Subgroup Discovery with Linguistic Rules

María José del Jesus1, Pedro González1, and Francisco Herrera2

[1] Department of Computer Science
University of Jaén, 23071 – Jaén, Spain
mjjesus@ujaen.es, pglez@ujaen.es
[2] Department of Computer Science and Artificial Intelligence
University of Granada, 18071 – Granada, Spain
herrera@decsai.ugr.es

Summary. Subgroup discovery can be defined as a form of supervised inductive learning in which, given a population of individuals and a specific property of individuals in which we are interested, find population subgroups that have the most unusual distributional characteristics with respect to the property of interest. Subgroup discovery algorithms aim at discovering individual rules, which must be represented in explicit symbolic form and which must be simple and understandable in order to be recognized as actionable by potential users.

A fuzzy approach for a subgroup discovery process, which considers linguistic variables with linguistic terms in descriptive fuzzy rules, lets us obtain knowledge in a similar way of the human thought process. Linguistic rules are naturally inclined towards coping with linguistic knowledge and to produce more interpretable and actionable solutions. This chapter analyzes the use of linguistic rules for modelling this problem, and shows a genetic extraction model for learning this kind of rules.

1 Introduction

Rule learning is an important form of *predictive* machine learning, aimed at inducing a set of rules to be used for classification and/or prediction [6], [31]. Developments in *descriptive induction* have recently also attracted much attention from researchers interested in rule learning. The objective of *descriptive machine learning* is to discover individual rules that define interesting patterns in data, and it includes approaches for mining association rules [2], for subgroup discovery [24], [35] and other non-classificatory induction approaches such as clausal discovery [34] or database dependency [15] among others.

Subgroup discovery is a form of descriptive supervised inductive learning. It aims to discover individual rules (or local patterns of interest, very frequent –hence typical– or very rare –hence atypical–) in relation to a specific property of interest, which must be represented in explicit symbolic form and which must be relatively simple in order to be recognized as actionable by potential users. There-

fore, the subgroups discovered in data are of a more explanatory nature and the interpretability of the extracted knowledge for the final user is a crucial aspect in this field.

As it was claimed by Dubois et al. in [14], the use of fuzzy sets to describe associations between data extends the types of relationships that may be represented, facilitates the interpretation of rules in linguistic terms, and avoids unnatural boundaries in the partitioning of the attribute domains. This is especially useful in medical, control or economic fields where the boundaries of a piece of information used may not be clearly defined. In fact, the use of linguistic variables and linguistic terms in a machine learning process has been thoroughly explored by various authors in predictive induction (see for instance Ishibuchi et al.'s book [22] for a complete and understandable up-to-date description of the design of classification and modelling fuzzy systems). There are some proposals using fuzzy logic in descriptive induction, for the extraction of fuzzy association rules [10], [20], and for subgroup discovery fuzzy rules [12], [13].

A fuzzy approach for a subgroup discovery process, which considers linguistic variables with linguistic terms in descriptive fuzzy rules, allows us to obtain knowledge in a similar way to the human thought process. In order to understand this it is enough to remember that much of the logic behind human reasoning is not traditional two-valued or even multivalued logic, but logic with fuzzy truths, fuzzy connectives and fuzzy rules of inference. Fuzzy rules are naturally inclined towards coping with linguistic knowledge, thereby producing more interpretable and actionable solutions in the field of subgroup discovery and in general in the analysis of data to establish relationships and identify patterns [21].

This chapter analyzes the use of linguistic rules in subgroup discovery. A genetic model for the extraction of fuzzy rules in subgroup discovery [12], [13] is described, analyzing its possibilities and limitations. To do so, the chapter is arranged in the following way: In Section 2, the subgroup discovery task is introduced. In Section 3 is described the use of linguistic rules in the subgroup discovery task. An evolutionary approach to obtain subgroup discovery descriptive fuzzy rules is explained in Section 4. Finally, in Section 5 the conclusions and further research are outlined.

2 Introduction to Subgroup Discovery

Subgroup discovery is a form of supervised inductive learning which is defined as follows [24], [35]: given a set of data and a property of interest to the user (target variable), an attempt is made to locate subgroups which are statistically "most interesting" for the user, e.g., are as large as possible and have the most unusual distributional characteristics with respect to the property of interest. The concept was initially formulated by Klösgen in EXPLORA [24] and by Wrobel in MIDOS [35].

Descriptive machine learning methods for subgroup discovery have the objective of discovering interesting properties of subgroups by obtaining *simple* rules (i.e.

with an understandable structure and with few variables), which are *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,

$$\text{Cond} \rightarrow \text{Class}$$

where the property of interest for subgroup discovery is class value *Class* that appears in the rule consequent, and the rule antecedent *Cond* is a conjunction of features (attribute-value pairs) selected from the features describing the training instances.

The subgroup discovery task relies on the following main properties:

- The description language specifying the subgroups which must be adequate to be applied effectively by the potential users. The subgroup description consists of a set of expressions. In the simplest case, each expression is one-valued; however negation or internal disjunctions are also possible.
- The quality function measuring the interest of the subgroup. A variety of quality functions have been proposed (see for instance [24], [25], [18]). The quality functions used is determined by the type of the target variable, the type of rules and the problem considered. In subsection 2.2 several quality measures used in subgroup discovery algorithms are described.
- The search strategy employed by the algorithm is very important, since the dimension of the search space has an exponential relation with respect to the number of features (or variables) and values considered.

Below related works and the quality measures used in subgroup discovery are shortly revised.

2.1 Related works in subgroup discovery

In the specialized bibliography, different methods have been developed which obtain descriptions of subgroups represented in different ways and using different quality measures:

- The first approach developed for subgroup discovery was EXPLORA [24]. It uses decision trees for the extraction of rules. The rules are specified by defining a descriptive schema and implementing a statistical verification method. The interest of the rules is measured using measures such as evidence, generality, redundancy and simplicity.
- MIDOS [35] applies the EXPLORA approach to multirelational databases. It uses optimistic estimation and minimum support pruning. The goal is to discover subgroups of the target relation which have unusual statistical distributions with respect to the complete population. The quality measure is a combination of unusualness and size.
- SubgroupMiner [26] is an extension of EXPLORA and MIDOS. It is an advanced subgroup discovery system which uses decision rules and interactive search in the space of the solutions, allowing the use of large databases, multirelational hypotheses, and the discovery of structures of causal subgroups.

This algorithm uses as quality function the classical binomial test to verify if the statistical distribution of the target is significantly different in the extracted subgroup.

- SD [17] is a rule induction system guided by expert knowledge: instead of defining an optimal measure to search and select automatically the subgroups, the objective is to help the expert in performing flexible and effective searches on a wide range of optimal solutions.
- CN2-SD [29] (a modified version of the CN2 algorithm [6]) induces subgroups in the form of rules using the relation between true positives and false positives as a quality measure. It uses a modified weighted relative accuracy as quality measure for the rule selection.
- RSD [30], *Relational Subgroup Discovery*, has the objective of obtaining population subgroups which are as large as possible, with a statistical distribution as unusual as possible with respect to the property of interest, and which are different enough to cover most of the target population. It is a recent upgrade of the CN2-SD algorithm which enables relational subgroup discovery.
- APRIORI-SD [23] is developed by adapting the association rule learning algorithm APRIORI [1] to subgroup discovery, including a new quality measure for the induced rules (the weighted relative accuracy) and using probabilistic classification of the examples. For the evaluation is used the support and significance of each individual rule, and the size, accuracy and area under the ROC curve of the set of rules.
- *Intensive Knowledge* [3] is a subgroup discovery approach which uses several types of application background knowledge to improve the quality of the results of the subgroup discovery task and the efficiency of the search method.
- SDIGA [13] is an evolutionary fuzzy rule induction system which uses as quality measures for the subgroup discovery task adaptations of the measures used in the association rules induction algorithms. Unlike all the other proposals, SDIGA uses linguistic rules as description language to specify the subgroups. This proposal is shown in section 4.

2.2 Quality measures in subgroup discovery

One of the most important aspects of any subgroup discovery algorithm -and a determining factor in the quality of the approach- is the quality measure to be used, both to select the rules and to evaluate the results of the process. *Objective* measures for descriptive induction evaluate each subgroup individually, but can be complemented by their variants to compute the mean of the induced set of descriptions of subgroups, allowing comparison between different subgroup discovery algorithms.

There are different studies about objective quality measures for the descriptive induction process [25], [32], [17] but it is difficult to reach an agreement about their use. Below, the more widely used objective quality measures in the specialized bibliography of subgroup discovery are described.

- *Coverage for a rule* [29]: this measures the percentage of examples covered on average by one rule of the induced set of rules.

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

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.

The *average coverage for the set of rules* finally obtained is calculated by the following expression:

$$COV = \frac{1}{n_r}\sum_{i=1}^{n_r} Cov(R^i) \tag{2}$$

where n_r is the number of induced rules.

- *Support for a rule*: In descriptive induction processes the support for a rule is a standard measure which considers, by means of an expression that can vary in different proposals, the number of examples satisfying both the antecedent and the consequent parts of the rule. Lavrac et al. compute in [29] the overall support as the percentage of target examples (positive examples) covered by the rules. The support of a rule is so defined as the frequency of correctly classified examples covered.

$$Sup_1(Cond^i \rightarrow Class_j) = p(Class_j.Cond^i) = \frac{n(Class_j.Cond^i)}{n_s} \tag{3}$$

where $n(Class_j.Cond^i)$ is the number of examples which satisfy the conditions for the antecedent ($Cond^i$) and also belong to the value for the target variable ($Class_j$) indicated in the consequent part of the rule. In (3), the support of a rule is computed dividing by the total number of examples. It can also be computed in other ways, such as dividing by the number of examples of the class or other variations.

The *support* for a set of rules is computed by:

$$SUP = \frac{1}{n_s}\sum_{j=1}^{n_c} n(Class_j \cdot \bigvee_{Cond^i \rightarrow Class_j} Cond^i) \tag{4}$$

where n_c is the number of values for the target variable considered. It must be noted that in this expression the examples which belong to many rules are considered only once.

- *Size (for a set of rules)*: The size of a set of rules is a complexity measure calculated as the number of induced rules (n_r). Complexity can also be measured as the mean number of obtained rules per class, or the mean of variables per rule.
- *Significance for a rule* [24]: indicates the significance of a finding, if measured by the likelihood ratio of a rule.

$$Sig(Cond^i \rightarrow Class_j) = 2 \cdot \sum_{j=1}^{n_c} n(Class_j.Cond^i) \cdot \log \frac{n(Class_j.Cond^i)}{n(Class_j) \cdot p(Cond^i)} \tag{5}$$

where $p(Cond_i)$, computed as $n(Cond_i)/n_s$, is used as a normalized factor.

It must be noted that, although each rule is for a specific class value, the significance measures impartially the novelty in the distribution, for all the class values.

The *significance for a set of rules* is computed as follows:

$$SIG = \frac{1}{n_r} \sum_{i=1}^{n_r} Sig(R^i) \tag{6}$$

- *Unusualness for a rule*: It is defined as the *weighted relative accuracy* of a rule [28].

$$WRAcc(Cond^i \rightarrow Class_j) = \frac{n(Cond^i)}{n_s} \cdot \left(\frac{n(Class_j.Cond^i)}{n(Cond^i)} - \frac{n(Class_j)}{n_s} \right) \tag{7}$$

The weighted relative accuracy of a rule can be described as the balance between the coverage of the rule ($p(Cond^i)$) and its accuracy gain ($p(Class_j.Cond^i) - p(Class_j)$). It must be noted that the higher a rule's unusualness, the more relevant is it.

The unusualness for a set of rules is computed as follows:

$$WRACC = \frac{1}{n_r} \sum_{i=1}^{n_r} WRAcc(R^i) \tag{8}$$

It must be noted that all the measures here described are crisp because in the majority of the proposals the rules used to represent the knowledge in subgroup discovery are not fuzzy.

3 Linguistic rules in subgroup discovery

As it has been described in the previous section many approaches have already been proposed for subgroup discovery task, usually based on non linguistic rules. Since human information processing is mainly based on linguistic information, in order to facilitate the human interpretability of the results, the use of linguistic rules must be considered.

In this section, the use of linguistic rules in subgroup discovery will be analyzed, and a kind of linguistic rules, DNF linguistic rules, and some quality measures for them are described.

3.1 The use of linguistic rules in subgroup discovery

In any Data Mining problem two main objectives are present:

- to obtain knowledge about patterns in data which must be fitted to the nature and reality of the problem, e.g., knowledge must be as precise as possible,
- to extract knowledge which must be simple, compact and understandable by the final user. That is to say, the obtained knowledge must be close to the form in which the expert represents his knowledge on the problem in order to be actionable by him.

The second objective becomes the most important in descriptive data mining, and specifically in the subgroup discovery task.

The way in which the knowledge is represented by a human expert is inherently qualitative and vague. In this sense the use of Fuzzy Logic in Data Mining allows us to model inaccurate and qualitative knowledge, as well as to handle uncertainty and deal naturally to a reasonable extent with human reasoning. Ever since it was proposed in 1965 by Zadeh [36], it has been applied to many areas of research, fundamentally because of its proximity to human reasoning and because it provides an effective way of capturing the approximate and inexact nature of the real world.

In rule induction processes, Fuzzy Logic is included in such a way that the models extracted are fuzzy rules. In the most interpretable type of fuzzy rules, linguistic fuzzy rules, and therefore the most appropriate for Data Mining, the continuous variables are defined as linguistic variables; that is, variables which take as possible values linguistic labels, the semantics of which are represented by an associated fuzzy set [37].

One of the fundamental aspects when working with linguistic rules is the definition of membership functions associated with the fuzzy sets used. There are several alternatives to determine this aspect:

- When the expert knowledge is not available, uniform partitions with triangular membership functions can be used, as it is shown in Fig. 1 for a variable with 5 linguistic labels.

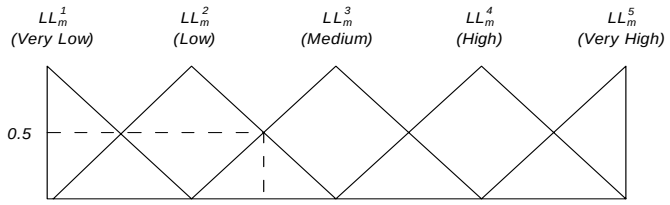

Fig. 1. Example of fuzzy partition for a continuous variable

- When expert knowledge about the problem is available or an analysis of the data can be realized, the definition for the fuzzy partition can be done in one of the following ways:

- In order to increase the interpretability of the results obtained in some proposals such as [4], [27] for the extraction of fuzzy rules the expert gives the algorithm the continuous variables and their corresponding membership functions. The quality of the results obtained depends on the suitability of the fuzzy sets.
- For many applications it is very difficult to know from the outset which fuzzy sets will be the most appropriate and so different algorithms which learn the fuzzy partitions have been proposed. In [16] the fuzzy sets and the membership functions are generated through clustering techniques. In [19] the definition for the linguistic labels is established by means a genetic algorithm.
- A fuzzy partition can be defined by a heuristic approach which places the fuzzy sets in such a way that each of them will cover approximately the same number of data, if the expert wants to. But it must be considered that, depending on the problem, the interpretation of the resulting fuzzy rules could be decreased.
- Moreover, if it is necessary, a preliminary data analysis which detects outliers in data can be done before the determination of the fuzzy partitions. This way a specific analysis of them can be realized and the fuzzy partition (without these outliers' data) is not biased by them.

3.2 DNF linguistic rules

The objective in subgroup discovery is to extract knowledge about a variable of interest for the user, in an easily interpretable way. In order to increase the interpretability of the extracted knowledge, the Disjunctive normal form (DNF) fuzzy rules can be used. A DNF fuzzy rule represents the knowledge in a flexible and compact way, allowing each variable to take more than one value, and facilitating the extraction of more general rules. Linguistic rules allow us to establish flexible limits between the different levels of meaning without ignoring or overemphasizing the elements closest to the edges in the same way as human perception does. In addition, linguistic DNF fuzzy rules allow us to make changes in the initial granularity in each rule in a descriptive way. The following is an example of linguistic DNF fuzzy rule:

IF Number of times pregnant *High* or *Medium* AND Body mass index is *Low*
THEN Diabetes is *Tested negative*

Below, the notation used to describe the DNF fuzzy rules is formally described. We consider a problem with:

- a set of features, discrete or continuous
$$\{X_m \ / \ m = 1,..,n_v\}$$
used to describe the subgroups, where n_v is the number of features;
- a set of values for the target variable

$$\left\{Class_j / j = 1,\ldots,n_c\right\}$$

where n_c is the number of values for the target variable considered;

- a set of examples

$$\left\{E^k = (e_1^k, e_2^k, \ldots, e_{n_v}^k, class^j) / k = 1,\ldots,n_s\right\}$$

where $class^j$ is the target variable value for the sample E^k (i.e., the class for this example) and n_s is the number of examples for the descriptive induction process;

- a set of linguistic labels for the continuous variables. The number of linguistic labels and the definition for the corresponding fuzzy sets depend on each variable

$$X_m : \left\{LL_m^1, LL_m^2, \ldots, LL_m^{l_m}\right\}.$$

In this expression, variable X_m, has l_m different linguistic labels to describe its domain in an understandable way.

A fuzzy rule R^i can be described as:

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

where the antecedent describes the subgroup.

Below is an example of a DNF fuzzy rule:

$$R^1 : \text{If } (X_1 \text{ is } LL_1^1 \text{ or } LL_1^3) \text{ and } (X_7 \text{ is } LL_7^1) \text{ then } Class_j \tag{9}$$

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

For these rules, we consider that

- an example E^k verifies the antecedent part of a rule R^i if

$$APC(E^k, R^i) = T(TC(\mu_{LL_1^1}(e_1^k),\ldots,\mu_{LL_1^{l_1}}(e_1^k)),\ldots,TC(\mu_{LL_{n_v}^1}(e_{n_v}^k),\ldots,\mu_{LL_{n_v}^{l_{n_v}}}(e_{n_v}^k))) > 0 \tag{10}$$

where:

- APC (Antecedent Part Compatibility) is the degree of compatibility 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,
- $LL_1^{l_1}$ is the linguistic label number l_1 of the variable *1*,
- $\mu_{LL_1^{l_1}}(e_1^k)$ is the degree of membership for the value of the feature *1* for the example E^k to the fuzzy set corresponding to the linguistic label l_1 for this feature,
- T is the t-norm selected to represent the meaning of the AND operator –the fuzzy intersection–, in our case the minimum t-norm, and

- TC is the t-conorm selected to represent the meaning of the OR operator – the fuzzy union–, in our case the maximum t-conorm.

- an example E^k is covered by a rule R^i if

$$APC(E^k, R^i) > 0 \quad AND \quad E^k \in Class_j \tag{11}$$

This means that an example is covered by a rule if the example has a degree of membership higher than 0 to the fuzzy subspace delimited by the antecedent part of the fuzzy rule, and the value indicated in the consequent part of the rule agrees with the value of the target feature for the example. For the categorical variables, the degrees of membership are just 0 or 1.

3.3 Quality measures for DNF linguistic rules

When using linguistic rules, it is necessary to define quality measures to manage this type of rules. Some of the quality measures used in the bibliography for the induction of fuzzy rules are next detailed:

- *Confidence of a fuzzy rule* [13]: The confidence of a rule determines the relative frequency of examples satisfying the complete rule among those satisfying only the antecedent. In our proposal the expression used for confidence reflects the degree to which the examples within the zone of the space marked by the antecedent verify the information indicated in the consequent part of the rule. To calculate this factor an adaptation of Quinlan's accuracy expression [33] is used in order to generate fuzzy classification rules [8]: the sum of the degree of membership of the examples of this class and the fuzzy input subspace determined by the antecedent, divided by 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:

$$Conf(R^i) = \frac{\sum_{E^k \in E / E^k \in Class_j} APC(E^k, R^i)}{\sum_{E^k \in E} APC(E^k, R^i)} \tag{12}$$

- *Support of a fuzzy rule*, defined in [13] as the degree of coverage that the rule offers to examples of that class:

$$Sup_2(R^i) = \frac{n(Class_j.Cond^i)}{n(Class_j)} \tag{13}$$

where $n(Class_j)$ is the number of examples of the class *j*. A variation of this measure will be detailed in next section.

4 A genetic algorithm for the induction of linguistic rules in subgroup discovery

In this section an evolutionary model for the extraction of linguistic rules for the subgroup discovery task, SDIGA (Subgroup Discovery Iterative Genetic Algorithm), which uses DNF rules is described [13]. The model follows the IRL approach –later explained– and works as follows:

- The core of the model is a genetic algorithm (GA) which uses a post-processing step based on a simple local search, a hill-climbing procedure. The hybrid GA extracts one simple and interpretable fuzzy rule with an adequate level of support and confidence. The post-processing step consists of a local search process increasing the generality of the rule.
- This hybrid GA is included in an iterative process for the extraction of a set of fuzzy rules for the description of subgroups supported by different areas (not necessarily disjuncts) of the instance space. In this way is obtained a set of *different* solutions generated in successive runs of the GA corresponding to the same value of the target feature. The method to guide the GA evolution over different –although may be overlapped– fuzzy rules is explained in detail in the next subsection.

The objective is to obtain a set of rules which describe subgroups for all the values of the target feature, and so the iterative process must be carried out as many times as different values the target feature has.

Once the basis of the proposal is outlined, the GA and the iterative rule extraction model are described in detail. The results of a comparison of the proposal with other subgroup discovery algorithms are also detailed.

4.1 Hybrid genetic algorithm for the induction of a fuzzy rule

The hybrid GA extracts a single fuzzy rule in an attempt to optimize the confidence and support. In the following subsections the elements of the hybrid GA are described.

4.1.1 Chromosome representation

The genetic representation of the solutions is the most determining aspect of the characteristics of any genetic learning proposal. The "*Chromosome = Rule*" approach (in which each individual codifies a single rule) is more suited in subgroup discovery because the objective is to find a reduced set of rules in which the quality of each rule is evaluated independently of the rest. This is the encoding approach used in the evolutionary proposal next described.

The GA discovers a single fuzzy rule whose consequent is prefixed to one of the possible values of the target feature. Only the antecedent is represented in the

chromosome and all the individuals in the population are associated with the same value of the target feature.

All the information relating to a rule is contained in a fixed-length chromosome with a binary representation in which, for each feature a bit for each one of the possible values of the feature is stored; in this way, if the corresponding bit contains the value 0 it indicates that the value is not used in the rule, and if the bit contains the value 1 it indicates that the corresponding value is included. If a rule contains all the bits corresponding to a feature with the value 1, or all of them contain the value 0, the feature is ignored and does not take part in the rule. In Fig. 2, V_0 and V_1 have 3 possible values, and V_2 and V_k have 2 possible values. In this example, neither V_2 nor V_k take part in the rule (V_2 does not take any of its values, and V_k takes all, and so both variables are irrelevant for the rule).

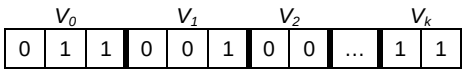

Fig. 2. Encoding model of a DNF rule

4.1.2 Fitness function

The objective of the rule discovery process is to obtain rules with high confidence, and which are understandable and general. It means that the problem has at least two objectives to maximize: the support and the confidence of the rule. To achieve this, the weighted sum method that weights a set of objectives into a single objective is the simplest approach, and lets us introduce the expert criteria related to the importance of the objectives for a specific problem in the rule generation process. So, this proposal uses a weighted lineal combination in the following way:

$$fitness\ (c) = \frac{\omega_1 \times Sup_3(c) + \omega_2 \times Conf(c)}{\omega_1 + \omega_2} \tag{14}$$

where confidence (*Conf*) and support (Sup_3) of the rule are defined as:

- *Confidence*. This determines the accuracy of the rule, in that it reflects the degree to which the examples within the zone of the space determined by the antecedent verify the information specified in the consequent of the rule, and it is computed as in (12).
- *Support*. This measures the degree of coverage that the rule offers to examples belonging to the class specified in the rule consequent. It is calculated in a different way than in (5) to promote different fuzzy rules being obtained in different runs of the hybrid GA. To do so, for the computation of the support it is only considered the examples not marked (i.e. the examples not covered by other fuzzy rules previously obtained by means of the past runs of the hybrid GA). Thus, the support is defined as the quotient between the examples of this partial set covered by the rule represented in the chromosome and the total number of examples of this partial set:

$$Sup_3(R^i) = \frac{Ne^+(R^i)}{Ne_{NC}} \tag{15}$$

where Ne_{NC} is the number of examples of the class specified in the consequent left uncovered by the previous rules, and $Ne^+(R^i)$ is the number of examples covered by the rule which are left uncovered by the previous rules, using (11) to determine when an example is covered by a rule.

This way of measuring support is sensible, when using the GA within an iterative process, in order to obtain different rules each time the GA is run. From the second iteration, rules which cover examples belonging to zones delimited by previously obtained rules are penalized, because the support factor only considers examples which have not been described by rules already obtained. No distance function is used as differences are penalized on a phenotypical level. This penalization does not eliminate the examples covered by previously obtained fuzzy and they take part in the computation of the confidence measure.

The overall objective of the evaluation function is to direct the search towards rules which maximize accuracy, minimizing the number of negative and examples not-covered. Whereas covered examples are used in the calculation of the confidence, they are not used in the calculation of the support, to prevent the obtaining of rules inconsistent with the examples previously penalized.

4.1.3 Reproduction model and genetic operators

The GA includes a steady-state reproduction model [5], in which the original population is only modified through the substitution of the worst individuals by individuals resulting from crossover and mutation. The recombination is carried out by means of a two-point crossover operator and a biased random mutation operator.

The crossover is applied over the two best individuals of the population, obtaining two new individuals, which will substitute the two worst individuals in the population. This strategy leads to a high selective pressure with the aim of getting a quick convergence of the algorithm.

Mutation is carried out as follows. First, according to the mutation probability, the chromosome and the gene of the chromosome to be muted are determined. Then, the biased random mutation operator is applied in two different ways, with probability 0.5 in each case. In the first way, the mutation causes the elimination of the variable to which the gene corresponds, setting to 0 all the values of this variable, as is shown in Fig. 3.a). The second type of mutation randomly assigns 0 or 1 to all the values of the variable, as can be seen in Fig 3.b). So, half the mutations have the effect of eliminating the corresponding variable, and the rest randomly set the values for the variable to be muted.

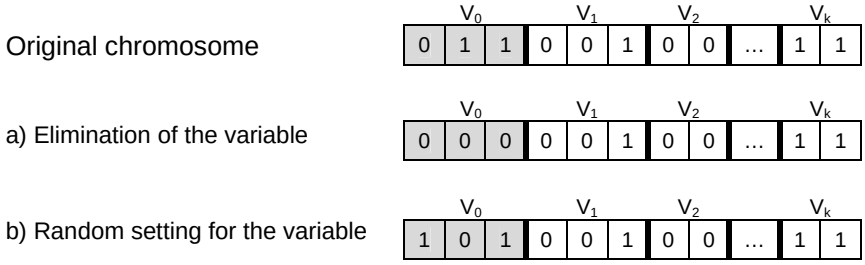

Fig. 3. Types of mutation for a variable in a DNF rule

The mutation is applied according to the mutation probability not only on the two best individuals in the population but on all the population. In order to obtain diversity by means of the application of this operator, a greater population size than the habitual in steady-state evolutionary models must be selected.

4.1.4 Hybrid GA post-processing phase: local search algorithm

The post-processing phase, which improves the obtained rule by a hill-climbing process, modifies the rule in order to increase the degree of support. To accomplish this, in each iteration a variable is selected such that when it is eliminated, the support of the resulting rule is increased; in this way more general rules are obtained. Finally, the optimized rule will substitute the original only if it overcomes minimum confidence.

The diagram of the post-processing phase is as shown in Fig. 4.

```
START
  Best_Rule ← R
  Best_support ← support(R)
  Better ← True
  REPEAT WHILE Better
    Better ← False
    FOR (m=1 to nv)
      R'm ← Best_Rule without considering variable m
      IF (support(R'm)>=support(R) AND
       confidence(R'm)>=confidence(R))
        Better ← True
        IF (support (R'm) > Best_support)
          Best_support ← support (R'm)
          Best_Rule ← R'm
    END FOR
```

```
   END WHILE
   IF (confidence(Best_Rule)>=min_conf)
      Return Best_Rule
   ELSE
      Return R
END
```

Fig. 4. The post-processing phase of the hybrid GA

4.2 Iterative rule extraction model

The fuzzy descriptive rule extraction model follows the Iterative Rule Learning (IRL) approach[9], 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. The objective of the model is to obtain a set of rules giving information on the majority of available examples for each value of the target feature.

The data mining process is carried out by means of an iterative algorithm allowing the generation of several rules (one for each GA run) whereas the generated rules reach a minimum level of confidence (previously specified) 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 repetition mechanism promotes the generating of different rules (in the sense that they give information on different groups of examples). This is achieved by penalizing –once a rule is obtained– the set of examples represented by the same one in order to generate future rules. It is important to point out that this penalization does not prevent the extraction of overlapped rules because the examples covered by previously obtained fuzzy rules are not eliminated and they take part in the computation of the confidence measure. In subgroup discovery algorithms, the possibility of extracting information on described examples is not eliminated since redundant descriptions of subgroups can show the properties of groups from a different perspective.

The confidence of the obtained rule in each iteration must be higher than a previously specified minimum value. In descriptive induction algorithms, one of the fundamental problems, and partially significant to the quality of the obtained results, is the specification of the minimum confidence required for the rules to be extracted. This value depends greatly on the problem to be solved and its solution is a problem which is still not completely resolved. In [38] a method based on fuzzy logic for the setting of the minimum confidence level is described.

4.3 Comparison between the proposal and other subgroup discovery algorithms

To verify the applicability of the proposal, we have compared the results of the model (and of the model with canonical rules, a particular case of DNF rules) with the results of other subgroup discovery algorithms.

For the experimental evaluation and comparison of the approach proposed, the datasets *breast-w* and *diabetes*, both of them containing medical data, and available in the UCI repository have been used. The diabetes dataset contains continuous variables, and is used to show the results of the fuzzy rules extracted by the proposal in comparison with other subgroup discovery algorithms. On the other hand, our proposal can also manage categorical variables, and the breast-w dataset is used to show the behaviour of this proposal with this kind of problems.

The experiments have been carried out in the same way as in [29] to allow the comparison: 10-fold cross validation for the error estimation.

Due to the proposal is a non-deterministic approach, we have carried out 5 runs of each training/test set. The results are the averages of the values obtained by the test partitions. After obtaining the rules with algorithm SDIGA, the measures of Coverage (*Cov*), Support (Sup_1), Size, Significance (*Sig*) and Unusualness (*WRAcc*), which are not used in other knowledge extraction processes, were calculated with the expressions indicated in Section 2 in order to make the comparison. The parameters used are:

- Population size: 100
- Maximum number of evaluations of individuals in each GA run: 10000
- Mutation probability: 0.01
- Number of linguistic labels for the continuous variables: 3
- Quality measure weights for the fitness function: $w_1 = 0.4$ and $w_2 = 0.3$

The specification of the weights for the fitness function depends on the expert knowledge of the characteristics and/or complexity of the problem to be solved. In this chapter, we use these values considering a slight promotion of the extraction of general rules.

Tables 1 and 2 show the results obtained. The tables include the results obtained with the two versions of the SDIGA algorithm (SDIGA, using canonical rules [12], and SDIGA-DNF using DNF rules) for 4 minimum confidence values (named "SDIGA CfMin 0.6" for the SDIGA algorithm with a minimum confidence value of 0.6, and so on), the results for the CN2 algorithm modifying the unusualness measure (CN2-WRAcc), and the results of the CN2-SD using different parameters for the weights (CN2-SD (γ=*x*) is the CN2-SD algorithm using multiplicative weights with γ=*x*, and CN2-SD (add.) is the CN2-SD algorithm using additive weights).

For each measure, the average value and the standard deviation (sd) are detailed. *"COV"* is the average coverage of the set of rules as measured in (2) *"SUP"* is the overall support of a set of rules as computed in (4), *"Siz"* is the number of rules in the induced set of rules, *"SIG"* is the average significance of a

set of rules as measured in (6), and *"WRACC"* is the average rule unusualness as computed in (8).

Table 1. Comparison of subgroup discovery algorithms for Breast-W dataset

Algorithm	*COV*	(sd)	*SUP*	(sd)	*Siz*	(sd)	*SIG*	(sd)	*WRACC*	(sd)
CN2 WRAcc	0.150	0.04	0.900	0.02	8.8	0.95	13.300	1.69	0.063	0.04
CN2-SD (γ=0.5)	0.208	0.05	0.890	0.09	7.9	0.50	27.100	3.37	0.095	0.02
CN2-SD (γ=0.7)	0.174	0.04	0.840	0.04	8.5	1.75	2.100	0.02	0.079	0.01
CN2-SD (γ=0.9)	0.218	0.05	0.930	0.02	9.0	0.24	20.500	2.45	0.093	0.07
CN2-SD (add.)	0.260	0.04	0.860	0.05	9.2	1.24	26.600	3.43	0.111	0.04
SDIGA CfMin 0.6	0.199	0.13	0.497	0.34	5.9	3.03	6.459	2.46	0.002	0.03
SDIGA CfMin 0.7	0.213	0.13	0.481	0.32	5.7	2.80	7.627	2.84	0.010	0.03
SDIGA CfMin 0.8	0.238	0.19	0.439	0.30	4.0	2.06	5.782	3.08	0.006	0.03
SDIGA CfMin 0.9	0.211	0.20	0.423	0.33	3.0	1.08	6.470	3.80	0.022	0.03
SDIGA-DNF CfMin 0.6	0.398	0.07	0.983	0.03	5.4	0.88	16.910	3.81	0.113	0.03
SDIGA-DNF CfMin 0.7	0.414	0.07	0.981	0.02	5.2	0.74	17.399	4.05	0.116	0.03
SDIGA-DNF CfMin 0.8	0.435	0.09	0.969	0.03	4.5	1.36	18.523	5.81	0.124	0.03
SDIGA-DNF CfMin 0.9	0.478	0.07	0.923	0.07	2.4	0.81	24.434	6.63	0.156	0.03

Table 2. Comparison of subgroup discovery algorithms for Diabetes dataset

Algorithm	*COV*	(sd)	*SUP*	(sd)	*Siz*	(sd)	*SIG*	(sd)	*WRACC*	(sd)
CN2 WRAcc	0.275	0.04	0.820	0.03	5.2	0.79	15.800	1.07	0.065	0.06
CN2-SD (γ=0.5)	0.296	0.06	0.920	0.06	6.0	0.68	14.900	1.95	0.085	0.07
CN2-SD (γ=0.7)	0.344	0.05	0.850	0.01	5.6	1.35	11.000	1.43	0.099	0.04
CN2-SD (γ=0.9)	0.299	0.05	0.950	0.01	5.4	0.30	15.200	1.85	0.086	0.07
CN2-SD (add.)	0.381	0.04	0.870	0.05	4.6	0.86	2.100	0.01	0.092	0.03
SDIGA CfMin 0.6	0.462	0.06	0.939	0.04	4.3	0.68	3.286	2.25	0.028	0.02
SDIGA CfMin 0.7	0.431	0.07	0.882	0.07	3.9	0.33	3.515	2.13	0.030	0.01
SDIGA CfMin 0.8	0.707	0.09	0.875	0.07	2.0	0.00	3.967	3.23	0.042	0.02
SDIGA CfMin 0.9	0.707	0.09	0.875	0.07	2.0	0.00	3.967	3.23	0.042	0.02
SDIGA-DNF CfMin 0.6	0.849	0.09	0.992	0.01	2.8	0.38	0.788	1.01	0.024	0.01
SDIGA-DNF CfMin 0.7	0.854	0.09	0.992	0.01	2.9	0.35	0.633	0.54	0.023	0.01
SDIGA-DNF CfMin 0.8	0.931	0.04	0.978	0.02	2.0	0.00	0.437	0.34	0.024	0.01
SDIGA-DNF CfMin 0.9	0.935	0.03	0.976	0.02	2.0	0.00	0.418	0.29	0.023	0.01

Both models of SDIGA (using canonical and DNF linguistic rules) perform better than the other non fuzzy algorithms for the measures coverage (COV), support (SUP) and size (Siz). This means that our proposal obtains a reduced set of rules with a high percentage of examples covered on average, a high number of examples satisfying both the antecedent and the consequent parts of the rules (i.e., a higher percentage of target positive examples leaving a smaller number of examples unclassified is covered), and with a low number of rules. On the other hand, the results for interest measures show different behaviour in the two problems: significance (SIG) and unusualness (WRACC) of SDIGA are similar to the other algorithms for the breast-w problem, but are worse for the diabetes one.

Analyzing the results it is observed that the use of different measures in the rule extraction process of CN2-SD with respect to SDIGA implies:

- the increase of the number of rules,
- the decrease of coverage and support, but
- the increase of the interest measurement values

The inclusion of these measures (or adaptation of them to the fuzzy rules) can be considered in the improvement of SDIGA by means of a multiobjective version of it.

The comparison between the results of the two models of linguistic rules extracted by SDIGA shows that the model which uses DNF linguistic rules obtains better results than the model which uses canonical linguistic rules. As main conclusions of this short comparison study we can conclude that SDIGA allows us to obtain subgroup discovering linguistic rules:

- with very high values of the measures of coverage and support, and so the linguistic rules can be considered very general and significantly representing the knowledge of the examples of the different values of the target variable;
- highly compact, because both the sizes of the set of rules and also the number of variables involved are small;
- highly descriptive, due to the use of DNF linguistic rules, allowing a representation of the knowledge near to human reasoning, and making the extracted knowledge very actionable, a main objective in any subgroup discovery algorithm;
- with a variable interest measure behaviour.

The use of DNF linguistic rules allows us to describe the extracted knowledge in a more flexible way and moreover, to make changes in the initial granularity in each rule in a descriptive way. In this kind of fuzzy rule, as defined in (9), fuzzy logic contributes to the interpretability of the extracted rules due to the use of a knowledge representation close to the expert, also allowing the use of continuous features without a previous discretization.

5. Conclusions

This chapter gives a survey about the use of linguistic rules in the data mining task of subgroup discovery. The subgroup discovery task has been defined, different proposals have been described, and the use of linguistic rules has been analyzed. Then an example of model using linguistic rules for the subgroup discovery task and its advantages has been described.

In summary, for the subgroup discovery task, that searches for unknown and interesting knowledge which can be used for the user, the use of linguistic rules allows the extraction of knowledge in a more natural way and improves it interpretability: Since words play a central role in human information processing, linguistic rules can be used for describe knowledge about subgroups in data which can be actionable by the user.

Finally, we point out some open problems in the development of a fuzzy approach for subgroup discovery:

- To consider the support measure based on /fuzzy set concepts.
- The definition of quality measures for subgroup discovery adapted to the use with linguistic rules.
- The use of multiobjective genetic algorithms [11] [7], analyzing the meaning of pareto-optimal solutions from the subgroup discovery point of view, can provide an interesting tool to get (for getting) set of rules with a trade-off among all the objectives used in the evolutionary model.

Acknowledgement

This work was supported in part by the Spanish Ministry of Education and Science (MEC) under Projects TIN-2005-08386-C05-03 and TIN-2005-08386-C05-01.

References

1. Agrawal R, Imielinski T, Swami AN (1993) Mining Association Rules between Sets of Items in Large Databases. In: Proceedings of the International Conference on Management of Data (ACM SIGMOD 1995). Washington, DC, pp 207–216
2. Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo I (1996) Fast Discovery of Association Rules. In: Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in Knowledge Discovery and Data Mining. AAAI Press, California, pp 307–328
3. Atzmueller M, Puppe F, Buscher H-P (2004) Towards Knowledge-Intensive Subgroup Discovery. In: Proceedings Lernen, Wissensentdeckung und Adaptivität Workshop (LWA'04). Berlin, pp 117–123
4. Au WH, Chan KCC (1998) An effective algorithm for discovering fuzzy rules in relational databases. In: Proceedings of the IEEE International Conference on Fuzzy Systems (Fuzz IEEE'98). Anchorage (USA), pp 1314–1319
5. Bäck T, Fogel D, Michalewicz Z (1997) Handbook of Evolutionary Computation. Oxford University Press, Oxford
6. Clark P, Niblett T (1989) The cn2 induction algorithm. Machine Learning 3(4): 261–283
7. Coello CA, Van Veldhuizen DA, Lamont GB (2002) Evolutionary Algorithms for Solving Multi-Objective Problems. Kluwer Academic Publishers, New York
8. Cordón O, del Jesus MJ, Herrera F (1998) Genetic Learning of Fuzzy Rule-based Classification Systems Co-operating with Fuzzy Reasoning Methods. International Journal of Intelligent Systems 13 (10/11): 1025–1053

9. Cordón O, Herrera F, Hoffmann F, Magdalena L (2001) Genetic fuzzy systems: evolutionary tuning and learning of fuzzy knowledge bases. World Scientific, Singapore
10. Chen G, Wei Q (2002) Fuzzy association rules and the extended mining algorithms. Information Sciences 147: 201–228
11. Deb K (2001) Multi-Objective Optimization using Evolutionary Algorithms. John Wiley & Sons, Chichester
12. Del Jesus MJ, González P, Herrera F, Mesonero M (2005) Evolutionary Induction of Descriptive Rules in a Market Problem. In Ruan D, Chen G, Kerre E, Wets G (eds) Intelligent Data Mining: Techniques and Applications. Springer Verlag, pp 267–292
13. Del Jesus MJ, González P, Herrera F, Mesonero M (Accepted) Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing. IEEE Trans. Fuzzy Systems
14. Dubois D, Prade H, Sudamp T (2005) On the representation, measurement, and discovery of fuzzy associations. IEEE Trans. on Fuzzy Systems 13: 250–262
15. Flach PA, Savnik I (1999) Database dependency discovery: a machine learning approach. AI Communications 12(3): 139–160
16. Fu AW, Wong MH, Sze SC, Wong WC, Wong WL, Yu WK (1998) Finding fuzzy sets for the mining of fuzzy association rules for numerical at-tributes. In: First International Symposium on Intelligent Data Engineering and Learning (IDEAL'98). Hong Kong, pp 263–268
17. Gamberger D, Lavrac N (2002) Expert-guided subgroup discovery: Methodology and application. Journal of Artificial Intelligence Research 17: 1–27
18. Gamberger D, Lavrac N, Krstacic G (2003) Active subgroup mining: a case study in coronary heart disease risk group detection. Artificial Intelligence in Medicine 28 (1): 27–57
19. Hong TP, Chen CH, Wu YL, Lee YC (2004) Using divide-and-conquer GA strategy in fuzzy data mining. In: Ninth International Symposium on Computers and Communications (ISCC 2004). Alexandria, EGYPT, pp 116–121
20. Hong TP, Liu KY, Wang SL (2003) Fuzzy data mining for interesting generalized association rules. Fuzzy sets and systems 138: 255–269
21. Hüllermeier E (2005) Fuzzy methods in machine learning and data mining: Status and prospects. Fuzzy Sets and Systems 156 (3): 387–407
22. Ishibuchi H, Nakashima T, Nii M (2004) Classification and modeling with linguistic information granules Springer-Verlag, New York
23. Kavsek B, Lavrac N, Jovanoski V (2003) APRIORI-SD: Adapting association rule learning to subgroup discovery. In: Proceedings of the 5th International Symposium on Intelligent Data Analysis (IDA 2003). Berlin, pp 230–241
24. Klösgen W (1996) Explora: A Multipattern and Multistrategy Discovery Assistant. In Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in Knowledge Discovery and Data Mining, AAAI Press, California, pp 249–271

25. Klösgen W (2002) Subgroup Discovery. In Klösgen W, Zytkow J (eds) Handbook of Data Mining and Knowledge Discovery. Oxford University Press, New York, pp 354–364
26. Klösgen W, May M (2002) Census Data Mining - An Application. In: 13th European Conference on Machine Learning (ECML'02) / 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02) workshop on on Mining Official Data. Helsinki, pp 65–79
27. Kuok C, Fu A, Wong ML (1998) Mining fuzzy association rules in databases. ACM SIGMOD Record 27: 41–46.
28. Lavrac N, Flach P, Zupan B (1999) Rule evaluation measures: A unifying view. In: Proceedings of the 9th International Workshop on Inductive Logic Programming (ILP'99). Bled, Slovenia, pp 174–185
29. Lavrac N, Kavsec B, Flach P, Todorovski L (2004) Subgroup discovery with CN2-SD. Journal of Machine Learning Research 5: 153–188
30. Lavrac N, Zelezny F, Flach P (2003) RSD: Relational subgroup discovery through first-order feature construction. In: Proceedings of the 13th International Conference on Inductive Logic Programming (ILP 2003). Szeged, Hungary, pp 149–165
31. Michie D, Spiegelhalter DJ, Taylor CC (1994) Machine learning, neural and estatistical classification. Ellis Horwood
32. Piatetsky-Shapiro G, Matheus, C (1994) The interestingness of deviation. In: Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases. Seattle, Washington, pp 25–36
33. Quinlan JR (1987) Generating Production Rules from Decision Trees. In: Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI'87). Milan, Italy, pp 304–307
34. Raedt LD, Dehaspe L (1997) Clausal discovery. Machine Learning 26: 99–146
35. Wrobel S (1997) An algorithm for multi-relational discovery of subgroups. In Proceedings of the First European Symposion on Principles of Data Mining and Knowledge Discovery (PKDD-97). Trondheim, Norway, pp 78–87
36. Zadeh LA (1965) Fuzzy sets. Information Control 8: 338–353
37. Zadeh LA (1975) The concept of a linguistic variable and its applications to approximate reasoning, parts I, II, III. Information Sciences 8-9: 199–249, 301–357, 43–80
38. Zhang S, Lu J, Zhang C (2004) A fuzzy logic based method to acquire user threshold of minimum-support for mining association rules. Information Sciences 164: 1–16

5. Web Intelligence

6. Computer Vision