# Multiobjective evolutionary induction of subgroup discovery rules in a market problem

Francisco Berlanga[1], María José del Jesus[1], Pedro González[1], Francisco Herrera[2]
[1] Department of Computer Science, University of Jaén, Jaén, Spain
{berlanga, mjjesus, pglez}@ujaen.es
[2] Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
herrera@decsai.ugr.es

*Abstract*— **In this paper we present a proposal for knowledge extraction in a market problem, the study of the influence that trade fair planning variables have on the attainment level of objectives previously planned. In this problem the main objective is the extraction of rules which describe subgroups contributing relevant information on them. The evolutionary algorithm proposed for the induction of descriptive rules follows a multiobjective approach in order to optimize in a suitable way the different quality measures used in this kind of problems. The obtained results show the adjustment of the multiobjective approach for the extraction of a set of rules that describe the knowledge extracted with a high level of confidence, support and interest.**

## I. INTRODUCTION

In the area of marketing, and specifically in the planning of trade fairs, it is important the extraction of conclusions from information of previous fairs in order to determine the relationship among the planning variables of a trade fair and the success of the stand. For this problem is suitable an algorithm of rule induction which describe subgroups.

In the area of subgroup discovery [31] any algorithm for rule induction must simultaneously optimize different objectives like the level of interest or novelty, the support or level of generality and the accuracy of the extracted knowledge, among others. The combination of these objectives in a single quality measurement usually produces compensation between such objectives causing that particular objectives are not optimized in a suitable form.

Genetic Algorithms (GAs) [22] [19] perform a global search that make them specially adapted in the resolution of different problems present in any knowledge discovery process [16]. In particular, multiobjective GAs are adapted to solve problems in which different objectives must be optimized. In the specialized bibliography can be found several evolutionary proposals for multiobjective optimization [9] [6], and recently the multiobjective GAs have been used in the extraction of knowledge in data mining [21] [25].

In this paper we present a proposal for the induction of rules which describe subgroups based upon a multiobjective GA, which combines the approximated reasoning method of the fuzzy systems with the learning capacities of the GAs.

To do so, the paper is arranged in the following way. Section II describes the marketing problem. Section III reviews the GAs proposals for induction of rules which describe subgroups. Section IV briefly describes the multiobjective GAs. Section V details our multiobjective evolutionary algorithm for the induction of subgroup discovery rules. Finally, section VI shows the experimentation carried out and the analysis of results, and section VII describes the conclusions obtained.

## II. THE EXTRACCION OF USEFUL INFORMATION ON TRADE FAIRS

Businesses consider trade fairs to be an instrument which facilitates the attainment of commercial objectives such as contact with current clients, the securing of new clients, the taking of orders, and the improvement of the company image amongst others. One of the main inconveniences in this type of trade fair is the elevated investment which they imply in terms of both time and money. This investment sometimes coincides with a lack of planning which emphasises the impression that trade fairs are no more than an "expense" which a business must accept for various reasons such as tradition, client demands, and not giving the impression that things are going badly among other factors [34]. Therefore convenient, is the automatic extraction of information about the relevant variables which permit the attainment of unknown knowledge, which partly determines the efficiency of the stands of a trade fair.

In the Machinery and Tools biennial held in Bilbao in March 2002, information was collected on all these aspects. To be precise, 104 variables of 228 exhibitors were analysed. Of these variables, 7 are continuous and the rest are categorical features, result of an expert discretization. Additionally, for each exhibitor, based on various marketing criteria, the stand's global efficiency was rated as *high*, *medium* or *low*, in terms of the level of achievement of objectives set for the trade fair.

For this real problem, the data mining algorithm should extract information of interest about each of the three efficiency groups of the stands. The rules generated will determine the influence which the different fair planning variables have over the results obtained by the exhibitor, therefore allowing fair planning policies to be improved.

## III. GENETIC ALGORITHMS IN RULE INDUCTION PROCESSES

A data mining algorithm can discover knowledge using different representation models and techniques from two different perspectives:

- Predictive induction, whose objective is the discovery of knowledge for classification or prediction.
- Descriptive induction, whose fundamental objective is the discovery of interesting knowledge from the data.

Considering the characteristics of the problem to be solved, the obtention of simple rules which provide conclusive information about the efficiency of the stands in trade fairs, the most suitable approach is descriptive induction.

In this paper we propose a GA for the descriptive induction of rules that describe subgroups, a task included in the area of data mining. So, in this section we will briefly describe the subgroup discovery task and the general trends in GAs for rule induction.

### *A. Subgroup discovery*

A subdivision of descriptive induction algorithms which has recently received a lot of attention from researchers is subgroup discovery. It is a form of supervised inductive learning of subgroup descriptions in which, given a set of data and having a property of interest to the user, attempts to locate subgroups which are statistically "most interesting" for the user.

Subgroup discovery has the objective of discovery interesting properties of subgroups obtaining simple rules (i.e. with an understandable structure and with few variables), highly significant and with high support (i.e. covering many of the instances of the target class).

The concept was initially formulated by Klösgen in his rule learning algorithm EXPLORA [28] and by Wrobel in the algorithm MIDOS [41]. Both use a rule-extraction model based on decision trees, in order to obtain the best subgroups among the population. In order to evaluate the subgroups, evaluation measurements are defined which determine the interest of an expression through a combination of unusualness and size. MIDOS tackles, within this same approach, the problem of discovery in multi-relational databases.

In addition to these proposals, different methods have been developed which obtain descriptions of subgroups represented in different ways and using different quality measurements:

- The SD algorithm [17] which induces rules guided by expert knowledge. In this proposal, instead of defining an optimal measurement for the search and the automatic selection of subgroups as the EXPLORA and MIDOS algorithms do, the objective is to help the expert to carry out flexible and effective searches over a wide range of optimal solutions.
- The CN2-SD algorithm [32] which induces subgroups in the form of rules using as quality measurement the relationship between true positive rate and false positive rate. This algorithm is applied to a marketing problem with information about interviews with potential clients. Based on the clients' responses, the objective is to discover which brands can potentially be more used by clients, and to direct a specific marketing campaign towards these clients.
- In data analysis with high uncertainty it is useful to present a subgroup of the population by listing its support factors, instead of using a subgroup discovery approach with obtains descriptions of subgroups in the form of rules. In [5], this approach is applied to a marketing problem in which the objective is to identify the characteristics of clients who do not recognise and/or use a given brand of non-alcoholic drink.
- SubgroupMiner [29] supports multirelational hypotheses, efficient data base integration, discovery of causal subgroup structures, and visualization based interaction options. The key point of this approach is the representation of spatial subgroups using an object-relational query language by embedding part of the search algorithm in a spatial database system.

These algorithms are adaptations of classification rule extraction models for the subgroup discovery task. Currently, interest is starting to be shown in the development of subgroup discovery approaches by modifying association rule extraction algorithms [31] [27].

### *B. Evolutionary rule induction*

Multiple proposals of GAs have been developed for the extraction of rules with different approaches: classification, association or functional dependencies. In our problem the objective is to generate rules whose consequent has one previously fixed variable, so in this section we will review GAs for the extraction of classification and association rules.

GAs are optimisation and search algorithms inspired in natural evolution processes and initially defined by Holland [22] [19]. Stated simply, they work as follows: the system starts with an initial population of individuals who encode, through a form of genetic representation, candidate solutions for the proposed problem. This population of individuals (called chromosomes) evolves in time through a process of competition and controlled variation. Each chromosome in the population is associated with a fitness function in order to determine which chromosomes will be selected to form part of the new population in the competition process. The new population will be created using genetic operators of crossover and mutation. Bäck, Fogel and Michalewicz in [3] give a complete description of GAs as well as other examples of Evolutionary Algorithms.

GAs have several advantages as a rule induction method:

- They tend to cope well with attribute interaction because they usually evaluate a rule as a whole via fitness function, rather than evaluating the impact of adding/removing one condition to/from a rule.
- They have the ability to scour a search space thoroughly and the ability to allow arbitrary fitness functions in the search. The fitness function can contain different criteria such as the ability to penalise overlap among rules or rule sets with too many rules or a problem-specific quality measure, etc.
- In addition, the genetic search performs implicit backtracking in its search of the rule space, thereby allowing it to find complex interactions that other non-backtracking searches would miss.
- An additional advantage over other conventional rule-learning algorithms is that the search is carried out among a set of competing candidate rules or rule sets.

However, this is not to say that GAs are inherently superior to rule induction algorithms as no rule discovery algorithm is superior in all cases [13] [33].

In KDD literature different GA proposals have been presented with predictive or descriptive aims. Rule induction algorithms for subgroup discovery (the aim of which is fundamentally descriptive) share characteristics with algorithms which guide the induction process using predictive quality measurements. In this section we will describe some of the main GA proposals for rule induction, no matter what is their final aim.

In the design of any rule induction GA, the genetic representation of the solutions is the most determining aspect of its characteristics. In this sense, different proposals in the bibliography are grouped around two approaches in order to encode rules within a population of individuals [8]:

- The *"Chromosome = Rule"* approach, in which each individual codifies a single rule.
- The *"Chromosome = Set of rules"*, also called the Pittsburgh approach, in which each individual represents a rule set. GIL [26], GA-MINER [15], or dAR [2], are examples of GAs of this type.

In turn, within the "Chromosome = Rule" approach, there are three generic proposals:

- The Michigan approach, in which each individual codifies a single rule. This kind of systems is usually called learning classifier systems. They are rule-based, message-passing systems that employ reinforcement learning and the GA to learn rules that guide their performance in a given environment [30]. XCS [40] is an example of GA of this type.
- The IRL (Iterative Rule Learning) approach, in which each chromosome represents a rule, but the GA solution is the best individual obtained and the global solution is formed by the best individuals obtained when the algorithm is run multiple times. SLAVE [20] and the proposal of Carvalho and Freitas [4] are GAs within this model.
- The "cooperative-competitive" approach, in which the complete population or a subset of it codifies the rule base. REGAL [18], GA-PVMINER [1] or GLOWER [12] are examples of this approach.

In algorithms for subgroup discovery the most suitable approach is the *"Chromosome = Rule"* approach because the objective is to find a reduced set of rules in which the quality of each rule is evaluated independently of the rest, and it is not necessary to evaluate jointly the rule set. This is the approach used in this evolutionary proposal.

## IV. MULTIOBJECTIVE GENETIC ALGORITHMS

As we have previously commented, in the area of subgroup discovery any rule induction algorithm must optimize simultaneously several objectives. The more suitable way to approach them is by means of multiobjective optimization algorithms in which we search a set of optimal alternative solutions (rules in our case) in the sense that no other solution within the search space is better than it in all the considered objectives. The expert will use the set of rules obtained to select all or a set of them for the description of the subgroups based on the particular preference information of the problem.

In a formal way, a multiobjective optimization problem can be defined in the following way:

$$\min/\max \vec{y} = f(\vec{x}) = f_1(\vec{x}), f_2(\vec{x}), \ldots, f_n(\vec{x})) \qquad (1)$$

where $\vec{x} = (x_1, x_2, \ldots x_m)$ is the decision vector and $\vec{y} = (y_1, y_2, \ldots, y_n)$ is the objective vector (a tuple with $n$ objectives). The objective of any multiobjective optimization algorithm is to find all the decision vectors for which the corresponding objective vectors can not be improved in a dimension without degrading another, which is denominated optimal Pareto front.

In the last two decades an increasing interest has been developed in the use of GAs for multiobjective optimization. There are multiple proposals of multiobjective GAs [9] [6] that can be grouped around three approaches:

- Aggregation methods that combine the objectives in an scalar function. They have the disadvantage of the possible compensation among objectives, the deep problem knowledge required and that, in general, they do not provide a family of solutions. RW-GA [24] is an algorithm of this type.
- Population based methods, in which search is guided in different directions to generate populations of non-dominated solutions. Within this approach are included algorithms like MOGA [14], NPGA [23], NSGA [39] and NSGA II [10].
- Elitism based methods, which maintain an elite population with non-dominated solutions which take part in different forms in the evolution depending on the proposal. Within this approach are included the algorithms SPEA [42], SPEA2 [43] and μ-λ MEA [37]. Our proposal is based in the SPEA2 algorithm.

## V. A MULTIOBJECTIVE EVOLUTIONARY APPROACH TO OBTAIN DESCRIPTIVE FUZZY RULES

In this section we describe a multiobjective GA for the extraction of rules which describe subgroups. The proposal extracts rules whose antecedent represents a conjunction of variables and whose consequent is fixed. The objective of this evolutionary process is to extract for each value of the target variable a variable number of different rules expressing information on the examples of the original set. This algorithm can generate fuzzy and/or crisp rules, for problems with continuous and/or nominal variables.

The multiobjective GA follows the SPEA2 approach [43], and so applies the concepts of elitism in the rule selection (using a secondary or elite population) and search of optimal solutions in the Pareto front (the individuals of the population are ordered according to if each individual is or not dominated using the concept of Pareto optimal).

Any multiobjective GA must be designed to achieve two purposes: to obtain good approximations to the Pareto front and to maintain the diversity of the solutions, with the objective of

correctly sample the solution space and not converging to a single solution or a limited section of the front. In order to preserve the diversity at a phenotypic level the algorithm uses a niches technique that considers the proximity in values of the objectives and an additional objective based on the novelty (to promote rules which give information on examples not described by other rules of the population). Algorithm 1 shows the scheme of operation of the proposed model.

*Step 1. Initialization*:
Generate an initial population $P_0$ and create an empty elite population $P'_0 = \emptyset$. Set t = 0.
Repeat
*Step 2. Fitness assignment*: calculate fitness values of the individuals in $P_t$ and $P'_t$.
*Step 3. Environmental selection*: copy all non-dominated individuals in Pt and $P'_t$ to $P'_{t+1}$. If size of $P'_{t+1}$ exceeds the number of individuals to store (N) reduce $P'_{t+1}$ by means of the truncation function; otherwise if size of $P'_{t+1}$ is less than N, fill $P'_{t+1}$ with dominated individuals in $P_t$ and $P'_t$.
*Step 4. Mating selection*: perform binary tournament selection with replacement on $P'_{t+1}$ applying later crossover and mutation operators in order to fill the mating pool (obtaining $P_{t+1}$).
*Step 5*. Increment generation counter (t = t+1)
While stop condition is not verified.
*Step 6*. Return the non-dominated individuals in $P'_{t+1}$.

Algorithm 1. Scheme of the proposed algorithm

In the next subsection we will detail the chromosome representation, the objectives, the fitness assignment, the environmental selection and the reproduction model.

### A. Chromosome representation

Each candidate solution is coded according to the "cooperative-competitive" approach representing only the antecedent in the chromosome and associating all the individuals of the population with the same value of the target variable. This representation of the target variable means that the evolutionary multiobjective algorithm must be run many times in order to discover the rules of the different classes, but it assures the knowledge extraction in all the classes.

The problem is described by means of nominal and continuous variables. The continuous ones are considered as linguistic variables with linguistic labels. The fuzzy sets corresponding to the linguistic labels are defined by expert information or by a uniform fuzzy partition with triangular membership functions, as shown in Figure 1.

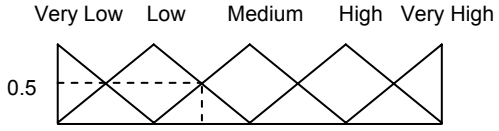


Figure 1. Example of fuzzy partition for a continuous variable

All the information relating to a rule is contained in a fixed-length chromosome for which we use an integer representation model (the i-th position indicates the value adopted by the i-th variable). The set of possible values for the categorical features is that indicated by the problem plus an additional value which, when it is used in a chromosome, indicates that the corresponding variable does not take part in the rule. For continuous variables the set of values is the set of linguistic terms determined heuristically or with expert information, plus the value indicating the absence of the variable.

### B. Definition of the objectives of the algorithm

In the rule induction process we try to get rules with high predictive accuracy, comprehensible and interesting. In our proposal, we have defined four objectives:

- *Confidence*. Determines the accuracy of the rule, and reflects the degree to which the examples within the zone of the space marked by the antecedent verify the information indicated in the consequent of the rule. In order to calculate this factor we use an adaptation of Quinlan's accuracy expression [36] to generate fuzzy classification rules [7]: the quotient between the sum of the degree of membership of the examples of this class to the zone determined by the antecedent, and the sum of the degree of membership of all the examples (irrespective of their class) to the same zone. In order to calculate these membership degrees, we use triangular membership functions and the minimum t-norm. In the case of non-fuzzy rules, the degrees of membership correspond to the classic sets, i.e. 0 or 1.
- *Support*. This is the measurement of the degree of coverage that the rule offers to examples of that class. It is calculated as the quotient between the number of examples belonging to the class which are covered by the rule and the total number of examples from the same class.
- *Interest*. The degree of interest is assessed objectively. We use the interest criteria provided by Noda, Freitas and Lopes [35] in a dependence modelling process. In our proposal we only use the term referring to the antecedent for the interest calculation, because the consequent is prefixed. The information measurement for the interest is as follows:

$$Interest = 1 - \left( \frac{\sum_{i=1}^{n} Gain(A_i)}{n \cdot \log_2 \left( \left| dom(G_k) \right| \right)} \right) \qquad (2)$$

where *n* is the number of variables which appear in the antecedent of the rule, *Gain (Ai)* is the information gain of the attribute $A_i$, and $|dom(G_k)|$ is the cardinality (number of possible values) of the target variable. To compute the information gain in the case of continuous variables we perform a discretization of the variable in so many intervals as linguistic labels are considered. The interpretation of this method is as follows:

variables with high information gain are suitable for predicting a class when they are considered individually. However, if the user knows the most predictive variables for a specific application domain, the rules containing these variables are less interesting. This way, the antecedent of a rule is more interesting if it contains attributes with a small quantity of information.

- *Original support*. This objective is a measurement of the originality level of the rule compared with the rest of rules. It is computed adding, for each example belonging to the antecedent of the rule, the factor *1/k*, where *k* is the number of rules of the population that describe information on that example. This measurement promotes the diversity at the population at a phenotypic level.

### C. Fitness assignment

The fitness assignment for the rules extracted is performed in the following way:

- For each individual in the population is computed the value for all the objectives.
- The values reached by each individual in the population (and in the elite population) are used to compute what individual dominate what other in the populations.
- The strength of each individual is computed as the number of individuals that it dominates.
- The raw fitness of each individual is determined as the sum of the strength of its dominators (even in the population as in the elite population).
- The computation of the raw fitness offers a niching mechanism based in the concept of Pareto dominance, but it can fail when much of the individuals are non-dominated. To avoid this, it is included additional information on density to discriminate between individuals with the same values of raw fitness. The density estimation technique used in SPEA2 is an adaptation of the method of the k-th nearest neighbour [38], where the density in a point is decreasing function of the distance to the k-th nearest point. In this proposal we use the inverse of the distance to the k-th nearest neighbour as density estimation.
- The fitness value of each individual is the sum of its raw fitness value and its density.

### D. Environmental selection

This algorithm establishes a fixed length for the elite population, so it is necessary to define a truncation and a fill function. The truncation function allows eliminating the non-dominated solutions of the elite population if it exceeds the defined size. For this purpose it is used a niche schema defined around the density measured by the distance to its k-th nearest neighbour, in which, in an iterative process, in each iteration it is eliminated from the elite population the individual that is nearest of others respect of the values of the objectives.

The fill function allows adding dominated individuals from the population and the elite population until the exact size of the set is reached (ordering the individuals according to their fitness values).

### E. Reproduction model and genetic operators

We use the following reproduction model:

- Join the original population with the elite population obtaining then the non-dominated individuals of the joining of these populations.
- Apply a binary tournament selection on the non-dominated individuals.
- Apply recombination to the resulting population by a two point cross operator and a biased uniform mutation operator in which half the mutations carried out have the effect of eliminating the corresponding variable, in order to increase the generality of the rules.

## VI. EXPERIMENTATION

To analyze the behaviour of the multiobjective proposal on the marketing problem, we have also run the evolutionary algorithm for the induction of subgroup discovery rules, AGI, developed by the authors [11], whose general characteristics are described next:

- Is an iterative model including a hybrid steady-state GA for the extraction of a subgroup discovery rule with the structure described in this paper.
- The iterative model allows new rules to be obtained while the generated rules reach a minimum level of confidence and give information on areas of search space in which examples which are not described by the rules generated by the previous iterations, remain.
- The GA uses the same chromosome representation and the genetic operators that the multiobjective proposal.
- The fitness function is a weighted lineal combination of the confidence, support and interest, computed according to the expressions described in section V.
- The rule obtained is improved, in a post-processing phase, by a hill-climbing process in which it eliminates variables of the rule while increasing the degree of confidence and obtaining more general rules (increasing the support).

For both algorithms, the experimentation is carried out performing 5 runs for each of the 3 classes of the target variable (*low*, *medium* and *high* efficiency) and the following common parameters:

- Population size: 100
- Maximum number of evaluations of individuals: 10,000
- Mutation probability: 0.01
- Number of linguistic labels for the continuous variables: 3

In addition, the multiobjective GA needs a size for the elite population (established to 25 in this experimentation) and a cross probability (fixed to 0.6). AGI algorithm needs a minimum confidence value under which the algorithm will stop evolving rules (established to 0.6). In the multiobjective GA, the final solution will be composed by all the solutions of the non-dominated set of solutions that surpass the same confidence level.

Tables I, II and III show the best results obtained with both algorithms for all the classes of the target variable (*low*, *medium* and *high* efficiency). The tables show the number of variables taking part in each rule (column *#V*) and the values for each of the objectives (*Supp* for the support, *Conf* for the confidence, *Int* for the interest, and *O.S.* for the original support).

TABLE I. RESULTS FOR LOW EFFICIENCY

| AG multiobjetivo | | | | | AGI | | | |
|---|---|---|---|---|---|---|---|---|
| *#V* | *Supp* | *Conf* | *Int* | *O.S.* | *#V* | *Supp* | *Conf* | *Int* |
| 9 | 5.26 | 100.00 | 0.58 | 0.05 | 3 | 5.26 | 100.00 | 0.59 |
| 11 | 15.79 | 66.67 | 0.58 | 0.18 | 4 | 2.63 | 100.00 | 0.58 |
| 7 | 42.10 | 61.54 | 0.57 | 0.60 | 5 | 2.63 | 100.00 | 0.56 |
| 6 | 44.74 | 64.92 | 0.54 | 0.85 | | | | |
| 11 | 21.05 | 100.00 | 0.56 | 0.24 | | | | |
| 10 | 18.42 | 87.50 | 0.56 | 0.42 | | | | |
| 8 | 5.26 | 76.92 | 0.59 | 0.05 | | | | |
| 9 | 36.84 | 73.68 | 0.57 | 0.49 | | | | |
| 11 | 34.21 | 87.16 | 0.56 | 0.41 | | | | |
| 7 | 23.68 | 60.00 | 0.58 | 0.52 | | | | |

TABLE II. RESULTS FOR MEDIUM EFFICIENCY

| AG multiobjetivo | | | | | AGI | | | |
|---|---|---|---|---|---|---|---|---|
| *#V* | *Supp* | *Conf* | *Int* | *O.S.* | *#V* | *Supp* | *Conf* | *Int* |
| 1 | 95.27 | 65.58 | 0.20 | 6.78 | 4 | 0.68 | 100.00 | 0.61 |
| 5 | 1.35 | 100.00 | 0.60 | 0.07 | 7 | 2.70 | 66.67 | 0.62 |
| 2 | 79.05 | 68.02 | 0.12 | 5.27 | 2 | 2.03 | 100.00 | 0.57 |
| 3 | 5.40 | 72.73 | 0.61 | 0.33 | 4 | 4.05 | 100.00 | 0.57 |
| 3 | 40.54 | 77.92 | 0.38 | 2.26 | 3 | 3.38 | 100.00 | 0.51 |
| 3 | 20.27 | 68.18 | 0.59 | 1.46 | 5 | 8.11 | 100.00 | 0.55 |
| 5 | 32.43 | 87.27 | 0.35 | 1.60 | 2 | 45.95 | 69.39 | 0.62 |
| 2 | 87.16 | 67.55 | 0.33 | 5.92 | | | | |
| 4 | 64.86 | 71.57 | 0.32 | 3.81 | | | | |
| 2 | 93.92 | 65.57 | 0.41 | 6.58 | | | | |
| 3 | 54.73 | 72.73 | 0.39 | 2.98 | | | | |
| 3 | 82.43 | 67.78 | 0.55 | 5.39 | | | | |
| 3 | 35.81 | 82.81 | 0.19 | 1.89 | | | | |
| 5 | 11.49 | 100.00 | 0.57 | 0.60 | | | | |
| 2 | 90.54 | 66.34 | 0.60 | 6.35 | | | | |
| 3 | 71.62 | 70.54 | 0.23 | 4.46 | | | | |
| 1 | 98.65 | 64.89 | 0.62 | 7.08 | | | | |
| 1 | 61.49 | 68.94 | 0.53 | 3.48 | | | | |
| 5 | 44.59 | 72.53 | 0.58 | 2.07 | | | | |
| 3 | 86.49 | 64.97 | 0.61 | 5.89 | | | | |
| 4 | 23.65 | 83.33 | 0.51 | 0.97 | | | | |
| 3 | 76.35 | 68.27 | 0.53 | 4.85 | | | | |
| 4 | 36.49 | 80.60 | 0.43 | 1.91 | | | | |
| 3 | 52.70 | 73.77 | 0.52 | 2.70 | | | | |
| 4 | 50.00 | 73.27 | 0.57 | 2.38 | | | | |

TABLE III. RESULTS FOR HIGH EFFICIENCY

| AG multiobjetivo | | | | | AGI | | | |
|---|---|---|---|---|---|---|---|---|
| *#V* | *Supp* | *Conf* | *Int* | *O.S.* | *3V* | *Supp* | *Conf* | *Int* |
| 5 | 2.38 | 100.00 | 0.61 | 0.08 | 4 | 7.14 | 75.00 | 0.58 |
| 9 | 28.57 | 93.75 | 0.55 | 0.37 | | | | |
| 7 | 50.00 | 69.93 | 0.55 | 0.73 | | | | |
| 9 | 23.81 | 71.43 | 0.56 | 0.34 | | | | |
| 11 | 19.05 | 100.00 | 0.56 | 0.22 | | | | |
| 11 | 9.52 | 76.92 | 0.56 | 0.11 | | | | |
| 7 | 26.19 | 71.43 | 0.56 | 0.55 | | | | |
| 6 | 57.14 | 61.24 | 0.55 | 0.91 | | | | |

The experimentation shows that the multiobjective GA allows the extraction of sets of rules with a high cardinality (greater number of rules) that the AGI algorithm. These rules, with adequate values for the confidence, support and interest, describe more information over the three subgroups (*low*, *medium* and *high* efficiency). This is because the multiobjective approach allows us to obtain a set of proper solutions according to the different objectives.

In the multiobjective GA the diversity in the population at a phenotypical level during the evolutionary process (and so in the final solution) is promoted through two ways:

- By means of the incorporation of a new objective that considers the original contribution (in the sense of covered examples) of a rule. This allows the extraction of rule sets with higher support and increases the possibilities of obtaining a set of rules that describe information over all the examples, and not only on the majority.
- By means of a scheme of niches implemented in the truncation function that, if it is necessary to reduce the elite population, eliminates rules with similar values for the different objectives.

In the AGI algorithm, the extraction of a set of rules with enough diversity is promoted by the inclusion of the GA in an iterative model that extracts rules while the rules extracted describe information on new examples (sequential niches at a phenotypical level), but the experimentation shows that, for this problem, the results obtained by the multiobjective GA are better.

The multiobjective GA eliminates the compensation among quality measures and allows the extraction of sets of rules with a high level of confidence, support and interest. The high level of support obtained in the different rules with the multiobjective GA is especially significant, even for the *high* and *low* efficiency classes, difficult to describe in this problem.

In this aspect, the results show that with the AGI algorithm sometimes are extracted rules with a higher level of confidence, especially for the *low* and *medium* efficiency classes (Tables I and II). Nevertheless, the high values of confidence achieved by the rules extracted by AGI bias the search and make difficult to attain up high support values for the *low* and *high* efficiency classes.

Both proposals allow the extraction of descriptive sets of rules due to the use of linguistic labels for the continuous variables and the low number of variables taking part in each

rule (less than 10% of the 104 variables). In this aspect we must emphasize that the extracted rules by the AGI algorithm are simpler that the extracted by the multiobjective proposal. The use in AGI of a hill-climbing algorithm that optimize each one of the extracted rules allows increasing their simplicity.

## VII. Conclusions

In this paper we describe an evolutionary multiobjective model for the descriptive induction of fuzzy rules which describe subgroups applied to a real knowledge extraction problem in trade fairs.

In spite of the characteristics of the problem (elevated number of variables and lost values, low number of examples and few continuous variables) this multiobjective approach to the problem allows to obtain sets of rules that are easily interpretable, and with a high level of confidence and support.

In future studies, we will examine the use of a more flexible structure for the rule using disjunctive normal form (DNF) rules and the study of an appropriate interest measurement for this structure.

## Acknowledgment

This work was supported by the Spanish Ministry of Science and Technology and by the European Fund. FEDER under Projects TIC-2002-04036-C05-01 and TIC-2002-04036-C05-04, and the nets TIN2004-20061-E and TIN2004-21343-E.

## References

[1] D.L.A. Araujo, H.S. Lopes, and A.A. Freitas, "A parallel genetic algorithm for rule discovery in large databases", in Proceedings IEEE Conference on Systems, Man and Cybernetics, vol. III, pp. 940-945, 1999.

[2] W.H. Au, and K.C.C. Chan, "An evolutionary approach for discovering changing patterns in historical data", in Proceedings of 2002 SPIE 4730, Data Mining and Knowledge Discovery: Theory, Tools and Technology, vol. IV, pp. 398-409, 2002.

[3] T. Bäck, D. Fogel, and Z. Michalewicz, "Handbook of Evolutionary Computation", Oxford University Press, Oxford, 1997.

[4] D.R. Carvalho, and A.A. Freitas, "A genetic algorithm for discovering small-disjunct rules in data mining", Applied Soft Computing, vol. 2, pp. 75-88, 2002.

[5] B. Cestnik, N. Lavrac, F. Zelezny, D. Gamberger, L. Todorovski and M. Kline, "Data mining for decision support in marketing: A case study in targeting a marketing campaign", in Proceedings of the ECML/PKDD-2002 Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning, pp. 25–34, 2002.

[6] C.A. Coello, D.A. Van Veldhuizen, and G.B. Lamont, Evolutionary algorithms for solving multi-objective problems, Kluwer Academic Publishers, 2002.

[7] O. Cordón, M.J. del Jesus, and F. Herrera, "Genetic learning of fuzzy rule-based classification systems co-operating with fuzzy reasoning methods", International Journal of Intelligent Systems, vol. 13 (10/11), pp. 1025-1053, 1998.

[8] O. Cordón, F. Herrera, F. Hoffmann, and L. Magdalena, Genetic fuzzy systems: evolutionary tuning and learning of fuzzy knowledge bases, World Scientific, 2001.

[9] K. Deb, Multiobjective optimization using evolutionary algorithms, Wiley, 2001.

[10] K. Deb, A. Pratap, A. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II", IEEE Transactions on Evolutionary Computation, vol. 6 (2), pp. 182-197, 2002.

[11] M.J. Del Jesus, P. González, F. Herrera, and M. Mesonero, "Evolutionary inducción of descriptive fuzzy rules in a market problem", in Proceedings of the First Workshop on Genetic Fuzzy Systems (GFS), pp. 57-63, Granada, 2005.

[12] V. Dhar, D. Chou, and F. Provost, "Discovering interesting patterns for investment decision making with Glower-a Genetic Learner Overlaid With Entropy Reduction", Data Mining and Knowledge Discovery, vol. 4, pp. 251-280, 2000.

[13] P. Domingos, "Occam's two razors: the sharp and the blunt", in Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98), pp. 37-43, 1998.

[14] C.M. Fonseca, and P.J. Fleming, "Genetic algorithms for multiobjective optimization: formulation, discussion and generalization", in Proceedings of the Fifth International Conference on Genetic Algorithms (ICGA), pp. 416-423, San Mateo, CA, 1993.

[15] I.W. Flockhart, and N.J. Radcliffe, GA-MINER: Parallel data mining with hierarchical genetic algorithms (Final Report by the University of Edimburgh, UK, EPCC-AIKMS-GA-Miner-Report 1.0), 1995.

[16] A.A. Freitas, Data mining and knowledge discovery with evolutionary algorithms, Springer, 2002.

[17] D. Gamberger, and N. Lavrac, "Expert-guided subgroup discovery: methodology and application", Artificial Intelligence Research, vol. 17, pp. 501–27, 2002.

[18] A. Giordana, and F. Neri, "Search-intensive concept induction", Evolutionary Computation, vol. 3 (4), pp. 375-416, 1995.

[19] D.E. Goldberg, Genetic algorithms in search, optimization and machine learning, Addison-Wesley, 1989.

[20] A. González, and R. Pérez, "SLAVE: a genetic learning system based on an iterative approach", IEEE Trans. Fuzzy Systems, vol. 7(2), pp. 176-191, 1999.

[21] A. Ghosh, and B. Nath, "Multi-objective rule mining using genetic algorithms", Information Sciences, vol. 163 (1-3), pp. 123-133, 2004.

[22] J.H. Holland, Adaptation in natural and artificial systems, University of Michigan Press, 1975.

[23] J. Horn, and N. Nafpliotis, Multiobjective optimization using the niched pareto genetic algorithms (IlliGAL Report 93005, University of Illinois, Urbana, Champaign), 1993.

[24] H. Ishibuchi, and T. Murata, "A multiobjective genetic local search algorithm and its application to flowshop scheduling", IEEE Trans. System, Man and Cybernetics, vol. 28 (3), pp. 392–403, 1998.

[25] H. Ishibuchi, and T. Yamamoto, "Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining", Fuzzy Sets and Systems, vol. 141 (1), pp. 59-88, 2004.

[26] C.Z. Janikow, "A knowledge-intensive genetic algorithm for supervised learning", Machine Learning, vol. 13, pp. 189-228, 1993.

[27] V. Jovanoski, and N. Lavrac, "Classification rule learning with APRIORI-C", in Proceedings of the Tenth Portuguese Conference on Artificial Intelligence (EPIA), pp. 44–51, Berlin, 2001.

[28] W. Klösgen, "Explora: a multipattern and multistrategy discovery assistant", in Advances in Knowledge Discovery and Data Mining, V. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds., pp. 249-271, MIT Press, 1996.

[29] W. Klösgen, Handbook of data mining and knowledge discovery, Oxford University Press, 2002.

[30] T. Kovacs, Strength or accuracy: credit assignment in learning classifier systems, Springer-Verlag, 2004.

[31] N. Lavrac, B. Cestnik, D. Gamberger, and P. Flach, "Decision support through subgroup discovery: three case studies and the lessons learned", Machine Learning, vol. 57 (1-2), pp. 115-143, 2004.

[32] N. Lavrac, B. Kavsec, P. Flach, and L. Todorovski, "Subgroup discovery with CN2-SD", Machine Learning Research, vol. 5, pp. 153-188, 2004.

[33] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, Machine learning, neural and estatistical classification, Ellis Horwood, 1994.

[34] S. Miller, Saque el máximo provecho de las ferias, Ediciones Urano, 2003.

[35] E. Noda, A.A. Freitas, and H.S. Lopes, "Discovering interesting prediction rules with a genetic algorithm", in Proceedings of the Congress on Evolutionary Computation (CEC), pp. 1322-1329, Washington D.C., 1999.

[36] J.R. Quinlan, "Generating production rules from decision trees", in Proceedings of the Tenth International Joint Conference on Artificial Intelligence (IJCAI), pp. 304-307. San Mateo, CA, 1987.

[37] R. Sarker, K.H. Liang, and C. Newton, "A new multiobjective evolutionary algorithm", European Journal of Operational Research, vol. 140, pp. 12-23, 2002.

[38] B.W. Silverman, Density estimation for statistics and data analysis, Chapman and Hall, 1986.

[39] N. Srinivas, and K. Deb, "Multiobjective optimization using nondominated sorting in genetic algorithms", Evolutionary Computation, vol. 2, pp. 221-248, 1995.

[40] S.W. Wilson, "Classifier fitness based on accuracy", Evolutionary Computation, vol. 3(2), pp. 149-175, 1995.

[41] S. Wrobel, "An algorithm for multi-relational discovery of subgroups", in Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD), pp. 78–87, Berlin, 1997.

[42] E. Zitzler, and L. Thiele, "Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach", IEEE Transactions on Evolutionary Computation, vol. 3(4), pp. 257-217, 1997.

[43] E. Zitzler, M. Laumanns, and L. Thiele, "SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimisation", in Evolutionary methods for design, optimisation and control, K. Giannakoglou, D. Tsahalis, F. Periaux, K. Papailiou, and T. Fogarty, Eds., pp. 95-100, CIMNE, 2002.