Research Journal of Applied Sciences, Engineering and Technology 4(2): 127-130, 2012
ISSN: 2040-7467

Submitted: September 23, 2011 Accepted: November 02 , 2011 Published: January 15, 2012

Rules Extraction by Clustering Artificial Fish-swarm and Rough Set

Yingwei Huang, Bo Fu, Xinchen Cai, Xin Xing, Xinxing Yuan and Lu Yu
School of Electrical and Electric Engineering, Hubei University of Technology, Wuhan, China

Abstract: Due to the ill-conditioned problem caused by inefficient discretization approaches, it is difficult for the traditional rough set theory to extract accurate rules. And the continuous value needs to be discretized in the process of rule extraction. Then in this paper, a method based on clustering Artificial Fish-Swarm Algorithm (AFSA) and rough set theory is proposed to extract decision rules. Firstly, the clustering algorithm is used to classify attribute values in accordance with decision attributes. Secondly, the artificial fish-swarm algorithm is used to discretize the continuous attributes and to reduce the decision table. The experimental results indicate that the decision rules derived from the proposed method are much simpler and more precise.

Key words: Artificial fish-swarm, clustering, discretization, rough set, rule extraction

INTRODUCTION

The research of decision rule extraction, which is based on rough set theory, is an important issue in the field of artificial intelligence. It can be used to dig out useful hidden information from a large number of raw data about decision rules, and provides a effective solution for the rule extraction.

Rough set theory proposed by Polish scholar Pawlak (1982) is a powerful mathematical tool to process default information in a large amount of data information. Due to its capability of analyzing and dealing with various incomplete data without prior knowledge or other additional conditions, the rough set theory has been widely applied to many fields of data mining. At present, the rough set theory is mainly focused on attribute reduction and rule extraction. In recent years, scholars have proposed many decision rule extraction methods based on artificial neural networks (Thuan and James, 2010), genetic algorithms (Quteishat *et al.*, 2010), and etc. Wong and Ziarko (1985) has proved that solving the minimum attribute reduction and attribute value reduction problems are all NP-hard problems (Wong and Ziarko, 1985). Therefore, we can hardly find the accurate solution and our goal is to find a more precise and available one. In order to resolve the above problems, a method based on rough set theory integrated with clustering Artificial Fish-Swarm Algorithm (AFSA) is introduced for decision rule extraction.

Description of rough set theory: Quadruple S = (U, R, V, f) represents a knowledge representation system, where S is an information system, U is the nonempty finite set of the objects and R = C∪D is the nonempty finite set of attribute, in which C is condition attribute and D is decision attribute set; V is set of attribute values; f: U×A→V is called an information function assigning a value of attributes for every state, where f (x, a) ∈Va.

Definition 1: If there exists an attribute r∈R and IND(R) = IND(R-{r}), we call r is dispensable in R. Otherwise, r is indispensable in R. If every r∈R is dispensable in R, we call R is independent. Otherwise, R is dependent.

Definition 2:

$$r_c(D) = \frac{card(POS_c(D))}{card(U)} \quad (1)$$

where rc(D) is the dependency of attribute C on decision attribute D, card(•) is called as cardinality, POSC(D) is C-positive region of D which contains all objects that can be classified to one class of the classification U/D by attributes C.

Clustering artificial fish-swarm algorithm: Artificial Fish-swarm Algorithm (AFSA) is a stochastic searching optimization algorithm based on the simulation of fish behaviors. By imitating the fish behaviors of prey, cluster, the approach can achieve global optimization. With the advantages of strong robustness and being non-sensitive to initial values and parameters, the algorithm has good capacity to strike a global extremum.

Introduction of AFSA: The state of artificial fish individuals is X = (x1, x2, ..., xn), where xi (i = 1, 2, ..., n) are optimization variables. The current food concentration is denoted as Y = f(x), where Y is the objective function value. The distance between two artificial fishes is dij = ||Xi-Xj||. The ken of artificial fish

Corresponding Author: Yingwei Huang, School of Electrical and Electric Engineering, Hubei University of Technology, Wuhan, China

is represented as visual. The maximum of artificial fish step is expressed as Step. δ is congestion factor and Np is the scale of fish swamp. The representative behavior is described as follows:

- **Behavior of prey:** If the food concentration of a direction in the field of ken is superior to anywhere else, the artificial fish moves in the direction. Otherwise, it randomly selects a state. If it is still not satisfied with the forward condition after number times, the fish moves a step randomly

$$\begin{cases} f(X_j) > f(X_i)\ X_{next} = X_i + rand(step) \times \dfrac{X_j - X_i}{d_{ij}} \\ f(X_j) < f(Xv)\ X_{next} = X_i + rand(step) \end{cases} \quad (2)$$

- **Behavior of cluster:** If the center position of fish swamp is not crowded and the food concentration there is superior to the current condition, then the artificial fish moves a step toward the center position. Otherwise, it executes the behavior of prey

$$X_{next} = X_i + rand(step) \times \frac{X_c - X_i}{d_{cj}} \quad (3)$$

Clustering algorithm: Clustering algorithm can classify a large number of m dimensions samples (there are n samples and each sample has l kinds of attributes) into k (k<n) categories according to a certain standard. By this way, the samples in same categories have the maximal similarity and the samples in different categories have the minimal similarity. Generally, we minimize the sum of the distance between each sample and category center as objective function.

$$J(w,c) = \min \sum_{j=1}^{k} \sum_{i=1}^{n^j} \sum_{p=1}^{l} \left\| x_{ip} - c_{jp} \right\|^2 \quad (4)$$

$$c_{jp} = \frac{\sum_{i=1}^{n^j} w_{ij} x_{ip}}{\sum_{i=1}^{n^j} w_{ij}} \left(j = 1,...,k; p = 1,...,l \right) \quad (5)$$

$$w_{ij} = \begin{cases} 1, sample\ i \in class\ j \\ 0, sample\ i \notin class\ j \end{cases} \left(j = 1,...,k;\ \ i = 1,...,n_j \right) \quad (6)$$

where xjp is the p-th attribute of the i-th sample and cjp is the p-th attribute of the j-th class center.

Clustering artificial fish-swarm algorithm: The paper combines the clustering and artificial fish-swarm algorithm into a hybrid clustering artificial fish-swarm algorithm. Firstly, by referring to the clustering algorithm, we pre-process the decision table data information and divide breakpoint interval according to decision attribute. Thus, the number of breakpoint is reduced. Secondly, the artificial fish algorithm is used to discretize the pre-processed decision table to obtain the simplified decision rules.

The extraction of the most simplified decision rules depends on two aspects. The number of the contained attributes and attribute values is as few as possible and the dependency of condition attribute on decision attribute is as large as possible. Therefore, the objective function is determined by the number of breakpoint and the dependence and can be defined as:

$$F(x) = \left(1 - \frac{l_X}{n}\right) + a \times r_{C_i}(D) \quad (7)$$

where n is the total number of breakpoint with attribute x, lx is the number of breakpoint contained in the simplified x, α is the weight and rci(D) is the dependency of decision attribute on condition attribute set in discrete decision table divided by breakpoints.

On the basis of the above clustering artificial fish model, we describe the algorithm as follows:

Step1: Initialize parameters of artificial fish;Enter decision table S = (U, R, V, f) and rank the clustering according to the decision attribute.

Step 2: Calculate the support rating rci(D) from condition attribute Ci to decision attribute D and repeat the calculation several times to keep rci(D) to be the maximum.

Step 3: If i<n, let i = i+1 and return step 2. Otherwise, the flow goes to step 4.

Step 4: Let us calculate the objective function value according to formula (7). If the value of objective function continuously keeps the maximum, the flow goes to step 5. Otherwise, we set i = 1 and re-direct the flow to step 2.

Step 5: Record the breakpoint set represented by each artificial fish and determine the discrete interval and discretize decision table according to current state.

Step 6: Record reduction attribute set represented by each artificial fish and reduce the discretized decision table.

EXPERIMENTAL RESULTS

Fault diagnosis of Hydro-power Unit is very complex. Literature (Peng *et al.*, 2006) summarizes the vibration frequency features of Hydro-power unit. Based on these conclusions, we use the data in literature (Sun, *et al.*, 2007) as the tested decision table in our experiment.

Table 1: Fault decision table

U	a	b	c	d	e	f	g	h	i	D
1	0.01	0.03	0.68	0.96	0.82	0.06	0.97	0.95	0.45	1
2	0.01	0.10	0.75	0.92	0.81	0.05	0.93	1.00	0.50	1
3	0.01	0.02	0.80	0.98	0.80	0.02	0.98	0.98	0.05	1
4	0.06	0.05	0.92	0.52	0.48	0.03	0.96	0.01	0.16	2
5	0.06	0.08	0.98	0.5	0.50	0.03	0.96	0.08	0.03	2
6	0.01	0.12	0.93	0.02	0.20	0.07	0.97	0.17	0.01	3
7	0.01	0.02	0.96	0.05	0.10	0.12	0.89	0.13	0.09	3
8	0.05	0.95	0.07	0.02	0.01	0.02	0.10	0.95	0.08	4
9	0.08	0.07	0.10	0.07	0.05	0.98	0.08	0.98	0.05	4
10	0.04	0.02	0.98	0.06	0.02	0.01	0.98	0.03	0.98	5

Table 2: Discretized decision table by the proposed algorithm

U	a	b	c	d	e	f	g	h	i	D
1	1	1	2	3	3	1	2	3	2	1
2	1	1	2	3	3	1	2	3	2	1
3	1	1	2	3	3	1	2	3	1	1
4	3	1	3	2	2	1	2	1	1	2
5	3	1	3	2	2	1	2	1	1	2
6	1	1	3	1	1	1	2	2	1	3
7	1	1	3	1	1	1	2	2	1	3
8	3	2	1	1	1	2	1	3	1	4
9	3	1	1	1	1	2	1	3	1	4
10	2	1	3	1	1	1	2	1	3	5

Table 3: Simplified rules by RBF-RS

U	d	h	i	D
1	3	3	2	1
2	3	3	1	1
3	2	1	1	2
4	1	1	1	3
5	1	3	1	4
6	1	1	3	5

Table 4: Simplified rules by clustering AFSA-RS

U	a	d	D
1	1	3	1
2	3	2	2
3	1	1	3
4	3	1	4
5	2	1	5

There are 10 sets of vibration data in the tested database (Sun *et al.*, 2007). There are 3 sets of rotator misalignment, 2 sets of movement collision, 2 sets of mass imbalance, 2 sets of draft tube eccentric vortex strip and 1 set of heterogeneous pole in the database. We select six vibration frequency spectrum features, which are 0.18-0.2f, 1/6-1/2f, 1f, 2f, ≥3f, 50 and 100 Hz (f is the unit rotator frequency), and three attribute relations between vibration and speed, load, flux respectively as condition attributes labeled from a to i. Five kinds of faults, which are rotor misalignment, movement rubbing, mass imbalance, draft tube eccentric vortex strip, heterogeneous pole, are selected as decision attributes labeled from 1 to 5 respectively in the decision table. Then the Hydro-power Unit fault decision table is shown as Table 1.

Firstly, we use the final breakpoints set to discretize Table 1 and obtain the discretized decision table shown as Table 2.

Secondly, the RBF-RS method (Sun *et al.*, 2007) is used to reduce Table 1 and the simplified rules are recorded in Table 3. We can conclude that in Table 3, the quality of general classification is 1, reducing attributes are {d, h, i} and the number of the extracted decision rules is 6.

Thirdly, the clustering AFSA-RS algorithm is used to reduce the decision attribute Table 2 and the simplified rules are recorded in Table 4. We can conclude that in Table 3, the quality of general classification is 1, reducing attributes are {a, d} and the number of the extracted decision rules is 5.

We can summarize the diagnostic rules as follows:

Rule 1: ald3→D rotator misalignment
Rule 2: a3d2→D movement collision
Rule 3: a1d1→D unbalance of rotator mass
Rule 4: a3d1→D vortex strip of draft tube eccentric
Rule 5: a2d1→D heterogeneous pole

Compared with the RBF-RS method mentioned in literature (Sun *et al.*, 2007), the clustering artificial fish-swarm algorithm proposed in the paper is more efficient.

CONCLUSION

Based on the characteristics of rough set theory and clustering behavior of clustering artificial fish, a new method of decision rule extraction is put forward in this paper. We apply this algorithm to the fault diagnosis of hydropower generating unit. The experimental results show that on the basis of maintaining the overall classification ability of decision table, this algorithm greatly simplifies the data decision table and is an available method of rules extraction.

ACKNOWLEDGMENT

This study is supported by the Project of National Natural Science Foundation of China (No.61072130), the State Key Lab of Digital Manufacturing Equipment and Technology Open Project (No. DMEFKF2008010).

REFERENCES

Peng, W.J., X.Q. Luo and D.L. Zhao, 2006. Vibrant fault diagosis of hydro-turbine generating unit based on specrum analysis and RBF network method. Proceedings CSEE, 26(9): 155-158.

Quteishat, A., C.P. Lim and K.S. Tan, 2010. A modified fuzzy min-max neural network with a genetic-algorithm-based rule extractor for pattern classification. IEEE Trans. Syst. Man Cybernetics part A. Syst. Humans, 40(3): 641-650.

Sun, Q.Y., H.G. Zhang and X.R. Liu, 2007. Fault diagnosis of hydroelectric units based on rough set and RBF network. Chinese J. Sci. Instrument, 28(10): 1806-1810.

Thuan, Q.H. and A.R. James, 2010. Guiding hidden layer representations for improved ruleextraction from neural networks. IEEE Trans. Neural Networks, 22(2): 264-275.

Wong, S.K.M. and W. Ziarko, 1985. On optimal decision rules in decision tables. Bull. Polish Acad. Sci., 33(11-12): 693-696.