Introduction to Subgroup Discovery. Some results on the evolutionary extraction of fuzzy rules for subgroup discovery

Pedro González García
Dept. Computer Science
University of Jaén
pglez@ujaen.es

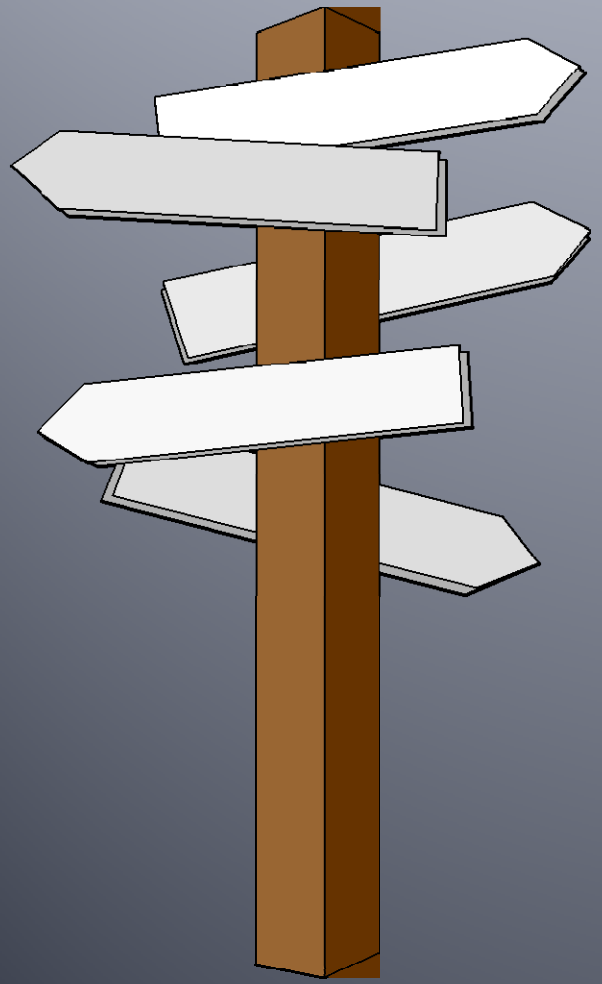
In collaboration with:

María José del Jesus
Dept. Computer Science
University of Jaén

Francisco Herrera
Dept. Computer Science and Artificial Intelligence
University of Granada

I Workshop on Knowledge Extraction based on Evolutionary Learning - Granada, 15-16 May 2008

Outline



1. Subgroup discovery
2. Genetic Algorithms for subgroup discovery
    - 2.1. SDIGA
    - 2.2. MESDIF
3. Conclusions and future work

Outline

1. Subgroup discovery

1.1. Definition

1.2. Emphasizing the differences

1.3. Quality measures for subgroup discovery

1.4. Non evolutionary proposals for subgroup discovery

1.5. Subgroup discovery and Soft Computing

1. Subgroup discovery

1.1. Definition

Given a population of individuals and a specific property of individuals in which we are interested, 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

W. Klösgen, "Explora: A multipattern and multistrategy discover assistant", in Advances in Knowledge Discovery and Data Mining, U. Fayyad, G. Piatetsky-Shapiro,P. Smyth and R. Uthurusamy (Eds.), Menlo Park, CA: AAAI, 1996, pp.249-271

S. Wrobel, "An Algorithm for multi-relational discovery of subgroups" in Proc. 1st European Symposium Principles Data Mining Discovery (PKDD'97), Berlin, Germany, 1997, pp.78-87

1. Subgroup discovery
1.2. Emphasizing the differences

Supervised vs. unsupervised learning

A rule learning perspective

- **Supervised learning:** Rules are induced from labeled instances (training examples with class assignment), usually used in **predictive induction**
- **Unsupervised learning:** Rules are induced from unlabeled instances (training examples with no class assignment), usually used in **descriptive induction**
- **Exception: Subgroup discovery**
  Discovers **individual rules** describing interesting regularities in the data from **labeled** examples

1. Subgroup discovery
1.2. Emphasizing the differences

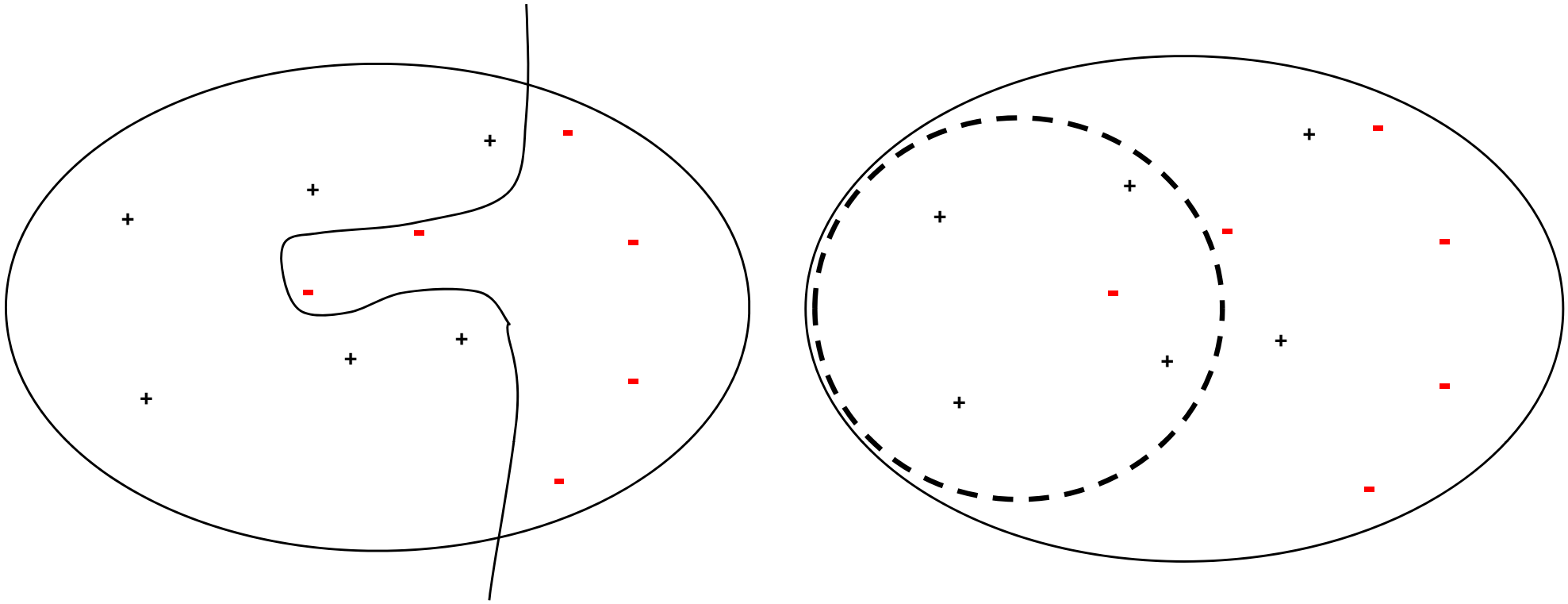
Classification vs. subgroup discovery

- Subgroup discovery is a type of descriptive induction aimed at generating (inducing) knowledge that is understandable (interpretable) by humans
- Main properties of descriptive induction for subgroup discovery:
  - Simple rules
  - Reasonable prediction quality (both on available and future cases)
- It is different from classification aimed induction where the main goal is high classification quality (but induced classification schemes are typically too complex for human interpretation)

1. Subgroup discovery

1.2. Emphasizing the differences

Classification vs. subgroup discovery



1. Subgroup discovery
1.2. Emphasizing the differences

In summary, subgroup discovery techniques:

- Are descriptive induction techniques
- Perform an exploratory analysis
- Aim to obtain
  - Simple and interpretable rules
  - Each rule representing an independent chunk of knowledge
  - Each rule must have a reasonable prediction quality
  - The generality is the main parameter of subgroup discovery and it depends on the problem

- An important aspect of the subgroup discovery algorithms is the quality measures used, both to select the rules and to evaluate the results

1. Subgroup discovery
1.3. Quality measures for subgroup discovery

- There are different proposals for quality measures in subgroup discovery processes

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

Most used objective quality measures in subgroup discovery:

- Coverage: percentage of examples satisfying the antecedent

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

  It measures the generality

- Support: percentage of examples satisfying both the antecedent and the consequent parts of the rule

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

  It measures the generality and the accuracy

- Confidence: relative frequency of examples satisfying the complete rule among those satisfying only the antecedent

1. Subgroup discovery
1.3. Quality measures for subgroup discovery

- Significance: indicates how significant is a finding, measured by the likelihood ratio of a rule

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

Although each subgroup description (i.e. each rule) is for a specific class value, the significance measures impartially the novelty in the distribution of all the class values

- Unusualness (or weighted relative accuracy): balance between the coverage of the rule and its accuracy gain

$$WRAcc(R^i) = \frac{n(Cond^i)}{n_s} \left(\frac{n(Class_j.Cond^i)}{n(Cond^i)} - \frac{n(Class_j)}{n_s} \right)$$

1. Subgroup discovery
1.4. Non evolutionary proposals for subgroup discovery

- **CN2-SD** induces subgroups in the form of rules using as quality measure for rule selection a modified weighted relative accuracy
  - N. Lavrac, B. Kavsek, P. Flach, L. Todorovski. "Subgroup Discovery with CN2-SD". Journal of Machine Learning Research 5 (2004) 153-188
- **APRIORI-SD** is developed adapting the Apriori-C classification rule learning algorithm, and uses the weighted relative accuracy as quality measure for and probabilistic classification of the examples
  - B. Kavsek, N. Lavrac. "APRIORI-SD: Adapting association rule learning to subgroup discovery." Applied Artificial Intelligence 20:7 (2006) 543-583
- **SD-Map** is an exhaustive subgroup discovery algorithm that uses the well-known FP-growth method for mining association rules with adaptations for the subgroup discovery task
  - M. Atzmueller, F. Puppe. "SD-Map - A fast algorithm for exhaustive subgroup discovery". 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006). Lecture Notes in Computer Science 4213, Springer-Verlag 2006, Berlin (Germany, 2006) 6-17

1. Subgroup discovery
1.4. Non evolutionary proposals for subgroup discovery

- Important issues in the development of an algorithm for the subgroup discovery task:
  - The trend is the adaptation of algorithms for rule induction for the subgroup discovery task
  - The standard classification rule learning algorithms can not be applied directly as subgroup discovery approaches because of the use of the coverage algorithm for the construction of the rule sets. A solution to this problem is the use of a weighted coverage algorithm
  - There is no consensus on the quality measures to be used for the task of subgroup discovery. There are a lot of quality measures for the evaluation and selection of rules, although it seems evident that we must take into account measures of both precision and generality of the rules

1. Subgroup discovery
1.5. Soft computing and subgroup discovery

Subgroup discovery aims to discover individual rules which must be represented in an explicit symbolic form and must be relatively simple in order to be recognized as actionable by potential users

Fuzzy logic

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

  - A fuzzy approach for subgroup discovery allow us to obtain knowledge in a similar way to human reasoning

- In particular, in our proposals for subgroup discovery we have used canonical and disjunctive normal function (DNF) fuzzy rules

1. Subgroup discovery

1.5. Soft computing and subgroup discovery

Evolutionary algorithms (in particular, genetic algorithms)

GAs offer a set of advantages for rule extraction processes:

- They tend to cope well with attribute interaction (evaluating a rule as a whole via a fitness function)
- They have the ability to scour a search space thoroughly and to handle a fitness function adapted to the problem to be solved
- The genetic search performs implicit backtracking in its search, thereby allowing it to find complex interactions that other searches would miss

Outline

2. Genetic algorithms for subgroup discovery

2.1. SDIGA: Hybrid GA for the induction of subgroup discovery fuzzy rules

- 2.1.1. Genetic rule extraction process
- 2.1.2. Local search
- 2.1.3. Iterative genetic rule extraction model
- 2.1.4. Experimentation
- 2.1.5. Real-world applications

M.J. Del Jesus, P. González, F. Herrera, M. Mesonero
Evolutionary fuzzy rule induction process for subgroup discovery: a case study in Marketing
IEEE Transactions on Fuzzy Systems, Vol. 15 (4), 2007, pp. 578-592.

2.1. SDIGA

Subgroup Discovery Iterative Genetic Algorithm

- SDIGA obtains fuzzy and/or crisp rules for subgroup discovery depending on the variables of the data base
- SDIGA can extract rules with two different structures:
  - Canonical:

$$R_1: \quad \text{IF } X_1 = LL_1^3 \text{ AND } X_7 = LL_7^1 \quad \text{THEN} \quad Class_j$$

  - DNF:

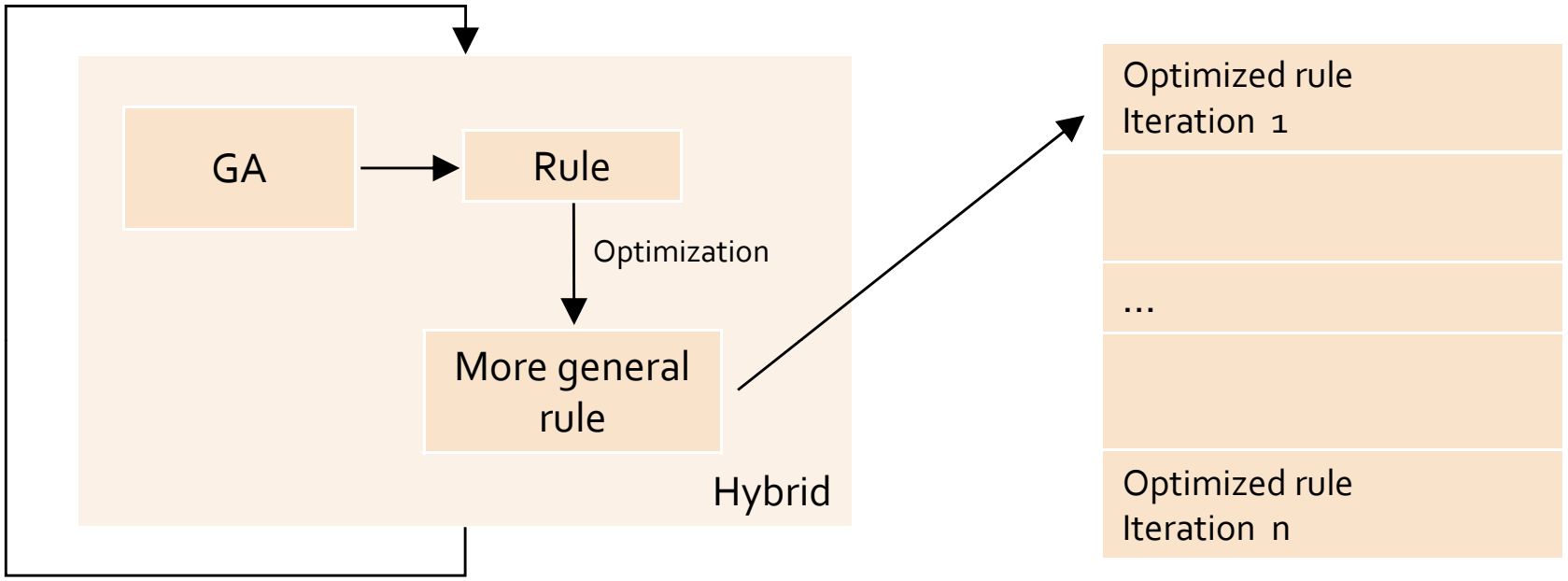
$$R_1: \quad \text{IF } (X_1 = LL_1^1 \text{ OR } LL_1^3) \text{ AND } (X_7 = LL_7^1) \quad \text{THEN} \quad Class_j$$

- The consequent of the rule consist of the target variable, whose value is fixed

2.1. SDIGA

Subgroup Discovery Iterative Genetic Algorithm

- The evolutionary model follows the IRL approach:
  - The GA retuns a single rule
  - The model returns the set of solutions obtained in successive runs of the GA



2.1. SDIGA

2.1.1. Genetic rule extraction process

Chromosome representation

- "*Chromosome = Rule*" approach
- Only the antecedent is represented in the chromosome
- *Canonical rules* (fixed-length integer representation)
  - IF X_1 = $Value_3$ AND X_3 = LL_3^1 THEN *Class 2*

X_1	X_2	X_3	X_4
3	4	1	5

- *DNF rules* (fixed-length binary representation)
  - IF X_1 = ($Value_1$ OR $Value_3$) AND X_3 = LL_3^1 THEN *Class 2*

X_1					X_2				X_3					X_4			
1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0

2.1. SDIGA

2.1.1. Genetic rule extraction process

Fitness function

- In the subgroup discovery process, the objective is the extraction of precise rules with a high descriptive capacity, comprehensible and interesting
- So the problem has several objectives to be maximized
- To obtain this objective, the direct approach is the use of a single objective obtained as the **weighted mean** of the set of objectives:
  - Allows the introduction in the rule generation process of the expert criteria respect of the importance of the objectives for a given problem
  - Has the difficulty of determining appropriate values for the weights
- For the Subgroup Discovery task seems appropriated the use of any measure of:
  - Precision
  - Generality
  - Novelty

2.1. SDIGA

2.1.1. Genetic rule extraction process

Fitness function

$$R_i: \quad \text{IF} \quad (X_1 = LL_1^1 \text{ OR...OR } LL_1^{l_1}) \text{ AND ...AND } (X_{n_v} = LL_{n_v}^1 \text{ OR...O } LL_{n_v}^{l_{n_v}}) \quad \text{THEN} \quad Class_j$$

- Precision measure: confidence

$$Conf\text{-}D(R_i) = \frac{\sum_{E^k \in E \,/\, E^k \in Class_j} APC(E^k, R_i)}{\sum_{E^k \in E} APC(E^k, R_i)}$$

  - where
    - canonical rules:

$$APC(E^k, R_i) = T(\mu_{LL_1}(e_1^k), \ldots, \mu_{LL_{n_v}}(e_{n_v}^k)) > 0$$

    - DNF rules:

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

2.1. SDIGA

2.1.1. Genetic rule extraction process

Fitness function

$$R_i: \quad \text{IF} \quad (X_1 = LL_1^1 \text{ OR...OR } LL_1^{l_1}) \text{ AND ...AND } (X_{n_v} = LL_{n_v}^1 \text{ OR...O } LL_{n_v}^{l_{n_v}}) \quad \text{THEN} \quad Class_j$$

- Precision measure: confidence

$$Conf\text{-}D(R_i) = \frac{\sum_{E^k \in E / E^k \in Class_j} APC(E^k, R_i)}{\sum_{E^k \in E} APC(E^k, R_i)}$$

- Generality measure: support

$$Sup_c\text{-}N(R_i) = \frac{Ne^+(R_i)}{Ne_{NC}}$$

- Novelty measure: interest

$$Int(R_i) = 1 - \left(\frac{\sum_{i=1}^{n_v} Gain(X_i)}{n_v \cdot \log_2 \left(\left| dom(G_k) \right| \right)} \right)$$

2.1. SDIGA

2.1.1. Genetic rule extraction process

Fitness function

- Canonical rules:

$$fitness\ (c) = \frac{\omega_1 \times Sup_c\text{-}N\ (c) + \omega_2 \times Conf\ \text{-}D\ (c) + \omega_3 \times Int\ (c)}{\omega_1 + \omega_2 + \omega_3}$$

- DNF rules:

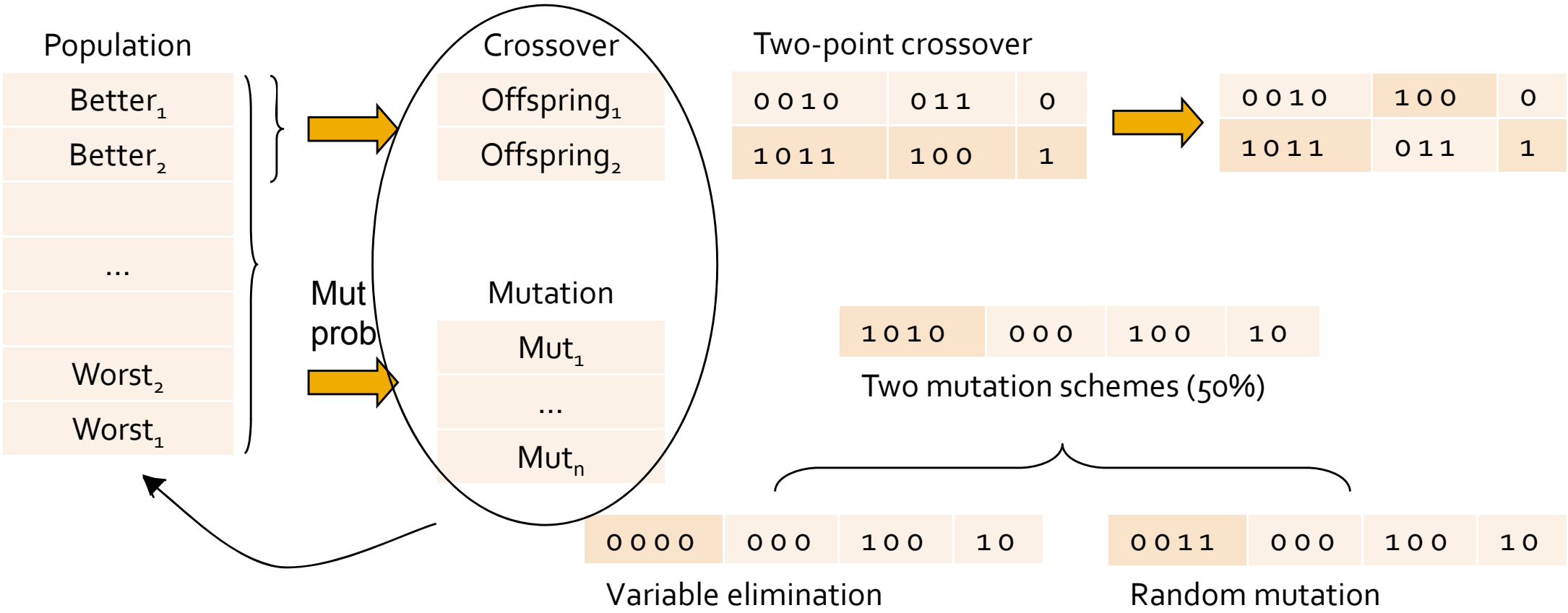
$$fitness\ (c) = \frac{\omega_1 \times Sup_c\text{-}N\ (c) + \omega_2 \times Conf\ \text{-}D\ (c)}{\omega_1 + \omega_2}$$

2.1. SDIGA

2.1.1. Genetic rule extraction process

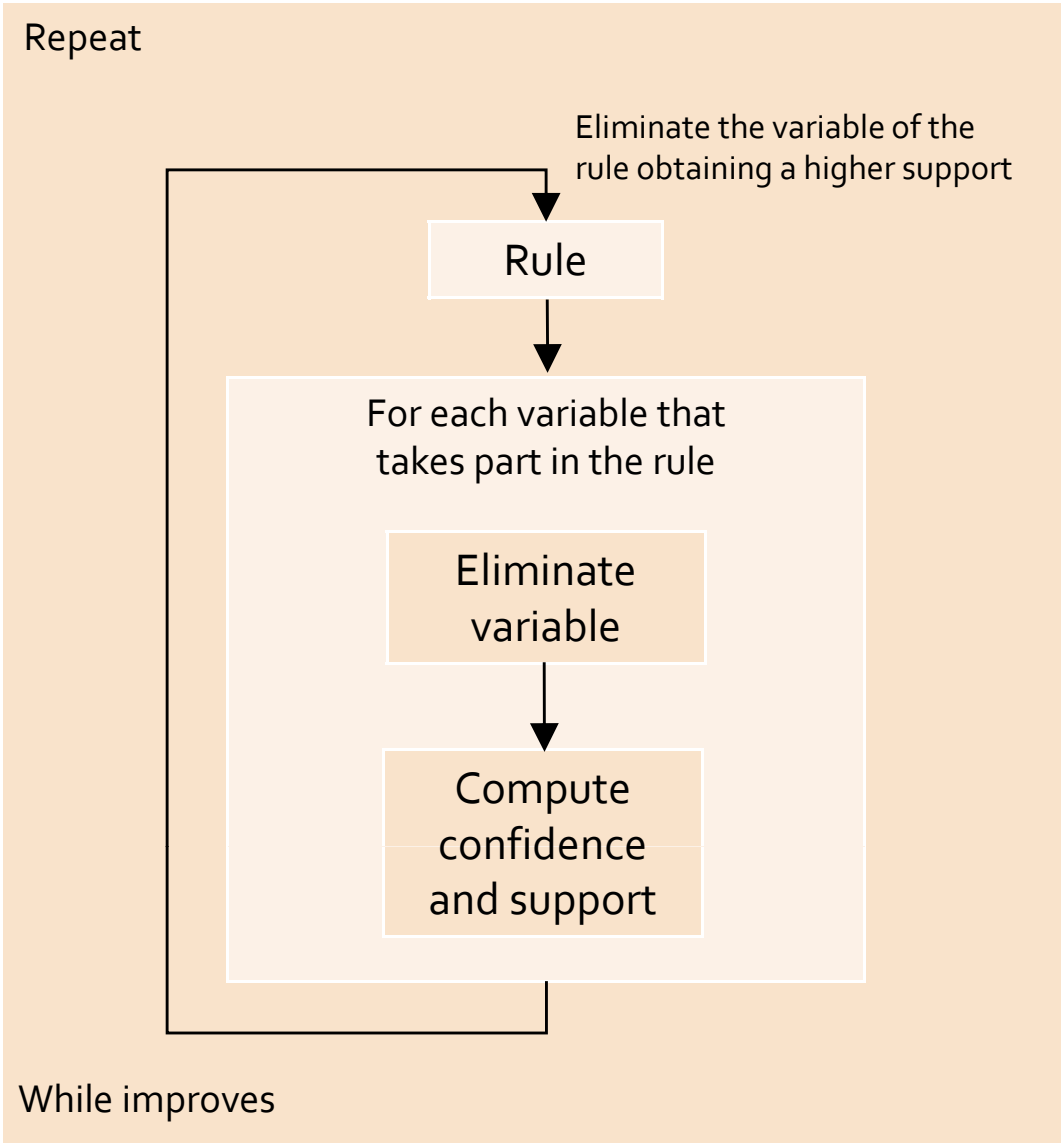
Reproduction model and genetic operators

- Modified steady-state reproduction model
- Genetic operators: a two-point crossover and a biased mutation



2.1. SDIGA

2.1.2. Local search



- Improves the obtained rule by a high-climbing process
- Modifies the rule in order to increase the degree of support

2.1. SDIGA

2.1.3. Iterative genetic rule extraction model

```
start
    Choose a target feature At_TAR
    Rule_Set ← ∅
    repeat
        Execute the GA(At_TAR) obtaining rule R
        Local_Search (R)
        If (confidence(R) ≥ minimum_confidence and
            R represents new examples)
                Rule_Set ← Rule_Set U R
                Mark the set of examples covered by R
        end_if
  while (Confidence (R) ≥ minimum_confidence and R
         represents new examples)
end
```

- Obtains rules while the generated rules:
  - reach the minimum confidence level, and
  - describe information of the zones in which there are examples not described by the previous rules

2.1. SDIGA

2.1.3. Iterative genetic rule extraction model

```
start
    Choose a target feature At_TAR
    Rule_Set ← ∅
    repeat
        Execute the GA(At_TAR) obtaining rule R
        Local_Search (R)
        If (confidence(R) ≥ minimum_confidence and
            R represents new examples)
                Rule_Set ← Rule_Set U R
                Mark the set of examples covered by R
        end_if
  while (Confidence (R) ≥ minimum_confidence and R
         represents new examples)
end
```

$$Sup_c\text{-}N(R_i) = \frac{Ne^+(R_i)}{Ne_{NC}}$$

- Contributes to the generation of different rules, penalizing the examples represented by the rule in the generation of new rules
- This penalization does not avoid the extraction of overlapped rules

2.1. SDIGA

2.1.4. Experimentation

- Results for the *Echo* dataset

Algorithm	Rules	Var	COV	SIGN	WRACC	SUP_c-N	CNF-D
SDIGA	3.02	2.87	0.325	0.943	0.026	0.668	0.637
SDIGA DNF	**2.02**	4.97	0.267	**1.456**	**0.043**	**0.744**	0.608
CN2-SD	17.30	3.23	**0.400**	1.181	0.019	0.490	0.667
Apriori-SD	7.20	**2.04**	0.252	1.141	0.040	0.309	**0.683**

6 variables (5 continuous)	131 examples	2 classes

2.1. SDIGA

2.1.4. Experimentation

- Results for the *Balance* data set

Algorithm	Rules	Var	COV	SIGN	WRACC	SUP_c-N	CNF-D
SDIGA	6.46	2.31	0.315	5.368	0.050	0.501	0.636
SDIGA DNF	**5.00**	2.80	**0.535**	6.978	**0.073**	**0.786**	0.522
CN2-SD	15.60	2.23	0.336	**8.397**	0.063	0.512	0.583
Apriori-SD	10.00	**1.01**	0.200	5.860	0.048	0.307	**0.705**

4 continuous variables	625 examples	3 classes

Conclusions

- Our proposals obtain rule sets:
  - Interpretable, due to the use of fuzzy logic
  - Compacts, composed by few rules
  - Representing information on set of examples not necessarily disjuncts, due to the rule extraction process implemented
  - Crisp and/or fuzzy, depending on the type of variable
  - With canonical or DNF structure, depending on the expert criteria
  - With appropriate values both in the measures used in the data mining process and in the rest of measures not considered in the knowledge extraction process

Conclusions

- It is not easy to determine which algorithm obtains better results when evaluating simultaneously so many quality measures, but results show that our proposal obtains globally the best results
  - Obtains a set of rules with a high level of generality, good accuracy and with a low number of rules

- Depending on the type of variables present in the data sets, it appears that SDIGA gets better results than APRIORI-SD and CN2-SD when all or most of the variables are continuous

- In any case, the results are more descriptive and interpretable in all the data bases

2.1. SDIGA

2.1.5. Real-world applications: marketing problem

- Data obtained from the *Bienal Máquina-Herramienta* (Bilbao, March 2002)
  - 228 exhibitors
  - 104 variables (continuous and categorical)
  - Efficiency of the stand established depending on the achievement of the objectives (*Low*, *Medium* and *High* Efficiency)
- Trade shows are a basic instrument in the marketing policies of the businesses:
  - Facilitate the achievement of business objectives
  - But require a major investment and planning needs
- **Objective:** Determine the contribution of the fair planning variables on the results obtained by the exhibitor, in order to improve future fairs

2.1. SDIGA

2.1.5. Real-world applications: marketing problem

- Low efficiency rules

#	Rule	Sup_c-N	Conf-D
1	IF (Employees = (Huge OR High OR Normal OR Very Few) AND Annual sales = (Very Huge OR Huge OR High OR Few) AND Gratefulness pamphlet = Only to quality contacts AND Bar = No AND Food/Drink = Yes) THEN Efficiency= Low	0.029	1.000
2	IF (Kind of tracking of contacts = All AND Thank-you letter = No AND Stand with different heights = No AND Stewardesses = Yes AND Bar = No) THEN Efficiency= Low	0.029	1.000
3	IF (Zone = (North OR South) AND Important improvement image of the company = Medium AND Thank-you letter = No AND Stand with different heights = No) THEN Efficiency= Low	0.114	1.000
4	IF (Zone = (East OR South) AND Employees = (Very High OR High OR Normal OR Few) AND Annual sales = (Very High OR Normal OR Few) AND Thank-you letter = NO AND Contact tracking = (No OR All) AND Carpet = No AND Bar = No) THEN Efficiency= Low	0.029	1.000

2.1. SDIGA

2.1.5. Real-world applications: marketing problem

- Medium efficiency rules

#	Rule	Sup_c-N	Conf-D
1	IF (Zone = (North OR Center OR South) AND Sector = (Starting OR Deformation OR Accessories OR CAD_CAM) AND Thank-you letter = All AND Thank-you pamphlet = (No OR Only Quality) AND Bar = Yes) THEN Efficiency = Medium	0.086	1.000
2	IF (Employees = (Huge OR Normal OR Very Few) AND Important quality contacts = (High OR Very High) AND Carpet = Yes AND Stewardesses = No AND Bar = No) THEN Efficiency = Medium	0.023	1.000
3	IF (Telephone calls = No AND Bar = Yes AND Food/Drink = Yes) THEN Efficiency = Medium	0.016	1.000
4	IF (Zone = Center AND Stewardesses = Yes) THEN Efficiency = Medium	0.008	1.000
5	IF (Important extracted information = (Very Low OR Low OR Medium OR High) AND Food/Drink = No) THEN Efficiency = Medium	0.578	0.667
6	IF (Zone = North AND Important improvement company image = (Medium OR High) AND Stewardesses = Yes) THEN Efficiency = Medium	0.047	1.000

2.1. SDIGA

2.1.5. Real-world applications: marketing problem

- High efficiency rules

#	Rule	Sup_c-N	Conf-D
1	IF (Employees = (High OR Normal) AND Annual sales = (Very Huge OR Few) AND Thank-you pamphlet = (No OR Only quality) THEN Efficiency = High	0.054	1.000
2	IF (Thank-you letter = (No OR Only quality) AND Columns = Yes AND Bar = No AND Food/Drink = Yes THEN Efficiency = High	0.027	1.000
3	IF (Zone = Center AND Thank-you pamphlet = No) THEN Efficiency = High	0.027	1.000
4	IF (Employees= (Huge OR Very High OR High OR Very Few) AND Satisfaction public relations = (Very Low OR Medium OR Very High) AND Columns = Yes AND Food/Drink = No) THEN Efficiency = High	0.027	1.000
5	IF (Satisfaction improvement company image = (Low OR Very High) AND Telephone calls = No) THEN Efficiency = High	0.081	1.000
6	IF (Employees = Huge OR Normal) AND Publicity in exhibitor's catalogue = Yes AND Bar = Yes AND Food/Drink = No) THEN Efficiency = High	0.027	1.000

2.1. SDIGA

2.1.5. Real-world applications: marketing problem

- Conclusions of the experts :
  - Exhibitors whose obtained worst results came from the South zone, not conducting follow-up of the contacts and could not optimize the contacts made in the trade fair
  - Exhibitors with better results came from the Central zone and do not send a thank-you pamphlet to all the contacts . These are large- or medium-sized companies, with either a very high or a low annual sales volume
  - Also, the exhibitors that obtained the best results have a very high or a small sales volume

Outline

2. Genetic algorithms for subgroup discovery

2.2. MESDIF: Multiobjective Evolutionary Approach to obtain Descriptive Fuzzy Rules Describing Subgroups

- 2.2.1. Multiobjective GAs and subgroup discovery
- 2.2.2. Multiobjective Evolutionary Approach to obtain Descriptive Fuzzy Rules Describing Subgroups
- 2.2.3. Experimentation
- 2.2.4. Real-world applications

M.J. del Jesus, P. González, F. Herrera
Multiobjective genetic algorithm for extracting subgroup discovery fuzzy rules
Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Multicriteria Decision Making (IEEE MCDM 2007), Honolulu (USA), 2007, pp. 50-57.

2.2. MESDIF

2.2.1. Multiobjective GAs and subgroup discovery

- 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
  - Multiobjective Evolutionary Algorithms (MOEAs) are adapted to solve this kind of problems
- 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 (Pareto-optimal front)
- Most multi-objective optimization algorithms use the concept of domination to obtain the Pareto-optimal front

2.2. MESDIF

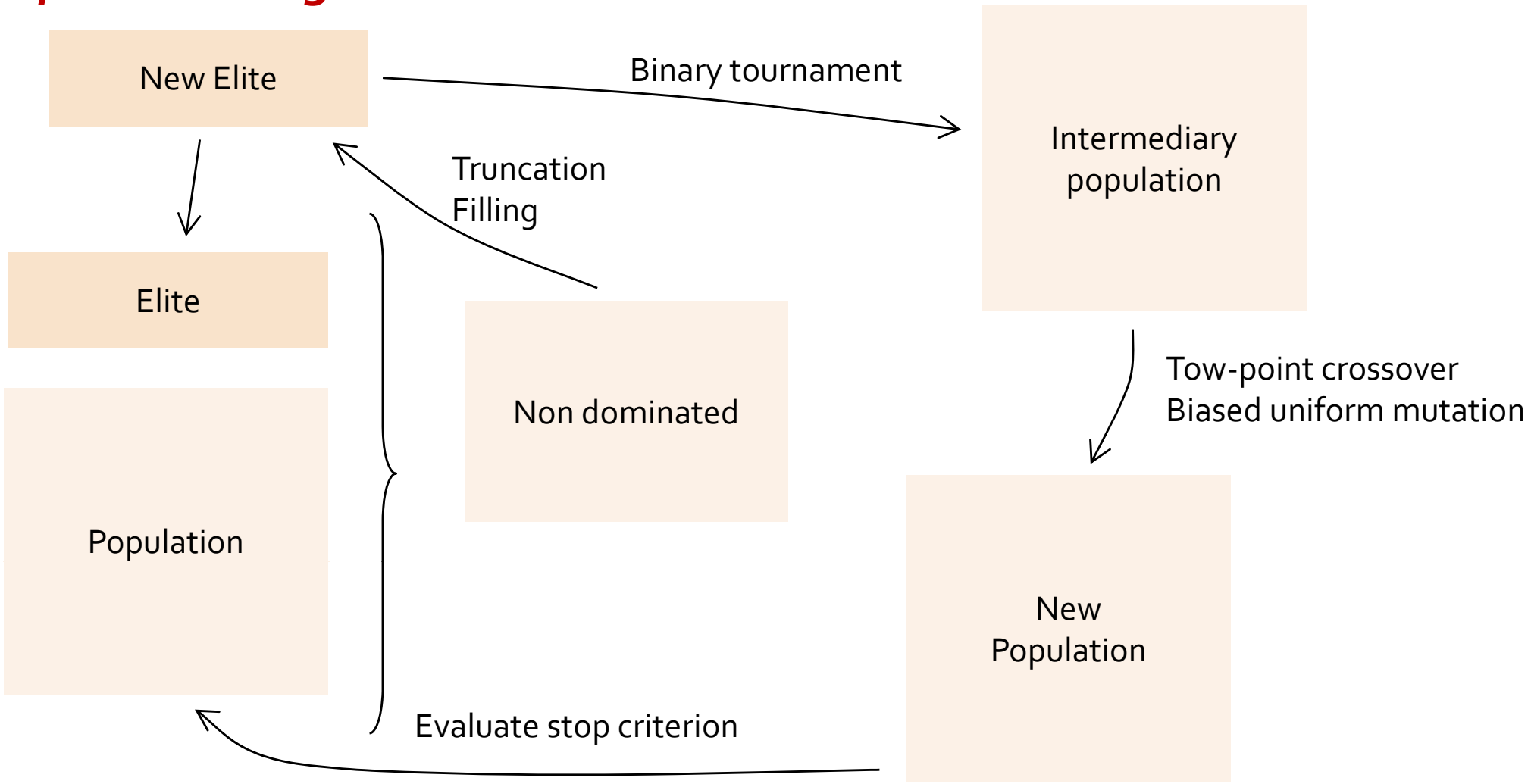
2.2.2. Multiobjective Evolutionary Approach to obtain Descriptive Fuzzy Rules Describing Subgroups

- **Objectives**:
  - Obtain a set of solutions in a single run of the algorithm
  - Avoid the need of indicate the weights of the defined objectives
  - Obtain a Pareto-front from which the expert can select the more useful or interesting rules
- The algorithm allows the representation of fuzzy and/or crisp rules for problems with continuous and/or categorical variables, as SDIGA does
- Follows the SPEA2 approach
  - Elitism
  - Search of optimal solutions in the Pareto front
- To preserve the diversity, the algorithm uses:
  - A niches technique that considers the proximity in values of the objectives
  - An additional objective based on the novelty, to promote rules which give information on examples not described by other rules

2.2. MESDIF

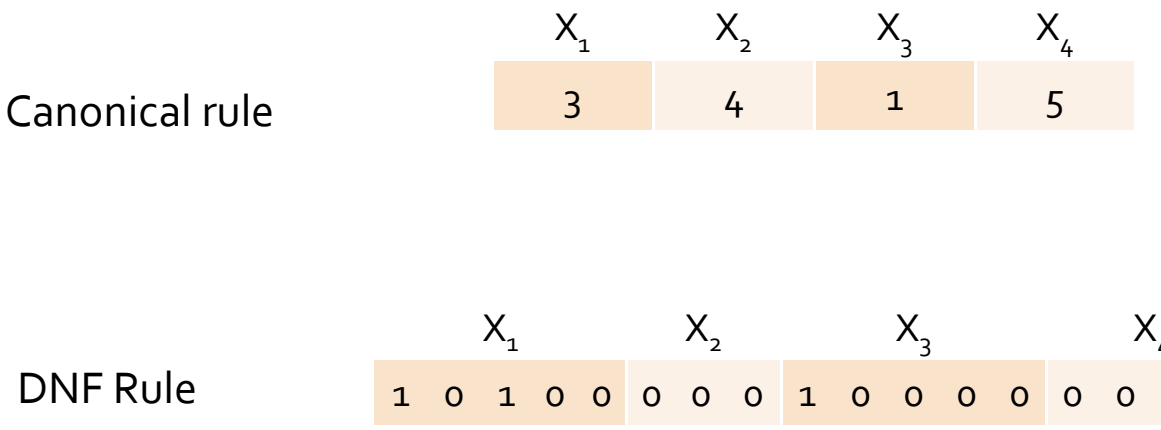
2.2.2. Multiobjective Evolutionary Approach to obtain Descriptive Fuzzy Rules Describing Subgroups

Operation diagram



Chromosome representation

- Each solution is codified according to the "Chromosome=Rule" approach
- All the individuals are associated to the same value of the target variable, and so only the antecedent is represented



Definition of the objectives of the multiobjective algorithm

- **Confidence**: The same used in SDIGA

$$Conf-D(R_i) = \frac{\sum_{E^k \in E / E^k \in Class_j} APC(E^k, R_i)}{\sum_{E^k \in E} APC(E^k, R_i)}$$

- **Support:** The same used in SDIGA

$$Sup_c-N(R_i) = \frac{n(Class.Cond_i)}{Ne_{NC}}$$

- **Original support** (originality level)

$$Sup-orig(R_i) = \sum_{\substack{\forall E^k \in E / \\ APC(E^k, R_i) > 0}} \frac{1}{k}$$

$$where \quad k = \#rule / APC(E^k, R_i) > 0$$

  - The use of this objective:
    - Allows to promote rules giving information on examples not described by other rules (obtaining rules belonging to different parts of the search space)
    - Is a restriction in the rules in order to obtain a Pareto-optimal front with a high degree of global coverage (important because the proposed algorithm is not a covering one and the obtained rules can be overlapped)

2.2. MESDIF

2.2.3. Experimentation

- Results for the *German* data set

Algorithm	Rules	Var	COV	SIG	WRACC	SUP_c-N	CNF-D
MESDIF	20.00	3.49	**0.328**	2.675	0.024	0.471	0.612
SDIGA	8.56	4.40	0.082	0.615	0.006	0.177	0.313
MESDIF DNF	19.78	4.03	0.246	2.860	0.026	**0.506**	0.632
SDIGA DNF	25.98	7.04	0.027	0.875	0.004	0.085	0.309
CN2-SD	25.70	5.90	0.321	**3.924**	0.024	0.405	0.700
Apriori-SD	**6.10**	**1.58**	0.308	3.856	**0.039**	0.366	**0.746**

20 variables (7 continuous)	1000 examples	2 classes

2.2. MESDIF

2.2.3. Experimentation

- Results for the *Hepatitis* data set

Algorithm	Rules	Var	COV	SIG	WRACC	SUP_c-N	CNF-D
MESDIF	19.82	4.31	0.381	1.340	0.043	0.640	0.659
SDIGA	5.16	4.63	0.187	0.813	0.017	0.415	0.582
MESDIF DNF	18.58	4.62	0.255	1.353	0.033	0.606	0.675
SDIGA DNF	**2.34**	**5.75**	0.286	0.864	0.027	**0.737**	0.663
CN2-SD	18.30	4.54	**0.431**	**2.175**	**0.059**	0.598	**0.805**
Apriori-SD	10.00	2.13	0.335	1.516	0.044	0.431	0.632

19 variables (6 continuous)	155 examples	2 classes

Conclusions

- The elite set size, allow to orient the algorithm to the extraction of a maximum number of rules
- Avoids the determination of the weights associated to each objective
- MESDIF overtakes SDIGA in the more complex data sets
- MESDIF improves in almost all the data sets the confidence level obtained by SDIGA
- MESDIF obtains sets of rules representing more completely the set of examples, due to the inclusion of two mechanisms to improve the diversity:
  - A niche schema included in the truncation operator and the density function
  - A special objective, the original support, that allows the extraction of rules describing information of examples of which the other rules did not obtained information

2.2. MESDIF

2.2.4. Real-world problems: e-learning problem

- We have data of the courses of the University of Córdoba using the Moodle system as a complement
  - 192 courses, from which we have selected 5, for a total of 293 students
- **Objective**: Study the relationship between the follow-up of complementary activities of the course in a e-learning system with the final mark
- Algorithms used:
  - Evolutionary: SDIGA and MESDIF for the extraction of canonical rules
  - Non-evolutionary: Apriori-SD and CN2-SD

2.2. MESDIF

2.2.4. Real-world problems: e-learning problem

- Results

Algorithm	CfMin / Weights	Rules	Var	COB	SIGN	SUP_c-N	CNF-D
MESDIF	0.6	7.8	1.95	0.383	20.248	0.292	0.735
	0.7	5.8	1.55	0.527	21.897	0.182	0.767
	0.8	5.4	1.49	0.590	19.675	0.117	0.836
	0.9	6.0	1.92	0.487	5.619	0.097	0.911
SDIGA	0.6	7.8	2.0	0.088	21.992	0.193	0.809
	0.7	6.2	2.1	0.077	16.792	0.179	0.750
	0.8	6.0	2.2	0.127	25.246	0.224	0.779
	0.9	4.8	2.0	0.129	33.835	0.265	0.755
CN2-SD	γ=0.5	13.0	5.5	0.415	44.949	0.326	0.716
	γ=0.7	17.0	5.5	0.398	48.438	0.283	0.719
	γ=0.9	16.0	5.3	0.388	50.281	0.274	0.729
	add	32.0	5.7	0.508	54.424	0.341	0.712
Apriori-SD	0.6	8.0	1.0	0.622	26.132	0.438	0.613
	0.7	9.0	1.3	0.668	29.541	0.452	0.613
	0.8	6.0	1.5	0.361	42.109	0.206	0.613
	0.9	5.0	2.0	0.225	36.810	0.176	0.631

2.2. MESDIF

2.2.4. Real-world problems: e-learning problem

Conclusions

- The proposed evolutionary algorithms are appropriated to solve the problem
- MESDIF obtains betters results than SDIGA
- Booth obtain sets of rules describing knowledge of interest for the expert in an interpretable way due to:
  - The reduced size (number of rules)
  - The structure:
    - Treatment of the continuous variables as linguistic variables
    - Low number of variables in the antecedent part of the rule
- Obtain appropriated values of the quality measures used for the evaluation of the rules
- The rules obtained allows the teacher to perform decisions about the activities of the course in order to improve the performance

2.2. MESDIF

2.2.4. Real-world problems: e-learning problem

- Extracted rules examples:

#	Rule	Sup_c-N	Conf-D
1	IF *Course* = 110 AND #*assignments* = High AND *Nº of posts* = High THEN *Mark* = Good	0,704	0,723
2	IF *Course* = 88 AND #*chat posts* = Very High THEN *Mark* = Fail	0,193	0,944
3	IF #*messages read* = Very Low THEN *Mark* = Fail	0,732	0,610
4	IF #*messages read* = High AND #*messages sent to teacher* = Very Low THEN *Mark* = Fail	0,117	0,750
5	IF #*quizzes completed* = Very Low THEN *Mark* = Fail	0,933	0,727

Outline

3. Conclusions and future work

3. Conclusions and future work

- Both SDIGA and MESDIF allow the extraction of Subgroup Discovery rules:
  - Highly descriptive, due to the use of the fuzzy logic for the representation of the knowledge
  - Canonical or DNF depending of the expert criteria
  - Crisp or fuzzy, depending on the variables of the problem
  - General an representatives of the knowledge of the examples of the different values of the target variable
  - Highly compacts

3. Conclusions and future work

- SDIGA allows the extraction of SD rules:
  - Describing knowledge on the different sets of examples of the problem to be solved due to the inclusion of a mechanism to penalize the examples covered
  - Occasionally overlapped, allowing the description of knowledge from several points of view
  - Also allowing the extraction of rules covering few examples, or *"small disjuncts"*

3. Conclusions and future work

- MESDIF:
  - Avoid the compensation between quality measures allowing the extraction of rule sets with a high level of confidence and support
  - The inclusion of the original support adds an additional level of promotion of the diversity
  - Has been shown appropriated for complex problems (high number of variables, of examples or of classes), improving the results of SDIGA
  - Rules obtained by SDIGA are simpler due to the application of a hill-climbing algorithm which optimizes the rules
  - Improves the confidence levels obtained, extracting more precise rules
  - Improves the results of SDIGA in the real-world problems analyzed
  - Allows the expert to select the more relevant rules among the set of rules generated

3. Conclusions and future work

Future work

- Development of new quality measures for the subgroup discovery problem
- Study of the inclusion of new interest or novelty measures for the evaluation and selection of rules, especially for DNF rules
- Development of new multiobjective models
- Study on the influence of the feature selection in subgroup discovery problems with high dimensionality