Achieving an appropriate balance between precision, support, and comprehensibility in the evolution of classification rules

Emiliano Carreño†, Guillermo Leguizamón†

†Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC)
Departamento de Informática
Universidad Nacional de San Luis
Ejército de Los Andes 950 - Local 106
(D5700HHW) - San Luis - Argentina
Tel: (02652) 420823 / Fax: (02652) 430224
e-mail: {ecarreno, legui}@unsl.edu.ar

Abstract

This article proposes a method for achieving an appropriate balance between the parameters of support, precision, and simplicity during the evolution of classification rules by means of genetic programming. The method includes an adaptive procedure in order to achieve such balance. This work lies within the data mining context, more precisely, it focuses on the extraction of comprehensible knowledge where the approach introduced plays a predominant role. Experimental results demonstrate the advantages of using the proposed method.

Key words: Data mining; classification rules; genetic programming; ranking; comprehensible knowledge.

1 INTRODUCTION

The application of genetic programming (GP) to the discovery of classification rules from a data set is not suitable when the size of trees (S-expressions) significantly increases. In such cases, the complexity of the model obtained makes it almost impossible to understand the underlying data generator process. Thus, if a model composed of many high complexity rules is obtained, it could be as hard to understand as a complex neural network. On the other hand, the measures of support and precision determine the predictive quality of a given hypothesis. Nevertheless, an appropriate model should provide an adequate balance between both parameters. For example, a rule with a 0.5 precision does not provide any information on whether an instance belongs or not to a given class; however, a rule with high precision and low support is not very useful either.

The approach proposed in this article aims to establish an appropriate balance between a rule's precision, support and complexity (directly related to comprehensibility) by incorporating an adaptive procedure which ranks individuals based on probabilities and considering their support, precision, and comprehensibility values. This procedure allows biasing the search towards hypothesis regions with high comprehensibility and an appropriate balance between support and precision.

The classification problem approached in this article consists in predicting the housing price (numerical value binned in three intervals) from information about its zone location (crime rate, etc.). The data set used, called *Boston Housing*, comes from the repository of the University of California at Irvine (UCI) [3]. It has 13 continuous attributes (including the "class" attribute "MEDV"), 1 binary-valued attribute and 506 instances. Instances reflect housing conditions in the suburbs of Boston.

The rest of this article is organized as follows. Section 2 describes the use of genetic programming for the discovery of classification rules (the basic GP system for the evolution of rules). In section 3 the approach proposed in this paper is presented and analyzed. In the same section a basic GP system, tailored with the proposed approach in order to discover comprehensible knowledge (rules), is described. Section 4 shows experimental results obtained for the Boston Housing data set. Finally, conclusions are given in Section 5.

2 RULE DISCOVERY USING GP

The main idea of GP is to evolve computer programs (S-expressions) which produce a solution for a particular problem where the candidate solutions are hierarchically structured computer programs represented as trees. Once a function and a terminal set are provided, the solution (model) is obtained by means of an evolutionary process. The function set (F) may contain arithmetic and logical operators, among other elements. The terminal set (T) contains the program's variables, and the random ephemeral constant $\Re$ which represents random numbers within some range and decimal precision. It is required that $F \cup T$ be sufficient to express a program that can solve the problem under consideration. The fitness function measures the capability of the individuals for solving the problem at hand. Several fitness measures may be adopted, some of which are: raw fitness, standardized fitness, normalized fitness, among others. These measures are explained in detail in [5].

After the initial population has been created, the algorithm is executed generation after generation until a certain termination criterion has been met. Then, the best solution found is selected. For example, a termination criterion may state that a run must terminate when a pre-specified maximum number G of generations have been run whereas a result designation criterion may be to chose the best individual in the population of the generation at termination time as the result of the whole run.

In each generation, each individual's fitness is evaluated, selecting probabilistically the best ones in the population, based on some selection method (proportionate, tournament, rank-based selection, etc), in order to apply reproduction, crossover, and mutation. Each operator is applied based on a certain probability. Reproduction is achieved by simply copying an individual from the current population into the next generation. In the crossover operation a crossover point is randomly chosen for each genetic tree. Then, both trees are split at these points creating four sub-trees that are combined to create the new individuals. When mutation operator takes place, a random point (node) is selected in a tree. The tree having this node as its root is substituted by a sub-tree generated randomly at that point. For a more detailed description of the genetic programming paradigm refer to [5].

2.1 Evolution of rules

Rules considered in this work are of the type *IF* $\langle antecedent \rangle$ *THEN* $\langle consequent \rangle$. The antecedent part of a rule is formed by logical combinations of conditions on the values of predictive attributes using the logical connectors *AND*, *OR*, and *NOT* whereas the consequent part indicates to which class a determined instance is assigned. However, each individual, represented as a tree, codes only the antecedent part of the rule. It is not necessary to code the consequent part since the genetic program is executed as many times as there are different classes. In each run a two class classification problem is solved and all rules evolved predict the same class.

The function set includes the logical operators *AND*, *OR*, and *NOT*, together with the equality operator which relates each attribute to some class. The equality operator is applied over binned attributes during the evolution of rules. The terminal set is conformed by predictive attributes and the ephemeral constant $\Re$. Figure 1 shows an example of the codification of the antecedent of a

rule. The equality operator takes an attribute as its first argument and an ordered nominal value as its second argument (representing an interval or bin). A logical operator may take as any of its arguments another logical operator or the equality operator. This structure is preserved by crossover and mutation operators.

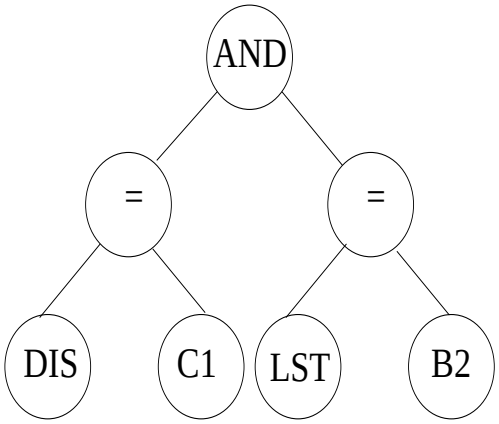

Figure 1: Representation of an individual for rule discovery using GP corresponding to the rule ***IF*** $(DIS = C1)$ ***AND*** $(LST = B2)$ ***THEN*** MEDV = High, where *DIS* and *LST* represent the intervals $(2, 5]$ and $(6.5, 9]$, respectively.

We evaluate the quality of a rule by using an estimation value for precision and support. The precision value is calculated as the ratio between the number of instances to which the rule is applied and predicts correctly over the number of instances to which the rule is applied, i.e., precision is the probability that a rule classifies correctly that instance to which it is applied. Accordingly, the support value is the ratio between the number of instances to which the rule is applied and predicts correctly over the total number of instances in that class. The fitness function can be established as an arithmetic equation including precision and support values. For example, we can use the measure F_β defined by equation 1, where β is a parameter that controls the relative importance between both values, precision and support.

$$F_\beta = \frac{(1+\beta^2)\ \textit{support}\ \cdot \textit{precision}}{\beta^2\ \cdot \textit{precision} + \text{support}} \tag{1}$$

So far, the basic GP system for rule discovery has been described. However, it must be taken into account that tree size could increase significantly. If we intend to obtain a set of comprehensible rules as a result, some kind of mechanism to control the size of solutions is required. This can be done by incorporating a procedure to favour comprehensible rule discovery, as it is presented in the next section.

3 THE PROPOSED APPROACH

In the literature we can find plenty of works where the process of knowledge discovery is aimed at obtaining comprehensible and interesting rules with a high predictive capacity. In [6], an approach is presented to discover interesting prediction rules by applying a genetic algorithm in which the adaptive function (fitness function) is divided into two parts. One part measures the degree of interest of rules, while the other measures their predictive capacity. In [1], GP is proposed for the discovery of

comprehensible rules, where a penalty for complexity is added in the adaptive function. This can also be achieved by means of a confusion matrix, or by applying a genetic algorithm with a multi-objective approach [2].

The proposal of this work includes the application of a stochastic and adaptive component which ranks solutions probabilistically considering the support, precision, and comprehensibility of individuals in the population in order to evaluate them (see figure 2).

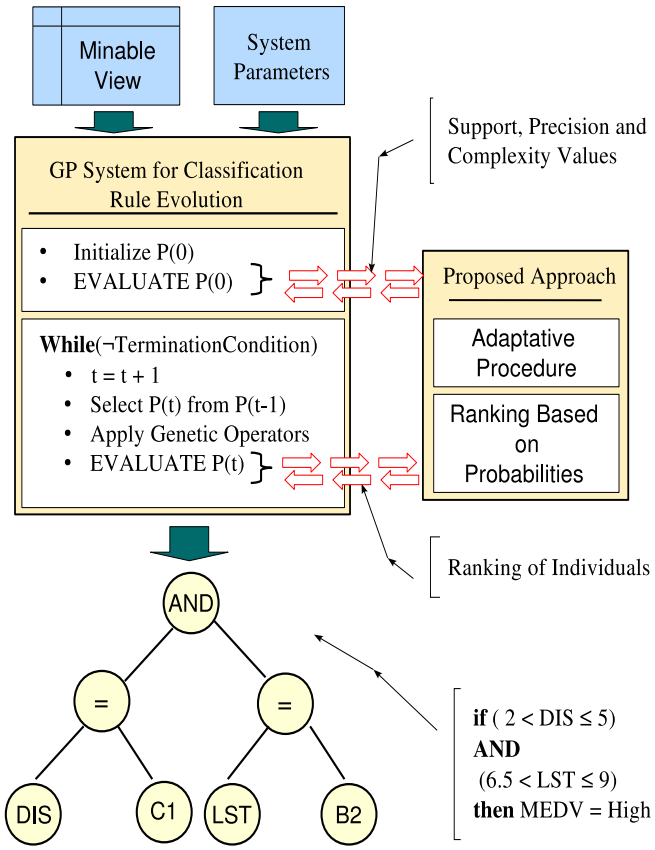

Figure 2: A GP system including the proposed approach for rule evolution

Thus, the aim is to bias the search towards hypothesis regions with the following features:

- High comprehensibility.
- An appropriate balance between support and precision measures.

Solutions of the population can be ranked by using a sorting algorithm (e.g., Hoare's quicksort) applying certain comparison criteria based on three probability values (see below) which are adaptively adjusted according to a function that gets feedback from the search process:

- *P_Sop*: is the probability of using the support factor to compare two solutions.
- *P_Conf*: is the probability of applying the precision factor to perform the comparison.
- *P_Long*: is the probability of applying the comprehensibility factor when comparing two individuals.

Where *P_Sop* + *P_Conf* + *P_Long* = 1. After adjusting these probabilities by applying the adaptive procedure, solutions are ranked. The comparison between two solutions is carried out in an excluding way, according to support, precision, and comprehensibility measures, based on such probabilities in the following way (***rnd*** is a random number in the interval [0, 1]):

1. If ***rnd*** $<$ *P_Sop* then the comparison is based on the support measure.
2. If *P_Sop* $\leq$ ***rnd*** $<$ *P_Sop* + *P_Conf* the comparison is based on the precision measure.
3. Otherwise, the comparison is carried out according to the complexity measure.

According to the selection criterion, the probability of a certain hypothesis (*X*) winning a comparison (against other hypothesis *Y*) in the sorting procedure is given by equation 2.

$$\begin{aligned} P(X) \;=\; P(\; X.Sop > Y.Sop \;) \cdot \; P\_Sop + \; P(\; X.Conf > Y.Conf \;) \cdot \; P\_Conf \\ + P(\; X.Long > Y.Long \;) \cdot \; P\_Long \end{aligned} \tag{2}$$

In the adaptive procedure, the values of *P_Sop*, *P_Conf* and *P_Long* are modified according to:

- **MaxLong**: a parameter defining the threshold from which the complexity of a solution starts influencing negatively the respective fitness value.
- **Population statistics**: the mean values of support (*Sop_Prom*), precision (*Conf_Prom*), and complexity (*Long_Prom*) of the solutions in the current population or in a subset of it.

Such procedure is illustrated in figure 3. In lines 1 to 6, if the reference parameter *Long_Prom* exceeds the **MaxLong** threshold, the value of *P_long* is established by using a quadratic function. Otherwise, the value of *P_Long* is set to 0, stating that this probability will have no influence on comparing two solutions when performing the sorting process. In line 5, an upper bound is set to the growth of *P_Long* to avoid making all comparisons based on comprehensibility, hence allowing those solutions with high predictive accuracy and moderate complexity to obtain an adequate position in the ranking. Preliminary studies show that better results are obtained by bounding this parameter. In this manner, the search is biased towards regions with the proper complexity (comprehensibility).

Next, the values of *P_Sop* and *P_Conf* are calculated by comparing the reference parameters of the population, *Sop_Prom* and *Conf_Prom*, in a way such that the probability with least value of reference parameter obtains an increase proportional to the absolute value of their difference ($0 \leq Sop\_Prom \leq 1 \quad and \quad 0 \leq Sop\_Conf \leq 1$). This step insures that an appropriate balance between support and precision measures is achieved (lines 8 to 19). To summarize, first the value of *P_Long* must be set, so as to then distribute the remaining probabilities between *P_Sop* and *P_Conf*, as reported previously.

4 EXPERIMENTAL RESULTS

This section presents preliminary results obtained with the proposed approach aiming at:

1. making a comparative analysis against the basic GP system for the evolution of classification rules explained in section 2. As stated in section 3, the approach proposed in this article incorporates to a basic GP system a procedure which ranks the population individuals for assesing the respective quality.

```
1   if ( Long_prom ≤ MaxLong )   P_Long = 0
2   else
3      {
4         P_Long = 1 − (MaxLong²/Long_prom²)
5        if(P_Long > 0.95)   P_Long = 0.95
6      }
7
8   if ( Sop_prom > Conf_prom )
9      {
10        P_Conf = P_Conf + ( Sop_prom − Conf_prom )
11        if( P_Conf > (1 − P_Long ))   P_Conf = 1 − P_Long
12        P_Sop = 1 − ( P_Conf + P_Long )
13     }
14  else
15     {
16        P_Sop = P_Sop + ( Conf_prom − Sop_prom )
17        if ( P_Sop > (1 − P_Long ))   P_Sop = 1 − P_Long
18        P_Conf = 1 − ( P_Sop + P_Long )
19     }
```

Figure 3: Adaptive procedure adjusting *P_Sop*, *P_Conf* and *P_Long* probabilities

2. studying the influence of the **MaxLong** parameter regarding predictive quality and comprehensibility of the models obtained. As mentioned in section 3, this parameter is the threshold from which the comprehensibility of solutions starts influencing their fitness.

In all experiments, the selection method based on linear ranking proposed by Baker [4] is applied. The individual with the highest value of F_β (see equation 1) in any generation is designated as the result of the whole run (result designation criterion). Runs are carried out with a population size of 300 individuals through 500 generations. Crossover probability is set to 0.95 and the mutation operator is applied to 1 out of 5 individuals which are the result of applying the reproduction and crossover operators. Statistical data is obtained by performing 30 independent runs.

Comprehensibility is of utter importance within the data mining context. Therefore, the main point of this work is to obtain rules that are comprehensible to the user. Even if comprehensibility is a very subjective concept, here it is measured by the syntactic complexity (length) of rules. Such complexity is obtained by counting the amount of nodes in the syntactic tree.

4.1 Application Problem

The classification problem chosen for our experimental study consists in predicting the housing value (numerical value binned into three intervals) from information reflecting housing conditions (amount of rooms, crime rate, etc.) The data set comes from the UCI repository (Housing Database) and has 13 continuous attributes (including the "class" attribute "MEDV"), 1 binary-valued attribute and 506 instances. Instances reflect housing conditions in the suburbs of the City of Boston. Some attributes are CRIM (crime rate), DIS (distance to the five working centers in Boston), etc.

Attribute selection, together with the binning process and selection of the training and test sets, are performed with the data mining tool WEKA (Waikato Environment for Knowledge Analysis) [7].

Attributes are binned into intervals of equal length allowing the binning method to find the optimum amount of bins, except for the objective attribute, which is binned arbitrarily into three intervals of equal length, representing high, medium, and low housing prices. Table 1 presents information regarding each objective attribute bin. The Boston Housing data set is divided as follows: 66% of the data are chosen randomly for training, the remaining data conform the testing set.

Table 1: Information regarding each objective attribute bin

Class	Lim.Inf	Lim.Sup.	#Inst.Training
Low	-INF	16.666	39
Medium	16.666	33.333	118
High	33.333	INF	17

4.2 Comparative Analysis

In this subsection, a comparative study between a basic system for the evolution of classification rules, named as system GP_B and the system incorporating the adaptive and ranking procedure proposed in section 3, named as system GP_{AR} , is presented. System GP_B uses F_β as the fitness function, whereas GP_{AR} incorporates a sorting algorithm based on probabilities for ranked solutions.

In preliminary experiments, best results from GP_B were obtained by setting $\beta = 0.8$. On the other hand, for GP_{AR} the best ones were achieved by assigning the same value to β with **MaxLong**$=$ 75. Therefore, these values for the β and **MaxLong** parameters were used in order to carry out a comparative study in the final experiments reported here. The results are shown in table 2 where

Table 2: Results for approach GP_B and GP_{AR} (**MaxLong =** 75 and $\beta = 0.8$)

Appro.	Class	Support	St.Dev.	Precision	St.Dev.	Length	St.Dev.	Time	St.Dev.
GP_B	Low	0.6411	0.033	0.7453	0.019	2567.33	112.61	237.88	22.13
	Med.	0.9112	0.024	0.7842	0.0254	9011.82	500.45	395.04	53.17
	High	0.8309	0.047	0.4755	0.0435	2924.19	201.15	310.29	20.34
GP_{AR}	Low	0.6440	0.012	0.7545	0.027	55.247	6.771	110.13	11.22
	Med.	0.7605	0.045	0.8701	0.031	42.931	8.486	75.72	9.46
	High	0.8361	0.028	0.5539	0.024	166.468	26.535	133.77	10.15

it can be observed that system GP_{AR} achieves an important reduction in the complexity of the rules obtained for each class, thus improving their comprehensibility. This improvement is obtained without compromising the predictive quality of the model. Furthermore, a slight improvement in support and precision values can be noticed regarding system GP_B (for certain classes).

Also, the reduction in CPU time taken for the evolution of rules is noticeable. It takes about one half of the time for classes **Low** and **High**, while for class **Medium**, it takes approximately 5 times less. Solution evaluation is the most time consuming task in the evolutionary process, so this improvement in CPU time is a direct consequence of the reduction of hypothesis complexity. Therefore, the overhead introduced by the proposed approach has been overcome by the improvement in run time.

An example of a result obtained by using $MaxLong = 5$ and $\beta = 0.8$ for the **Medium** class is:

$$((NOT (AND (< \ -INF \quad DIS) (\leq \ DIS \quad 2,02))))$$

The above solution has a support value of 0.89 and a precision value of 0.75, whereas its complexity is 4. Regarding the results shown in table 2, complexity has been reduced noticeably as the value of support increased, however, there is also an important decrease in the precision measure. On evolving classification rules we must recall that best solutions as regards to predictive quality may be found in other regions of the search space than those where hypothesis have the desired comprehensibility. This is why it is necessary to reach a consensus between comprehensibility on the one hand, and predictive quality on the other. Thus, the value of **MaxLong** parameter must be tuned in order to achieve such balance in the obtained model.

4.3 Analysis of the MaxLong parameter

In this section we study the influence of the **MaxLong** parameter over the predictive quality and the complexity of the result obtained, and its influence on the CPU time required to obtain it. Figure 4 shows results for support and precision measures. On the left side of figure 4 we can observe that

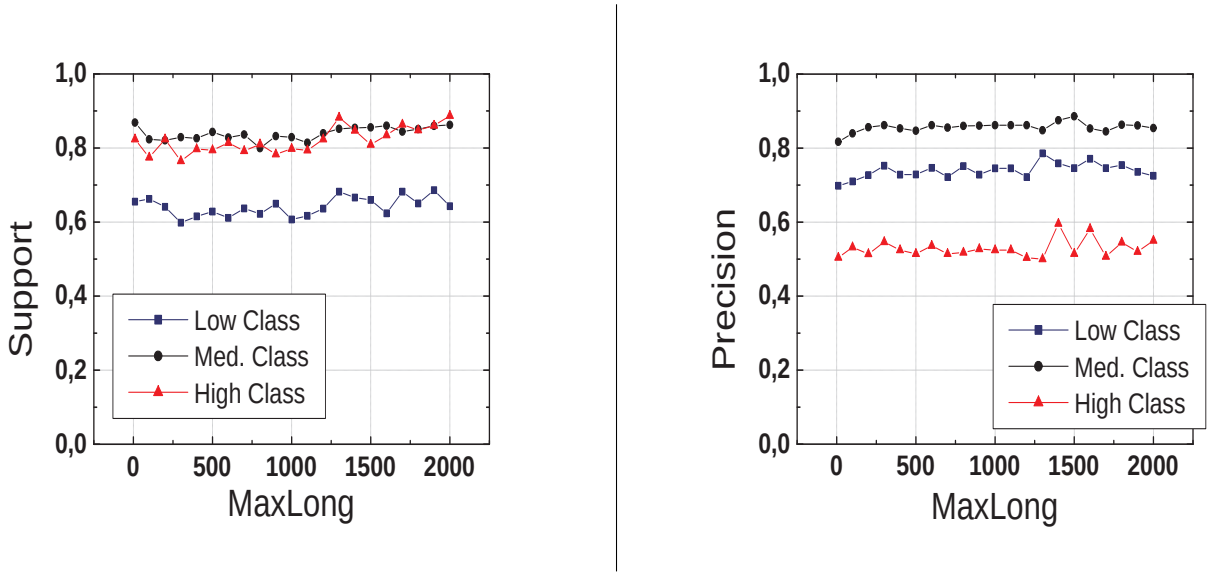

Figure 4: Influence of **MaxLong** on the support and precision values. Left: Support Vs. **MaxLong**. Right: Precision Vs. **MaxLong**

neither support nor precision values vary significantly for different values of the **MaxLong** parameter. In both cases, there is a significant difference between classes, showing that some classes are more difficult to predict than others.

Figure 5 shows the results for complexity and CPU time. With respect to classes **Medium** and **High**, it can be seen (left side of figure 5) a clear increment on the complexity of the model obtained by incrementing the values of **MaxLong**. However, for **Low** class, the variation in the complexity of solutions obtained is not meaningful. This is due to low complexity solutions found in earlier generations exceeding in predictive quality the more complex solutions found in later generations of evolutionary process. However, in all cases, average population complexity increments. Thus, this parameter biases the search towards regions where there are hypotheses with a certain complexity.

On the right (figure 5) it can be observed a clear tendency regarding an increase of CPU time when incrementing **MaxLong**. This last result can be explained by the increment in the average structural complexity of the whole population.

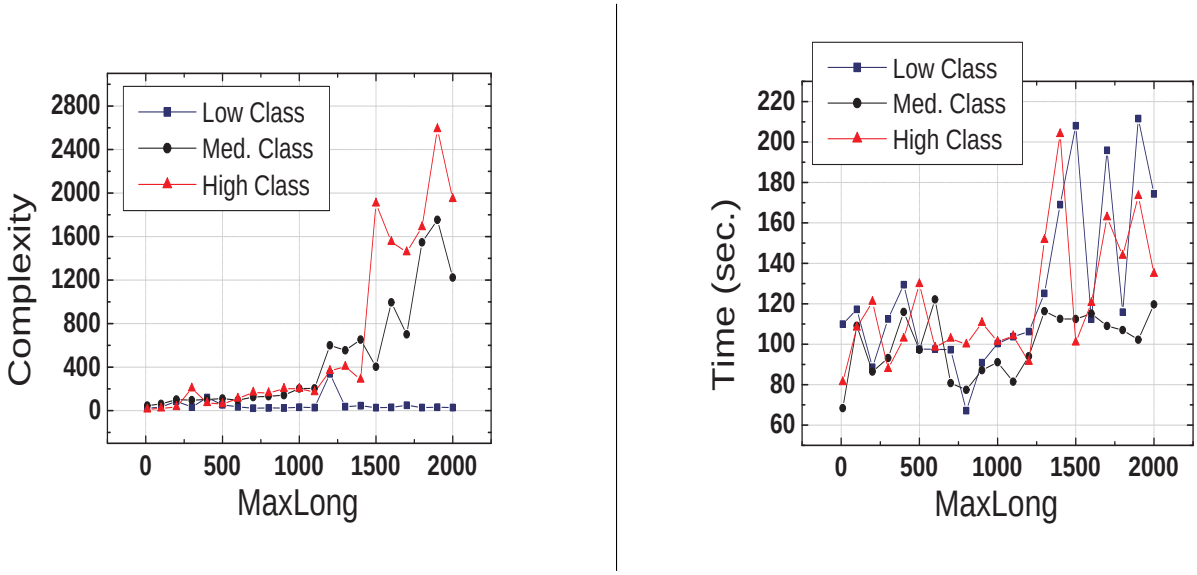

Figure 5: Influence of **MaxLong** on the solution complexity and CPU time required to obtain it. Left: Complexity Vs. **MaxLong**. Right: Time Vs. **MaxLong**

5 CONCLUSIONS AND FUTURE WORK

The proposed method intends to balance support, precision and comprehensibility measures in the evolution of classification rules. The aim is to direct the search towards regions with the desired characteristics, i.e., comprehensible hypothesis with a high predictive accuracy.

According to the results presented in section 4, it can be concluded that the proposed approach is capable of focusing the search on regions where hypothesis have a structural complexity that allows an appropriate understanding. Improvements achieved with respect to GP that does not apply this approach are significant.

Regarding support and precision values, the global quality of the results obtained is equal to or better than that obtained without applying the proposed approach; i.e., the quality of the solutions is not affected by the decreased structural complexity (which, in addition, increases understanding). A good balance between support and precision measures was also achieved, and the execution time for the evolution of rules decreases considerably when applying GP_{AR}.

Even if it is possible to evaluate separately the quality of the rules by using the support and precision measures, it would be advantageous to evaluate the predictive quality of the set of rules as a whole (the set conformed by rules from all classes). Also, it would be convenient to make a comparative analysis against a different algorithm for rule discovery (e.g., **C5.0**). Additionally, it would be possible to evolve more than one rule for each class. However, if a large number of rules from each class are evolved, the complexity of the model will increase significantly. Also, improvements in the adaptive process and result designation criterion can be proposed for improving the solutions quality. Finally, all the results presented suggest directions for future work from theoretical and experimental perspectives, which can lead to improvements of models both in their understanding and predictive quality.

REFERENCES

[1] Celia C. Bojarczuk, Heitor S. Lopes, and Alex A. Freitas. Genetic programming for knowledge discovery in chest-pain diagnosis. *IEEE Engineering in Medicine and Biology Magazine*, 19(4):38–44, July-August 2000.

[2] Kalyanmoy Deb and Deb Kalyanmoy. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 2001.

[3] C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI repository of machine learning databases, 1998.

[4] J. E. Baker J. Adaptive selection methods for genetic algorithms. In *Proc. ICGA 1*, pages 101–111, 1985.

[5] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.

[6] Edgar Noda, Alex A. Freitas, and Heitor S. Lopes. Discovering interesting prediction rules with a genetic algorithm. In Peter J. Angeline, Zbyszek Michalewicz, Marc Schoenauer, Xin Yao, and Ali Zalzala, editors, *Proceedings of the Congress on Evolutionary Computation*, volume 2, pages 1322–1329, Mayflower Hotel, Washington D.C., USA, 6-9 1999. IEEE Press.

[7] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, second edition, 2005.