Front. Mech. Eng. China (2006) 2: 162–167
DOI 10.1007/s 11465-006-0017-z

RESEARCH ARTICLE

YANG Ping

Data mining diagnosis system based on rough set theory for boilers in thermal power plants

Abstract Large amounts of data in the SCADA systems' databases of thermal power plants have been used for monitoring, control and over-limit alarm, but not for fault diagnosis. Additional tests are often required from the technology support center of manufacturing companies to diagnose faults for large-scale equipment, although these tests are often expensive and involve some risks to equipment. Aimed at difficulties in fault diagnosis for boilers in thermal power plants, a hybrid-intelligence data-mining system based only on acquired data in SCADA systems is structured to extract hidden diagnosis information directly from the SCADA systems' databases in thermal power plants. This makes it possible to eliminate additional tests for fault diagnosis. In the system, a focusing quantization algorithm is proposed to discretize all variables in the preparation set to improve resolution near the change between normal value to abnormal value. A reduction algorithm based on rough set theory is designed to find minimum reducts from all discrete variables in the preparation set to represent diagnosis rules succinctly. The diagnosis rules mining from SCADA systems' database are expressed directly by variables in the database, making it easy for engineers to understand and use in industry applications. A boiler fault diagnosis system is designed and realized by the proposed approach, its running results in a thermal power plant of Guangdong Province show that the system can satisfy fault diagnosis requirement of large-scale boilers and its accuracy rangers from 91% to 98% in different months.

Keywords fault diagnosis, data mining, focusing quantization, reduction

YANG Ping (✉)
Electric Power Collage,
South China University of Technology,
Guangzhou 510640, China
Email: eppyang@scut.edu.com

1 Introduction

With developments in modern science and technology, competition has become very fierce. Equipments in industrial production are becoming larger, more intelligent and more complex. The structure of larger-scale equipment is so complex that there are strong nonlinearity and randomness in the relationship between their faults and fault symptoms [1]. Wide applications of large-scale equipment bring huge economic benefits, but the investment and maintenance of these equipments are costly and the accidents arising from their use are serious. This makes fault diagnosis of large-scale equipment attract more attention to ensure their safety and reliability in modern industrial production.

Fault diagnosis of large-scale equipment includes four steps [2]: fault signal acquisition, fault symptoms extraction, fault orientation and diagnosis decision. These four steps constitute a diagnosis circle. Usually, the diagnosis circle should be executed several times to find out a complex fault of large-scale equipment. Every diagnosis circle will make the diagnosis problem clearer than the last one and the diagnosis problem can be solved by several diagnosis circles. Nowadays, computer monitoring and control system is widely used in industrial engineering and more and more data about equipments are collected everyday through this system. The large databases are usually very difficult to handle even by human experts due to their high-dimensionality and noise. This makes it hard to execute a certain diagnosis circle. The most important problem is how to extract the hidden diagnosis knowledge from vast amounts of collected data.

Data mining is a method of extracting meaningful and interesting information directly from large amounts of data. In recent years, data mining applications have been successfully applied in many areas [3, 4] such as astronomy, molecular biology, medicine, geology etc. Data mining applications can provide an effective way for fault diagnosis of large-scale equipment by extracting potential diagnosis knowledge from large amounts of collected data. In order to improve the efficiency of data mining, lots of algorithms have been presented. As a method of knowledge discovery, rough set theory is one of the theories in common use for data mining. Reduction algorithm based on rough set theory can simplify the knowledge extracted from large amounts of data.

Based on the analysis of the relation between a fault and collected data from SCADA systems' databases in a thermal power plant, a hybrid-intelligence data-mining system is structured to extract hidden diagnosis information for boilers directly from SCADA systems' databases. This makes it possible to eliminate additional tests for fault diagnosis. In this paper, a focusing quantization algorithm is proposed to discretize all variables in a preparation set collected from SCADA systems' databases. The focusing quantization algorithm improves the resolution near the variables' change from normal value to abnormal value. Then a reduction algorithm based on rough set theory is designed to find minimum reducts from all discrete variables in the preparation set to represent diagnosis rules succinctly. The diagnosis rules mining from SCADA systems' database are expressed directly by variables in the database, so it is easy for engineers to understand and use in industry applications. Based on focusing quantization algorithm and the reduction algorithm, a fault diagnosis system for boilers in thermal power plants is designed and realized, its running results in a thermal power plant of Guangdong Province show that the system can satisfy fault diagnosis requirement of large-scale boilers.

2 Structure of data mining diagnosis system for boilers

The boiler is one of the most important equipment in thermal power plants and is monitored and controlled by a SCADA system. Large amounts of variables in a SCADA systems' database provide useful information on boilers. These variables include analog input variables, digital input variables, calculated value variables, enter value variables, pulse input variables, and S.O.E. (sequence of events) input variables, etc. Generally, process variables of boilers are sampled per second and stored to the SCADA systems' database once per minute, totalling 1 440 real numbers that represent the readings of one variable from 0 hour 0 minutes to 23 hour 59 minutes. They are called slow-varying data. According to characteristics of boilers, some mathematical relationships exist between the variables' data and boiler states, the abnormal states can be represented by the change of some variables. If relationships between faults and variables' data can be found directly by mining diagnosis knowledge from SCADA systems' database, plant operators can solve fault problems by controlling the variables' value in a certain range.

As a method of extracting meaningful and interesting information directly from large amounts of data, it is possible for data mining to get useful information related to boilers' faults directly from the SCADA systems' database in thermal power plants. In order to mine diagnosis rules directly from SCADA systems' database, the process of data mining should be translated into practical steps. According to the characteristics of data mining technology, five steps can be designed for data mining.

The first step is problem description. This step is to define the target of data mining and organize raw data. It involves description of fault type, fault process, fault diagnosis process, understanding of SCADA systems' database in thermal power plant, and basic knowledge of thermal power plant.

The second step is data selection. In this step, data mining experts and domain experts analyze the investigated faults according to fault process and recognize data that are related to faults possibly from raw data.

The third step is data pre-processing. Selected data in the second step contain many variables that are not related to the investigated faults. In order to reduce the calculation of data mining algorithm, data pre-processing should be done first. Generally, correlation analysis is applied to calculate correlation coefficients of data to faults to get a preparation set of variables from selected data.

The fourth step is data mining. Preparation set of variables still contains many variables that are not tightly related to the investigated faults. A data mining algorithm should be used to mine the minimal set, which expresses the investigated faults from the preparation set of variables.

The fifth step is evaluation of data mining results. This step is to confirm data mining results and decide if the data mining process should be redone or given up.

In the entire process of data mining, the first three steps reduce the amount of raw data by up to two orders of magnitude, still presenting the original characteristics of raw data to investigated faults, but contains too many variables that are not tightly related to the investigated faults. The fourth step reduces data to a manageable set of variables then faults can be represented by a small quantity of variables. Then we can eliminate faults by controlling these variables' value in a certain range. Based on the five steps of data mining and the characteristic of process variables in SCADA systems' database, a hybrid-intelligence data-mining system is structured to extract hidden diagnosis information for boilers directly from SCADA systems' database in thermal power plants. The system structure is shown in Fig 1.

3 Realization of data mining diagnosis system for boilers

Based on the system's structure proposed in the last section, a hybrid-intelligence data-mining system is realized to diagnose the boilers' faults in a thermal power plant. The system contains five modules.

The first module is raw data organization, which organizes raw data from the SCADA systems' database according to the data mining target. In order to organize raw

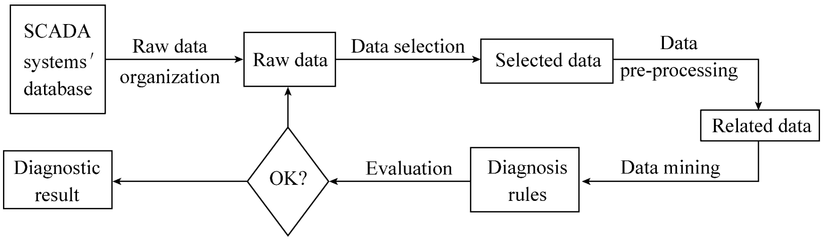

Fig .1 Structure of data mining diagnosis system for boilers in thermal power plant

data, fault type, fault process, acquired data and the basic knowledge of boilers are presented first, then data mining experts analyze faults according to fault process and recognize raw data possibly related to faults. In a thermal power plant, large numbers of variables are stored in its SCADA systems' database, including analog input variables, digital input variables, calculated value variables, enter value variables, pulse input variables, and sequence of events (S.O.E.) input variables, etc. In this system, raw data are collected variables in the SCADA systems' database. We choose all process variables from the SCADA systems' database every other month as a data mining set (raw data), while the remaining data are regarded as a test data set for evaluation of data mining results.

The second module is data selection, which selects data possibly related to faults from raw data according to the experiences of data mining experts and domain experts. According to boiler characteristic, selected data are all analog input (AI) variables and all calculated value (CV) variables chosen from the SCADA systems' database.

The third module is data pre-processing, which chooses data related to faults closely from the selected data. This module includes four steps as follows [5]:

1. Correlation analysis. The correlation coefficients of all variables to faults in the SCADA systems' database are calculated. All variables whose correlation coefficients to faults are less than a certain value are deleted, the remaining variables are called the related set.
2. Principal component analysis. All variables are mapped to another linear space to get principal components. When the proportion of information is determined, a principal component set of variables is obtained to express all variables.
3. Domain expert experience. An experiential set of variables is obtained according to domain experts.
4. Finally, a preparation set for data mining is obtained from the union of related set, principal component set and experiential set.

The fourth module is data mining. After data pre-processing, a preparation set of variables is obtained, but it still has redundant variables and contains many variables that are not tightly related to faults. Thus the data mining module should be applied to find a minimum set of variables from the preparation set. The main core of this module is a focusing quantization algorithm and a reduction algorithm. The focusing quantization algorithm is proposed to discretize all variables in the preparation set to improve resolution near their change from normal value to abnormal value, and then diagnostic accuracy can be improved. Reduction algorithm based on rough set theory is designed to find a minimum set of variables (reducts) that represent diagnosis rules.

The fifth module is evaluation of data mining results to confirm the results and then visualize them.

In these five modules, the fourth module of data mining is the most important part of the system. It contains a focusing quantization algorithm and a reduction algorithm. They are presented in detail here.

3.1 Focusing quantization algorithm

Reduction algorithm based on rough set theory should be applied to discrete-valued variables, so continuous-valued variables must therefore be discretized first. Focusing quantization algorithm is used to discretize all variables in the preparation set, which includes five steps. First of all, a suitable focusing factor F should be chosen according to the actual situation of a variable, where $F \in (0,1)$, then focusing quantization of a variable can be done by using F. The five steps are as follows:

1. Define the domain area of each variable and its quantization number, not losing the generality, let the quantization number be $2n+1$.
2. To make sure of the variables' type. The first type is the high-limited variable whose value is the smaller, the better; the second type is the low-limited variable whose value is the bigger, the better; the third type is the bi-limited variable whose value is appreciated to be mezzo.
3. For the first type of variable, the intersection point M between its normal value and abnormal value is regarded as the focus, the lowest value 0 is regarded as the staring point, the maximal value of the variable is A, and then this variable is quantized to $2n+1$ parts in value domain $[0, A]$, the division point is as follows:
   $M \times F$
   $M \times F + M \times F^2$
   ……
   $M \times F + M \times F^2 + M \times F^3 + \ldots + M \times F^n$
   $N \times F - N \times F^2 - N \times F^3 - \ldots - N \times F^n + M$
   ……

$N \times F - N \times F^2 + M$

$N \times F + M$

Where $N = A - M$, then the variable is quantized to {0,1,2,…,2n}.

4. For the second type of variables, the intersection point M between its normal value and abnormal value is also regarded as the focus, the maximal value B of the variable is regarded as the staring point, the lowest value is 0, and then the variable is quantized to 2n+1 parts in value domain [0, B], the division point is as follows:

$N \times F + M$

$N \times F - N \times F^2 + M$

$N \times F - N \times F^2 - N \times F^3 + M$

……

$N \times F - N \times F^2 - N \times F^3 - \ldots - N \times F^n + M$

$M \times F + M \times F^2 + M \times F^3 + \ldots + M \times F^n$

……

$M \times F + M \times F^2 + M \times F^3$

$M \times F + M \times F^2$

$M \times F$

Where $N = B - M$, then the variable is quantized to {0,1,2,…,2n}.

5. For the third type of variables, the intersection point between its normal value and abnormal value is also regarded as the focus. Here, there are two focuses, M_1 and M_2, because the normal value is in the middle part of value domain [0, D], let the best value of variables be C. Because a boiler is abnormal when these variables' value is bigger than the upper limit or less than the lower limit of normal value, the best value of the variables is regarded as a center and then two parts are derived to be quantized respectively.

If the value of variables is in [0, C], they can be quantized by the same method as the second type and the variables in [0, C] are quantized to 2n+1 parts {0,1,2,…,2n}. If the value of variables is in [C, D], they can be quantized by the same method as the first type and the variables in [C, D] are quantized to 2n+1 parts {0,1,2,…,2n}.

For example, let F=0.5, n=2, the results of focusing quantization are shown in Fig. 2.

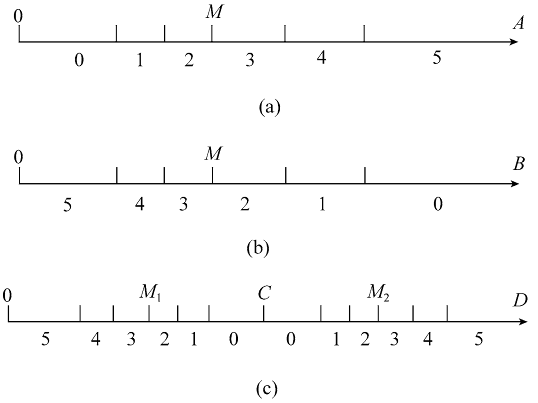

Fig. 2 The results of focusing quantization. **a** To the first type of variabies **b** To the second type of variables. **c** To the third type of variables

After all variables in the preparation set have been quantized by focusing quantization algorithm, the variables are converted to a unilateral value. A value of 0 represents the best value of each variable. Then it is convenient to apply reduction algorithm to mine the minimal set of variables from the preparation set. In addition, transition area of all variables from normal value to abnormal value is more particular than the conventional quantization method. So the focusing quantization algorithm can improve the resolution near the focus (the change from normal value to abnormal value) so that diagnostic accuracy can be improved obviously [6].

3.2 Reduction algorithm

After all variables in the preparation set have been quantized by focusing quantization algorithm, a reduction algorithm based on rough set theory is proposed to mine the minimal set from these discrete-valued variables. In order to obtain diagnosis rules for fault diagnosis, variables in the preparation set are regarded as condition attribute and faults are regarded as decision attribute, and then a decision table is established. Reduction algorithm based on rough set theory can generate a minimal variable set (minimal reduct) directly from the preparation set, representing faults classification at the highest accuracy. Steps of the reduction algorithm are as follows:

1. After quantization, a 2-dimension table is established by a month's data from the preparation set; all the real numbers of a variable line up as a condition column, a fault status lines up as a decision column, then the table has n+1 columns and 1 440 × d rows, n is the number of variables in the preparation set, d is the number of days in the assigned month.
2. According to rough set theory, calculate all reducts of condition attributes by global search and then find the minimal reduct.
3. Sum up the number of duplicate rows to get the reliability of this row, then eliminate the duplicate rows.
4. If all the condition attributes are the same but the corresponding decision attribute is different, eliminate the rows whose summation is less than 30 percent of this condition attributes' summation.
5. Eliminate the superfluous values of attributes, the minimal variable set representing the fault classification at the highest accuracy in this month is obtained.

In practical application, conflict data always exist in a SCADA systems' database. Step 5 is necessary to avoid the adverse influence by fluctuating data.

Based on another month's data, step 1 to 5 should be repeated until all data in the SCADA systems' database have been considered. Then the intersection of all the minimal variable sets of all months is calculated to get the final minimal set of variables.

After data mining, evaluation should be carried out to confirm data mining results by using the test data set. The evaluation algorithm is as follows:

1. Establish the final decision table by the final minimal set of variables.
2. Calculate the accuracy of the final decision table according to a month's data in the test data set.
3. Based on another month's data, repeat Step2 until all data in the test data sets have been checked.

If data mining is performed on enough data and the final decision table satisfies the test data set, it can be used in industry applications.

4 Experimental results

Using a focusing quantization algorithm and reduction algorithm based on rough set theory, an intelligent data-mining system is designed to extract hidden diagnosis information directly from data in a SCADA systems' database. Five years of data in a SCADA system's database of a thermal power plant in Guangdong Province are collected to test the proposed approach.

In a thermal power plant, overall thermal efficiency is the most important financial index. The temperature on both sides of the boiler's super-heater outlet should be balanced, otherwise, the high temperature of one side should be dropped to the same low temperature on the other side using cold water. Some thermal efficiency should be lost in this process. In the investigated thermal power plant of Guangdong Province, overall thermal efficiency was not good, because the temperature on the left side and right side of the boiler's super-heater outlet was not balanced. In order to find the cause of the temperature difference of the boiler's super-heater, we collect five years' data from this thermal power plant. We select data from the SCADA system's database every other month and use them as data mining set, while the remaining data is regarded as the test data set.

The super-heater circuits of the boiler in the investigated thermal power plant are shown in Fig. 3. In this thermal power plant, the temperature of T_1 and T_2 is not balanced. Let $D=T_A-T_B$. According to domain experts' experience, if D is less than 10℃, it is considered good. If D is greater than 10℃ but not greater than 20℃, it is not good but acceptable. If D is greater than 20℃, it is a fault.

The collected data includes analog input variables, digital input variables, calculated value variables, enter value variables, pulse input variables, S.O.E. input variables, totaling 6 082 variables. These are raw data. In order to reduce calculation work, we select data possibly related to temperature difference from the raw data. According to the behavior of the super-heater, the selected data are analog input points and calculated value points, totaling 2 364 variables. So we derive analog input variables and calculated value variables from the SCADA system's database. After data pre-processing, we can get a preparation set of variables which contains 465 variables.

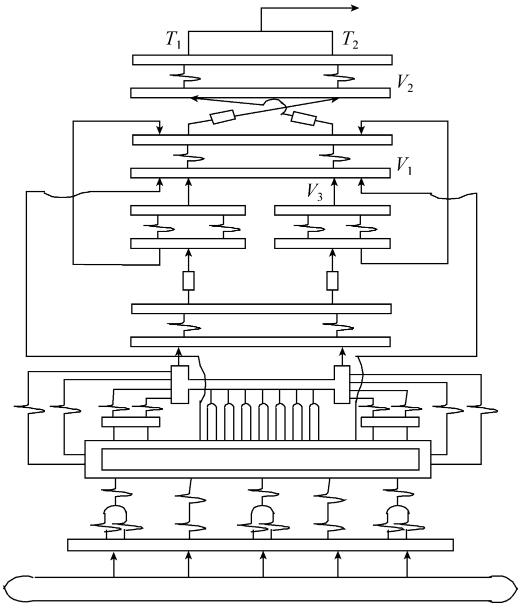

Fig. 3 The super-heater circuits

Applying the proposed focusing quantization algorithm and reduction algorithm in Sect. 3, let F=0.5, n=2, a minimal variable set for a certain month is obtained by the designed system. Then the intersection of minimal variable sets of all months in the data mining set is calculated to obtain the final minimal variable set, which contains three condition attributes. The final decision table consisting of these three condition attributes is shown in Table 1. The accuracy of diagnosis rules expressed by the final decision table varies from 91 percent to 98 percent in various months.

Table 1 Decision table for February, 2002

V_1	V_2	V_3	D
0,1,2	0,1	0,1,2,3	2,3
0,1,2	0,1	0,1,2,3	0,1
3,4	0,1	0,1,2,3	4
*	2,3,	0,1,2,3	2,3
*	4,5	0,1,2,3	4,5
0,1,2,3	4,5	0,1,2	4,5
0,1,2,3	2,3	0,1,2	2,3

In Table 1, '*' means don't care. V_1 is metal temperature of the tertiary super-heater on the right side, V_2 is metal temperature of the fourth super-heater outlet on the right side, V_3 is steam temperature of the tertiary super-heater inlet on the right side, D is the temperature difference.

5 Conclusions

Based only on acquired data in the SCADA system's database, a hybrid-intelligence data-mining system is structured to extract hidden diagnosis information for a large-scale boiler in a thermal power plant. This makes it possible to eliminate additional tests or experiments for fault diagnosis, which are often expensive and involve risks to boilers. In the system, a focusing quantization method is adopted to discretize all variables in the preparation set, so the transition area of all variables from normal value to abnormal value is more particular than the conventional quantization method. This quantization algorithm improves the resolution near the focus of each variable so that the diagnostic accuracy can be improved obviously. The main core of the system is a reduction algorithm based on rough set theory for finding a minimum set of variables. The diagnosis rules mining from the SCADA system's database are expressed directly by variables in the database, so it is easy for engineers to understand and use them in industry applications.

A fault diagnosis system of boilers is designed and realized by the proposed approach, its running results in a thermal power plant of Guangdong Province shows that the system can satisfy fault diagnosis requirements of a large-scale boiler. Its accuracy varies from 91 percent to 98 percent in different months. It can be expected that data mining should be applied to diagnose faults for large-scale equipment in thermal power plants in the near future.

References

1. Yang Ping, Study on intelligent fault diagnosis of complex system[D], South China University of Technology, Guangzhou, 1998, 40–67
2. Yang Ping, Liu Sui-sheng, Fault diagnosis for boilers in thermal power plant by data mining, Proceedings of the 8th International Conference on Control, Automation, robotics and vision, Kunming, China, 2004, 2, 2175–2179
3. Tony Ogilvie, E. Swidenbank, B. W. Hogg, Use of data mining techniques in the performance monitoring and optimization of a thermal power plant, IEE Colloquium on Knowledge Discovery and Data Mining, London, UK：IEEE Press, 1998：7/1–7/4
4. Mejía-Lavalle Manue, Rodríguez-Ortiz Guillermo, Obtaining expert system rules using data mining tools from a power generation databases,. Expert systems with application, 1998, 14(1-2)：37–42
5. Lebrevelec C. Cholley P. Quenet J.F. Wehenkel, L., A statistical analysis of the impact on security of a protection scheme on the French power system 1998 International Conference on Power System Technology, Beijing, China：IEEE Press, 1998, 2：1102–1106
6. Yang Ping, Liu Sui-sheng, Microwave measurement system of density and moisture content based on focus fuzzy clustering algorithm, Proceedings of 3rd international symposium on instrumentation science and technology, Xi'an, China, 2004, 1, 922–928