Multiobjective Evolutionary Induction of Subgroup Discovery Fuzzy Rules: A Case Study in Marketing

Francisco Berlanga[1], María José del Jesus[1], Pedro González[1], Francisco Herrera[2], and Mikel Mesonero[3]

[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
[3] Department of Organization and Marketing,
University of Mondragón, Spain
mmesoner@eteo.mondragon.edu

Abstract. This paper presents a multiobjective genetic algorithm which obtains fuzzy rules for subgroup discovery in disjunctive normal form. This kind of fuzzy rules lets us represent knowledge about patterns of interest in an explanatory and understandable form which can be used by the expert. The evolutionary algorithm follows a multiobjective approach in order to optimize in a suitable way the different quality measures used in this kind of problems. Experimental evaluation of the algorithm, applying it to a market problem studied in the University of Mondragón (Spain), shows the validity of the proposal. The application of the proposal to this problem allows us to obtain novel and valuable knowledge for the experts.

Keywords: Data mining, descriptive induction, multiobjective evolutionary algorithms, genetic fuzzy systems, subgroup discovery.

1 Introduction

Knowledge Discovery in Databases (KDD) is defined as the non trivial process of identifying valid, original, potentially useful patterns which have comprehensible data [1]. Within KDD process the data mining stage is responsible for high level automatic knowledge discovery from information obtained from real data.

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 [2].
- *Descriptive induction*, whose fundamental objective is the discovery of interesting knowledge from the data. In this area, attention can be drawn to the discovery of association rules following an unsupervised learning model [3], subgroup discovery [4], [5] and other approaches to non-classificatory induction.

P. Perner (Ed.): ICDM 2006, LNAI 4065, pp. 337 – 349, 2006.
© Springer-Verlag Berlin Heidelberg 2006

In subgroup discovery the objective is, given a population of individuals and a specific property of individuals we are interested in, find population subgroups that are statistically "most interesting", e.g., are as large as possible and have the most unusual distributional characteristics with respect to the property of interest.

This paper describes a new proposal for the induction of rules which describe subgroups based upon a multiobjective evolutionary algorithm (MOEA) which combines the approximated reasoning method of the fuzzy systems with the learning capacities of the genetic algorithms (GAs).

The induction of rules describing subgroups can be considered as a multi-objective problem rather than a single objective one, in which the different measures used for evaluating a rule can be thought of as different objectives of the subgroup discovery rule induction algorithm. In this sense, MOEAs 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 [6], [7]. Recently the MOEAs have been used in the extraction of knowledge in data mining [8], [9].

The multiobjective algorithm proposed in this paper defines three objectives. One of them is used as a restriction in the rules in order to obtain a set of rules (the pareto front) with a high degree of coverage, and the other objectives take into account the support and the confidence of the rules. The use of this mentioned objective allows us the extraction of a set of rules with different features and labels for every property of interest.

The paper is arranged in the following way: Section 2 describes some preliminary concepts. The multiobjective evolutionary approach to obtain subgroup discovery descriptive fuzzy rules is explained in Section 3. Finally, Section 4 shows the experimentation carried out and the analysis of results and section 5 outlines the conclusions and further research.

2 Preliminaries

2.1 Subgroup Discovery

Subgroup discovery represents a form of supervised inductive learning in which, given a set of data and having a property of interest to the user (target variable), attempts to locate subgroups which are statistically "most interesting" for the user. In this sense, a subgroup is interesting if it has an unusual statistical distribution respect of the property of interest. The methods for subgroup discovery have the objective of discover 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).

An induced subgroup description has the form of an implication, R^i: $Cond^i \rightarrow Class_j$, where the property of interest for subgroup discovery is the class value $Class_j$ that appears in the rule consequent, and the rule antecedent $Cond^i$ is a conjunction of features (attribute-value pairs) selected from the features describing the training instances.

The concept of subgroup discovery was initially formulated by Klösgen in his rule learning algorithm EXPLORA [4] and by Wrobel in the algorithm MIDOS [5]. In the specialized bibliography, different methods have been developed which obtain descriptions of subgroups represented in different ways and using different quality measures, as SD [10], CN2-SD [11] or APRIORI-SD [12] among others.

One of the most important aspects of any subgroup discovery algorithm is the quality measures to be used, both to select the rules and to evaluate the results of the process. We can distinguish between objective and subjective quality measures. Some of the most used objective quality measures for the descriptive induction process are:

- *Coverage for a rule* [11]: measures the percentage of examples covered on average by one rule R^i of the induced rule set.

$$Cov(R^i) = Cov(Cond^i \to Class_j) = p(Cond^i) = \frac{n(Cond^i)}{n_s} \tag{1}$$

  where $n(Cond^i)$ is the number of examples which verifies the condition $Cond^i$ described in the antecedent (independently of the class to which belongs), and n_s is the number of examples.

- *Support for a rule*: considers the number of examples satisfying both the antecedent and the consequent parts of the rule. Lavrac et al. compute in [11] the support as:

$$Sup(R^i) = Sup(Cond^i \to Class_j) = p(Class_j.Cond^i) = \frac{n(Class_j.Cond^i)}{n_s} \tag{2}$$

  where $n(Class_j.Cond^i)$ is the number of examples which satisfy the conditions for the antecedent ($Cond^i$) and simultaneously belong to the value for the target variable ($Class_j$) indicated in the consequent part of the rule.

- *Significance for a rule* [4]): indicates how significant is a finding, if measured by the likelihood ratio of a rule.

$$Sig(R^i) = Sig(Cond^i \to Class_j) = 2 \cdot \sum_{j=1}^{n_c} n(Class_j.Cond^i) \cdot \log \frac{n(Class_j.Cond^i)}{n(Class_j) \cdot p(Cond^i)} \tag{3}$$

  where n_c is the number of values for the target variable and $p(Cond^i)$, computed as $n(Cond^i)/n_s$, is used as a normalized factor.

- *Unusualness for a rule*: is defined as the *weighted relative accuracy* of a rule [13].

$$WRAcc(Cond^i \to Class_j) = \frac{n(Cond^i)}{n_s} \cdot \left(\frac{n(Class_j.Cond^i)}{n(Cond^i)} - \frac{n(Class_j)}{n_s} \right) \tag{4}$$

  The WRAcc of a rule can be described as the balance between the coverage of the rule ($p(Cond^i)$) and its accuracy gain ($p(Class_j.Cond^i)$ - $p(Class_j)$).

2.2 Disjunctive Normal Form Fuzzy Rules

In the proposal presented in this paper, we use fuzzy rules in disjunctive normal form (DNF fuzzy rules) as description language to specify the subgroups, which permit a disjunction for the values of any variable present in the antecedent part.

We can describe a fuzzy rule R^i as:

$$R^i : Cond^i \rightarrow Class_j$$

where the antecedent describes the subgroup in disjunctive normal form, and the consequent is a value of the target variable.

So, the DNF fuzzy rule can be expressed as:

$$R^i : \text{If } X_1 \text{ is } LL_1^1 \text{ or } LL_1^3 \text{ and } X_7 \text{ is } LL_7^1 \text{ then } Class_j \tag{5}$$

where $LL_{n_v}^{k_{n_v}}$ is the linguistic label number k_{n_v} of the variable $\mathrm{n_v}$.

The fuzzy sets corresponding to the linguistic labels ($LL_v^1 \ldots LL_v^{k_v}$) are defined by means of the corresponding membership functions which can be defined by the user or defined by means of a uniform partition if the expert knowledge is not available. In this algorithm, we use uniform partitions with triangular membership functions, as it is shown in Fig. 1 for a variable *v* with 5 linguistic labels.

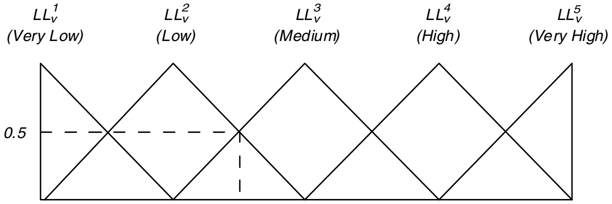

Fig. 1. Example of fuzzy partition for a continuous variable

It must be noted that any subset of the complete set of variables can take part in the rule antecedent, with any combination of linguistic labels related with the operator OR. In this way a subgroup is a compact and interpretable description of patterns of interest in data.

2.3 Multiobjective Genetic Algorithms

GAs are general purpose search algorithms which use principles inspired by natural genetics to evolve solutions to problems [14]. 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:

$$\min/\max \vec{y} = f(\vec{x}) = f_1(\vec{x}), f_2(\vec{x}), \ldots, f_n(\vec{x})) \tag{6}$$

where $\vec{x} = (x_1,x_2,...x_m)$ is the decision vector and $\vec{y} = (y_1,y_2,...,y_n)$ 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 [6], [7] as the algorithms MOGA [15], NSGA II [16] or SPEA2 [17] for instance.

The genetic representation of the solutions is the most determining aspect of the characteristics of any GA proposal. In this sense, the proposals in the specialized literature follow different approaches in order to encode rules within a population of individuals. In [18] a detailed description of these approaches is shown. Our proposal follows the "*Chromosome = Rule*" approach, in which each individual codifies a single rule, and a set of rules is codified by a subset of the complete population [19].

3 A Multiobjective Evolutionary Approach to Obtain Descriptive Fuzzy Rules

In this section we describe *MESDIF* (Multiobjective Evolutionary Subgroup DIscovery Fuzzy rules), 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. As the objective is to obtain a set of rules which describe subgroups for all the values of the target feature, the algorithm must be carried out so many times as different values has the target feature. This algorithm can generate fuzzy and/or crisp DNF rules, for problems with continuous and/or nominal variables.

The multiobjective GA is based on the SPEA2 approach [17], 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).

In order to preserve the diversity at a phenotypic level our 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. Therefore, in a run we obtain a set of rules that provide us knowledge on a property of interest.

Figure 2 shows the scheme 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 $P'_{t+1}$. As the size of $P'_{t+1}$ must be exactly the number of individuals to store (N), we may have to use a truncation or a filling function.
    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 $P'_{t+1}$.
```

Fig. 2. Scheme of the proposed algorithm

Once outlined the basis of the model, we will describe in detail some more important topics.

3.1 Chromosome Representation

In a subgroup discovery task, we have a number of descriptive features and a single target feature of interest. As we mentioned previously the multiobjective GA discovers a DNF fuzzy rule whose consequent is prefixed to one of the possible values of the target feature and each candidate solution is coded according to the "*Chromosome = Rule*" 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.

All the information relating to a rule is contained in a fixed-length chromosome with a binary representation in which, for each feature it is stored a bit for each of the possible values of the feature; in this way, if the corresponding bit contains the value 0 it indicates that the bit is not used in the rule, and if the value is 1 it indicates that the corresponding value is included. If a rule contains all the bits corresponding to a feature with the value 1, this indicates that this feature has no relevance for the information contributed in the rule (all the values or the feature verify the rule condition), and so this feature is ignored. This takes us to a binary representation model with so many genes by variable as possible values exist for the same one. The set of possible values for the categorical features is that indicated by the problem, and for continuous variables is the set of linguistic terms determined heuristically or with expert information.

3.2 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 three objectives, and the algorithm tries to maximize all the defined objectives.

- *Confidence*. Determines the relative frequency of examples satisfying the complete rule among those satisfying only the antecedent. In this paper we use an adaptation of Quinlan's accuracy expression in order to generate fuzzy classification rules [20]: the sum of the degree of membership of the examples of this class (the examples covered by this rule) to the zone determined by the antecedent, divided the sum of the degree of membership of all the examples that verifies the antecedent part of this rule (irrespective of their class) to the same zone:

$$Conf\,(\mathrm{R}^{\mathrm{i}}) = \frac{\sum\limits_{E^S \in E \,/\, E^S \in Class_j} APC(E^S, R^i)}{\sum\limits_{E^S \in E} APC(E^S, R^i)} \tag{7}$$

  where *APC* (Antecedent Part Compatibility) is the compatibility degree between an example and the antecedent part of a fuzzy rule, i.e., the degree of membership for the example to the fuzzy subspace delimited by the antecedent part of the rule.
- *Support*. This is the measure of the degree of coverage that the rule offers to examples of that class, 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:

$$Sup1\,(R^i) = \frac{n(Class_j.Cond^i)}{n(Class_j)} \tag{8}$$

- *Original support*. This objective is a measure 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 measure promotes the diversity at the population at a phenotypic level.

The last objective defined, the original support, is a restriction in the rules in order to obtain a set of rules, the pareto front, with a high degree of coverage, and is related with the cooperation between rules; the other objectives take into account the support and the confidence.

3.3 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 both the population and the elite population are used to compute what individual dominate what other.

- 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, 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.

3.4 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).

3.5 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.

4 A Case Study in Marketing: Knowledge Discovery in Trade Fairs

In the area of marketing, and specifically in the trade fairs planning, it is important to extract conclusions of the information on previous trade fairs to determine the relationship between the trade fair planning variables and the success of the stand. This problem over the extraction of useful information on trade fairs has been analyzed in the Department of Organization and Marketing of the University of Mondragón, Spain [21].

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 [22]. 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, amongst other factors [23]. Therefore convenient, is the automatic extraction of information about the relevant variables which permit the attainment of unknown data, which partly determines the efficiency of the stands of a trade fair.

A questionnaire was designed to reflect the variables that better allow explaining the trade fair success containing 104 variables (7 of them are continuous and the rest are categorical features, result of an expert discretization). Then, the stand's global efficiency is rated as *high*, *medium* or *low*, in terms of the level of achievement of objectives set for the trade fair. The data contained in this dataset were collected in the Machinery and Tools biennial held in Bilbao in March 2002 and contain information on 228 exhibitors.

For this real problem, the data mining algorithm should extract information of interest about each efficiency group. 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.

4.1 Results of the Experimentation on the Marketing Dataset

As our proposal is a non-deterministic approach, the experimentation is carried out with 5 runs for each class of the target variable: *low*, *medium* and *high* efficiency. The parameters used in this experimentation are:

- Population size: 100.
- Elite population size: 5.
- Maximum number of evaluations of individual in each GA run: 10000.
- Mutation probability: 0.01.
- Number of linguistic labels for the continuous variables: 3

We have experimentally verified that the approach have a better behaviour using an elite population size of 5 individuals.

Table 1 shows the best results obtained for all the classes of the target variable (*low*, *medium* and *high* efficiency). In this table, it is shown for each rule obtained, the number of variables involved *(# VAR*), the *Support* (*SUP1*) as defined in (8) and used in our proposal, the *Confidence* (*CONF*) of each rule as defined in (7), the *Coverage (COV)* as defined in (1), the *Support* (*SUP2*) as defined in (2), the *Significance* (*SIG)* as defined in (3) and the *Unusualness* (*WRACC*) of the rule as computed in (4).

Table 1. Results for *Low*, *Medium* and *High* efficiency

Efficiency	*# VAR.*	*SUP1*	*CONF*	*COV*	*SUP2*	*SIG*	*WRACC*
Low	8	0.079	0.820	0.026	0.013	5.026	0.007
	4	0.026	1.000	0.004	0.004	3.584	0.001
	5	0.395	0.724	0.127	0.066	25.684	0.042
	6	0.289	0.759	0.088	0.048	19.672	0.031
Medium	6	0.088	0.892	0.658	0.057	6.623	0.008
	1	0.959	0.657	0.947	0.623	0.605	0.004
	2	0.574	0.802	0.469	0.373	12.104	0.065
	2	0.845	0.676	0.811	0.548	3.447	0.017
	4	0.182	0.750	0.158	0.118	2.441	0.011
High	5	0.095	0.595	0.031	0.017	6.565	0.010
	3	0.024	1.000	0.004	0.004	3.383	0.001
	4	0.047	0.722	0.013	0.009	3.812	0.004

It must be noted that high values in support (*SUP1*, expression (8)) means that the rule covers most of the examples of the class, and high values in confidence (*CONF*, expression (7)) means that the rule has few negative examples.

The rules generated have adequate values of confidence (*CONF*, expression (7)) and support (*SUP1*, expression (8)). The algorithm induces set of rules with a high confidence (higher than the minimum confidence value). The rule support, except for some rules, is low. The market problem used in this work is a difficult real problem in which inductive algorithms tend to obtain small disjuncts (specific rules which represent a small number of examples). However, the small disjunct problem is not a determining factor in the induction process for subgroup discovery because partial relations, i.e., subgroups with interesting characteristics, with a significant deviation from the rest of the dataset, are sufficient. The results show that *Low* and *High* efficiency classes are the more interesting for the subgroup discovery task, but also the more difficult.

The knowledge discovered for each one of the target variable values is understandable by the user due to the use of DNF fuzzy rules, and the low number of rules and conditions in the rule antecedents (below 10% of the 104 variables). Moreover, the rules obtained with the *MESDIF* algorithm are very simple.

Tables 2, 3 and 4 show the extracted rules for the three levels of efficiency (*low*, *medium* and *high*).

Marketing experts from Department of Organization and Marketing of the University of Mondragón (Spain) analysed the results obtained and indicated that:

- The exhibitors who obtained worse results were those with a medium or high size of the stand, not using indicator flags in it and with a low or medium valuation of the assembly and disassemble services.
- The companies which obtain medium efficiency are those with none or high satisfaction with the relation maintained with the clients, and medium, high or very high global satisfaction.
- Finally, the exhibitors who obtained better results (high efficiency) are big or huge companies using telemarketing with the quality contacts.

Table 2. Rules for *Low* efficiency

#Rule	*Rule*
1	IF (Publicity utility = None OR Medium OR High) AND (Number of annual fairs = 2-5 OR 6-10 OR 11-15 OR >15) AND (Use of consultants = NO) AND (Importance improvement image of the company = None OR Low OR Medium) AND (Addressees if only clients = NO) AND (Stand size = Medium OR High) AND (Valuation assembly/disassembly = Low OR Medium) AND (Indicator flags = NO) THEN Efficiency = Low

2	IF (Stand size = Medium OR High) AND (Telemarketing = ALL OR Only quality) AND (Gifts = NO) AND (Indicator flags = NO)
	THEN Efficiency = Low

3	IF (Use of consultants = NO) AND (Importance improvement image of the company = None OR Low OR Medium) AND (Stand size = Medium OR High) AND (Valuation assembly/disassembly = Low OR Medium) AND (Indicator flags = NO) THEN Efficiency = Low
4	IF (Publicity utility = None OR Low OR High) AND (Importance improvement image of the company = None OR Low OR Medium) AND (Addressees if only clients = NO) AND Stand size = Medium OR High) AND (Valuation assembly/disassembly = Low OR Medium) AND (Indicator flags = NO) THEN Efficiency = Low

Table 3. Rules for *Medium* efficiency

#Rule	*Rule*

1	IF (Satisfaction relation clients = None OR High) AND (Importance public relations = Very high) AND (Global satisfaction = Medium OR High OR Very high) AND (Quality visitors valuation = Low OR High) AND (Gifts = NO) AND (Inserts = NO)
	THEN Efficiency = Medium

2	IF (Previous promotion = YES)
	THEN Efficiency = Medium

3	IF (Satisfaction relation clients = None OR High) AND (Global satisfaction = Medium OR High OR Very high)
	THEN Efficiency = Medium

4	IF (Global satisfaction = Medium OR High OR Very high) AND (Inserts = NO)
	THEN Efficiency = Medium

5	IF (Satisfaction relation clients = None OR High) AND (Previous promotion = YES) AND (Company advertising mention = YES) AND (Inserts = NO)
	THEN Efficiency = Medium

Table 4. Rules for *High* efficiency

#Rule	*Rule*
1	IF (Importance new contacts = Low OR Medium OR Very High) AND (Visitor information valuation = Medium OR High) AND (Gratefulness letter = All OR Only quality) AND (Telemarketing = None OR Only quality) AND (Little gifts before fair = YES) THEN Efficiency = High

2	IF (Employees = 251-500 OR >500) AND (Follow-up modality = Only quality) AND (Telemarketing = NO OR Only quality)
	THEN Efficiency = High

3	IF (Employees =251-500 OR >500) AND (Visitor information valuation = Medium OR High) AND (Gratefulness letter = All OR Only quality) AND (Telemarketing = NO OR Only quality)
	THEN Efficiency = High

5 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.

The use of a subgroup discovery algorithm for this problem is well suited because in subgroup discovery task the objective is not to generate a set of rules which cover all the dataset examples, but individual rules that, given a property of interest of the data, describe in an interpretable way the more interesting subgroups for the user.

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, with an appropriate balance between the quality measures specified in the algorithm that are easily interpretable, and with a high level of confidence and support.

DNF fuzzy rules contribute a more flexible structure to the rules, allowing each variable to take more than one value, and facilitating the extraction of more general rules. In this kind of fuzzy rules, fuzzy logic contributes to the interpretability of the extracted rules due to the use of a knowledge representation nearest to the expert, also allowing the use of continuous features without a previous discretization.

As future work, we will study the inclusion in the *MESDIF* algorithm of different quality measures (and combinations of them) as objective functions in order to obtain fuzzy subgroup discovery rules with better properties.

Acknowledgment

This work was supported by the Spanish Ministry of Science and Technology and by the European Fund. FEDER under Projects TIC-2005-08386-C05-01 and TIC-2005-08386-C05-03, and the nets TIN2004-20061-E and TIN2004-21343-E.

References

1. Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P., From Data Mining to Knowledge Discovery: An Overview, in Advances in Knowledge Discovery and Data Mining, U. Fayyad, et al., Editors, AAAI Press (1996) 1–30
2. Michie, D., Spiegelhalter, D.J., and Taylor, C.C.: Machine learning, neural and estatistical classification. Ellis Horwood, (1994)
3. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, I., Fast Discovery of Association Rules, in Advances in Knowledge Discovery and Data Mining, U. Fayyad, et al., Editors, AAAI Press: Menlo Park, Calif. (1996) 307–328
4. Klösgen, W., Explora: A Multipattern and Multistrategy Discovery Assistant, in Advances in Knowledge Discovery and Data Mining, U. Fayyad, et al., Editors, AAAI Press: Menlo Park, Calif. (1996) 249–271
5. Wrobel, S., An algorithm for multi-relational discovery of subgroups, in Principles Of Data Mining And Knowledge Discovery (1997) 78-87
6. Deb, K.: Multi-Objective Optimization using Evolutionary Algorithms. John Wiley & Sons, (2001)

7. Coello, C.A., Van Veldhuizen, D.A., and Lamont, G.B.: Evolutionary Algorithms for Solving Multi-Objective Problems. Kluwer Academic Publishers, (2002)
8. Ghosh, A. and Nath, B.: Multi-objective rule mining using genetic algorithms. Information Sciences. 163 (2004) 123-133
9. Ishibuchi, H. and Yamamoto, T.: Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining. Fuzzy Sets and Systems. 141 (2004) 59-88
10. Gamberger, D. and Lavrac, N.: Expert-guided subgroup discovery: Methodology and application. Journal Of Artificial Intelligence Research. 17 (2002) 1-27
11. Lavrac, N., Kavsec, B., Flach, P., and Todorovski, L.: Subgroup discovery with CN2-SD. Journal of Machine Learning Research. 5 (2004) 153-188
12. Kavsek, B., Lavrac, N., and Jovanoski, V., APRIORI-SD: Adapting association rule learning to subgroup discovery, in Advances In Intelligent Data Analysis V (2003) 230-241
13. Lavrac, N., Flach, P., and Zupan, B., Rule evaluation measures: A unifying view, in Inductive Logic Programming (1999) 174-185
14. Goldberg, D.E.: Genetic algorithms in search, optimization and machine learning. Addison-Wesley, (1989)
15. Fonseca, C.M. and Fleming, P.J. Genetic algorithms for multiobjective optimization: formulation, discussion and generalization. in Fifth International Conference on Genetic Algorithms (ICGA). 1993. San Mateo, CA
16. Deb, K., Pratap, A., Agarwal, A., and Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation. 6 (2002) 182-197
17. Zitzler, E., Laumanns, M., and Thiele, L., SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimisation, in Evolutionary methods for design, optimisation and control, K. Giannakoglou, et al., Editors, CIMNE (2002) 95-100
18. Cordón, O., Herrera, F., Hoffmann, F., and Magdalena, L.: Genetic fuzzy systems: evolutionary tuning and learning of fuzzy knowledge bases. World Scientific, (2001)
19. Wong, M.L. and Leung, K.S.: Data Mining using Grammar Based Genetic Programming and Applications. Kluwer Academics Publishers, (2000)
20. Cordón, O., del Jesus, M.J., and Herrera, F.: Genetic Learning of Fuzzy Rule-based Classification Systems Co-operating with Fuzzy Reasoning Methods. International Journal of Intelligent Systems. 13 (1998) 1025-1053
21. Mesonero, M., Hacia un modelo efectivo de planificación ferial basado en algoritmos genéticos, in Departamento de Organización y Marketing, Universidad de Mondragón: Mondragón (2004)
22. Gopalakrishna, S., Lilien, G.L., Williams, J.D., and Sequeira, I.K.: Do trade shows pay off. Journal of Marketing. 59 (1995) 75-83
23. Millar, S.: How to get the most of the trade shows. NTC Publishing Group, (2003)