IEEE SSCI 2011

APRIL 11-15, 2011
PARIS, FRANCE

SYMPOSIUM SERIES ON COMPUTATIONAL INTELLIGENCE

www.ieee-ssci.org

GEFS 2011

2011 IEEE 5th International Workshop on Genetic and Evolutionary Fuzzy Systems



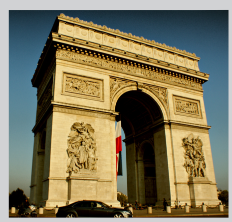

Table of Contents

Technical Sessions

Author Index

ISBN: 978-1-61284-048-2
IEEE Catalog Number: CFP1195A-CDR

Technical Support:
Chris Dyer
Conference Catalysts, LLC
Phone: +1 785 341 3583
cdyer@conferencecatalysts.com

IEEE

IEEE Computational Intelligence Society

Organized and sponsored by the IEEE Computational Intelligence Society

2011 IEEE 5th International Workshop on Genetic and Evolutionary Fuzzy Systems

(GEFS 2011) Proceedings

Additional copies may be ordered from:
IEEE Service Center
445 Hoes Lane
Piscataway, NJ 08855-1331 USA

+1 800 678 IEEE (+1 800 678 4333)
+1 732 981 1393
+1 732 981 9667 (FAX)
email: customer-service@ieee.org

IEEE Catalog Number: CFP1195A-CDR
ISBN: 978-1-61284-048-2

TABLE OF CONTENTS

11:00 - 12:00

S109: GEFS - Keynote
Chair: Rafael Alcalá (University of Granada, Spain)

Multi-objective Evolutionary Learning of Fuzzy Rule-based Systems for Regression Problems
Francesco Marcelloni (University of Pisa, Italy)

14:00 - 16:00

S110: Regression and Control
Chair: Yusuke Nojima (Osaka Prefecture University, Japan)

16:30 - 17:30

S111: Applications
Chair: Yusuke Nojima (Osaka Prefecture University, Japan)

Analysis of the Impact of Using Different Diversity Functions for the Subgroup Discovery Algorithm NMEEF-SD

Cristóbal J. Carmona, Pedro González, Maria José del Jesus
Department of Computer Science
University of Jaen
Jaen, Spain
ccarmona@ujaen.es, pglez@ujaen.es, mjjesus@ujaen.es

Francisco Herrera
Department of Computer Science
and Artificial Intelligence
University of Granada
Granada, Spain
herrera@decsai.ugr.es

***Abstract*—A main purpose of a multi-objective evolutionary algorithm is to find a good relationship between convergence and diversity of the population. Convergence guides the algorithm to search the optimal solution and diversity tries to avoid a premature stagnation of the search. In multi-objective evolutionary algorithms, diversity has been promoted using different techniques.**

In this paper, several diversity functions were implemented in NMEEF-SD, an algorithm for the extraction of fuzzy rules in a subgroup discovery task, to analyse the influence of these functions in the evolutionary process. The results show the advantages of the different measures, depending on the intended objective.

***Index Terms*—Subgroup Discovery, Evolutionary Fuzzy System, NMEEF-SD, NSGA-II.**

I. Introduction

Within the Knowledge Discovery in Databases (KDD) process, the data mining stage is responsible for the automatic discovery of high level knowledge obtained from real data [1]. A data mining algorithm can discover knowledge using different representation models and techniques from two different perspectives:

- Predictive induction, whose objective is the discovery of knowledge for classification or prediction [2].
- Descriptive induction, whose main objective is the discovery of interesting knowledge from the data.

Subgroup Discovery (SD) [3], [4] is a descriptive data mining task including some features of predictive data mining which has recently received a lot of attention from researchers. The goal of SD is the discovery of interesting individual patterns in relation to a specific property which is of interest to the user, in form of rules.

The SD task has been successfully tackled [5], [6] using evolutionary fuzzy systems (EFS) [7]–[9], a hybridisation between evolutionary algorithms [10] and fuzzy logic [11]. A genetic algorithm (GA) is a type of EFS which performs a thorough exploration of the search space, also handling appropriately the relations between variables. Therefore, GAs develop searches particularly suited to rule extraction. The use of fuzzy logic by means of descriptive fuzzy rules allows the representation and use of knowledge in a similar way to human reasoning, and the obtaining of more interpretable and actionable solutions in the field of SD, and in general in the analysis of data to establish relationships and identify patterns [12].

As in many other optimization problems, the induction of fuzzy rules describing subgroups involves several objectives to be considered simultaneously. In SD, these objectives are expressed by means of different quality measures which can be used for the evaluation of a rule. However, these objectives are conflicting, and it is not possible to obtain a single better solution with respect to all the objectives. Therefore, the induction of SD rules can be considered a multi-objective problem rather than a single objective one. Multi-objective evolutionary algorithms (MOEAs) are adapted to solve problems in which different objectives must be optimized [13], [14]. A high-quality exponent of this type of MOEAs is NSGA-II [15], which has been widely used in EFSs [16].

The main factor in the search performed by a MOEA is the relationship between convergence and diversity of the population [17]. Convergence guides the algorithm to search the optimal solution and diversity tries to avoid a premature stagnation of the search, hence preventing the algorithm from falling into a local maximum. Therefore, diversity is crucial to the ability of the algorithm to continue the exploration of the search space [18]. If the population loses diversity too early, the search is likely to be trapped in a region not containing the global optimum. This problem is called premature convergence [19]. In MOEAs based in the NSGA-II approach, the use of the crowding distance promotes the diversity in the individuals of the population, guiding the selection process of the algorithm towards an uniformly spread-out Pareto optimal front. However, alternative diversity measures have been presented for the NSGA-II approach [20].

The main objective of this work is to analyse the impact of using different diversity functions in the results of the Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery (NMEEF-SD) [6], an EFS for the extraction of fuzzy rules for the SD task.

978-1-61284-048-2/11/$26.00 ©2011 IEEE

To do so, the paper is organised as follows: SD task and EFSs used for SD are presented in Section II and Section III respectively. In Section IV, NMEEF-SD algorithm is briefly described together with the different measures defined to promote the diversity in the algorithm. Section V analyses the results obtained by the algorithm using the different measures. Finally, conclusions are outlined in Section VI.

II. Subgroup discovery

The concept of SD was initially introduced by Kloesgen [3] and Wrobel [4], and more formally defined by Siebes [21] (using the name Data Surveying for the discovery of interesting subgroups). It can be defined as [22]:

> *"In subgroup discovery, we assume we are given a so-called population of individuals (objects, customer, ...) and a property of those individuals we are interested in. The task of subgroup discovery is then to discover the subgroups of the population that are statistically "most interesting", i.e., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest."*

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:

$$R : Cond \ \rightarrow \ Class$$

where the property of interest is the $Class$ that appears in the consequent part of the rule, and the antecedent part of the rule, $Cond$, is a conjunction of features (attribute-value pairs) selected from the features describing the training instances [23], [24].

A model corresponding to a subgroup in a problem with two classes (x and o) can be found in Fig. 1, where the rule defining the subgroup corresponds to class x. The model defined by the rule is very simple (represented in the figure as a circle) and therefore it is very interpretable. However, the model covers a high number of objects of class x, but not all of them, also including objects corresponding to the other class, o. This illustrates one of the main features of the SD task: it is normally preferred to obtain a simple model rather than a completely precise one.

The interested reader can find in [25] a recent review describing the main properties of the SD task, its most used quality measures, the available approaches in the literature to approach this problem, and the main applications in real-world problems.

III. Evolutionary fuzzy systems applied to subgroup discovery

EFSs are essentially fuzzy systems enhanced by a learning process based on an evolutionary algorithm [8], [9]. Currently, EFSs are being applied to a wide range of real-world problems.

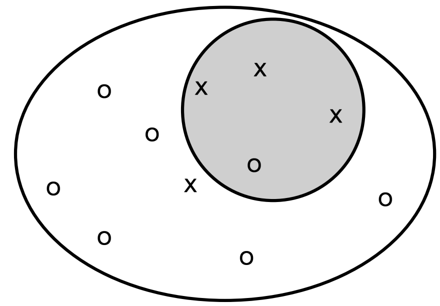

Fig. 1. Representation of a SD rule for the class x

The research related to this area is growing, and a number of open problems and future directions can be found in [16], [26], [27].

Evolutionary algorithms [28] are employed because they are ideal techniques to solve search and optimization problems. Nowadays they are robust, flexible and tend to cope well with attribute interactions. On the other hand, fuzzy systems are one of the most important areas for the application of the fuzzy set theory [11], [29]. Fuzzy sets correspond to linguistic labels which are defined by means of their corresponding membership functions. These can be specified by the user or defined through uniform partitions.

There is a large body of literature which focuses on the extraction of fuzzy rules in descriptive data mining. This has been widely applied to association rule extraction [30]–[35]. The use of fuzzy sets in fuzzy rules 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 attribute domains.

Proposals for the extraction of fuzzy rules in a SD task through EFSs include several works:

- New SD algorithms presented such as SDIGA [5], a genetic algorithm based on the IRL approach, MESDIF [36], a multi-objective genetic algorithm based on SPEA-2 (a MOEA approach), and NMEEF-SD [6], a new multi-objective SD algorithm explained in the next section.
- Applications to real-world problems in marketing [5], [37], e-learning [38], [39], and psychiatric emergencies [40].
- The use of canonical or disjunctive normal form rules for the NMEEF-SD algorithm [41].
- The representation of fuzzy models for SD [42].

IV. NMEEF-SD: Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery

Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery (NMEEF-SD) [6] is an EFS 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 includes some quality measures in order to obtain rules with suitable values not only in the quality measures used but also in the rest of the most used quality measures in SD. The best way to obtain solutions with a good compromise between several quality measures for SD is through a MOEA approach. In this sense, NMEEF-SD has a multi-objective approach based on NSGA-II [15], a MOEA based on a non-dominated sorting approach, and on the use of elitism. NMEEF-SD is oriented towards the SD task and uses specific operators to promote the extraction of simple, interpretable and high quality SD rules. The algorithm permits a number of quality measures to be used both for the selection and the evaluation of rules within the evolutionary process.

As the general objective of NMEEF-SD is to obtain a set of general and accurate rules, the algorithm includes components to enhance these characteristics. In particular, diversity is enhanced in the population using a new operator which performs a re-initialisation based on coverage. In addition, the algorithm employs a niching technique, the crowding distance, for the selection of the rules. In this study, a comparison among different measures promoting the diversity of the population is presented, in order to obtain the best compromise between the objectives of the MOEA. On the other hand, to promote generalisation, as well as the objectives considered in the evolutionary approach, the algorithm includes operators of biased initialisation and biased mutation. Finally, to ensure accuracy, in addition to the objectives, NMEEF-SD returns as its final solution those rules which reach a predetermined confidence threshold.

With respect to the rule structure, NMEEF-SD uses fuzzy logic to represent the continuous variables, by means of linguistic variables. In data mining processes, this allows the use of numerical features without the need of a previous discretisation, so increasing the interpretability of the extracted knowledge. Continuous variables are considered linguistic ones and the fuzzy sets corresponding to the linguistic labels can be specified by the user or defined by means of a uniform partition, if the expert knowledge is not available. In this paper, uniform partitions with triangular membership functions are used, as shown in Fig. 2 for a variable m with five linguistic labels: $X_m : \{LL_m^1, LL_m^2, \ldots, LL_m^5\}$.

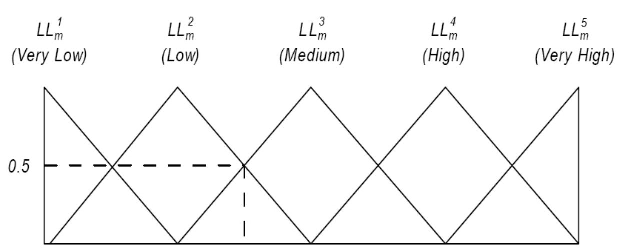

Fig. 2. Example of fuzzy partition for a continuous variable with five labels

A fuzzy rule describing a subgroup is represented in NMEEF-SD as:

$$R: \; If \; X_1 \; is \; LL_1^2 \; and \; X_7 \; is \; LL_7^1 \; then \; Class_j \quad (1)$$

where variable X_1 takes the second linguistic label (LL_1^2) as its value, and variable X_7 takes the first linguistic label (LL_7^1). In addition, 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 class. Therefore, in order to obtain subgroups describing knowledge on all the values of the target variable, the algorithm must be executed as many times as different values the target variable contains.

NMEEF-SD uses an integer representation model with as many genes as variables are contained in the original data set, not including the target variable (as it is prefixed in the algorithm). The set of possible values for the categorical features is that indicated by the problem, and for numerical variables it is the set of linguistic terms determined heuristically or with expert information. In Fig. 3 we can observe the representation of a rule with continuous and discrete variables for the class $Positive$.

$$Genotype$$

$$\begin{array}{|c|c|c|c|} v_1 & v_2 & v_3 & v_4 \\ 1 & 0 & 2 & 0 \end{array}$$

$$\Downarrow$$

$$Phenotype$$

IF ($v_1 = Low$) AND ($v_3 = 14$) THEN ($Class = Positive$)

Fig. 3. Representation of a fuzzy rule with continuous and categorical variables in NMEEF-SD

The objectives considered in NMEEF-SD are defined by means of the following quality measures:

- *Support based on examples of the class*, is a crisp measure defined as the coverage of the rule on the examples of that class [5]:

$$Sup_c(R_i) = Sup_c(Cond_i \rightarrow Class_j) = \frac{n(Class_j \cdot Cond_i)}{n(Class_j)} \quad (2)$$

  where $n(Class_j)$ is the number of examples of the class, and $n(Class_j \cdot Cond_i)$ is the number of examples which satisfy the conditions and also belong to the class.
- *Unusualness of a rule*, defined as the weighted relative accuracy of a rule [43]:

$$Unus(R_i) = Unus(Cond_i \rightarrow Class_j) = \frac{n(Cond_i)}{n_s}\left(\frac{n(Class_j \cdot Cond_i)}{n(Cond_i)} - \frac{n(Class_j)}{n_s}\right) \quad (3)$$

  where $n(Cond_i)$ is the number of examples which satisfy the antecedent, and n_s is the number of examples of the data set. The weighted relative accuracy of a rule can be described as the balance between the coverage of the rule and its accuracy gain.

The performance of NMEEF-SD algorithm relies on two measures when comparing individuals, which come to the approach NSGA-II: the non-dominated ranking and the crowding distance measure. NSGA-II uses the crowding distance to guide the selection process at the various stages of the algorithm towards a uniformly spread-out Pareto optimal front. This measure promotes the diversity in the individuals included in the main population of the next generation. However, in [20] two other diversity measures were presented for the NSGA-II approach in order to find knees in the Pareto front, by modifying the crowding distance measure: angle measure and utility-based measure. Below, the crowding distance measure and these new measures are described.

A. Crowding distance measure

The crowding distance is defined as the average distance of two points on either side of this point along each one of the objectives. This quantity serves as an estimate of the perimeter of the cuboid formed by using the nearest neighbours as vertices. In Fig. 4, the crowding distance of the ith solution in its front (marked with circles) is the average side length of the cuboid (shown with a dashed box), using support (2) and unusualness (3) as objectives on which to calculate the distances.

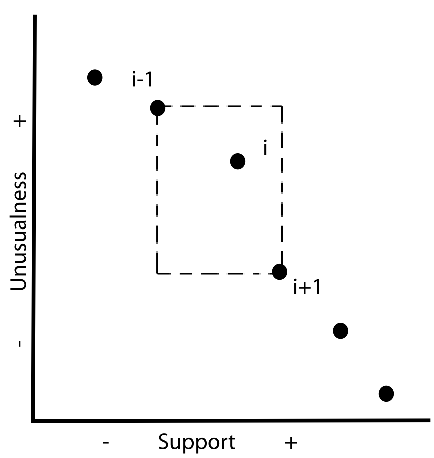

Fig. 4. Crowding distance calculation. Points marked in filled circles are solutions of the same non-dominated front

The computation of the crowding distance requires sorting the population according to each value of the objective function in ascending order of magnitude. Thereafter, for each objective function, the boundary solutions (solutions with smallest and largest function values) are assigned an infinite distance value. All other intermediate solutions are assigned a distance value equal to the absolute normalized difference in the function values of two adjacent solutions. The overall crowding distance value is calculated as the sum of the individual distance values corresponding to each objective. Each objective function is normalized before calculating the crowding distance.

B. Angle measure

The search of an angle could represent the best compromise between the quality measures defined as objectives in the multi-objective evolutionary approach. The tradeoff between two objectives can be estimated by the slopes of the two lines through an individual and its two neighbours. The angle between these slopes can be regarded as an indication or whether the individual is at a knee or not [20]. In Fig. 5 an angle calculation can be observed, where greater degree angles are preferred.

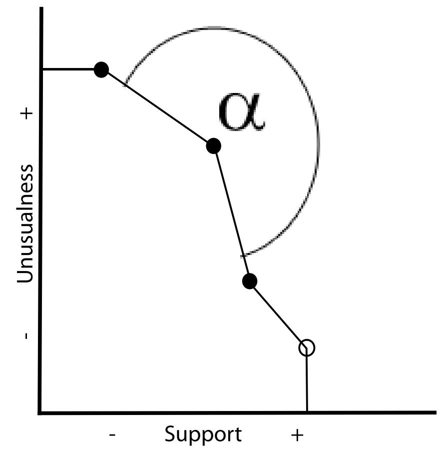

Fig. 5. Calculation of the angle measure

The angle of an individual is calculated through the angle between the individual and its two neighbours. These three individuals have to be pairwise linearly independent, thus duplicate individuals are treated as one and are assigned the same angle-measure. Individuals at the end of the Pareto, i.e. individuals without neighbours, are complemented with an horizontal or vertical line to calculate the angle.

The optimal calculation of this measure is achieved using two objectives such as the ones used by NMEEF-SD: support based on the examples of the class (2) and unusualness (3). The values of these quality measures must be normalized in order to calculate fair angle values. The search of the angle measure for more than two objectives becomes impractical, even finding the neighbours.

C. Utility-based measure

This measure was presented in [20] and defined as the marginal utility that a solution provides to a decision maker, assuming linear utility functions of the form $U(C, \lambda) = \lambda f_1(C) + (1 - \lambda) f_2(C)$, with all $\lambda \in [0, 1]$ being equally likely.

An individual's marginal utility, $U'(C, \lambda')$, is defined as the additional cost the decision maker would have to accept if that particular individual would not be available and he should have to settle for the second best, i.e.: $U'(C^i, \lambda') = min_{j \neq i} U(C^j, \lambda') - U(C^i, \lambda')$, where $i = argmin U(C^j, \lambda')$; otherwise $U'(C^i, \lambda') = 0$.

This measure can be calculated for all individuals for a number of randomly chosen utility functions, taking the average as the expected marginal utility. Sampling can be done either randomly or, as was proposed in [20] in order to reduce variance, in a systematic manner. The number of utility functions used for approximation is called the precision of the measure. Authors recommend a precision of at least the number of individuals of the population. Naturally, individuals with the largest overall marginal utility are preferred. In our case, the measure have been computed by sampling, considering equidistant values for λ and a precision of exactly the number of individuals in the population. As for the angle measure, the quality measures must be normalized in order to calculate fair utility values.

V. EXPERIMENTATION

The main objective of this experimentats is to study the influence in the results of the NMEEF-SD algorithm of using different diversity measures introduced in the bibliography for the NSGA-II based algorithms. To do this, a comparative study of the NMEEF-SD algorithm using the different measures defined in Section IV (crowding measure, angle measure and utility-based measure) is performed, using a set of real data sets.

Thus, the experimental framework can be observed in Section V-A, and Section V-B presents the results obtained by the algorithm with the different diversity measures, and the analysis of these results.

A. Experimental framework

The experimentation was undertaken with real data sets from UCI repository [44]. The main properties of the data sets used (number of variables (n_v), number of discrete variables (n_{vD}), number of continuous variables (n_{vC}), number of classes of the data set (n_c) and number of examples (n_s)) are presented in Table II.

TABLE II
PROPERTIES OF THE DATA SETS USED FROM THE UCI REPOSITORY

$Name$	n_v	n_{vD}	n_{vC}	n_c	n_s
Australian	14	8	6	2	690
Breast	9	9	0	2	699
Bridges	7	4	3	2	102
German	20	13	7	2	1000
Heart	13	6	7	2	270
Hepatitis	19	13	6	2	155
Hypothyroid	25	18	7	2	3163
Ionosphere	34	0	34	2	351

A ten fold cross-validation (10-fcv) procedure has been used to perform the comparisons for each data set. For each fold, the following parameters are used in the NMEEF-SD algorithm: *Executions*=5, *Population size*=50, *Evaluations*=10000, *Crossover probability*=0.60, *Mutation probability*=0.1, *Re-initialisation based on coverage with 50% of biased*, *Minimum confidence*=0.7, *Representation of the rule*=Canonical and *Linguistic labels*=3.

As NMEEF-SD is a non-deterministic algorithm, 5 executions are carried out for each experiment. The results shown in Table I are the average of the results obtained for each data set for the different executions, i.e. the results shown are the average of the 50 results obtained (5 executions $\star$ 10 fold).

B. Analysis of the results obtained

Table I shows the results obtained by NMEEF-SD with the different diversity measures used (crowding measure, angle measure and utility-based measure), where $Dataset$ is the name of the data set, $Diversity$ is the name of the diversity measure employed, $\sharp Rules$ is the average number of rules, $\sharp Variables$ is the average number of variables for each rule, $SIGNIF$ is the significance measure, $UNUSUAL$ is the unusualness measure, $SUPPORT_c$ is the support based on the examples of the class, and $CONFID$ is the fuzzy confidence obtained.

To analyse the results, two different aspects have been considered. On the one hand, the interpretability obtained using different diversity functions has been analysed; on the other hand, the values obtained by the most commonly used quality measures in SD are studied:

- Interpretability is considered in this study as the relationship between the average number of rules and the average number of variables per rule obtained by the algorithm. This is because in SD is considered that a good subgroup is that with a low number of rules and variables, i.e. interesting subgroups are described by general rules with few variables. Related to this aspect, both angle and utility-based measures obtain a good relationship between these values, but the use of the utility-based measure obtains the best results in the majority of the data sets, where in 75% of them it obtains more interpretable rules. In particular, it should be noted the results obtained in the $German$ data set, where the use of both angle and utility-based measures obtains a high difference with respect to the use of crowding distance. Only in the $Bridges$ data set the use of the crowding measure allows to obtain a good interpretability in relation to the rest of the measures analysed.
- In this experiments, significance, unusualness, support based on the examples of the class (also called sensitivity) and fuzzy confidence quality measures have been analysed. Analysing these quality measures can be noted that:
  - In significance, the average result obtained by the crowding distance is very good in comparison with the results obtained by the other measures. Crowding distance obtains the best results in significance in 75% of the data sets studied.
  - In unusualness, the diversity measure of crowding distance obtains the best average results too, where in 75% of the data sets obtains the best results with important differences over the other measures.
  - For sensitivity, the best results are obtained using the angle measure, although the three diversity measures

TABLE I
RESULTS OF NMEEF-SD ALGORITHM WITH SEVERAL DIVERSITY FUNCTIONS

$Dataset$	$Diversity$	$\sharp Rules$	$\sharp Variables$	$SIGNIF$	$UNUSUAL$	$SUPPORT_c$	$CONFID$
Australian	Crowding	3.200	2.947	**21.4090**	**0.1749**	0.8385	**0.8804**
	Angle	2.200	**2.717**	18.3935	0.1653	0.8441	0.8374
	Utility	**2.100**	2.800	19.1742	0.1699	**0.8518**	0.8469
Breast	Crowding	4.800	2.215	**18.0465**	**0.1337**	**0.8039**	0.9051
	Angle	4.700	2.195	16.4346	0.1238	0.7559	0.9049
	Utility	**4.200**	**2.112**	16.9521	0.1242	0.7522	**0.9077**
Bridges	Crowding	**4.000**	**1.967**	0.7893	0.0309	**0.5590**	0.7018
	Angle	4.300	2.043	0.8054	0.0315	0.5549	**0.7172**
	Utility	4.200	2.003	**0.8074**	**0.0316**	0.5566	0.7115
German	Crowding	9.600	2.832	**2.9228**	**0.0395**	0.7541	**0.7790**
	Angle	**5.200**	**2.536**	1.6463	0.0318	**0.8761**	0.7485
	Utility	**5.200**	**2.536**	1.6463	0.0318	**0.8761**	0.7485
Heart	Crowding	**3.300**	2.667	**3.6831**	0.1038	**0.7833**	**0.7757**
	Angle	3.900	**2.602**	3.6382	**0.1053**	0.7747	0.7691
	Utility	3.900	2.660	3.5207	0.1013	0.7406	0.7571
Hepatitis	Crowding	10.900	3.402	1.3225	0.0428	0.7166	**0.7915**
	Angle	**10.400**	3.347	**1.3524**	**0.0435**	**0.7212**	0.7912
	Utility	**10.400**	**3.324**	1.2424	0.0410	0.7189	0.7853
Hypothyroid	Crowding	3.100	2.350	**11.7487**	**0.0243**	0.9966	**0.9814**
	Angle	**2.100**	2.033	9.4424	0.0221	**0.9973**	0.9788
	Utility	**2.100**	**2.017**	9.3961	0.0214	0.8973	0.8827
Ionosphere	Crowding	3.700	3.315	**6.5118**	**0.1305**	**0.9532**	**0.8681**
	Angle	3.800	3.053	2.6163	0.0789	0.9328	0.7738
	Utility	**3.300**	**2.875**	2.3416	0.0755	0.9345	0.7626
AVERAGE	Crowding	5.325	2.712	**8.5525**	**0.0869**	0.8003	**0.8390**
	Angle	4.575	2.565	7.2569	0.0772	**0.8040**	0.8228
	Utility	**4.425**	**2.541**	7.4976	0.0802	0.7996	0.8290

studied obtain similar results with small differences.

– In confidence, the crowding distance obtains the best average result in 75% of the data sets studied.

In conclusion, the best results of NMEEF-SD with respect to the values of the quality measures are obtained using the crowding distance. This diversity measure obtains the best relation between sensitivity and confidence. Furthermore, crowding distance allows the obtaining of very good results in very specific quality measures of SD: significance and unusualness. It has to be noted that both measures are specially important in SD.

In summary, if the experts have to cope with a SD problem using NMEEF-SD algorithm in which the interpretability of the results is a main issue, the best choice is to use the utility-based diversity measure. Otherwise, in problems in which good results for the specific quality measures used in the SD task are required, the crowding distance measure is the most suitable one. Finally, the best option to obtain a compromise between interpretability and the results of the specific SD measures, is the use of the utility-based diversity measure.

VI. CONCLUSION

In this paper, an analysis with different diversity measures for the NMEEF-SD algorithm is performed. NMEEF-SD is a recent algorithm for the extraction of descriptive fuzzy rules for the SD task, based on the NSGA-II approach. The main concepts of this multi-objective approach are the non-dominated ranking and the crowding distance.

An experimental study has been performed to analyse the results of NMEEF-SD using several diversity measures: crowding distance, angle measure and utility-based measure. The main objective is to check the impact of the use of these diversity measures, which guide the selection process of the algorithm towards a uniformly spread-out Pareto optimal front, promoting the diversity in the individuals included in the population of the next generation.

NMEEF-SD using the original diversity function employed in the NSGA-II approach (the crowding distance) obtains the best results with respect to the quality measures both for specific SD measures, such as significance and unusualness, and for the relationship between support and confidence. However, the use of the utility-based measure allows the obtaining of better results with respect to the interpretability.

Therefore, depending on the needs of the experts, any of the diversity measures studied can be used. However, it is important to note that the interpretability is a key factor for the experts within the SD task, as their objective is to obtain general and interpretable subgroups, i.e. subgroups with few variables and rules. In this way, a suitable SD algorithm must obtain few rules with few variables, also obtaining high values for the quality measures.

ACKNOWLEDGMENT

This work was supported by the Spanish Ministry of Education, Social Policy and Sports under projects TIN-2008-06681-C06-01 and TIN-2008-06681-C06-02, and by the Andalusian Research Plan under project TIC-3928.

REFERENCES

[1] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: an overview," in *Advances in knowledge discovery and data mining*. AAAI/MIT Press, 1996, pp. 1–34.

[2] D. Michie, D. J. Spiegelhalter, and C. C. Tayloy, *Machine Learning*. Ellis Horwood, 1994.

[3] W. Kloesgen, "Explora: A Multipattern and Multistrategy Discovery Assistant," in *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, 1996, pp. 249–271.

[4] S. Wrobel, "An Algorithm for Multi-relational Discovery of Subgroups," in *Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery*, ser. LNAI, vol. 1263. Springer, 1997, pp. 78–87.

[5] 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*, vol. 15, no. 4, pp. 578–592, 2007.

[6] C. J. Carmona, P. González, M. J. del Jesus, and F. Herrera, "NMEEF-SD: Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 5, pp. 958–970, 2010.

[7] O. Cordón, F. Herrera, F. Hoffmann, and L. Magdalena, *Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases*. World Scientific, 2001.

[8] 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*, vol. 14, pp. 5–31, 2004.

[9] F. Herrera, "Genetic fuzzy systems: taxomony, current research trends and prospects," *Evolutionary Intelligence*, vol. 1, pp. 27–46, 2008.

[10] D. E. Goldberg, *Genetic Algorithms in search, optimization and machine learning*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[11] L. A. Zadeh, "The concept of a linguistic variable and its applications to approximate reasoning. Parts I, II, III," *Information Science*, vol. 8-9, pp. 199–249,301–357,43–80, 1975.

[12] E. Hüllermeier, "Fuzzy methods in machine learning and data mining: Status and prospects," *Fuzzy Sets and Systems*, vol. 156, no. 3, pp. 387–406, 2005.

[13] C. A. Coello, D. A. V. Veldzhuizen, and G. B. Lamont, *Evolutionary Algorithms for Solving Multi-Objective Problems*, 2nd ed. Kluwer Academic Publishers, 2007.

[14] K. Deb, *Multi-Objective Optimization using Evolutionary Algorithms*. John Willey & Sons, 2001.

[15] K. Deb, A. Pratap, S. Agrawal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[16] H. Isibuchi, "Multiobjective genetic fuzzy systems: review and future research directions," in *IEEE International Conference on Fuzzy Systems*, 2007, pp. 913–918.

[17] D. Whitley, "The GENITOR algorithm and selection pressure: why rank-based allocation of reproductive trials is best," in *Proceedings of the 3rd International Conference on Genetic Algorithms*, 1989.

[18] T. H. Li, C. Lucasius, and G. Kateman, "Optimization of calibration data with the dynamic genetic algorithm," *Analytica Chimica Acta*, vol. 2768, pp. 123–134, 1992.

[19] L. J. Eshelman and J. D. Schaffer, "Preventing premature convergence in genetic algorithms by preventing incest," in *Proceedings of the 4th International Conference on Genetic Algorithms*, 1991, pp. 115–122.

[20] J. Branke, K. Deb, H. Dierolf, and M. Osswals, "Finding knees in multi-objective optimization," in *Proc. Parallel Problem Solving from Nature Conf.*, ser. LNCS, vol. 3242, 2004, pp. 722–731.

[21] A. Siebes, "Data Surveying: Foundations of an Inductive Query Language," in *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1995, pp. 269–274.

[22] S. Wrobel, *Inductive logic programming for knowledge discovery in databases*. Springer, 2001, ch. Relational Data Mining, pp. 74–101.

[23] D. Gamberger and N. Lavrac, "Expert-Guided Subgroup Discovery: Methodology and Application," *Journal Artificial Intelligence Research*, vol. 17, pp. 501–527, 2002.

[24] N. Lavrac, B. Cestnik, D. Gamberger, and P. A. Flach, "Decision Support Through Subgroup Discovery: Three Case Studies and the Lessons Learned," *Machine Learning*, vol. 57, no. 1-2, pp. 115–143, 2004.

[25] F. Herrera, C. J. Carmona, P. González, and M. J. del Jesus, "An overview on Subgroup Discovery: Foundations and Applications," *Knowledge and Information Systems*, vol. In press, 2011.

[26] 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*, vol. 15, no. 4, pp. 533–535, 2007.

[27] J. Casillas and B. Carse, "Special issue on Genetic Fuzzy Systems: Recent Developments and Future Directions," *Soft Computing*, vol. 13, no. 5, pp. 417–418, 2009.

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

[29] L. A. Zadeh, "Information Control," *Fuzzy sets*, vol. 8, pp. 338–353, 1965.

[30] C. H. Chen, T. P. Hong, and V. S. Tseng, "An improved approach to find membership functions and multiple minimum supports in fuzzy data mining," *Expert Systems with Applications*, vol. 36, no. 6, pp. 10 016–10 024, 2009.

[31] 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*, vol. 160, no. 7, pp. 905–921, 2009.

[32] C. H. Chen, T. P. Hong, V. S. Tseng, and C. S. Lee, "A genetic-fuzzy mining approach for items with multiple minimum supports," *Special Issue on Genetic Fuzzy Systems: Recent Developments and Future Directions. Soft Computing - A Fusion of Foundations, Methodologies and Applications*, vol. 13, no. 5, pp. 521–533, 2009.

[33] M. Kaya, "MOGAMOD: Multi-objective genetic algorithm for motif discovery," *Expert Systems with Applications*, vol. 36, no. 2, pp. 1039–1047, 2009.

[34] M. Kaya, "Automated extraction of extended structured motifs using multi-objective genetic algorithm," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2421–2426, 2010.

[35] M. Kaya, "Autonomous classifiers with understandable rule using multi-objective genetic algorithms," *Expert Systems with Applications*, vol. 37, no. 4, pp. 3489–3494, 2010.

[36] M. J. del Jesus, P. González, and F. Herrera, "Multiobjective Genetic Algorithm for Extracting Subgroup Discovery Fuzzy Rules," in *Proceedings of the IEEE Symposium on Computational Intelligence in Multicriteria Decision Making*. IEEE Press, 2007, pp. 50–57.

[37] 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," in *Proceedings of the 6th Industrial Conference on Data Mining*, ser. LNCS, vol. 4065. Springer, 2006, pp. 337–349.

[38] 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*, vol. 36, pp. 1632–1644, 2009.

[39] C. J. Carmona, P. González, M. J. del Jesus, C. Romero, and S. Ventura, "Evolutionary algorithms for subgroup discovery applied to e-learning data," in *Proceedings of the IEEE International Education Engineering*, 2010, pp. 983–990.

[40] 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 Special Issue on "Genetic Fuzzy Systems"*, vol. In Press, 2011.

[41] C. J. Carmona, P. González, M. J. del Jesus, and F. Herrera, "An Analysis of Evolutionary Algorithms with Different Types of Fuzzy Rules in Subgroup Discovery," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, 2009, pp. 1706–1711.

[42] M. J. del Jesus, P. González, and F. Herrera, *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models*. Springer, 2007, vol. 220, ch. Subgroup Discovery with Linguistic Rules, pp. 411–430.

[43] N. Lavrac, P. A. Flach, and B. Zupan, "Rule Evaluation Measures: A Unifying View," in *Proceedings of the 9th International Workshop on Inductive Logic Programming*, ser. LNCS, vol. 1634. Springer, 1999, pp. 174–185.

[44] A. Asuncion and D. J. Newman, "UCI Machine Learning Repository," 2007. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html