two performance criteria, the classification accuracy and the simplicity of the discovered rule sets.

The results comparing the algorithms precision are presented in table 2. FGS obtained the higher accuracy in five of the six analyzed data sets.

Table 1: Summary of the databases' characteristic.

Database	# examples	# atrib.		classes
		continuous.	simbolic.	
Wisconsin breast cancer	683	9	0	2
Credit card approval	1044	6	9	2
Car evaluation	1728	0	6	2
Íris plant	150	4	0	3
Glass identification	214	9	0	7
Wine recognition	178	13	0	3

Table 2: Accuracy rate on the testing set.

Database	Accuracy (%)				
	Ant-miner	IFRAIS	FGS	CN2	C4.5
Wisconsin breast cancer	95.5	95.3	**99.4**	94.9	93.0
Credit card approval	82,3	87,6	**90.4**	89.9	86.4
Car evaluation	75.1	83.9	82.9	83.4	**90.1**
Iris plant	94.4	92.2	**97.4**	92.6	94.0
Glass identification	71.7	68.3	**76.2**	70.4	71.6
Wine recognition	91.3	89.8	**98.8**	84.6	90.1

The average number of discovered rules and the average number of terms (conditions) in the rules are used to verify the simplicity of the discovered rule set. Those averages were computed over the 10 fold cross-validation. The values summarizing the simplicity of the discovered rules are presented in table 3. It can be observed that IFRAIS presented the smaller number of rules in three of all analyzed data sets, although those results change regarding the total number of conditions. In the six datasets analyzed, the best results concerning the number of conditions are distributed among three algorithms Ant-Miner, IFRAIS and CN2.

7 Conclusions

A comparison among three bio-inspired metaheuristics to perform data classification were done employing six databases of public domain. The results showed that, regarding the classification accuracy, the algorithms IFRAIS and Ant-miner presented good performances with comparison to CN2 and C4.5, although they couldn't overcome the algorithm FGS. Using fuzzy logic to deal with linguistic terms and for discovering fuzzy prediction rules certainly

influences the performance IFRAIS and FGS. On the other hand, Ant-Miner, presented much smaller execution time than the IFRAIS and FGS that employ fuzzy logic.

Regarding the simplicity of discovered rules, it was observed that with most of the used databases the algorithms Ant-Miner and IFRAIS had outperform algorithm C4.5. Comparing only algorithms Ant-Miner and IFRAIS, the amount of rules generated by the second is always smaller than the number of rules generated by the first one, while Ant-Miner shows to be very competitive when concerning the number of terms. This behavior certainly is related to the employed rule pruning technique in Ant-miner that improved the simplicity of the rules.

The presented discussions on the relations involving the computation time and the simplicity of rules should be extended. This study can further provide useful suggestions to further improve this most recent nature-inspired metaheuristics

Table 3: Simplicity of the discovered rule set.

Database	# rules					Total conditions		
	Ant-	IFRAIS	FGS	CN2	C4.5	Ant	IFRAIS	CN2
Wisconsin breast cancer	5.6	**4.0**	9.5	18.6	11.1	12.5	**9.4**	44.5
Credit card	6.1	**5.0**	7.3	15.6	15.6	51.9	75.0	**41.4**
Car evaluation	9.2	7.1	**6.9**	21.2	51.3	**56.1**	71.1	64.3
Iris plant	5.1	**3.1**	4.5	4.7	4.5	**12.2**	17.4	18.4
Glass identific.	14.1	12.0	10.8	8.3	**8.2**	38.3	43.8	**28.3**
Wine recogn.	14.2	12.8	**8.0**	10.4	8.9	31.8	**29.8**	36.1

References

[1] A.A. Freitas. Data Mining and Knowledge Discovery with Evolutionary Algorithms. Springer-Verlag, 2002.

[2] Marco Dorigo and Thomas Stützle, Ant Colony Optimization, MIT Press, Cambridge, 2004.

[3] Parpinelli, R.S., Lopes, H.R. & Freitas, A.A. Data mining with an ant colony optimizaton algorithms, v. 6, n. 4, pp. 321-332, 2002.

[4] Alves, R.T., Delgado, M.R., Lopes, H.S. & Freitas, A.A., An Artificial Immune System for Fuzzy-Rule Induction in Data Mining, Lecture Notes in Computer Science, v. 3242, pp. 1011-1020, 2004.

[5] Castro, L. N. & Timmis, J., Artificial Immune Systems: A New Computation Intelligence Approach, Springer-Verlag, Berlin, 2002.

[6] Zadeh, L.A., Fuzzy Sets Inform. Control 9, pp. 338-352, 1965.

[7] Pedrycz, W. & Gomide, F., An Introduction to Fuzzy Sets. Analysis and Design, MIT Press, Cambridge, 1998.
[8] Ishibuchi, H. & Nakashima, T., Effect of Rule Weights in Fuzzy Rule-based Classification, Proc. Congress on Evolutionary Computation, pp. 506-515, 2001.
[9] Quinlan, J.R., C4.5: Programs for Machine Learning, Morgan Kaufmann, San Francisco, CA, 1993.
[10] Goldberg, D. E., Genetic Algorithms in Search Optimization and Machine Learning, Addison Wesley – 1989.
[11] Espindola, R.P. & Ebecken, N.F.F., Data classification by a fuzzy genetic system approach, Data Mining IV, MIT Press, Cambridge, 2003.
[12] Clark, P., Niblett, T., The CN2 induction algorithm, Machine Learning, vol. 3, pp. 261-283, 1989.