Article ID: 1000 - 8152(2004)06 - 0927 - 05

Fault diagnosis for large-scale equipments in thermal power plant by data mining

YANG Ping[1], LIU Sui-sheng[1], ZHANG Hao[2]

(1. Electric Power College, South China University of Technology, Guangzhou Guangdong 510640, China;
2. Automation Engineering Research and Manufacturing Center, Guangdong Academy of Science, Guangzhou Guangdong 510070, China)

Abstract: This paper proposes a new approach to diagnose frequent faults for large-scale equipments in thermal power plants. Based on the acquired data in SCADA (Supervisory control and data acquisition) systems, a hybrid-intelligence data-mining framework is developed to extract hidden diagnosis information. The hard core of the hybrid-intelligence data-mining framework is an algorithm in finding minimum size reduction which is based on rough set approach, which makes it possible to eliminate additional test or experiments for fault diagnosis which are usually expensive and involve some risks to the equipment. This approach is also tested by all the data in a SCADA system's database of a thermal power plant for boilers fault diagnosis. The decision rules' accuracy varied from 92 percent to 95 percent in different months.

Key words: fault diagnosis; data mining; rough set; attribute reduction; decision tree

CLC number: TM621 **Document code**: A

火电厂大型设备故障诊断的数据挖掘方法

杨 苹[1], 刘穗生[1], 张 昊[2]

(1.华南理工大学 电力学院,广东 广州 510640, 2.广东省科学院 自动化工程研制中心,广东 广州 510070)

摘要: 针对火电厂大型设备的常见故障,提出一种新的诊断方法——数据挖掘方法.该方法通过建立一个智能化的数据挖掘工具,直接从火电厂 SCADA 系统历史数据库的大量实时数据中获取故障诊断知识进行故障诊断.数据挖掘工具的核心是,采用粗糙集的约简方式,将数据库中抽取的故障诊断规则简化为基于最小变量集的决策表.该方法避免了为诊断故障而附加的专门测试或试验,降低了费用,同时减少了试验对设备造成的潜在危险.将这一方法应用于火电厂锅炉的一个复杂故障事例,结果表明其诊断的精度在92%以上,可以满足现场应用的要求.

关键词: 故障诊断; 数据挖掘; 粗糙集; 属性约简; 决策树

1 Introduction

With the development of modern science and technology, the equipments of thermal power plants are getting larger and more complex. The wide application of large-scale equipments produce huge economic benefit; but, at the same time, it causes a series of problems. The investment and maintenance of large-scale equipments are unbearable and the accidents caused by these equipments are serious. So fault diagnosis of large-scale equipments has gained increased attention during the last few years. However, due to the complex interaction among the fault symptoms, the mechanisms of faults and their characteristics are very complex, it is very difficult to get high accuracy for large-scale equipments' fault diagnosis.

In practical application, engineers in power plants handle day-to-day maintenance, additional test and expert advice are often required from the technology supporting the center of manufacturing companies for more complex fault diagnosis and maintenance, although these additional tests are often expensive and involve some risks to equipments, and the domain experts' knowledge is very important at that time.

In recent years, fault diagnosis methods for large-scale systems have been widely developed. Model-based methods, fault tree approaches, fuzzy comprehensive evaluation systems, pattern recognition techniques and neural networks(NN) are in common use for such tasks. However, due to the complexity of large scale equipments in thermal power plants, methods based on process data, not on model or expert knowledge, would be more adequate. Neural networks are based on data only, but the accuracy of trained NN is not satisfied to large-scale equipments' fault diagnosis because data from practical industry always contain conflicted the data. Therefore, it is

Received date: 2004 - 02 - 05; Revised date: 2004 - 08 - 03.

Foundation item: Supported by 973 Project (G1998020308); Natural Science Foundation of Guandong Province (003049); The Fifteenth-Plan of Science and Technology of Guangdong Province (A1050202).

necessary to develop another technique which is based on process data and supposed to have high accuracy. In order to develop a new method based on process data for fault diagnosis, this paper proposes an intelligent data-mining framework to extract hidden information directly from data in SCADA systems' database, no additional tests or experiments are needed.

2 Overview of data mining in power engineering

Data mining is a process of discovering new, meaningful and interesting information directly from large amounts of data[1]. In recent years, data mining applications have been successfully applied in many areas, such as astronomy, molecular biology, medicine, and geology etc. Recently, some researchers are trying to apply this new method to power engineering applications. In [2], data mining techniques are applied to fault diagnosis in power transformers. Reference [3] mines association rules from historical data of a thermal power plant to derive accurate models of the behavior of plant component. In [4], data mining tools are used to extract rules from the power-generation database in the Mexican system. [5] extracts hidden information in power company database, which is related to customer billing. [6] studies the work on load profiling through data mining. Up to now, due to the complexity of large-scale equipments in thermal power plants, data mining techniques have not been used in these equipments' fault diagnosis.

Based on the properties of data mining, it is possible to get useful information related to frequent faults directly from SCADA systems' database.

3 Data mining for fault diagnosis

Data in SCADA systems' database provide useful information on equipment states. A data mining technique based on fuzzy rough set theory for large-scale equipments' fault diagnosis has been proposed in this section.

3.1 Process data in SCADA systems' database

Large numbers of process variables' value are stored in SCADA systems' database, including analog input points, digital input points, calculated value points, enter value points, pulse input points, and sequence of events (S.O.E.) input points, etc. Generally, the process variables' value is stored in SCADA systems' historic database once a minute, totally 1440 real numbers which represent the readings of one variable from 0 hour 0 minutes to 23 hour 59 minutes.

According to the features of thermal power plants' equipments, some mathematical relationships exist among equipments' states. The abnormal states of equipments can be represented by the change of some variables' values.

3.2 Data mining process for large scale equipment's fault diagnosis

In order to find the relationships among equipment's faults and related variables states, the process of data mining should be translated into practical steps. Figure 1 represents the data mining process in five steps.

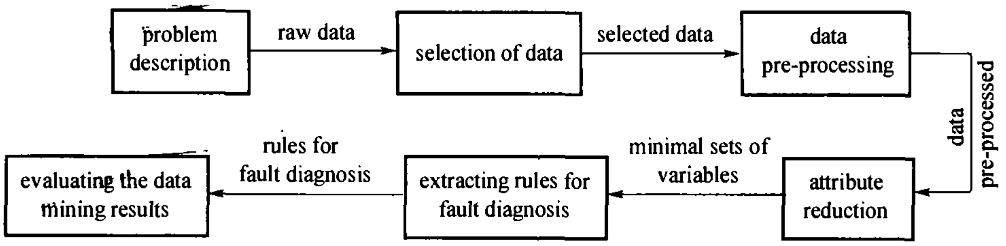
problem description
raw data
selection of data
selected data
data pre-processing
pre-processed data
attribute reduction
minimal sets of variables
extracting rules for fault diagnosis
rules for fault diagnosis
evaluating the data mining results

Fig. 1 Data mining process for fault diagnosis

Generally, the first step of data mining is to organize the raw data and the last step is to visualize the results. In the entire process, the first three steps reduce the amount of data by up to two orders of magnitude, still presenting the original characteristics of the raw data. But the pre-processed data contains too many variables to industrial application. The core of the data mining program is an attribute reduction module based on rough set theory. This step reduces data to a manageable set of variables.

3.2.1 Problem description and data selection

This step involves the description of fault type, fault process, fault diagnosis process, the understanding of historic database in thermal power plant's SCADA system, and the basic knowledge of thermal power plant. Then, the data mining expert analyzes the problem according to fault process and recognizes the data which are related to the faults possibly.

Based on the analysis of fault process and features of related equipment, the data mining experts and domain experts can recognize all the possible data related to the faults.

3.2.2 Data pre-processing

The selected data contain many variables that are not related to the faults. In order to reduce the calculation, data pre-processing should be done first. In particular, the data pre-processing include three steps:

Step 1 correlation analysis. The correlation coefficients of all analog input variables and calculated value variables related to the fault are calculated first. Based on the calculation results, all the variables whose correlation coefficients related to the fault are less than a certain value are deleted, then a related variable set is obtained.

Step 2 principal component analysis. All variables are mapped to another linear space to get principal components, then a principal component set of variables is obtained.

Step 3 domain expert experience. We can get an experiential set of variables from domain experts.

Finally we can get the preparation set of variables for data mining from the union of related set, principal component set and experiential set.

3.2.3 Data mining

The preparation set of variables still contains many variables that are not tightly related to faults. An attribute reduction technique based on rough set theory is used to mine the minimal set of variables in the preparation set.

A) Background.

One of the theories specially developed for data mining is the rough set theory [7]. It has been used to discover structural relationships from imprecise or noisy data.

Rough set theory is based on the establishment of equivalence classes within the given training data. A rough set definition for a given class C is approximated by two sets: a lower approximation of C and an upper approximation of C. Based on rough set theory, knowledge is expressed by a decision table which is defined in terms of *LR*-systems as follows. Let $L = (U, A)$ be a knowledge representation system and let $C, D \subset A$ be two subsets of attributes, called condition attributes and decision attributes respectively. *KR*-system with distinguished condition and decision attributes is called a decision table and is denoted $T = (U, A, C, D)$ or in short *CD*-decision table. In decision tables, some condition attributes are redundancy, so they are removed and decision tables are simplified. The reduction can be loosely defined as a minimal subset of attributes uniquely identifying all objects, it expresses an altemative and simplified way to represent a set of objects.

In the reduced decision table, the same decision is made based on a smaller number of condition attributes. This kind of reduction eliminates the need for checking unnecessary condition attributes to arrive at a conclusion.

The approach of attribute reduction in a decision table generally consists of the following steps:

1) Based on rough set theory, compute reductions of condition attributes which is equivalent to elimination of some columns from the decision table;

2) When two rows have the same values in the decision table, eliminate the duplicate rows;

3) Elimination of superfluous values of attributes.

B) Attribute reduction algorithm.

In order to mine decision rules for fault diagnosis, the variables in preparation set are regarded as condition attribute and the fault status is regarded as decision attribute. The attribute reduction algorithm can generate a minimal variable set (minimal reduction) directly from preparation variable set, representing the fault classification at the highest accuracy. The variables in minimal variable set can be used to obtain decision tree for fault diagnosis.

The attribute reduction algorithm based on rough set theory is applied to discrete-valued attributes, continuous-valued attributes must therefore be discretized prior to its use. The steps of attribute reduction algorithm are as follows.

Step 1 Establish a 2-dimension table by one month's data from data sample sets, all the real numbers of a variable in preparation set line up as a condition column, the fault status as a decision column, then the table has $n + 1$ columns and $1440 \times d$ rows, n = the number of variables in preparation set, d = the number of days in the assigned month.

Step 2 Discretize the variable values to five parts according to equal frequency principle, discretize the fault status values to three parts according to expert's experience. These form the decision table.

Step 3 According to rough set theory, calculate all reductions of condition attributes by global search and then find the minimal reduction.

Step 4 Summarize the number of duplicate rows to get the reliability of this row, then eliminate the duplicate rows.

Step 5 If all the condition attributes are the same while the corresponding decision attribute is different, eliminate the rows whose summation is less than 30 percent of this condition attributes' summation.

Step 6 Eliminate the superfluous values of attributes, the minimal variable set representing the fault classification at the highest accuracy in this month is obtained.

Step 7 Based on another month's data, repeat Step 1 to Step 6 again until all data samples in historic database have been considered.

Step 8 Calculate the intersection of all the minimal variable sets of all month, the final minimal variable set is obtained.

In practical application, conflicted data always exist in historic database of SCADA system, Step 5 is necessary to avoid the adverse influence by the fluctuation data.

3.2.4 Extracting rules for fault diagnosis

The variables in the final minimal set are regarded as condition attributes and the fault state is regarded as decision attribute, then decision rules for fault diagnosis are generated for each class (one row expresses one class) in the final minimal variable set. In this paper, the condition attributes' value is discretized to five parts and the decision attribute's value is discretized to three parts, the scale of decision tree is very large if the number of condition attributes is more than 5, then it is very difficult for engineers to apply. So we do further attribute reduction as follows:

Step 1 Arrange the condition variables, let A_i express the ith variable of the condition attributes. Let $i = 1$. The decision table contains N condition variables.

Step 2 Eliminate the ith attributes, then calculate the accuracy of the decision tree based on the left attributes. If the descent of accuracy is not more than 0.7 percent, the elimination is confirmed, otherwise cancel the elimination.

Step 3 If the number of the left condition variables is less than 3 or $i = N$, then go to Step 5, otherwise, go to Step 4.

Step 4 Let $i = i + 1$, go to Step 2.

Step 5 Stop.

Finally, we can get the sub-optimal variables set for fault diagnosis, the accuracy of this final decision table is high (not highest) enough for actual application and the scale of decision tree is not too large.

3.2.5 Evaluating the data mining results

Evaluation has been carried out to confirm the data mining results by using the test data sets.

The evaluation algorithm is as follows:

Step 1 Establish a two dimension table by one month's data from test data sets, as the data mining algorithm, to form a table with $n + 1$ columns and $1440 \times d$ rows, n = the number of variables in preparation set, d = the number of days in the assigned month.

Step 2 Discretize the variables value to five parts according to equal frequency principle, discretize the fault state value to three parts based on expert's experience. Then the decision table is formed.

Step 3 Calculate the accuracy of the final decision tree according to this month's data.

Step 4 Based on another month's data, repeat Step 1 ~ 3 until all data in test data sets have been checked up.

If the data mining is performed on enough data samples and the final decision tree satisfy the test data sets, it can be applied to industrial application.

4 Case study

The proposed approach is tested by three years' data in a SCADA system's database of a thermal power plant. One case that occurred more than 100 times in three years has been studied. The research result is discussed in this section.

4.1 Case description

In the thermal power plant, the temperature on A (T_A) and B side (T_B) of boilers' 4th Superheater outlet should be balanced. Imbalance is considered to be abnormal. Let $D = T_A - T_B$. According to domain experts' experience, if D is not larger than 10 ℃, it is good. If D is larger than 10 ℃ but not larger than 20 ℃, it is not good but acceptable. If D is larger than 20 ℃, it is a fault.

For the purpose of the case study, we collect three years' data from a thermal power plant. We select data samples from the historic database of SCADA system every other month as a mining data set, the left data is regarded as a testing data set.

4.2 Preparation of data

The collected data include analog input points, digital input points, calculated value points, enter value points, pulse input points, and S.O.E. input points, totally 6082 points. According to the behavior of boiler, the related data are analog input points and calculated value points, totally 2834. Based on these 2834 variables, we get the preparation set by data pre-processing in Section 3.2.2, it has 497 variables.

4.3 Data mining for temperature difference diagnosis

The preparation variables is regarded as condition attributes and the temperature difference is regarded as decision attribute. Applying the proposed attribute reduction algorithm in 3.2.3, a minimal variable set for one certain month is obtained. Then the intersection of minimal variable sets of all months in mining data set is calculated to obtain the final minimal variable set, it contains 12 condition attributes. Because of the large scale of this decision table, it is difficult for engineers to utilize. Further reduction has been done by the algorithm in Section 3.2.4, then two condition variables are left, the scale of this decision table is acceptable, it is shown in Table 1, its accuracy is varied from 92% to 95% in various months.

Table 1　Final decision table

V_7	V_{10}	D
1	1	1
1	2	3
1	3,4	2
2	1,2,3,4	2
3	1,2,3	1
3	4	2
4	1,2,3,4,5	1
5	1,2,3,4,5	1
1,2,3,4,5	5	1

V_7 is the metal temperature of tertiary superheater on A side, V_{10} is the metal temperature of fourth superheater outlet on E side, D is the temperature difference. The final decision table can be expressed by a decision tree simply as shown in Fig.2.

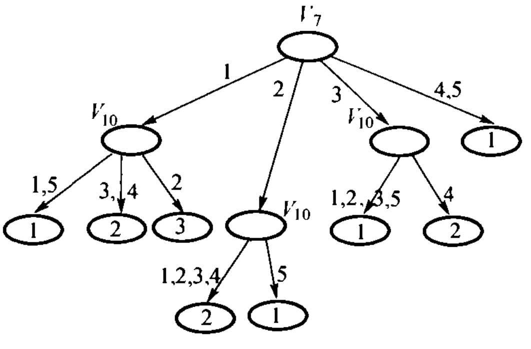

Fig. 2　Final decision tree

The decision accuracy of the final decision tree for February, 2000 is 94.76%, a little less than the accuracy based on the decision table contained 12 condition attributes, but it is much more compendious.

4.4 Evaluation of the data mining results

The data mining result has been verified by the testing data set, the classification accuracy is varied from 91% to 95% in different months. For example, for the data of May 2000, the classification accuracy of the final decision table is 94.47%. By data mining results, the temperature difference is due to the abnormal metal temperature of tertiary and fourth superheaters.

5 Conclusion

The research reported in this paper proposes a new approach for fault diagnosis of large scale equipments in complex industry application with SCADA systems. Based on the acquired data in historic database, this paper develops a hybrid-intelligence data-mining framework to extract hidden knowledge directly from the SCADA system. This new technique is based on process data only and eliminates additional tests or experiments that are often expensive and involve risks to equipments.

On the other hand, the variables themselves in historic database are regarded as condition attributes in decision table, the rules mining from historic database are expressed directly by the variables. Then it is acceptable for engineers to understand and apply to industry application.

References:

[1] HAN Jiawei, KAMBER Micheline. *Data Mining: Concepts and Techniques* [M]. San Francisco: Morgan Kaufmann Publishers, 2001.

[2] MEJIA-LAVALLE M, RODRIGUEZ-ORTIZ G. Obtaining expert system rules using data mining tools from a power generation databases [J]. *Expert Systems with Application*, 1998, 14(1/2): 37 - 42.

[3] OGILVIE T, SWIDENBANK E, HOGG B W. Use of data mining techniques in the performance monitoring and optimization of a thermal power plant [C]// *IEE Colloquium on Knowledge Discovery and Data Mining*. London, UK: IEEE Press, 1998: 7/1 - 7/4.

[4] SWIDENBANK E, GARCIA J A, FLYNN D, et al. On-line optimization of power plant performance through machine learning techniques [C]// *UKACC Int Conf on Control*'98. Swansea, UK: IEE Press, 1998: 257 - 262.

[5] PITT B D, KIRSCHEN D S. Application of data mining techniques to load profiling [C]// *IEEE Conf Proceedings on Power Industry Computer Applications*. Santa Clara, CA: IEEE Press, 1999: 131 - 136.

[6] WEHENKEL L, MACK P. Artificial intelligence toolbox for planning andoperation of power systems [C]// *IEEE Power Engineering Society Winter Meeting*. Singapore: IEEE Press, 2000, 2: 1057 - 1062.

[7] LEBREVELEC C, CHOLLEY P, QUENET J F, et al. A statistical analysis of the impact on security of a protection scheme on the French power system [C]// *1998 Int Conf on Power System Technology*. Beijing, China: IEEE Press, 1998, 2: 1102 - 1106.

作者简介：

杨　苹　（1967 —），女，华南理工大学电力学院副教授，博士，主要研究领域：电力电子电路的建模与控制、人工智能系统及其应用，E-mail: eppyang@scut.edu.cn；

刘穗生　（1979 —），男，华南理工大学电力学院硕士研究生，主要研究领域：电力电子电路的建模与控制，人工智能系统及其应用；

张　昊　（1969 —），男，广东省科学院自动化工程研制中心副研究员，主要研究领域：人工智能系统及其应用。