# Decision Rule Generation Using Data Mining Approach

Yann-Chang Huang, Huo-Ching Sun, Heng-Jiu Lu

*Department of Electrical Engineering, Cheng Shiu University*
*Kaohsiung, TAIWAN*
huangyc@csu.edu.tw

***Abstract*—This paper presents a novel data mining approach for fault diagnosis of turbine-generator units. The proposed rough set theory based approach generates the diagnosis rules from inconsistent and redundant information using genetic algorithm and process of rule generalization. In this paper, a fault diagnosis decision table is obtained from discretization of continuous symptom attributes in the data set. Then, the proposed genetic algorithm is used to achieve the minimal reduct from the discretized symptom attributes. In addition, a set of maximal generalized decision rules is obtained from the proposed rule generalization process.**

*Keywords*—**Data Mining, Fault Diagnosis System**

## 1. Introduction

The work on earlier detection of the incipient fault of the fundamental equipment in the power system, especially the steam turbine-generator units, has attracted quite much attention. In power system operation, the fault of the steam turbine-generator unit will lead itself to a very wide range of outage of the system. Due to the increasing capacity and structure complexity of steam turbine-generator units, the relations among components of the unit become closer than before. These causes resulting in the machine vibration become more various and complicated, but the vibration signal is still a very important information to evaluate the operating conditions of the machine.

In addition, the vibration fault diagnosis of steam turbine-generator unit is a complicated system diagnosis problem, and the real-time operating status of the unit provides useful information for machine condition monitoring and fault diagnosis. However, the raw data collected from the machine may exist in inconsistent and redundant information. Moreover, many fault symptom attributes extracted are needed to completely describe the fault condition of the machine. In contrast, the importance of different attributes is widely varied and some of them are even inconsistent and redundant. Therefore, this paper presents a rule extraction approach for fault diagnosis of turbine-generator units via rough set theory. The proposed approach can extract efficient and efficient diagnosis rules from inconsistent and redundant raw data set.

## 2. Proposed Data Mining Approach

This paper presents rough set theory (RST) to handle vagueness and uncertainty inherent in making decisions. RST has been applied to some branches of artificial intelligence and cognitive sciences, such as machine learning, knowledge discovery from databases, expert systems, inductive reasoning, pattern recognition and learning [1-3]. The basic definitions of the RST are briefly stated as follows.

An inform system *S* is an ordered pair *S* = (*U*,*A*), where *U* is a nonempty, finite set called the universe, *A* is a nonempty, finite set of attributes, i.e., *a*: $U \to V_a$, for $a \in A$, where $V_a$ is the set of values of *a*, called the domain of *a*. In the RST, the elements of the universe are referred to as objects which are characterized through their attribute values.

In the RST, an approximation space is a pair (*U*,*R*), where *R* is an equivalence relation defined on the universe *U*. Let *X* be a subset of *U*. The lower approximation of *X* by *R* in *S* is defined as:

$$\underline{R}X = \{x \in U : [x] \subseteq X\} \qquad (1)$$

and, the upper approximation of *X* by *R* in *S* is defined as:

$$\overline{R}X = \{x \in U : [x] \cap X \neq \text{Ø}\}, \qquad (2)$$

where [*x*] denotes the equivalence class containing *x*. A subset *X* of *U* is said to be *R*-definable in *S* if and only if $\underline{R}X = \overline{R}X$. The boundary set $B_R(X)$ is defined as $\overline{R}X - \underline{R}X$. The pair $(\underline{R}X, \overline{R}X)$ defines a rough set in *S*, which is a family of subsets of $U$ with the same lower and upper approximations as $\underline{R}X$ and $\overline{R}X$.

Let $S$ = ($U$,$A$) be an information system with $k$ objects. The discernibility matrix of $S$ is a $k \times k$ matrix with entries $c_{ij}$ consisting of the set of attributes from A on which objects $x_i$ and $x_j$ differ, i.e.,

$c_{ij} = \{a \in A : a(x_i) \neq a(x_j)\}$, for $i, j$ = 1, 2, …, $k$. (3)

A discernibility function $f_S$ for $S$ is a propositional formula of $n$ Boolean variables, $a_1^*, ..., a_n^*$, with respective to the attributes $a_1, ..., a_n$, defined as bellows.

$$f_S(a_1^*, ..., a_n^*) = \underset{1 \le j < i \le n}{\wedge} \; \underset{c \in c_{ij}^*, c_{ij} \neq 0}{\vee} c \qquad (4)$$

where $c_{ij}^* = \{a^* : a \in c_{ij}\}$.

The discernibility function $f_S$ describes constraints which must hold to preserve discernibility between all pair of discernible objects from $S$. It requires keeping at least one attribute from each non-empty element of the discernibility matrix corresponding to any pair of discernible objects. It can be shown that for any information system $S$ = ($U$,$A$) the set of all prime implicants of $f_S$ determines the set of all reducts of $S$. Moreover, the problem of finding a minimal reduct of a given information system is NP-hard was proved in [4].

A decision system $A$ = ($U$, $C${$d$}) is an information system for which the attributes are separated into disjoint sets of condition attributes $C$ and decision attributes $D$, $C \cap D = \emptyset$. A decision system is represented in the form of a decision table, in which its rows contain some objects and columns contain the values of attributes describing the objects, and the decision table contains rules specifying what decisions should be made when certain conditions are satisfied. Note that some redundant and inconsistent attributes may exist in the decision rules, and the reduction of the decision table must be further processed to eliminate these attributes.

## 3. FAULT DIAGNOSIS DECISION TABLE

A decision table must be established before using rough set for fault diagnosis of turbine-generator unit. Seven typical frequency bands of the vibration signal based on the running frequency $f$, i.e., (0.3-0.44)$f$, (0.45-0.6)$f$, $f$, 2$f$, 3$f$, 4$f$, $\geq 4f$, can be used to identify the machinery vibration fault [5]. The peak value of power spectrum in each band indicates a fault symptom at a particular frequency. A fault symptom attribute set is represented by $C = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7\}$, $V_{ci} = \{0,1\}$, $i$ = 1,2, …, 7. The decision attribute set is denoted $D = \{d\}$, $V_d = \{d_1, d_2, d_3, d_4\}$. In this paper, three typical vibration fault types are studied, oil-film oscillation $d_1$, unbalance $d_2$, asymmetry $d_3$, and normal state $d_4$.

After seven fault symptoms are normalized, the boundary values selected by experience are used to discretize the domains of the attributes into intervals according to different conditions [5].

If $c_i \in$ (0.35, 1], then $c_i$ = 1; otherwise $c_i$ = 0,
for $i$ = 1,2,3.
If $c_j \in$ (0.2, 1], then $c_j$ = 1; otherwise $c_j$ = 0,
for $i$ = 4,5,6,7.

Table 1 lists 300 cases of a turbine-generator unit vibration fault symptom after attribute discretization, where $n$ represents the number of cases with the same attributes.

**TABLE 1**

**FAULT DIAGNOSIS DECISION TABLE**

| ***U*** | ***n*** | **condition** | | | | | | | **decision** |
|---|---|---|---|---|---|---|---|---|---|
| | | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $c_7$ | ***D*** |
| 1 | 10 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| 2 | 20 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 3 | 10 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 4 | 30 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 5 | 10 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 6 | 10 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 7 | 10 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 8 | 10 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 2 |
| 9 | 10 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 2 |
| 10 | 10 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 2 |
| 11 | 10 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 2 |
| 12 | 10 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 |
| 13 | 10 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 2 |
| 14 | 20 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 15 | 10 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| 16 | 10 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 3 |
| 17 | 10 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 3 |
| 18 | 10 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 3 |
| 19 | 10 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 3 |
| 20 | 10 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 3 |
| 21 | 10 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 3 |
| 22 | 30 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 3 |
| 23 | 10 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 3 |
| 24 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |

## 4. DECISION RULE GENERATION

In this paper, attribute reduction is a process to obtain a subset from the original set of symptoms of the given fault patterns. The attribute reduction is solved by an optimal process guided by the proposed genetic algorithm (GA) [6,7], and the proposed GA is used to compute the minimal reduct.

Let DM be the discernibility matrix of the decision table, as described in Table 1. $C = \{c_1, c_2, \ldots, c_7\}$ is the set of all symptom attributes, $S$ is a set and consists of all the s sets of attributes combination in DM, except for repeated item. Therefore, $S$ can be represented as $B_m \in S$, $B_n \in S$, $B_m \neq B_n$ for $m$, $n$ =1, 2, …, $s$. The proposed GA for process of computing the minimal reduct can be briefly described as follows.

The decision variables stands for a minimal reduct, and are encoded as a $i$-bit chromosome in a string, where $i$ is the number of symptom attributes (in this paper, $i$ = 7). In each bit representation, '1' means that the corresponding attribute is present, and '0' means it is not present.

Each string $B$ describes one tentative minimal reduct. As described in [8, 9], the proposed GA has support for both cost information and approximate solutions. The fitness function $f$ in the GA is defined as

$$f(B) = \beta \times \frac{|A| - |B|}{|A|} + (1-\beta) \times \min\left\{h, \frac{|[s \ in \ S \mid S \cap B \neq \phi]|}{|S|}\right\} \quad (5)$$

where $\beta$ is a weighting between subset cost and hitting fraction; $S$ is the set of sets corresponding to the discernibility function; $h$ is a hitting fraction and is relevant in the case of approximate solutions [10]. The subsets $B$ of A that are found through the evolutionary search of the GA driven by the fitness function and that are good enough hitting sets, i.e., have a hitting fraction of at least h, are collected in a keep list. The size of the keep list can be specified. The function cost specifies the cost of an attribute subset. In this paper, $|Y|$ represents the cardinality of set $Y$.

As described in (5), the first term rewards the little elements in the reduct $B$, and the second term intend to ensure that the rewards a good enough hitting set with a hitting fraction of at least $h$. Approximate solutions are controlled through two parameters, $h$ and $k$. The parameter $h$ signifies a minimal value for the hitting fraction, while $k$ denotes the number of extra keep lists in use by the algorithm. If $k$ = 0, then only minimal hitting sets with a hitting fraction of approximately $h$ are returned. If k > 0, then $k$+1 groups of minimal hitting sets are returned, each group having an approximate (but not smaller) hitting fraction evenly spaced between $h$ and 1. Note that $h$ = 1 implies proper minimal hitting sets.

The paper presents GA to optimize the fitness function through reproduction, crossover, and mutation of the GA operation. The optimization process stop if average population fitness of the pre-setting number of generations to wait for fitness to improve does not improve or if keep list does not change.

The minimal reduct $\{c_2, c_3, c_4, c_5\}$ of the symptom attribute set in Table 1 can be achieved. As shown in Table 2, there are 13 decision rules can be generated by the minimal reduct $\{ c_2, c_3, c_4, c_5\}$ from the proposed GA; however, inconsistent cases exist in the decision rules. Table 2 shows that rules 1, 2, 3, 6 and 7 are inconsistent rules, and each rule has the same values of symptom attributes, but possess different decision attributes.

**TABLE 2**

**DECISION RULES USING REDUCT OF ATTRUBITES**

| No. | Decision rules | | | | |
|---|---|---|---|---|---|
| | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $D$ |
| 1 | 1 | 1 | * | * | 1, 2 |
| 2 | 1 | * | 1 | * | 1, 3 |
| 3 | * | 1 | * | 1 | 1, 3 |
| 4 | 0 | 1 | * | 0 | 2 |
| 5 | 1 | 0 | * | 1 | 1 |
| 6 | * | 1 | 1 | * | 2, 3 |
| 7 | * | 1 | * | 1 | 2, 3 |
| 8 | 0 | * | 0 | 1 | 2 |
| 9 | 0 | 1 | * | 1 | 3 |
| 10 | 0 | 0 | 1 | * | 3 |
| 11 | | * | 1 | 0 | 3 |
| 12 | 0 | 0 | * | 1 | 3 |
| 13 | 0 | 0 | 0 | * | 4 |

*: don't care

## 5. DECISION RULE GENERALIZATION

A set of decision rules can be generated by the minimal reduct obtained from the GA. However, the rules may be inconsistent. This paper develops a process for maximum generalized decision rules from imprecise, incomplete and inconsistent diagnosis rules [11]. In the process of rule generalization, the maximal number of symptom attribute values is removed without losing essential information, and the maximum generalized rules can be achieved. The process for obtaining the maximum generalized rules is stated as follows.

1) A attribute is removed from a rule, and then the rule is checked for consistency with the other rules in the rule set.

2) If the rule is consistent with the other rules, then the attribute in the rule can be removed; otherwise, the attribute must be retained.

3) The process of attribute simplification is repeated until all attributes of the rule have been tested.

4) If each rule in the rule set has been proceed, then a maximum generalized decision rules are achieved.

As listed in Table 3, 11 maximal generalized decision rules extracted from the original decision table agree well with the information shown in Table 1. This demonstrates the effectiveness of the proposed approach.

**TABLE 3**

**MAXIMAL GENERALIZED DECISION RULES**

| **No.** | **Decision rules** | | | | |
|---|---|---|---|---|---|
| | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $D$ |
| 1 | * | 0 | * | 0 | 1 |
| 2 | 1 | * | * | 0 | 1 |
| 3 | 1 | 0 | * | * | 1 |
| 4 | * | 1 | * | 0 | 2 |
| 5 | 0 | 1 | * | * | 2 |
| 6 | 0 | * | 0 | 1 | 3 |
| 7 | * | 0 | 1 | * | 3 |
| 8 | 0 | * | 1 | * | 3 |
| 9 | * | 0 | 0 | 1 | 3 |
| 10 | 0 | 0 | * | 1 | 3 |
| 11 | 0 | 0 | 0 | * | 4 |

## 6. CONCLUSIONS

This paper has presented an effective and efficient data mining approach to generate diagnosis rules from inconsistent and redundant data set of turbine-generator units using rough set theory. The generated diagnosis rules can effectively reduce space of symptom attribute, simplify knowledge representation and fault diagnosis process. In this paper, the machine fault diagnosis decision table is first built using discretized attributes. Then, the GA based optimization process is used to obtain the minimal reduct of symptom attributes. Finally, the rule generalization process is used to achieve the maximal generalized decision rules, which can be derived from inconsistent and redundant information.

## REFERENCES

[1] Z. Pawlak, Rough Sets, International Journal of Information and Computer Science, Vol. 11, No. 5, pp. 341-356, 1982.

[2] Z. Pawlak, J. Grzymala-Busse, R. Slowinski and W. Ziarko, Rough Sets, Communications of the ACM, Vol. 38, No. 11, pp. 89-95, 1995.

[3] Z. Pawlak, Rough Sets and Intelligent Dada Analysis, Information Sciences, Vol. 147, No. 1, pp. 1-12, 2002.

[4] A. Skowron, C. Rauszer, The Discernibility Matrices and Functions in Information Systems, in: R. Slowinski (Ed.), Intelligent Decision Support—Handbook of Applications and Advances of the Rough Sets Theory, System Theory, Knowledge Engineering and Problem Solving, Vol. 11, pp. 331-362, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992.

[5] W. Huang, X. Zhao, W. Wang, and L. Dai, "Rough Set Model for Vibration Fault Diagnosis of Steam Turbine-Generator Set," Automation of Electric Power Systems, Vol. 28, No. 15, pp. 80-84, 2004.

[6] D. E. Goldberg, Genetic Algorithm in Search, Optimization and Machine Learning, Addison-Wesley, 1989.

[7] C. T. Lin, and C. S. Lee, Neural Fuzzy Systems, Prentice Hall, 1999.

[8] S. Vinterbo and A. Øhrn, Minimal Approximate Hitting Sets and Rule Templates. In Predictive Models in Medicine: Some Methods for Construction and Adaptation, Department of Computer and Information Science, NTNU report, Dec. 1999.

[9] S. Vinterbo and A. Øhrn, "Minimal Approximate Hitting Sets and Rule Templates," International Journal of Approximate Reasoning, Vol. 25, No. 2, pp. 123–143, 2000.

[10] A. Øhrn, Discernibility and Rough Sets in Medicine: Tools and Applications, PhD thesis, Norwegian University of Science and Technology, Department of Computer and Information Science, NTNU report, Dec. 1999.

[11] X. Hu and N. Cercone, "Discovering Maximal Generalized Decision Rules through Horizontal and Vertical Data Reduction," Computational Intelligence, Vol. 17, No. 4, pp. 685-702, 2001.