ATLANTIS PRESS Journal of Robotics, Networking and Artificial Life, Vol. 4, No. 3 (December 2017) 239–242

A Rule-Based Classification System Enhanced by Multi-Objective Genetic Algorithm

Shingo Mabu

Graduate School of Sciences and Technology for Innovation, Yamaguchi University, Tokiwadai 2-16-1

Ube, Yamaguchi 755-8611, JAPAN

Kenzoh Azakami

Graduate School of Science and Engineering, Yamaguchi University, Tokiwadai 2-16-1

Ube, Yamaguchi 755-8611, JAPAN

Masanao Obayashi

Graduate School of Sciences and Technology for Innovation, Yamaguchi University, Tokiwadai 2-16-1 Address
Ube, Yamaguchi 755-8611, JAPAN

Takashi Kuremoto

Graduate School of Sciences and Technology for Innovation, Yamaguchi University, Tokiwadai 2-16-1 Address
Ube, Yamaguchi 755-8611, JAPAN
E-mail: {mabu, v001vk, m.obayas, wu}@yamaguchi-u.ac.jp

Abstract

Recent years, data mining techniques have been developed for extracting rules from big data. However, there are some problems to be considered, for example, it is difficult to judge which rules are important and which are not important; and even in simple classification problems with the small number of classes, a various sub-patterns to be considered potentially exist in each class. To solve the above problems, a rule clustering algorithm using multi-objective genetic algorithm is proposed.

1. Introduction

Data mining is a technique of extracting effective rules from big data. However, data mining has some problems to be considered. That is, a large number of rules are extracted, so it is difficult to judge which are the important rules and which are not important; and even in simple classification problems, e.g., two-class problems, a variety of patterns potentially exist in each class, which makes the problems more difficult. In the conventional method [1], a genetic algorithm (GA)-based clustering was applied to the class association rules extracted by genetic network programming (GNP) [2] to solve the above problems and enhance the classification system. In this paper, a rule clustering algorithm using the multi-objective genetic algorithm (MOGA) is proposed to enhance the conventional classification system. To confirm the effectiveness of the proposed method, the accuracy of the conventional method and proposed method is compared. In addition, the accuracy of the proposed method and other classification methods: Tree-based classifier (J4.8), Multi-Layer Perceptron (MLP), Support Vector Machine (SVM) is also compared. The data used for the comparisons are Credit approval, Pima Indian, and German, which were downloaded from UCI machine learning repository [3].

In [4], a multi-objective evolutionary algorithm is applied to making rule-based classifiers. It is important for rule-based classifiers to implement information granulation and obtain effective input variables, which is effectively solved by multi-objective optimization. In [5], a fuzzy-rule-based credit classification system is proposed. The accuracy and interpretability are important in credit classification; thus, this method introduces three objectives related to accuracy and interpretability and uses multi-objective genetic optimization. In [6], an integrated algorithm for simultaneous feature selection and designing of classifiers using multi-objective genetic programming. This method minimizes three objectives such as false positives, false negatives and the number of leaf nodes in the tree of genetic programming. Comparing to the above methods, the objective of the proposed method in this paper is different. In detail, after GNP extracts a large number of rules, the extracted rules are distributed to several clusters using MOGA, which means that the proposed method automatically generates sub-classes that potentially exist in the original classes. The proposed method aims to improve classification accuracy by adapting to various situations.

2. Association rule extraction by GNP

In the conventional method, a clustering using GA was applied to the rules extracted by GNP. In this section, the method of extracting class association rules is explained.

2.1 Class association rule

As X is the antecedent, and Y is the consequent, an association rule is denoted by X⇒Y. For example, a rule "people who buy milk and egg also buy bread" , is denoted by {milk, egg}⇒{bread}. A class association rule is denoted by $(A_1=1)\wedge(A_2=1)\Rightarrow$class1, where the consequent is changed to a class label.

2.2 Rule extraction by GNP

Class association rules are extracted by GNP from a database with attributes of binary values and class labels (Fig. 1). GNP has a structure as shown in Fig. 2. GNP consists of initial nodes and attribute nodes for rule extraction. According to the transition of attribute nodes, many candidate rules are generated. The node transition is repeated by a predefined number of times. The candidate rules that satisfy the evaluation criteria are stored in a rule pool. The evaluation criteria are support, confidence, and χ^2.

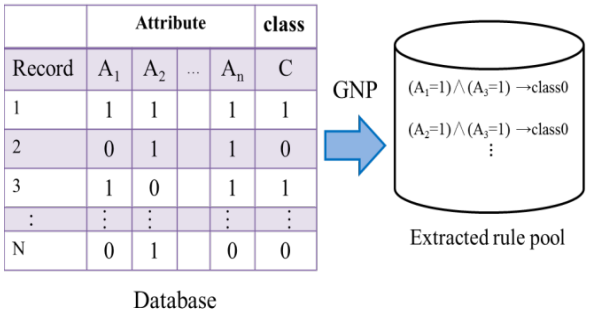

Fig. 1. Flow of rule extraction by GNP

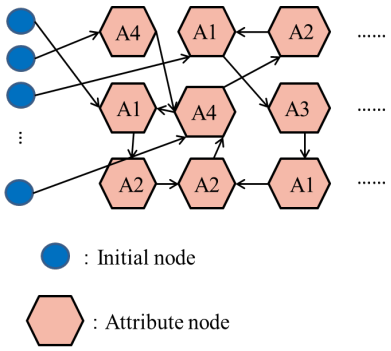

Fig. 2. Structure of GNP in rule extraction

2.3 Evolution of GNP

After storing rules, crossover and mutation are performed. In GNP, the crossover is performed by selecting two individuals, and the selected individuals exchange their node connections with each other. The mutation selects one individual, and connections of nodes are randomly changed. Fitness is calculated by Eq. (1). Eq. (1) means individuals that extract important rules are highly evaluated.

$$\text{Fitness}=\sum_{r=1}^{|R|}\left[\chi^2(r)+10\left(n(r)-1\right)\right]+\alpha_{new} \qquad (1)$$

$\chi^2(r)$: χ^2 value of class association rule r
$n(r)$: Length of rule r
α_{new} : Reward when a new rule is found
R : Set of the extracted rule numbers

The crossover and mutation are repeated by a predefined number of times.

3. Conventional rule clustering by GA

In this section, the conventional GA clustering for the class association rules is explained.

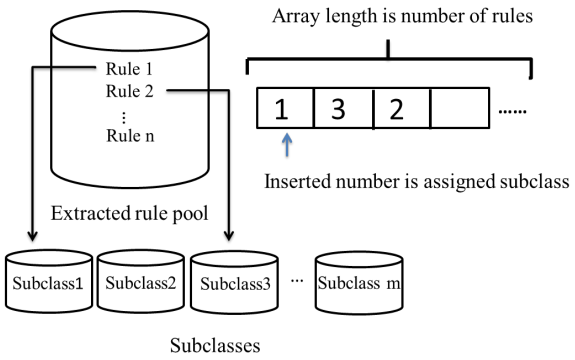

Fig. 3. Rule clustering by GA

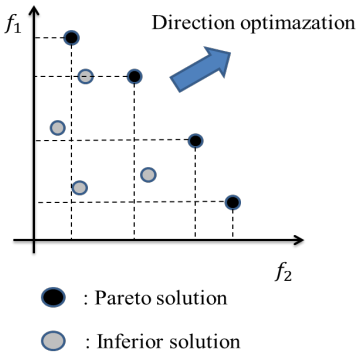

Fig. 4. Example of Pareto solutions

3.1 Individual initialization

First, several integer arrays are prepared, where the number of arrays is the same as the number of classes, and the length of each array is the same as the number of rules in the corresponding classes. All the arrays are initialized by random integer values ranging from 1 to *n*, where *n* means the number of subclasses. The random values in the arrays show the cluster numbers to which each rule is assigned. The flow of GA clustering is shown in Fig. 3.

3.2 Evolution of individuals

The fitness of GA is defined as the classification accuracy for training data, and the individuals for crossover and mutation are selected by tournament selection. Then, the crossover exchanges the genes of the selected individuals, and the mutation changes the genes randomly.

4. Proposed rule clustering by Multi-Objective Genetic Algorithm (MOGA)

In this section, a class association rule clustering using MOGA is explained. In the conventional method, the fitness is based only on the classification accuracy for the training data. Therefore, the direction of the evolution is biased toward the correct classification for the training data, which loses the generalization performance for new data. The proposed method adds two new criteria, optimizes these three criteria simultaneously, and aims to enhance the clustering accuracy. The two new criteria are inter-cluster variance and intra-cluster variance. Inter-cluster variance D_{ij}^2 and intra-cluster variance S_i^2 are calculated by

$$D_{ij}^2 = \sum_{k=1}^{a} \left(ce_{ik} - ce_{jk} \right)^2 , \tag{2}$$

$$S_i^2 = \sum_{r=1}^{n_i} \sqrt{\sum_{k=1}^{a} \left(ce_{ik} - e_{rik} \right)^2} , \tag{3}$$

$$ce_{ik} = \frac{1}{n_i} \sum_{r=1}^{n_i} e_{rik} , \tag{4}$$

where,
e_{rik} : value of k^{th} attribute of r^{th} rule in cluster i,
a : the total number of attributes in a database,
n_i : the number of rules in cluster i.

Inter-cluster variance, intra-cluster variance and accuracy are used as evaluation functions ($f_1 = accuracy$, $f_2 = D_{ij}^2$, $f_3 = 1/S_i^2$). The evaluation is higher as the values of the evaluation functions are larger.

4.1. Multi-objective algorithm

Considering the trade-off between the evaluation functions, MOGA is applied to the optimization.

4.2.1. Pareto solution

An example of Pareto solutions of the individuals in MOGA are shown in Fig. 4, where each axis corresponds to each evaluation function. The individuals that are not inferior to any other individuals are called Pareto solutions (non-dominated solutions).

Table 1. Comparison of the classification accuracy

Dataset	Classification accuracy [%]				
	Conventional method	Proposed method	J4.8	MLP	SVM
German	72.5	75.3	71.5	72.5	76.0
Pima Indian	77.1	80.4	75.8	77.1	77.1
Credit approval	86.6	88.0	86.4	82.4	85.9

4.2.2. Evolution in the proposed method

MOGA obtains the plural number of Pareto solutions (elite individuals), not one elite individual unlike standard GA. The evolution of MOGA is performed by crossover and mutation. For selecting crossover and mutation individuals, individuals are ranked, where the individual rank is defined by the number of superior individuals to itself, i.e., the rank of Pareto solution is 0.

5. Simulation results

The datasets named German, Pima Indian and Credit approval are used for evaluating the classification accuracy. The results of the accuracy obtained by the proposed method, conventional method, J4.8, MLP and SVM are shown in Table 1. The accuracy of the proposed method is the best one obtained by the Pareto solutions. From Table 1, the proposed method shows higher accuracy than the conventional method and also shows comparable or higher accuracy than J4.8, MLP, and SVM.

As for the computational time, the proposed method takes larger training time than other classifiers. Thus, it does not fit to online learning. However, we suppose that the proposed method is trained offline like typical evolutionary algorithms to obtain better classification accuracy. From the simulation results, the proposed method shows comparative or better accuracy than other methods and even a small improvement is effective to increase the number of accurate classifications when applied to big databases.

6. Conclusion

To enhance the performance of the classification system using class association rules, MOGA was applied to the rule clustering, and its classification ability was evaluated using three datasets. The results showed that the proposed method with MOGA generated solutions with high accuracy. The future work will develop a method selecting the best solutions from the generated Pareto solutions. In addition, the Pareto solutions should be analyzed in detail to find the importance of each evaluation function.

References

1. K. Azakami, S. Mabu, M. Obayashi, A. Kuremoto, Performance Enhancement of Classification Systems by Clustering Class Association Rules with Pruning and Its Evaluation, *The Conference Record of 66th Chugoku-branch Joint Convention of Institutes of Electrical and Information Engineers*, (2015)
2. K. Shimada, K. Hirasawa, and J. Hu, Genetic network programming with acquisition mechanisms of association rules, *Journal of Advanced Computational Intelligence and Intelligent Informatic*, **10**(1) (2006) 102-111.
3. UCI Machine Learning Repository, URL:http://archive.ics.uci.edu/ml/datasets.html/ (accessed 2016-5-20)
4. M. Antonelli, P. Ducange, B. Lazzerini, and F. Marcelloni, Multi-objective evolutionary design of granular rule-based classifiers, *Granular Computing*, **1**(1) (2016) 37-58.
5. M. B. Gorzałczany, F. Rudziński, A multi-objective genetic optimization for fast, fuzzy rule-based credit classification with balanced accuracy and interpretability, *Applied Soft Computing*, **40** (2016) 206-220
6. K. Nag and N. R. Pal, A multiobjective genetic programming-based ensemble for simultaneous feature selection and classification, *IEEE transactions on cybernetics*, **46**(2) (2016) 499-510.