Chapter 17
Research of the Subgroup Discovery Algorithm NMEEF-SD

Haichun Xie, Yong Zhang, Limin Jia and Yong Qin

Abstract Subgroup discovery (SD) is a data mining technique which could find the most interesting individual patterns from a population of individuals for the user. Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discover (NMEEF-SD) which is based on non-dominated sorting genetic algorithm II (NSGA-II) is a kind of algorithm for SD. First, the concept of subgroup discovery is introduced. Then NMEEF-SD algorithm and its main properties are researched. Finally, the algorithm is applied to analyze the concrete comprehensive strength dataset from UCI database, the result of experiment shows that the NMEEF-SD algorithm is able to extract fuzzy rules with interesting characteristics and is easy to understand.

Keywords Subgroup discovery · NMEEF-SD · Fuzzy rules · UCI

17.1 Introduction

As data volumes explode, it becomes important to find the useful knowledge from a large number of complex data. Knowledge Discovery in Database (KDD) is aimed at assisting humans in extracting useful information from the rapidly growing volumes of data [6]. Knowledge is usually expressed in the form of rules, descriptive and predictive are the two standards to measure the quality of the rules. Subgroup discovery (SD) can obtain descriptive and predictive rules that make it attracts a lot of attention from researchers.

H. Xie · Y. Zhang
School of Automation,
Nanjing University of Science and Technology, Nanjing 210094, China

L. Jia · Y. Qin (✉)
State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University,
Beijing 100044, China
e-mail: qinyong2146@126.com

Z. Deng and H. Li (eds.), *Proceedings of the 2015 Chinese Intelligent Automation Conference*, Lecture Notes in Electrical Engineering 336,
DOI 10.1007/978-3-662-46469-4_17

Subgroup discovery was initially proposed by Klogen [8] and Wrobel [10]. Lavrac and Kavsek proposed CN2-SD [9] which was developed by modifying parts of the CN2 [3] classification rule learner, CN2-SD obtains rules through its covering algorithm, search heuristic, probabilistic classification of instances, and evaluation measures. Herrera processed SDIGA [5] which is a genetic fuzzy system for data mining task of subgroup discovery, using fuzzy rules describing the inductive knowledge.

This paper will apply Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery (NMEEF-SD) to the strength datasets of concrete. The remainder of this paper is organized as follows. Section 17.2 shows the basic concept of subgroup discovery and NMEEF-SD. Section 17.3 discusses an experimentation of NMEEF-SD. Finally, some conclusions are offered in Sect. 17.4.

17.2 Analysis of NMEEF-SD

17.2.1 Subgroup Discovery

The task of subgroup discovery is to discovery groups that are statically most unusual from a given dataset and target property. Simple rules with highly significant and high support are used to describe those groups. Rules have the form: Cond > Class.

Class is the target property, which appears in the rule consequent. Cond (the rule antecedent) is a conjunction of features (attribute-values).

"The probability of coronary heart disease is higher in smokers who have a family disease" is a rule, and the rule can be defined as:

$$\begin{aligned} &\text{if } (\text{smoker} = \text{true and family history} = \text{positive}) \\ &\quad \text{then coronary} - \text{heart} - \text{disease} = \text{true} \end{aligned}$$

The subgroup of smokers who have a family disease is described in this rule; coronary heart disease is the target property. The population described by this rule has a higher probability for the target property.

Target attribute, description language of subgroup, quality measures, search strategy are the four main aspects of subgroup discovery algorithm. Target attributes may be binary, nominal, or continuous [7]; language is the representation of the subgroups which must be suitable for obtaining interesting rules; quality measures are used to measure the obtained rules. Coverage, significance, unusualness, and support are often chosen as the quality measurements to extract and evaluate the rules; search strategy is important for the dimension of the search space. Different strategies have been used in subgroup discovery, and top-down search strategy is the usual choice.

17.2.2 NMEEF-SD: Non-dominated Multi-objective Evolutionary Algorithm for Extracting Fuzzy Rules in Subgroup Discovery

Subgroup discovery algorithm selects a number of quality measures to measure the quality of rules obtained, so subgroup discovery can be considered as a multiobjective problem. Different quality measures in the evolutionary process of rules population can be regarded as different evolutionary goals of genetic algorithm, therefore multiobjective evolutionary algorithm (MOEAs) is suitable to solve multiobjective optimization problems in subgroup discovery.

Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery (NMEEF-SD) [2] is based on hybridization between fuzzy logic and genetic algorithms. NMEEF-SD uses NSGA-II [4] to exact interesting, novel, and interpretable fuzzy rules. In NMEEF-SD, each candidate solution is coded according to the "Chromosome = Rule" approach, where 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.

17.2.2.1 Objective Function

The objective is to obtain rules with high confidence, understandable, and generality in the process of rule discovery. To do so, support, confidence, and accuracy are selected as quality measures.

- *Support*: the frequency of correctly classified examples covered by rule.

$$\mathrm{Sup}(R) = \frac{n(\mathrm{Class} \cdot \mathrm{Cond})}{n(\mathrm{Class})}$$

  where *n(*Class•Cond*)* is the number of examples which satisfy the conditions for antecedent and *n(*class*)* is the number of examples for target variable indicated.
- *Confidence*: standard measure that determines the relative frequency of examples satisfying the complete rule among those satisfying only the antecedent.

$$\mathrm{Conf}(R) = \frac{\sum_{E_k \in E / E_k \in \mathrm{Class}} \mathrm{APC}(E_K, R)}{\sum_{E_k \in E} \mathrm{APC}(E_K, R)}$$

  where APC is the degree of compatibility between an example and the antecedent part of a fuzzy rule.

- *Accuracy*: the membership of the examples covered by the antecedent part of rule and satisfying consequent of the *rule*.

$$\mathrm{Accu}(R) = \frac{n(\mathrm{Cond} \cdot \mathrm{Cond}_i)}{n(\mathrm{Cond}) + k}$$

where K is the number of the objective variables.

17.2.2.2 Main Properties of the Algorithm

NMEEF-SD consists of initialization, genetic operators, fast non-dominated sort, re-initialization based on coverage and stop condition [2], a single operation scheme of the algorithm can be seen in Fig. 17.1.

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

Fuzzy logic is used to process the continuous variables, by means of linguistic variable. This allows the use of numerical features without the need of a previous discretization [1].

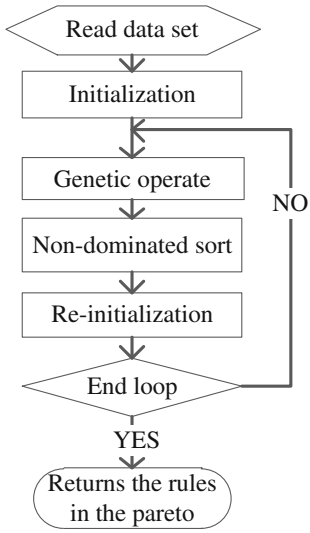

Fig. 17.1 The NMEEF-SD algorithm

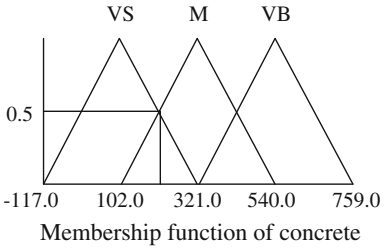

Fig. 17.2 Fuzzy partition for a numerical variable

17.3 Experimentation

Compressive strength of concrete dataset in UCI repository is selected as the experimental data. Experimental data consists of 1,030 samples, each sample includes 8 input variables and 1 output variable (compressive strength of concrete). Output variable is selected as the target variable in the dataset.

Compressive strength of concrete in the raw data is continuous variables, in order to apply to use subgroup discovery algorithm, compressive strength of concrete between 0 and 20 MPa is classified as low, 20–55 MPa as medium, and 55–80 MPa as high. Rearrange samples depend on target variables, where the compressive strength of 1–197 samples is low, the compressive strength of 198–883 samples is medium, and the compressive strength of 884–1030 samples is high.

All input variables are continuous variable, in order to apply the fuzzy rule, input variables are processed using fuzzy treatment. Triangular membership function is used to obtain rules with higher explanatory as seen in Fig. 17.2.

First, process the compressive strength of concrete dataset ten times by cross-validation, then execute algorithm. The algorithm is executed three times. The optimal rule set contains nine rules, which describe group of low, medium, and high, as can be seen in Table 17.1.

Results in Table 17.1 show that rules with simple structure contain less variables but have higher support, confidence, and accuracy.

Table 17.1 Best rules obtained by NMEEF-SD

Rule	Target property	Variable	Support	Confidence	Accuracy
R_1	High	4	0.578947	0.618307	0.500000
R_2	High	5	0.646617	0.811371	0.333333
R_3	High	4	0.691729	0.682891	0.333333
R_4	Medium	2	1.000000	0.747430	0.817204
R_5	Medium	4	0.941653	0.805049	0.940000
R_6	Medium	4	0.948136	0.801996	0.940000
R_7	Medium	6	0.478120	0.847051	0.875000
R_8	Medium	5	0.687196	0.840319	0.727273
R_9	Low	6	0.280899	0.614202	0.333333

Table 17.2 Rules of high compressive strength of concrete

Rule	Rule description
R_1	if (cement = VB and fly ash = VS and water = VS and coarse aggregate = M) then concrete compressive strength = high
R_2	if (cement = VB and fly ash = VS and superplasticizer = M and coarse aggregate = M and age = (N ORVL)) then concrete compressive strength = high
R_3	if(cement = (M OR VB) and fly ash = VS and water = VS and age = N) then concrete compressive strength = high

Select rules of high compressive strength of concrete from Table 17.1, description of the rules are shown in Table 17.2.

Analysis rules obtained indicate the following: The concrete that contains lots of cement, a small amount of fly ash, right quantity of coarse-aggregate and normal coagulation time performance for the high compressive strength of concrete.

Figure 17.3 shows target variables and Pareto sequence when algorithm was running, the red curve represents the average support, the blue curve represents the average confidence, and the green curve represents the average accuracy.

Figure 17.3 shows that during the process of extracting rules, the overall quality of the population tends to be stable with the evolution of the rules of population. There is still some fluctuation in shall scope mainly because NMEEF-SD algorithm uses random parent population to crossover operation. The diversity in the genetic population is increased with re-initialization based on coverage. The change curve of number of rules in Pareto shows there is a large fluctuation in the early evolution, with the increasing of number of rules, Pareto will maintain a stable number.

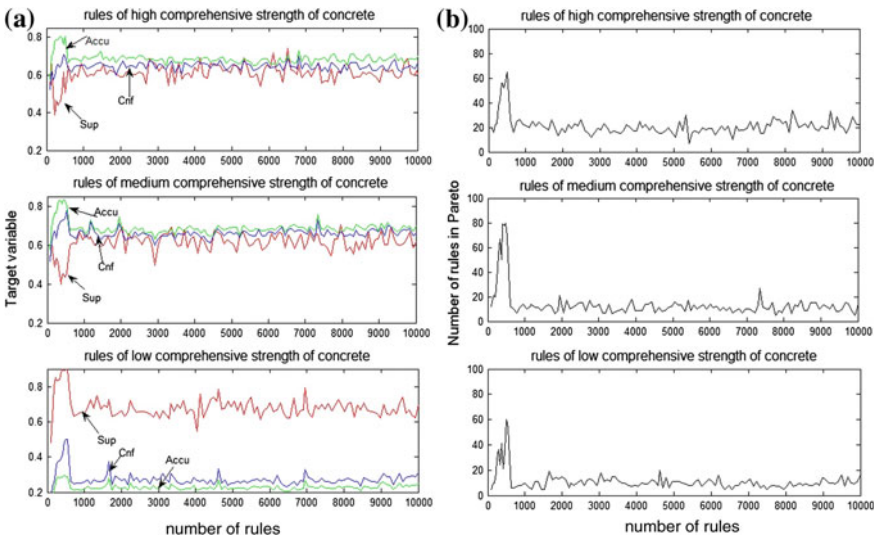

Fig. 17.3 Target variable during process of extracting and change curve of Pareto

The results show that the application of subgroup discovery algorithm in processing compressive strength of concrete dataset, on the one hand can extract simple, effective rules to describe different compressive strength of concrete. These rules can provide an effective reference for quality optimization of concrete in the manufacturing process; on the other hand the rules obtained with high classification accuracy can be used to predict the compressive strength of concrete and provide an effective tool for detecting compressive strength of concrete.

17.4 Conclusion

In this paper, Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery (NMEEF-SD) is researched. We realize the NMEEF-SD algorithm and apply to Compressive strength of concrete dataset in UCI repository. The experimental results indicate that subgroup discovery could efficiently extract rules that are interesting, understandable, and these rules can provide an effective basis for compressive strength of concrete detection.

Acknowledgments This research is supported by Independent Subject of Y. Qin (No. RCS2014ZT24) and Research Fund for the Doctoral Program (No. 20120009110035). The supports are gratefully acknowledged.

References

1. Carmona CJ, Chrysostomou C, Seker H, del Jesus MJ (2013) Fuzzy rules for describing subgroups from Influenza A virus using a multi-objective evolutionary algorithm. Appl Soft Comput 13(8):3439–3448
2. Carmona CJ, González P, del Jesus MJ, Herrera F (2009) Non-dominated multi-objective evolutionary algorithm based on fuzzy rules extraction for subgroup discovery. In: Proceeding of 4th international conference on hybrid artificial intelligence Systems, vol 5572. Springer, LNAI, pp 573–580
3. Clark P, Niblett T (1989) The CN2 induction algorithm. Mach Learn 3(4):261–283
4. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast elitist multi-objective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6(2):182–197
5. Del Jesus MJ, González P, Herrera F, Mesonero M (2007) Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing. IEEE Trans Fuzzy Syst 15 (4):578–592
6. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Mag 17(3):37
7. Herrera F, Carmona CJ, Gonzalez P, del Jesus MJ (2001) An overview on subgroup discovery: foundations and applications. Knowl Inf Syst 29:495–525
8. Klosgen W (1996) Explora: a multipattern and multistrategy discovery assistant. In: Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, pp 249–271

9. Lavrac N, Kavsek B, Flach PA, Todorovski L (2004) Subgroup Discovery with CN2-SD. J Mach Learn Res 5:153–188
10. Wrobel S (1997) An algorithm for multi-relational discovery of subgroup. In: Proceeding of the 1st European symposium on principles of data mining and knowledge discovery, vol 1263. Springer, LNAI, pp 78–87