Non-dominated Multi-objective Evolutionary algorithm based on Fuzzy rules extraction for Subgroup Discovery

C.J. Carmona[1], P. González[1], M.J. del Jesus[1], and F. Herrera[2]

Department of Computer Science of University of Jaen[1]
{ccarmona,pglez,mjjesus}@ujaen.es
Deparment of Computer Science and AI of University of Granada[2]
herrera@decsai.ugr.es

Abstract. A new multi-objective evolutionary model for subgroup discovery with fuzzy rules is presented in this paper. The method resolves subgroup discovery problems based on the hybridization between fuzzy logic and genetic algorithms, with the aim of extracting interesting, novel and interpretable fuzzy rules. To do so, the algorithm includes different mechanisms for improving diversity in the population. This proposal focuses on the classification of individuals in fronts, based on non-dominated sort. A study can be seen for the proposal and other previous methods for different databases. In this study good results are obtained for subgroup discovery by this new evolutionary model in comparison with existing algorithms.

Key words: Data mining, Subgroup Discovery, Multi-Objective Evolutionary Algorithms, Fuzzy Rules, Genetic Fuzzy Systems.

1 Introduction

Knowledge Discovery in Databases (KDD) is defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [8]. Within the KDD process the data mining stage is responsible for high level automatic knowledge discovery using real data. In the KDD process two different tasks can be distinguished: predictive induction, whose objective is the discovery of knowledge for classification or prediction [16]; and descriptive induction, whose main objective is the extraction of interesting knowledge from the data. In descriptive induction, attention can be drawn to the discovery of association rules following an unsupervised learning model [1], subgroup discovery [14][19] and other approaches to non-classificated induction.

Subgroup discovery (SD) is a descriptive induction task [9] whose goal is the discovery of interesting individual patterns in relation to a specific property of interest for the user. The development of new models in this task are focusing in the use of soft computing techniques: genetic algorithm and fuzzy logic.

Genetic Algorithms (GAs) [10] are beginning to be used to solve SD problems [3][7][18] because they offer a set of advantages for knowledge extraction

and specifically for rule induction processes, although they were not specifically designed for learning.

In an SD algorithm, a fuzzy approach [20], which considers linguistic variables expressed in linguistic terms through descriptive fuzzy rules, allows us to obtain knowledge in a similar way to human reasoning. The use of fuzzy rules allows us to obtain more interpretable and actionable solutions in the field of SD, and in general in the analysis of data in order to establish relationships and identify patterns [12].

The hybridization between fuzzy logic and GAs, called genetic fuzzy systems (GFSs) [5], has attracted considerable attention in the computational intelligence community. GFSs provide an useful tools for pattern analysis and for the extraction of new types of useful information.

In [7] a mono-objective GFS within the iterative rule learning approach for SD is presented with proper results. In spite of that, the induction of rules describing subgroups can be considered as a multi-objective problem rather than a single objective one, since there are different quality measures which can be used for SD. The different measures used for evaluating a rule can be thought of as different objectives of the SD rule induction algorithm. In this sense, multi-objective evolutionary algorithms (MOEAs) are adapted to solve problems in which different objectives must be optimized [4]. In [3] a multi-objective GFS for SD based on SPEA II algorithm is proposed.

This paper describes a new proposal based on the NSGA-II algorithm for the induction of rules which describe subgroups, the Non-dominated Multi-objective Evolutionary algorithm based on Fuzzy rules extraction for Subgroup Discovery, NMEF-SD, which combines the approximated reasoning capacity of the fuzzy systems with the learning capacities of the MOEAs. This proposal tries to obtain a set of general and interesting fuzzy rules. The generality is obtained both with an operator which performs a biased initialization process and with biased genetic operators, while the diversity in the genetic population is increased with re-initialization based on coverage.

The paper is organized as follows: In Section 2, subgroup discovery is described. The new evolutionary approach to obtain fuzzy rules for SD is explained in Section 3. In Section 4 the results obtained are analyzed. Finally, the conclusions and further research are outlined.

2 Subgroup discovery

The concept of SD was initially formulated by Klösgen [14] and Wrobel [19], and was defined as: *Given a population of individuals and a property of those individuals we are interested in, find population subgroups that are statistically "most interesting", e.g., are as large as possible and have the most unusual statistical characteristics with respect to the property of interest.*

Therefore, the objective in SD is to discover characteristics of the subgroups by constructing simple individual rules with high support and significance. These rules have the form: $Cond \rightarrow Class$.

One of the most important aspects of any rule induction approach that describe subgroups is the selection of the quality measures to use. Although there is no consensus about which measures are more adapted for SD, the most common in the literature include: coverage [15], significance [14], unusualness [15], support [15] and confidence [7].

Some of the more interesting models which obtain description of subgroups represented in different forms are the classical deterministic algorithms like Apriori-SD [13] and CN2-SD [15] (available in KEEL[1] software tool [2]), and the evolutionary algorithms SDIGA [7] and MESDIF [3].

3 NMEF-SD: Non-dominated Multi-objective Evolutionary algorithm based on the extraction of Fuzzy rules for Subgroup Discovery

In this section a new evolutionary model, NMEF-SD, is described. This algorithm extracts descriptive fuzzy or crisp rules -depending on the nature of the features of the problem (continuous and/or nominal variables)- which describe subgroups.

The objective of this evolutionary process is to extract a variable number of different rules describing information of the examples belonging to the original set for each value of the target variable. As the objective is to obtain a set of rules which describe subgroups for all the values of the target variable, the algorithm must be executed as many times as the number of different values the target variable contains.

Each candidate solution is codified according to the "Chromosome = Rule" approach [5], representing only the antecedent part of the rule in the chromosome. The antecedent of a rule is composed of a conjunction of value-variable pairs. A special value is used to indicate that the variable is not considered for the rule. Fig. 1 shows a chromosome and the rule it codifies, for a problem with four features and three possible values for each one.

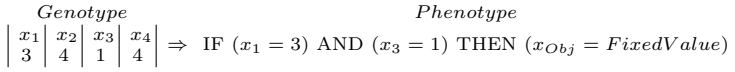

Fig. 1. Representation of a rule in NMEF-SD

When the features are continuous, the model uses fuzzy rules, and the fuzzy sets corresponding to the linguistic labels are defined by means of the corresponding membership functions. These 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.

[1] http://www.keel.es

In this extraction process the objective is to obtain interpretable rules with high quality, precision and generality. To do so, two quality measures are selected as objectives:
Support [15]: Is defined as the frequency of correctly classified examples covered by the rule.

$$Sup_cN(R_i) \;=\; \frac{n(Class \cdot Cond_i)}{n(Class)} \tag{1}$$

where $n(Class \cdot Cond_i)$ is the number of examples which satisfy the conditions for the antecedent and $n(Class)$ is the number of examples for the target variable indicated in the consequent part of the rule.
Unusualness [15]: Measures the balance between the coverage of the rule and its accuracy gain.

$$WRAcc(R_i) \;=\; \frac{n(Cond_i)}{N}\left(\frac{n(Class \cdot Cond_i)}{n(Cond_i)} - \frac{n(Class)}{N}\right) \tag{2}$$

where $n(Cond_i)$ is the number of example which satisfy the antecedent part of the rule, N is the number of examples of the data set, and the weighted relative accuracy of a rule can be described for the coverage using the first part of the expression $\frac{n(Cond_i)}{N}$ and the accuracy gain using the second part $\frac{n(Class \cdot Cond_i)}{n(Cond_i)} - \frac{n(Class)}{N}$.

NMEF-SD is based on the NSGA-II approach [6], and its main purpose is to evolve the population based on the non-dominated sort of the solutions in fronts of dominance. The first front is composed of the non-dominated solutions of the population (the Pareto front), the second is composed of the solutions dominated by one solution, the third of solutions dominated by two, and so on.

The operating scheme of NMEF-SD, can be seen in Fig. 2.

```
BEGIN
    Create P_0 with biased initialization
    REPEAT
        Q_t ← Ø
        Tournament Selection (P_t)
        Q_tc ← Multi-point Crossover (P_t)
        Q_tm ← Biased Mutation (Q_tc)
        Q_t ← Q_tc + Q_tm
        Q_t ← Q_t+ descendants
        R_t ← Join(P_t,Q_t)
        Fast-non-dominated-sort(R_t)
        IF F_1 evolves
            Introduce fronts in P_t+1
        ELSE
            Re-initialization based on coverage P_t+1
    WHILE (num-eval < Max-eval)
    RETURN F_1
END
```

Fig. 2. The NMEF-SD algorithm

NMEF-SD tries to obtain a set rules with high generality (one of the main objectives of SD) by introducing diversity in the population with different ope-

rators, since the diversity in the MOEAs is a handicap for these algorithms. The generality is obtained both with an operator which performs a biased initialization process and with biased genetic operators, while the diversity is introduced with the crowding distance [6] and with re-initialization based on coverage.

In the following subsections the different parts of the algorithm are defined.

3.1 Initialization

The first step of the algorithm is to create an initial population (P_0) whose size is prefixed by an external parameter.

The purpose of this initialization is to generate part of the individuals of the population (75% of the total) using only a maximum percentage of the variables which form part of each rule (25% of the rule). The rest of the variables of the rule and the rest of the individuals of the population are randomly generated.

This operator allows the algorithm to obtain a set of rules with high generality because most of the generated individuals are rules with a low percentage of variables.

3.2 Genetic operators

The model obtains the descendant population (Q_t), with the same size as the original one, by means of the Tournament Selection [17], Multi-point Crossover [11] and Biased Mutation operators.

Biased Mutation [7] is applied to the gene selected considering the mutation probability. This operator can be applied in two different ways: The first causes the elimination of the variable of the individual, in order to generate a more general rule; and the second randomly mutates the value of the variable. Either one of these two ways can be applied in each mutation, with the same probability.

3.3 Fast non-dominated sort

The algorithm joins the populations (P_t and Q_t) in a new population R_t, subsequently applying the non-dominated sort [6] to the new population R_t in order to obtain the classification in fronts of dominance.

This proposal achieves diversity in the population through the crowding distance [5][6] used in the sorting of the individuals belonging to the last front introduced in the main population (P_{t+1}) of the next generation.

3.4 Re-initialization based on coverage

The last step of the model is the obtaining of the population for the next generation (P_{t+1}). Before carrying out this step a check is needed on the pareto to see whether or not it evolves. We consider that the pareto evolves if it covers at least one example more than the pareto of the previous generation. If the pareto does not evolve during more than five percent of the evolutive process (quantified through number of evaluations) the re-initialization is performed.

Re-initialization based on coverage performs an elimination of those individuals repeated in the pareto which cover the same examples of the data set. The rest of the individuals are copied in the population of the next generation (P_{t+1}). Repeated individuals which have been eliminated are replaced with new individuals generated through re-initialization based on coverage, introducing individuals which cover previously uncovered examples.

3.5 Stop condition

The evolutionary process ends when the number of evaluations is reached. Then the algorithm returns the rules in the pareto which reach a predefined confidence [7] value threshold.

4 Experimentation

In order to analyze the behaviour of the proposed model an experimentation with different data sets available in UCI repository[2] has been carried out. The selected data sets have different numbers of features and classes, and different types of features (discrete and continuous). These data sets are: Australian, Balance, Echo and Vote.

As the NMEF-SD model is non-deterministic, it has been run five times, and the mean values of these runs are computed. In addition a 10 cross-validation is performed and the results are compared with those obtained by other SD algorithms: CN2-SD [15], Apriori-SD [13], SDIGA [7] and MESDIF [3].

In this experimentation, the parameters used in NMEF-SD are: a population size of 25 individuals, a maximum number of evaluations of 5000, crossover probability of 0.6 and mutation probability of 0.1.

Table 1 shows the average values obtained by the analyzed methods: number of rules (Rul), number of variables (Var), coverage (COV) [15], significance ($SIGN$) [14], unusualness ($WRAcc$) [15], support (SUP_cN) [15] and fuzzy confidence ($FCNF$) [7]. These measures have been chosen because in previous studies they have been shown to be the most suitable measures for the SD task. The best results are shown in bold characters.

Results in table 1 show that the NMEF-SD model obtains the best results for the quality measures in almost all data sets with respect to the other algorithms. Considering the different characteristics of the data sets, NMEF-SD is able to obtain better results in generality and precision than CN2-SD, Apriori-SD and SDIGA. The results obtained by NMEF-SD are usually the best for the different measures for the databases used.

The subgroups obtained for NMEF-SD (rules and variables), are good, useful and representative. These characteristics, together with the previously mentioned one, make this new model a promising approach for the SD task.

[2] http://www.ics.uci.edu/∼mlearn/MLRepository.html

Table 1. Results of the experimentation

Database	*Algorithm*	*Rul*	*Var*	*COV*	*SIGN*	*WRAcc*	SUP_cN	*FCNF*
Australian	NMEF-SD	3,58	2,92	**0,454**	**23,178**	**0,171**	0,783	**0,930**
	MESDIF	10,00	3,52	0,311	7,594	0,060	0,577	0,807
	SDIGA	**2,68**	3,28	0,310	16,348	0,120	**0,803**	0,591
	CN2-SD	30,50	4,58	0,400	15,350	0,055	0,649	0,830
	AprioriSD	10,00	**2,02**	0,377	16,998	0,074	0,654	0,863
Balance	NMEF-SD	**2,30**	2,00	**0,362**	5,326	**0,070**	**0,530**	**0,698**
	MESDIF	28,10	3,08	0,163	3,516	0,022	0,318	0,557
	SDIGA	7,40	2,39	0,291	5,331	0,049	0,487	0,664
	CN2-SD	15,60	2,23	0,336	**8,397**	0,063	0,512	0,583
	AprioriSD	10,00	**1,20**	0,333	5,444	0,058	0,480	0,649
Echo	NMEF-SD	3,62	2,35	**0,428**	**1,293**	**0,043**	**0,628**	**0,757**
	MESDIF	19,74	3,30	0,164	0,877	0,017	0,355	0,591
	SDIGA	**2,32**	2,27	0,394	1,165	0,013	0,566	0,590
	CN2-SD	17,30	3,23	0,400	1,181	0,019	0,490	0,667
	AprioriSD	9,80	**1,70**	0,194	0,901	0,034	0,226	0,510
Vote	NMEF-SD	**1,10**	2,05	**0,577**	**21,974**	**0,217**	**0,946**	**0,979**
	MESDIF	7,86	3,44	0,429	19,937	0,187	0,827	0,957
	SDIGA	3,06	3,19	0,422	18,243	0,180	0,802	0,891
	CN2-SD	8,00	1,79	0,438	18,830	0,176	0,858	0,932
	AprioriSD	10,00	**1,44**	0,428	17,060	0,147	0,800	0,930

5 Conclusions

In this paper a new multi-objective evolutionary model for the induction of fuzzy rules which describe subgroups is presented. NMEF-SD hybridizes soft-computing techniques like fuzzy logic and the GAs in a MOEA, which is able to obtain high quality results.

The model allows us to obtain small interpretable rule sets which may be fuzzy or crisp depending on the problem. These rules are obtained with a multi-objective model which considers only two quality measures used in SD, with quite good results. Furthermore, the results obtained are better than the results of the classical models considered in this paper.

The proposed model improves on the results obtained for every quality measures with other evolutionary models such as SDIGA and MESDIF. The combination of the NSGA-II approach makes NMEF-SD improves over other.

Moreover, NMEF-SD algorithm tries to obtain a set rules with high generality (one of the main objectives of SD) by introducing diversity in the population with different operators. The generality is obtained both with an operator which performs a biased initialization process and with biased genetic operators, while the diversity in the genetic population is increased with re-initialization based on coverage. All these characteristics make it a good proposal for SD regardless the quality measures the expert considers important for the process.

As future work we will study the use of other rule representation, such as DNF rules, in order to obtain more expressive and understandable results.

Acknowledgments

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. R. Agrawal, T. Imieliski, and A. Swami, *Mining association rules between sets of items in large databases*, SIGMOD '93 (New York, NY, USA), 1993, pp. 207–216.
2. 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.
3. F. 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*, LNCS, vol. 4065, Springer, 2006, pp. 337–349.
4. C.A. Coello, D.A. Van Veldzhuizen, and G.B. Lamont, *Evolutionary Algorithms for Solving Multi-Objective Problems*, 2nd ed., Kluwer Academic Publishers, 2007.
5. O. Cordón, F. Herrera, F. Hoffmann, and L. Magdalena, *Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases*, 2001.
6. 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.
7. 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.
8. U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, *From data mining to knowledge discovery: an overview*, Advances in knowledge discovery and data mining, 1996, pp. 1–34.
9. D. Gamberger and N. Lavraĉ, *Expert-Guided Subgroup Discovery: Methodology and Application*, Journal Artificial Intelligence Research **17** (2002), 501–527.
10. D.E. Golberg, *Genetic Algorithms in search, optimization and machine learning*, Addison-Wesley (1989).
11. J.H. Holland, *Adaptation in natural and artificial systems*, University of Michigan Press (1975).
12. E. Hüllermeier, *Fuzzy methods in machine learning and data mining: Status and prospects*, Fuzzy Sets and Systems **156** (2005), no. 3, 387–406.
13. B. Kavŝek and N. Lavraĉ, *APRIORI-SD: Adapting association rule learning to subgroup discovery*, Applied Artificial Intelligence **20** (2006), 543–583.
14. W. Klösgen, *Explora: A Multipattern and Multistrategy Discovery Assistant*, Advances in Knowledge Discovery and Data Mining, Fayyad, U., et. al. Editors, 1996, pp. 249–271.
15. N. Lavraĉ, B. Kavŝek, P.A. Flach, and L. Todorovski, *Subgroup Discovery with CN2-SD*, Journal of Machine Learning Research **5** (2004), 153–188.
16. D. Michie, D.J. Spiegelhalter, and C.C. Tayloy, *Machine Learning*, Ellis Horwood, 1994.
17. B.L. Miller and D.E. Goldberg, *Genetic Algorithms, Tournament Selection, and the Effects of Noise*, Complex System **9** (1995), 193–212.
18. 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.
19. S. Wröbel, *An Algorithm for Multi-relational Discovery of Subgroups*, PKDD '97: Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery, LNCS, vol. 1263, Springer, 1997, pp. 78–87.
20. 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.