International Journal of Computational Intelligence Systems, Vol. 5, No. 2 (April, 2012), 355-367

# Genetic lateral tuning for subgroup discovery with fuzzy rules using the algorithm NMEEF-SD

**C.J. Carmona , P. González , M.J. Gacto , M.J. del Jesus**

*Department of Computer Science, University of Jaen,*
*Campus Las Lagunillas, s/n,*
*23071, Jaen, Spain*
*E-mail: ccarmona@ujaen.es, pglez@ujaen.es, mgacto@ujaen.es, mjjesus@ujaen.es*

Received 15 December 2010
Accepted 1 June 2011

**Abstract**

The main objective of subgroup discovery is to discover interesting and interpretable patterns with respect to a specific property. The use of evolutionary fuzzy systems provides good algorithms to approach this problem. In this sense, NMEEF-SD algorithm –one of the most representative evolutionary fuzzy systems for subgroup discovery– obtains precise and interpretable subgroups. However in the majority of the evolutionary fuzzy systems, the membership functions of the linguistic labels are usually fixed to static values and the partitions are not adapted to the context of each variable. In this paper, a post-processing tuning step to improve the results of the subgroup discovery algorithm NMEEF-SD is proposed, allowing the partitions to be adapted to the context the variables. The application of this tuning step is a novelty in subgroup discovery and consist of a genetic algorithm which allows the lateral displacement of the membership functions of a label considering a unique parameter, using the 2-tuples linguistic representation. The results obtained using different data sets of the KEEL repository show the improvement in the performance of the NMEEF-SD algorithm with lateral displacement. The study is supported by statistical tests to improve the analysis performed.

*Keywords:* Subgroup discovery, evolutionary fuzzy system, fuzzy rules, 2-tuples linguistic representation

## 1. Introduction

Subgroup Discovery (SD) [34,44] is a data mining task whose objective is the discovery of interesting individual patterns (rules in this case) in relation to a specific property which is of interest to the user. SD is broadly applicable such as in medicine [11] or e-learning [41] among others, and focus its interest on partial relations instead of complete ones. The discovered subgroups should be interpretable and interesting according to the criteria of the user. In [29], a recent review describing the SD task, the quality measures used, the approaches and the applications can be found. The SD task is somehow between descriptive and predictive induction, and different algorithms adapting classical algorithms of both classification –as CN2-SD [38]– and association rule learning –as Apriori-SD [33] or SD-MAP [8]– have been proposed. Nowadays, one of the most important aspect in SD is the measures to be used to evaluate the quality of the subgroups extracted.

Basically, evolutionary fuzzy systems (EFSs) [15,16,28] use evolutionary algorithms (EA) [21] for learning or tuning fuzzy systems. EFSs have been successfully applied to SD because EAs handle ap-

propriately the relations between variables, and the use of fuzzy logic by means of descriptive fuzzy rules allows the representation of knowledge in a similar way to human reasoning, leading to the obtaining of more interpretable and actionable solutions in SD. In addition, several EFSs have been proposed for the SD task –as SDIGA [19] or MESDIF [9]–. The latter EFS proposed so far for the SD task is the Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery (NMEEF-SD) [10], whose objective is the extraction of descriptive fuzzy and/or crisp rules for SD, depending on the type of variables present in the problem.

One of the advantages in the use of EFSs to solve the SD task is the interpretability of the rules obtained; in fact, algorithms for SD based on EFSs consider the interpretability as a main issue. But in addition, EFSs can be useful for the context adaptation of the labels of the linguistic variables, adjusting the number of labels for each linguistic variable or the definition of the fuzzy membership functions among others [28]. When the membership functions are obtained by a normalisation process, the fuzzy partitions may not be adapted to the context of each variable, and the cooperative behaviour of the rules may not be optimal. This contextualisation can be performed using an a priori learning process or a posteriori tuning process.

In this paper, a lateral tuning in the membership functions that improves their global interaction, with the main aim of inducing a better cooperation among rules [16,28] is introduced through a genetic tuning process. This process of contextualizing the membership functions enables to achieve a better covering degree while maintaining the original shapes. The main objective is to analyse the use of this process on the rules obtained by the SD algorithm NMEEF-SD [10]. This algorithm has been used to obtain the rule base, and then a lateral tuning of the membership functions has been performed. This optimisation has been carried out by a genetic lateral tuning using a linguistic rule representation model proposed in [2]. It is based on the 2-tuples linguistic representation [31] which allows the lateral displacement of a label considering a unique parameter. The majority of SD algorithms are not capable of handling continuous variables, and they need to perform a previous step to discretise them. However, the use of fuzzy logic in NMEEF-SD avoids the need for previous discretisation. Moreover, the lateral tuning provides the experts an optimisation of the results in continuous variables.

To do so, the paper is organised as follows: an introduction to the SD task and the most used quality measures in SD are presented in Section 2. An introduction to the EFSs for the SD task, together with the NMEEF-SD algorithm are presented in Section 3. In Section 4, the genetic lateral tuning and the application of the tuning process to the results of the NMEEF-SD algorithm are described. Finally, an analysis of the results of the genetic lateral tuning applied to the NMEEF-SD algorithms in a wide number of quality measures employed in SD is performed in Section 5, and in Section 6 the conclusions are outlined.

## 2. Subgroup discovery

First, a description of the main concepts of SD are shown in Section 2.1. Then, a summary of the quality measures most used in SD can be observed in Section 2.2.

### *2.1. Introduction to subgroup discovery*

SD is a data mining task in which given a population of individuals and a property of those individuals we are interested in, the objective is to find population subgroups that are statistically "most interesting", i.e. are as large as possible and have the most unusual statistical characteristics with respect to the property of interest [34,44]. The main goal in SD is to discover characteristics of the subgroups by constructing rules with high support and significance. As SD focusses its interest on partial relations instead of complete ones, small subgroups with interesting characteristics can be sufficient.

In SD, a rule $R$ can be described as:

$$R : Cond \rightarrow Target_{value}$$

where the property of interest is the $Target_{value}$ that appears in the consequent part of the rule, and

the antecedent part of the rule, *Cond*, is a conjunction of features (usually attribute-value pairs) selected from the features describing the training instances [25,36].

For further information, the interested reader can find in [29] a recent review describing the main properties of the SD task, its most used quality measures, the available approaches in the literature to solve this problem, and the main applications in real-world problems.

The lateral tuning approach presented in this paper provides novelty within the data mining task of SD allowing the tuning of the linguistic labels employed in the subgroups for the algorithms based on EFSs, thus adapting the membership functions to the context of the variables and improving the quality of the results.

### 2.2. Quality measures

One of the most important aspects in SD is the choice of the quality measures employed to extract and evaluate the rules. There is a wide number of measures for the SD task presented throughout the bibliography, but it is difficult to establish which is the most suitable one as it often depends on the problem. Due to this fact, it is really complicated to analyse the behaviour of a SD algorithm. Hence, experts can choose in general the most suitable quality measures in terms of the problem.

Below is a definition of different quality measures for SD, focusing not only in quality measures employed in the tuning process, but also in other important quality measures presented in Section 5:

- *Coverage*: It measures the percentage of examples covered on average [38]. This can be computed as:

$$Cove(R) = \frac{n(Cond)}{n_s}, \quad (1)$$

  where $n_s$ is the number of total examples and $n(Cond)$ is the number of examples which satisfy the conditions determined by the antecedent part of the rule.
- *Crisp Support*: It measures the frequency of correctly classified examples covered by the rule [38]. This can be computed as:

$$CSup(R) = \frac{n(Target_{value} \cdot Cond)}{n_s}, \quad (2)$$

  where $n(Target_{value} \cdot Cond)$ is the number of examples which satisfy the conditions and also belong to the value for the target variable in the rule.
- *Fuzzy Support*: It is defined as the degree of coverage that the rule offers to examples of that target value, and it is computed as:

$$FSup(R) = \frac{n(Target_{value} \cdot Cond)}{n_s}, \quad (3)$$

  where $n(Target_{value} \cdot Cond)$ is the number of examples which satisfy the conditions with fuzzy properties and also belong to the value of the target variable.
- *Crisp Confidence*: It measures the relative frequency of examples satisfying the complete rule among those satisfying only the antecedent. This can be computed with different expressions [19], and in this paper is computed as:

$$CCnf(R) = \frac{n(Target_{value} \cdot Cond)}{n(Cond)}, \quad (4)$$

  This quality measure can also be found as *accuracy* in the specialised bibliography.
- *Fuzzy confidence*: It measures the relative frequency of examples satisfying the complete rule among those satisfying only the antecedent, and it is defined as [19]:

$$FCnf(R) = \frac{\sum_{E^k \in E / E^k \in Target_{value}} APC(E^k, R)}{\sum_{E^k \in E} APC(E^k, R)}, \quad (5)$$

  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.
- *Significance*: This measure indicates the significance of a finding, if measured by the likelihood ratio of a rule [34]. It can be computed as:

$$Sign(R) = 2 \cdot \sum_{k=1}^{n_{Tv}} n(Target_{value_k} \cdot Cond) \cdot log \frac{n(Target_{value_k} \cdot Cond)}{n(Target_{value_k}) \cdot p(Cond)}, \quad (6)$$

where $p(Cond)$, computed as $n(Cond)/n_s$, is used as a normalized factor, and $n_{Tv}$ is the number of values of the target variable. It must be noted that although each rule is for a specific $Target_{value}$, the significance measures the novelty in the distribution impartially, for all the values.

- *Sensitivity*: This measure represents the the proportion of actual matches that have been classified correctly [34]. It can be computed as:

$$Sens(R) = \frac{n(Target_{value} \cdot Cond)}{n(Target_{value})}, \quad (7)$$

where $n(Target_{value})$ are all the examples of the target variable. This quality measure was defined in [19] as *Support based on the examples of the class*, and used to evaluate the quality of the subgroups in the Receiver Operating Characteristic (ROC) space. Sensitivity combines precision and generality related to the target variable.

- *Unusualness*: This measure is defined as the weighted relative accuracy of a rule [37]. It can be computed as:

$$WRAcc(R) = \frac{n(Cond)}{n_s} \cdot \left( \frac{n(Target_{value} \cdot Cond)}{n(Cond)} - \frac{n(Target_{value})}{n_s} \right). \quad (8)$$

The unusualness of a rule can be described as the balance between the coverage of the rule $p(Cond_i)$ and its accuracy gain $p(Target_{value} \cdot Cond) - p(Target_{value})$.

- *Accuracy*: It is the percentage of positive examples of a rule. This quality measure is called "confidence" in descriptive data mining references. It can be computed as:

$$Accu(R) = \frac{n(Target_{value} \cdot Cond) + 1}{n(Target_{value} + n_{Tv})}. \quad (9)$$

## 3. Evolutionary fuzzy systems for SD

This section presents an introduction to EFSs and their application to the SD task (Section 3.1). In addition, a brief description of the NMEEF-SD algorithm, used in the experiments, is presented in Section 3.2.

### 3.1. Introduction of the evolutionary fuzzy systems

EFSs are basically fuzzy systems augmented by a learning process based on evolutionary computation, which includes genetic algorithms, genetic programming, and evolutionary strategies, among other evolutionary algorithms [21]. Currently, EFSs are being applied to a wide range of real-world problems. The research related to this area is growing, and a number of open problems and future directions can be found in [12,13,39].

Fuzzy systems are one of the most important areas for the application of the fuzzy set theory [45]. Usually this kind of systems consider a model structure in the form of fuzzy rules. They are called fuzzy rule based systems (FRBSs), which have demonstrated their ability with respect to different problems like control problems, modeling, classification or data mining in a large number of applications. The pioneering works in application of FRBSs to these types of problems can be found in [32,35]. FRBSs provide us a comprehensible representation of the extracted knowledge and moreover a suitable tool for processing the continuous variables.

There are two different processes in an EFS, tuning and learning. It is difficult to make a clear distinction between both processes, since establishing a precise borderline becomes as difficult as defining the concept of learning itself. It is important to take into consideration the existence or not of a previous Knowledge Base (KB), including Data Base (DB) and Rule Base (RB). The following distinction can be considered in EFSs:

- Genetic tuning. If there exists a KB, a genetic tuning process is applied to improve the system performance without changing the existing RB.
- Genetic learning. To learn KB components (where an adaptive inference engine can even be included). That is, to involve the learning of KB components among other components of the algorithm.

The task of SD has been successfully tackled using different EFSs, such as SDIGA [19] (an evolutionary fuzzy rule induction system based on the iterative rule-learning (IRL) proposal [14], which evaluates the quality of the rules by means of a weighted average of the measures selected), MESDIF [9,18] (a multi-objective genetic algorithm for the extraction of fuzzy rules which describes subgroups based on the multi-objective SPEA2 [46] approach, and which can use several quality measures at a time to evaluate the rules obtained) or NMEEF-SD [10] (a multi-objective genetic algorithm based on the NSGA-II algorithm [17], which is explained in the following section).

In this paper a tuning process performed to the rules obtained by the SD algorithm NMEEF-SD is presented, to ease the genetic optimization of the membership functions of the data base through a new linguistic rule representation model proposed in [2]. It is based on the 2-tuples linguistic representation [31] that allows the lateral displacement of a label considering a unique parameter.

### *3.2. NMEEF-SD algorithm*

NMEEF-SD [10] is an evolutionary algorithm whose objective is to extract descriptive fuzzy and/or crisp rules for the SD task, depending on the type of variables present in the problem.

This algorithm is based on the NSGA-II algorithm [17], which is a multi-objective evolutionary algorithm with a non-dominated sorting approach. NMEEF-SD is oriented towards SD and uses specific operators to promote the extraction of simple, interpretable and high quality SD rules.

With respect to the representation of the rules, each candidate solution is coded according to the "*Chromosome = Rule*" approach, where only the antecedent is represented in the chromosome and the consequent is prefixed to one of the possible values of the target variable in the evolution. Therefore, the algorithm must be executed as many times as the number of different values the target variable contains. In this paper, a canonical representation with as many genes as variables contained in the original data set without considering the target variable is used with NMEEF-SD.

This proposal permits a number of quality measures to be used both for the selection and the evaluation of the rules within the evolutionary process. In this study, the algorithm uses two quality measures in the process: unusualness (Eq. 8) and sensitivity (Eq. 7). With these quality measures the algorithm follows to obtain rules with a good balance between support and confidence, with also high values of unusualness.

NMEEF-SD uses a new operator to enhance the diversity, the re-initialisation based on coverage together with the crowding distance in the selection operator. On the other hand, the algorithm includes operators of biased initialisation and biased mutation to promote generalisation. In addition, only the final solutions which reach a predetermined confidence threshold are returned.

A complete study of this algorithm with respect to other SD approaches using different quality measures, including a statistical analysis, can be observed in [10].

## 4. Genetic lateral tuning

The main objective of this work is to improve the performance of the SD algorithm NMEEF-SD by means of a tuning approach based on the 2-tuples linguistic representation. This methodology consists of refining a previous definition of the DB once the RB has been obtained [2]. The tuning introduces a lateral tuning in the membership functions that improves their global interaction, with the main aim of inducing a better cooperation among the rules [16,28]. In this way, the aim of the tuning process is not only to find specific membership functions in an independent way but to find the best global configuration of the membership functions.

In next subsections, the lateral tuning process and the genetic algorithm used for the tuning are introduced.

### *4.1. Lateral tuning*

In this approach, a rule representation model based on the 2-tuples linguistic representation [31] is used. This representation allows the lateral displacement of the labels considering only one parameter (slight

displacements to the left/right of the original membership functions). This involves a simplification of the search space that eases the derivation of optimal models. Furthermore, this process of contextualizing the membership functions enables them to achieve a better covering degree while maintaining the original shapes, which results in accuracy improvements without a loss in the interpretability of the fuzzy labels. In the specialised literature, the 2-tuples representation has been used to tackle different problems as regression [4] or classification [1,2,40].

The symbolic translation of a linguistic term is a number within the interval [-0.5, 0.5) that expresses the domain of a label when it is moving between its two lateral labels as can be observed in Fig. 1. Let us consider a set of labels S representing a fuzzy partition. Formally, we have the pair, $(s_i, \alpha_i), s_i \in S, \alpha_i \in [-0.5, 0.5)$.

There are two different possible methods to perform the lateral tuning: the most interpretable method, the global tuning of the semantics, and the most accurate one, the local tuning of the rules. For SD task the first one is used because the interpretability of the models must be preserved. This method is applied to the level of linguistic partition, where the pair ($s_i$, $\alpha_i$) takes the same tuning value in all the rules where it is considered.

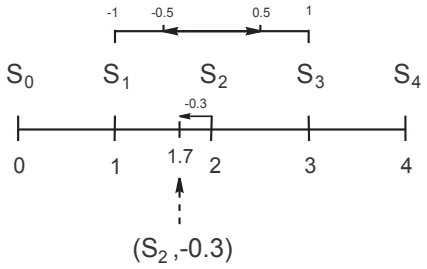


Fig. 1. Symbolic translation of a label

### 4.2. CHC algorithm: the genetic algorithm for tuning

The CHC algorithm [22] is used to design the proposed learning method. It is a genetic algorithm that presents a good trade-off between exploration and exploitation, making it a good choice in problems with complex search spaces. This genetic model makes use of a mechanism of "Selection of Populations": $M$ parents and their corresponding offspring are put together to select the best $M$ individuals to take part in the next population (with $M$ being the population size). In this paper, a single-objective algorithm has been employed because using a multi-objective algorithm mean that the expert would have to choose the best tuning between the different solutions in the Pareto front obtained by the algorithm. On the contrary, a single solution that can be extracted automatically is needed, just what is obtained using a single-objective algorithm. In the specialised literature, some proposals have employed the CHC algorithm to approach this problem [5,3,24].

In CHC, diversity is preserved through an incest prevention mechanism and a restarting approach, instead of using the well-known mutation operator. This incest prevention mechanism is considered in order to apply the crossover operator, i.e., two parents are recombined if their distance (considering an adequate metric) divided by two is above a predetermined threshold, $L$. This threshold value is initialized as the maximum possible distance between two individuals divided by four. Following the original CHC scheme, $L$ is decremented by one when there are no new individuals in the population in one generation. When $L$ is below zero the algorithm restarts the population.

The components needed to design this process are: the coding scheme, the initial gene pool, the chromosome evaluation, the crossover operator and the restarting approach, which are explained below:

- A real coding is considered where the number of variables $m$ which involve in some rules extracted multiplied for the $n$ number of linguistic labels is the size of the chromosome. Then, a chromosome has the form (where each gene is associated with the lateral displacement of the corresponding label in the DB),
  $C_T = (c_1^1, \ldots, c_1^n, \ldots, c_m^1, \ldots, c_m^n)$.
  The values of the tuning parameter must be a real number in $[-0.5, 0.5)$.
- For the evaluation of the chromosome, an aggregation function have been performed with different SD quality measures: Sensitivity (*Sens*, Eq.

$$Fitness(R_i, chrome) = \frac{(Sens(R_i, chrome) * \omega_1) + (Unus(R_i, chrome) * \omega_2) + (FCnf(R_i, chrome) * \omega_3)}{\omega_1 + \omega_2 + \omega_3}. \quad (10)$$

7), Unusualness (*Unus*, Eq. 8) and Fuzzy confidence (*FCnf*, Eq. 5). These measures have been selected to improve the accuracy without affecting the support. The aggregation function is shown in Equation 10. The average of the set of rules for the tuning chromosome is the final fitness value for the chromosome of the CHC algorithm. With respect to the weights for each quality measure, $\omega_1 = 0.2$, $\omega_2 = 0.4$ and $\omega_3 = 0.4$ have been considered. It has been experimentally determined that these weights provide the best combination for the fitness.

- The crossover operator considered is the Parent Centric BLX (PCBLX) operator [30], which is based on the BLX-$\alpha$.
- With respect to the restarting approach, the mechanism presented in [23] is employed when the threshold value L is lower than zero. In this case, all the chromosomes are generated at random within the interval $[-0.5, 0.5)$. Furthermore, the best global solution found is included in the population to increase the convergence of the algorithm.

## 5. Experimental study

In this experimental study the aim was to analyse the performance of using a tuning based in the 2-tuples linguistic approach for the NMEEF-SD algorithm in data sets with continuous variables. The experimentation was undertaken with real data sets from KEEL repository [6,7].

Firstly, the experimental framework is presented in Section 5.1. In Section 5.2 the results obtained by the NMEEF-SD algorithm, with and without the tuning based in the 2-tuples linguistic representation, can be observed. Section 5.3 shows the statistical study performed on these results. Finally, in Section 5.4 an example of lateral tuning in a data set is presented.

### 5.1. Experimental framework

The data sets employed in the experiments performed to analyse the NMEEF-SD algorithm with 2-tuples linguistic approach can be observed in Table 1, where the *Name* of the data set, the number of discrete variables ($n_D$), the number of continuous variables ($n_C$) and the number of *Instances* are shown.

As can be observed in Table 1 the data sets selected are diverse, where data sets with only continuous variables and data sets with both continuous and discrete variables can be found. In addition, data sets with a high number of instances and others with low number of instances have been selected.

Table 1. Data sets employed in the experimental study

| *Name* | $n_D$ | $n_C$ | *Instances* |
|---|---|---|---|
| Appendicitis | 0 | 7 | 106 |
| Australian | 8 | 6 | 690 |
| Balance | 0 | 4 | 625 |
| Echo | 1 | 5 | 131 |
| Glass | 0 | 9 | 214 |
| Haberman | 0 | 3 | 306 |
| Ionosphere | 0 | 34 | 351 |
| Iris | 0 | 4 | 150 |
| Phoneme | 0 | 5 | 5404 |
| Pima | 0 | 8 | 768 |
| Wdbc | 0 | 30 | 569 |
| Wine | 0 | 13 | 178 |

The parameters employed by the algorithms are presented in Table 2. Due to the non-deterministic nature of the NMEEF-SD, the algorithm is executed five times for each data set with a 10-fold cross validation for each data set. The results shown are the average of the results obtained for each data set for the different executions, i.e. the average of the 50 executions.

For the study of the genetic lateral tuning through 2-tuples linguistic representation in NMEEF-SD algorithm, several linguistic labels have been em-

ployed, such as 3 and 5. So, it is interesting to show what is the improvement of the genetic lateral tuning with respect to different number of linguistic labels in a SD algorithm.

Table 2. Parameters of the algorithms employed

| NMEEF-SD |
|---|
| Population size=50 |
| Evaluations=10000 |
| Crossover probability=0.60 |
| Mutation probability=0.1 |
| Re-initialisation based on coverage (50% of biased) |
| Minimum confidence=0.6 |
| Representation of the rule=Canonical |
| **CHC** |
| Population size=50 |
| Evalutions=10000 |
| $\alpha$=0.5 |

### 5.2. Results obtained

A complete study with respect to several quality measures of SD is performed in this paper. Therefore, the quality measures most used for SD are analysed. For each quality measure, the average results obtained by the algorithms in the data sets mentioned in the previous section are shown.

The quality measures shown in Table 3 are the average results for the rule sets, where the quality measures of coverage ($COVE$), significance ($SIGN$), unusualness ($WRACC$), accuracy ($ACCU$), sensitivity ($SENS$), crisp support ($CSUP$), fuzzy support ($FSUP$), crisp confidence ($CCNF$) and fuzzy confidence ($FCNF$) can be observed. Furthermore, the number of linguistic labels ($LL$) used and the name of the $Algorithm$ are presented.

The results in Table 3 show an improvement of the results using the NMEEF-SD algorithm with Lateral Tuning based on 2-tuples (NMEEF-SD-LT), obtaining higher average values in most of the quality measures than for the algorithm without the tunning. Furthermore, it is really interesting the improvement of the values for both the support and the confidence measures, as it is usual the degradation of a measure when improving the other. However, the results obtained for the unusualness measure are better for the algorithm without the genetic lateral tuning; this is because this quality measure has different properties like coverage, accuracy and novelty, and it is very difficult to optimise it through an aggregation function.

This is a preliminary analysis of the results. In next Section, statistical tests are used in order to support this analysis, performing a complete statistical analysis.

### 5.3. Statistical analysis

An analysis was performed in order to find significant differences between both approaches through a non-parametric test, following the recommendations made in [20], providing a set of simple, safe and robust methods for statistical comparisons. In this analysis the Wilcoxom signed-ranks test [42,43] is selected to do the comparison. Detailed information related to this statistical test is available in [26,27] and on the Website `http://sci2s.ugr.es/sicidm/`.

In all the experiments, a level of significance of $\alpha = 0.05$ has been used. The results for the Wilcoxom test are presented in Table 4. This test is performed with respect to the linguistic label ($LL$), quality measure ($Qua_{mea}$) and the comparison NMEEF-SD-LT versus NMEEF-SD algorithm. In addition, in this table positives range ($R^+$), negative ranges ($R^-$) and the $p-Value$ for the test are shown.

The results obtained by the algorithm with genetic lateral tuning are better in accuracy and crisp confidence regardless of the number of labels as can be observed in Table 4, where significant differences are shown. Furthermore, the results obtained by the 2-tuples linguistic representation with 3 linguistic labels gives also significant differences for the quality measure of significance.

As mentioned above, several quality measures have been proposed for the SD task, and usually the more suitable ones depend on the problem. Hence, depending on the objectives to be analysed and the problem to be tackled, the experts must select the most appropriated quality measures. In the experimental study performed, the results obtained by several quality measures for SD have been presented to

Table 3. Results for the simple NMEEF-SD algorithm and with genetic lateral tuning.

| *LL* | *Algorithm* | *COVE* | *SIGN* | *WRACC* | *ACCU* | *SENS* | *CSUP* | *FSUP* | *FCNF* | *CCNF* |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | NMEEF-SD | 0.499 | 3.292 | **0.103** | 0.696 | 0.876 | 0.622 | 0.700 | 0.729 | 0.765 |
| | NMEEF-SD-LT | **0.517** | **4.327** | 0.075 | **0.711** | **0.887** | **0.636** | **0.718** | **0.741** | **0.782** |
| 5 | NMEEF-SD | 0.445 | 3.585 | **0.082** | 0.696 | 0.828 | **0.656** | **0.724** | 0.792 | 0.813 |
| | NMEEF-SD-LT | **0.469** | **3.930** | 0.080 | **0.709** | **0.830** | 0.653 | 0.718 | **0.804** | **0.834** |

analyse the validity of the approach presented.

Table 4. Statistical analysis between NMEEF-SD and NMEEF-SD-LT. The comparison performed is NMEEF-SD-LT Vs. NMEEF-SD

| *LL* | $Qua_{mea}$ | $R^+$ | $R^-$ | $p-Value$ | *Result* |
|---|---|---|---|---|---|
| 3 | *COVE* | 47 | 31 | 0.530 | Accepted |
| | *SIGN* | 66 | 12 | 0.034 | Rejected by NMEEF-SD-LT |
| | *WRACC* | 16 | 62 | 0.071 | Accepted |
| | *ACCU* | 73.5 | 4.5 | 0.007 | Rejected by NMEEF-SD-LT |
| | *SENS* | 27 | 18 | 0.594 | Accepted |
| | *CSUP* | 39 | 27 | 0.594 | Accepted |
| | *FSUP* | 13 | 15 | 0.866 | Accepted |
| | *FCNF* | 52 | 26 | 0.308 | Accepted |
| | *CCNF* | 72 | 6 | 0.010 | Rejected by NMEEF-SD-LT |
| 5 | *COVE* | 44 | 22 | 0.328 | Accepted |
| | *SIGN* | 46 | 20 | 0.248 | Accepted |
| | *WRACC* | 17 | 49 | 0.155 | Accepted |
| | *ACCU* | 57 | 9 | 0.033 | Rejected by NMEEF-SD-LT |
| | *SENS* | 32 | 23 | 0.646 | Accepted |
| | *CSUP* | 25 | 20 | 0.767 | Accepted |
| | *FSUP* | 15 | 21 | 0.674 | Accepted |
| | *FCNF* | 49 | 17 | 0.155 | Accepted |
| | *CCNF* | 59 | 7 | 0.021 | Rejected by NMEEF-SD-LT |

The results of the experimental study performed show that the use of genetic lateral tuning in NMEEF-SD allows an improvement in the accuracy of the model without degrading the support. In addition, in other quality measures like significance, coverage and sensitivity, the average values obtained for the approach with genetic lateral tuning are better than for the algorithm without this tuning. As a consequence, the use of the genetic lateral tuning approach in the algorithm facilitates the usual objective of experts to obtain accurate results.

### *5.4. Case study: Genetic lateral tuning in the data set Iris*

The main objective of this study is to show graphically the conclusions obtained in the previous analysis. Therefore, a complete example of rules and results for the approach presented in this paper is shown for the data set *Iris*.

It should be noted that, as a global tuning of the semantic is used in order to preserve the interpretability of the model, the same pair "tuning parameter-variable" is used for the same variable in all the subgroups generated by the model.

In Table 5 the complete set of rules (for an execution) extracted by the NMEEF-SD algorithm can be observed. In this case, only the variables *petalLength* and *petalWidth* involve in the rules. In addition, the tuning parameter ($T$) obtained in the genetic lateral tuning using CHC is also shown.

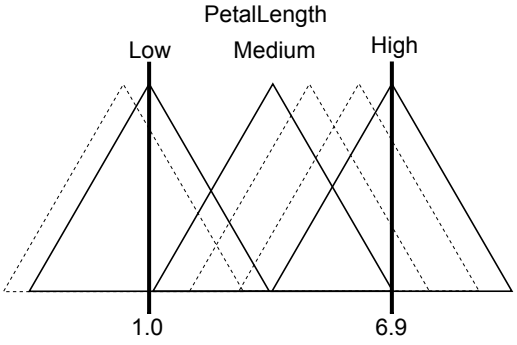


Fig. 2. Initial and lateral tuning for the variable PetalLength in the iris data set

Graphically, the tuning parameter can be observed in figures: Fig. 2 shows the use of genetic lateral tuning with global tuning semantics in the *petalLenght* variable and Fig. 3 in the *petalWidth*. The membership functions are contextualized for each one of the activated variables in the set of subgroups, adapting the fuzzy system to the problem,

Table 5. Rules extracted for the NMEEF-SD algorithm and the tuning parameter obtained in the genetic lateral tuning for the data set *Iris*.

| *Rule* | *Parameter* |
|---|---|
| 1: $IF\ petalLength = LL_1 \rightarrow Iris - setosa$ | $T_{pL} = (LL_1, -0.152)$ |
| 2: $IF\ petalWidth = LL_1 \rightarrow Iris - setosa$ | $T_{pW} = (LL_1, -0.063)$ |
| 3: $IF\ petalLength = LL_1\ AND\ petalWidth = LL_1 \rightarrow Iris - setosa$ | $T_{pL} = (LL_1, -0.152)$, $T_{pW} = (LL_1, -0.063)$ |
| 4: $IF\ petalLength = LL_2\ AND\ petalWidth = LL_2 \rightarrow Iris - versicolor$ | $T_{pL} = (LL_2, 0.274)$, $T_{pW} = (LL_2, -0.319)$ |
| 5: $IF\ petalLength = LL_3 \rightarrow Iris - virginica$ | $T_{pL} = (LL_3, -0.212)$ |
| 6: $IF\ petalWidth = LL_3 \rightarrow Iris - virginica$ | $T_{pW} = (LL_3, -0.135)$ |

Table 6. Results for the simple NMEEF-SD algorithm and with genetic lateral tuning for the rules extracted in the data set *Iris*.

| | *Algorithm* | *COVE* | *SIGN* | *WRACC* | *ACCU* | *SENS* | *CSUP* | *FSUP* | *FCNF* | *CCNF* |
|---|---|---|---|---|---|---|---|---|---|---|
| *Rule* 1 | NMEEF-SD | 0.333 | 4.771 | 0.222 | 0.750 | 1.000 | 0.333 | 0.333 | 0.954 | 1.000 |
| | NMEEF-SD-LT | 0.333 | 4.771 | 0.152 | 0.750 | 1.000 | 0.333 | 0.228 | 0.999 | 1.000 |
| *Rule* 2 | NMEEF-SD | 0.333 | 4.771 | 0.222 | 0.750 | 1.000 | 0.333 | 0.333 | 0.916 | 1.000 |
| | NMEEF-SD-LT | 0.333 | 4.771 | 0.183 | 0.750 | 1.000 | 0.333 | 0.284 | 0.936 | 1.000 |
| *Rule* 3 | NMEEF-SD | 0.333 | 4.771 | 0.222 | 0.750 | 1.000 | 0.333 | 0.333 | 0.953 | 1.000 |
| | NMEEF-SD-LT | 0.333 | 4.771 | 0.152 | 0.750 | 1.000 | 0.333 | 0.228 | 0.999 | 1.000 |
| *Rule* 4 | NMEEF-SD | 0.466 | 3.042 | 0.177 | 0.600 | 1.000 | 0.333 | 0.333 | 0.600 | 0.714 |
| | NMEEF-SD-LT | 0.333 | 2.597 | 0.100 | 0.625 | 1.000 | 0.266 | 0.333 | 0.701 | 0.800 |
| *Rule* 5 | NMEEF-SD | 0.066 | 0.954 | 0.044 | 0.500 | 1.000 | 0.066 | 0.333 | 0.753 | 1.000 |
| | NMEEF-SD-LT | 0.333 | 4.771 | 0.103 | 0.750 | 1.000 | 0.333 | 0.205 | 0.669 | 1.000 |
| *Rule* 6 | NMEEF-SD | 0.133 | 1.908 | 0.088 | 0.600 | 1.000 | 0.133 | 0.333 | 0.848 | 1.000 |
| | NMEEF-SD-LT | 0.266 | 3.816 | 0.115 | 0.714 | 1.000 | 0.266 | 0.200 | 0.785 | 1.000 |
| *Average* | NMEEF-SD | 0.277 | 3.369 | 0.162 | 0.658 | 1.000 | 0.866 | 1.000 | 0.837 | 0.952 |
| | NMEEF-SD-LT | 0.322 | 4.249 | 0.134 | 0.723 | 1.000 | 0.933 | 1.000 | 0.848 | 0.967 |

obtaining in this way better results.

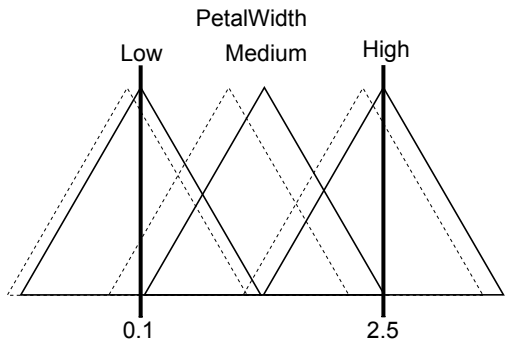


Fig. 3. Initial and lateral tuning for the variable PetalWidth in the iris data set

Finally, the results associated to the previously shown rules can be observed in Table 6, where the results for the NMEEF-SD algorithm and NMEEF-SD with genetic lateral tuning are shown. In addition, the average results for the set of rules are shown.

The results presented in Table 6 for data set *Iris* support the conclusions previously mentioned in the analysis: NMEEF-SD algorithm with genetic lateral tuning is able to improve the accuracy of the model without degrading the support. In addition, improvements in coverage and significance can be observed.

## 6. Conclusions

In this paper, the application of the genetic lateral tuning based on 2-tuples linguistic representation to a SD algorithm, NMEEF-SD, to improve the performance of this algorithm has been introduced. As far

as we know, this is the first application of this widely known technique in the data mining task of SD. The process introduces a lateral tuning in the membership functions of the linguistic labels of the variables to improve their global interaction. Moreover, this tuning allows to maintain the interpretability of the subgroups, which is one of the most important dimensions in SD.

The genetic tuning using 2-tuples linguistic representation has been implemented with a CHC algorithm using the most interpretable lateral tuning: the global tuning of the semantics. This technique has been applied because in the SD task the interpretability is one of the most important aspects. It is applied to the linguistic partitions level where the tuning value for the label is the same in all the rules where it is considered.

The results obtained by the NMEEF-SD with the tuning using the 2-tuples linguistic representation show in general better results with respect to the initial partitions, highlighting the improvement in accuracy without degrading the support. This improvement in both dimensions shows a good behaviour of this technique. This statement is also supported by the statistical results obtained, which show in conclusion that the genetic tuning through 2-tuples linguistic representation improves the performance of the NMEEF-SD algorithm.

In addition, the final results have also demonstrated that, regardless of the number of linguistic labels employed per variable, the results are improved. Nevertheless, the choice of a lower level of granularity performs better.

## Acknowledgments

This paper was supported by the Spanish Ministry of Education, Social Policy and Sports under projects TIN-2008-06681-C06-02, and by the Andalusian Research Plan under project TIC-3928.

## References

1. R. Alcalá, J. Alcalá-Fdez, M. J. Gacto, and F. Herrera, *Rule base reduction and genetic tuning of fuzzy systems based on the linguistic 3-tuples representation*, Soft Computing **11** (2007), no. 5, 401–419.
2. R. Alcalá, J. Alcalá-Fdez, and F. Herrera, *A proposal for the genetic lateral tuning of linguistic fuzzy systems and its interaction with rule selection*, IEEE Transactions on Fuzzy Systems **15** (2007), no. 4, 616–635.
3. R. Alcalá, J. Alcalá-Fdez, F. Herrera, and J. Otero, *Genetic learning of accurate and compact fuzzy rule based systems based on the 2-tuples linguistic representation*, International Journal of Approximate Reasoning **44** (2007), no. 1, 45–64.
4. R. Alcalá, P. Ducange, F. Herrera, B. Lazzerini, and F. Marcelloni, *A Multi-Objective Evolutionary Approach to Concurrently Learn Rule and Data Bases of Linguistic Fuzzy Rule-Based Systems*, IEEE Transactions on Fuzzy Systems **17** (2009), no. 5, 1106–1122.
5. J. Alcalá-Fdez, R. Alcalá, M. J. Gacto, and F. Herrera, *Learning the Membership Function Contexts for Mining Fuzzy Association Rules by Using Genetic Algorithms*, Fuzzy Sets and Systems **160** (2009), no. 7, 905–921.
6. J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, *KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework*, Journal of Multiple-Valued Logic and Soft Computing **17** (2011), no. 2-3, 255–287.
7. J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, and F. Herrera, *KEEL: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems*, Soft Computing **13** (2009), no. 3, 307–318.
8. M. Atzmueller and F. Puppe, *SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery*, Proceedings of the 17th European Conference on Machine Learning and 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, LNCS, vol. 4213, Springer, 2006, pp. 6–17.
9. F. J. Berlanga, M. J. del Jesus, P. González, F. Herrera, and M. Mesonero, *Multiobjective Evolutionary Induction of Subgroup Discovery Fuzzy Rules: A Case Study in Marketing*, Proceedings of the 6th Industrial Conference on Data Mining, LNCS, vol. 4065, Springer, 2006, pp. 337–349.
10. C. J. Carmona, P. González, M. J. del Jesus, and F. Herrera, *NMEEF-SD: Non-dominated Multiobjective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery*, IEEE Transactions on Fuzzy Systems **18** (2010), no. 5, 958–970.
11. C. J. Carmona, P. González, M. J. del Jesus, M. Navío, and L. Jiménez, *Evolutionary Fuzzy Rule Extraction for Subgroup Discovery in a Psychiatric Emergency Department*, Soft Computing, in press, doi: 10.1007/s00500-010-0670-3.
12. J. Casillas and B. Carse, *Special issue on Genetic*

*Fuzzy Systems: Recent Developments and Future Directions*, Soft Computing **13** (2009), no. 5, 417–418.
13. O. Cordón, R. Alcalá, J. Alcalá-Fdez, and I. Rojas, *Special Issue on Genetic Fuzzy Systems: What's Next?*, Editorial, IEEE Transactions on Fuzzy Systems **15** (2007), no. 4, 533–535.
14. O. Cordón, M. J. del Jesus, F. Herrera, and M. Lozano, *MOGUL: A Methodology to obtain genetic fuzzy rule-based systems under the iterative rule learning approach*, Internation Journal of Intelligent Systems **14** (1999), 1123–1153.
15. O. Cordón, F. A. C. Gomide, F. Herrera, F. Hoffmann, and L. Magdalena, *Ten years of genetic fuzzy systems. Current framework and new trends*, Fuzzy Sets and Systems **14** (2004), 5–31.
16. O. Cordón, F. Herrera, F. Hoffmann, and L. Magdalena, *Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases*, World Scientific, 2001.
17. K. Deb, A. Pratap, S. Agrawal, and T. Meyarivan, *A fast and elitist multiobjective genetic algorithm: NSGA-II*, IEEE Transactions Evolutionary Computation **6** (2002), no. 2, 182–197.
18. M. J. del Jesus, P. González, and F. Herrera, *Multiobjective Genetic Algorithm for Extracting Subgroup Discovery Fuzzy Rules*, Proceedings of the IEEE Symposium on Computational Intelligence in Multicriteria Decision Making, IEEE Press, 2007, pp. 50–57.
19. M. J. del Jesus, P. González, F. Herrera, and M. Mesonero, *Evolutionary Fuzzy Rule Induction Process for Subgroup Discovery: A case study in marketing*, IEEE Transactions on Fuzzy Systems **15** (2007), no. 4, 578–592.
20. J. Demsar, *Statistical comparisons of classifiers over multiple data sets*, Journal Machine Learning Research **7** (2006), 1–30.
21. A. E. Eiben and J. E. Smith, *Introduction to evolutionary computation*, Springer, 2003.
22. L. J. Eshelman, *Foundations of genetic algorithms*, ch. The CHC Adaptive Search Algorithm: How to have Safe Search When Engaging in Nontraditional Genetic Recombination, pp. 265–283, 1991.
23. L. J. Eshelman and J. D. Schaffer, *Preventing premature convergence in genetic algorithms by preventing incest*, Proceedings of the 4th International Conference on Genetic Algorithms, 1991, pp. 115–122.
24. A. Fernández, M. J. del Jesus, and F. Herrera, *On the 2-Tuples Based Genetic Tuning Performance for Fuzzy Rule Based Classification Systems in Imbalanced Data-SetsInformation Sciences 180:8 (2010) 1268-1291*, Information Sciences **180** (2010), no. 8, 1268–1291.
25. D. Gamberger and N. Lavrac, *Expert-Guided Subgroup Discovery: Methodology and Application*, Journal Artificial Intelligence Research **17** (2002), 501–527.
26. S. García, A. Fernández, J. Luengo, and F. Herrera, *Study of Statistical Techniques and Performance Measures for Genetics-Based Machine Learning: Accuracy and Interpretability*, Soft Computing **13** (2009), no. 10, 959–977.
27. S. García, A. Fernandez, J. Luengo, and F. Herrera, *Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental Analysis of Power*, Information Sciences **180** (2010), 2044–2064.
28. F. Herrera, *Genetic fuzzy systems: taxomony, current research trends and prospects*, Evolutionary Intelligence **1** (2008), 27–46.
29. F. Herrera, C. J. Carmona, P. González, and M. J. del Jesus, *An overview on Subgroup Discovery: Foundations and Applications*, Knowledge and Information Systems, in press, doi: 10.1007/s10115-010-0356-2.
30. F. Herrera, M. Lozano, and A.M. Sánchez, *A taxonomy for the crossover operator for real-coded genetic algorithms: an experimental study*, International Journal of Intelligent Systems **18** (2003), 309–338.
31. F. Herrera and L. Martínez, *A 2-tuple fuzzy linguistic representation model for computing with words*, IEEE Transactions on Fuzzy Systems **8** (2000), no. 6, 746–752.
32. H. Ishibuchi, T. Nakashima, and M. Nii, *Classification and modeling with linguistic information granules: Advanced approaches to linguistic data mining*, Springer, 2004.
33. B. Kavsek and N. Lavrac, *APRIORI-SD: Adapting association rule learning to subgroup discovery*, Applied Artificial Intelligence **20** (2006), 543–583.
34. W. Kloesgen, *Explora: A Multipattern and Multistrategy Discovery Assistant*, Advances in Knowledge Discovery and Data Mining, American Association for Artificial Intelligence, 1996, pp. 249–271.
35. L. Kuncheva, *Fuzzy classifier design*, Springer, 2000.
36. N. Lavrac, B. Cestnik, D. Gamberger, and P. A. Flach, *Decision Support Through Subgroup Discovery: Three Case Studies and the Lessons Learned*, Machine Learning **57** (2004), no. 1-2, 115–143.
37. N. Lavrac, P. A. Flach, and B. Zupan, *Rule Evaluation Measures: A Unifying View*, Proceedings of the 9th International Workshop on Inductive Logic Programming, LNCS, vol. 1634, Springer, 1999, pp. 174–185.
38. N. Lavrac, B. Kavsek, P. A. Flach, and L. Todorovski, *Subgroup Discovery with CN2-SD*, Journal of Machine Learning Research **5** (2004), 153–188.
39. Y. Nojima, R. Alcalá, H. Ishibuchi, and F. Herrera, *Special Issue on Evolutionary Fuzzy Systems*, SoftComputing, in press, doi: 10.1007/s00500-010-0663-2.
40. I. Robles, R. Alcalá, J. M. Benítez, and F. Herrera, *Evolutionary Parallel and Gradually Distributed Lat-*

*eral Tuning of Fuzzy Rule-Based Systems*, Evolutionary Intelligence **2** (2009), 5–19.
41. C. Romero, P. González, S. Ventura, M. J. del Jesus, and F. Herrera, *Evolutionary algorithm for subgroup discovery in e-learning: A practical application using Moodle data*, Expert Systems with Applications **36** (2009), 1632–1644.
42. D. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, 2 ed., Chapman and Hall/CRC, 2006.
43. F. Wilcoxom, *Individual comparisons by ranking methods*, Biometrics **1** (1945), 80–83.
44. S. Wrobel, *An Algorithm for Multi-relational Discovery of Subgroups*, Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery, LNAI, vol. 1263, Springer, 1997, pp. 78–87.
45. L. A. Zadeh, *The concept of a linguistic variable and its applications to approximate reasoning. Parts I, II, III*, Information Science **8-9** (1975), 199–249,301–357,43–80.
46. E. Zitzler, M. Laumanns, and L. Thiele, *SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimization*, International Congress on Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems, 2002, pp. 95–100.