KNOWLEDGE EXTRACTION FROM PROCESS DATA: A ROUGH SET APPROACH TO DATA MINING ON TIME SERIES

Charlotte Skourup* and Jan Komorowski[1]**
*Department of Engineering Cybernetics
**Department of Information and Computer Science
Norwegian University of Science and Technology
Phone: +47-73594567, Fax: +47-73594460
email: Charlotte.Skourup@itk.ntnu.no, Jan.Komorowski@idi.ntnu.no

ABSTRACT: The complexity of process data and the limitations in humans' ability in pattern recognition suggest that there is a need for improved methods to analyse this type of data. The overall purpose of this paper is to study how data mining techniques in form of Pawlak's rough set approach can be adapted for knowledge discovery from collections of process plant data. The pilot study is based on a set of simulated process data containing 49 objects logged from a coal-fired power plant. Knowledge extracted from process data in the experiments can mainly be used in two ways; to classify new situations and to extend the operator's process knowledge. The classification performance of the rules is generally very good. The generated rule sets are typically small and consists of short, readable rules. The generated reducts and rules have furthermore proven to include "new" knowledge that extends and improves operator's process knowledge.

KEYWORD: Process plants, time series, data mining, rough sets approach, operator support.

INTRODUCTION

In process plants, hundreds of process variables show the process mode during plant operation. These variables interrelate in complex cause-and-effect relationships. Process data is mainly represented as time series and is automatically stored in process databases. Such process data produces a large amount of stored information that contains vital knowledge about the operation of the plant. During plant operation, operators generally monitors the process regarding the trends of the process variables in addition to some state descriptions and numerical values. Operators look for deviations in these variables to detect and classify abnormal situations to secure optimal and stable plant operation. In case of an abnormal situation, domain experts normally inspect the logged process data to determine the cause of the situation and to establish procedures to prevent similar situations to occur in the future. Operators and domain experts generate individual mental models of the process and its relationships between inputs and outputs. Later on, they use such mental models to simulate and test potential actions before these are implemented in the real world. New operational experiences improve the operator's mental model. However, humans in general have limitations in recognising complex patterns such as relationships between process variables. Furthermore, the fact that a cause and its effect may take place at different times due to time delays complicates the discovery of knowledge. Therefore, knowledge may remain unrecognised in the process databases.

The complexity of process data and the limitations in human's ability in pattern recognition suggest that there is a need for improved methods to analyse this type of data. Such methods exist within the fields of data mining and knowledge discovery as presented in Pawlak (1991) and Komorowski (1999a). Moreover, time series, unlike discrete data sets, include the time as an additional dimension. This dimension is very important in process plants in order to monitor the process and to describe situations such as an increasing temperature and an unstable pressure. Data mining techniques have, however, mainly been used on data sets with a limited extension in time such as using a static time window back in time as presented in Szladow (1992) and defining values of the condition and decision attributes at time moments, *t*,

[1] Address for correspondence

as presented in Mrozek (1992). Berndt (1996) has presented a study of detecting patterns in time series for knowledge discovery based on the dynamic time warping technique used in speech recognition.
The overall purpose of this paper is to study how data mining techniques in form of the rough set approach can be adapted for knowledge discovery in process plant data. A set of simulated process plant data containing 49 objects has been used for a pilot study using the ROSETTA tool-kit. Each object is represented by a set of parameters primarily describing trends of the time series. Practical experiments using the ROSETTA tool-kit show that this approach results in sets of short, readable rules that classify objects with a high classification performance. Furthermore, according to a domain expert, these rules contain valuable knowledge about plant operation in form of unrecognised reducts and knowledge regarding exact relationships between process variables.
The next section reviews the methods used in this work including the rough sets theory. The pilot study is then introduced followed by a presentation of the used data set and the results of the pilot study. The paper thereafter analyses and discusses the classification performance and the extracted knowledge in form of rules. Finally, the work presented in this paper is concluded.

METHOD

We use Pawlak's rough set approach Pawlak (1991) to data mining and knowledge discovery. Rough sets seem to be rather suitable for this project since they produce white box models, that is, sets of induced classification rules that may be easily inspected by humans provided that the number of rules is small which clearly is the case in our application. Rough sets are also attractive because Boolean reasoning methods which are used to compute reducts provide an effective way of dealing with important attributes only. A tutorial introduction to rough sets is provided in, for example, Komorowski (1999a). All the experiments are performed using Rosetta, a software toolkit for analysing data within the framework of rough sets. The tool is a joint product of two groups: Section of Logic, Warsaw, Poland and Knowledge Systems Group, Trondheim, Norway. For a description of the system and its availability see Øhrn (1998), Komorowski (1999b) and Øhrn (1999).

PILOT STUDY: KNOWLEDGE EXTRACTION FROM PROCESS DATA

This paper presents a pilot study using ROSETTA tool-kit to test whether it is possible to represent time series for the rough sets approach and furthermore, to extract essential knowledge in form of reducts and rules from such time series. In the evaluation of the rules, a domain expert has been consulted for a deeper analysis and discussion.

THE DATA SET

The pilot study is based on a set of simulated process data containing 49 objects logged from a coal-fired power plant. The simulator is developed at Institute for Process and Production Control Technology, Technical University of Clausthal, Germany, in cooperation with ABB. The simulator includes 157 process variables on which a domain expert has defined 25 parameters as representatives for describing various situations. These parameters represent primarily gradients of trends such as *strongly increasing*, *increasing*, *constant*, *decreasing* and *strongly decreasing*. This representation is in accordance with the operators' interpretation of process variables. Besides, few parameters represent state descriptions and numerical values. Process variables are logged every second second and hence, form approximately time series as in real plant operation. The logged situations are primarily abnormal situations in that they reflect unwanted behaviour such as a leakage in the high-pressure preheater. Furthermore, the logged situations also include some normal situations in which a few changes of set points affect several other process variables such as changing the load. For an inexperienced operator, such situations can be mistaken for critical situations.

Class no	Class description	No of initial classified situations	Total no of situations
1	Change of load	9	14
2	High-pressure preheater valve fails	4	30
3	Feed water pump breaks down	3	3
4	Leakage in high-pressure preheater	20	20
5	Break down of spindle in high-pressure preheater	8	8
6	Coal mill breaks down	5	5

Table I: The six classes and the number of classified situations

The objects are initially classified into six classes, but most of the objects (31) belong to more than one class as listed in Table I. Due to the small data set and the objects' membership in more than one class, we run separate experiments for each decision class. Thus, objects are classified as either belonging to or not belonging to that specific class and the total number of objects belonging to each class therefore increases. This method generates separate rules for each decision class.
The rule generation for each class uses three separate random splits. The training set consists of approximately 2/3 of the objects (33) whereas the rest of the objects (16) form the test set that is used to evaluate the classification performance of the generated rules. The Johnson-algorithm with object-related discernibility and default parameters has been used for all experiments.

RESULTS

The classification of objects has been done using a "no fallback" scheme. This means that if the rules are not able to classify an object, the object remains unclassified. This choice is due to the fact that the intention of the generated rules is to assist the operator in diagnosing, that is, to classify situations. Hence, if the rules are unable to classify an object, we would like the operator to be aware of this reality. Moreover, operators may be more qualified to reason about the classification of the new situation. Therefore, the confusion matrices also include unclassified objects. The entries of the confusion matrix, *C(i, j)*, counts the number of objects that really belong to class *i*, but were classified by class *c* as belonging to class *j*. Table II lists the confusion measurements for all decision classes.

Class no	Split no	*C(c, c)*	*C(c, Ø c)*	*C(Ø c, c)*	*C(Ø c, Ø c)*	Unclassified	Sensitivity	Specificity
1	1	5	1	0	10	0	0,83	1,0
	2	4	2	0	8	2	0,67	0,8
	3	5	1	0	9	1	0,83	0,9
2	1	3	2	0	5	6	0,3	0,83
	2	7	0	4	3	2	0,78	0,43
	3	3	3	2	6	2	0,38	0,75
3	1	1	0	0	15	0	1,0	1,0
	2	0	1	0	13	2	0,0	0,93
	3	0	1	0	14	1	0,0	0,93
4	1	7	0	0	8	1	1,0	0,89
	2	4	0	1	9	2	0,8	0,62
	3	4	0	0	12	0	1,0	1,0
5	1	2	0	0	14	0	1,0	1,0
	2	3	0	0	13	0	1,0	1,0
	3	4	0	0	11	1	1,0	0,92
6	1	0	1	0	12	3	0,0	0,8
	2	1	1	0	11	3	0,5	0,79
	3	1	0	4	11	0	1,0	0,73

Table II: The confusion matrices and evaluation functions for all decision classes

The classification performance of the rules expressed by the sensitivity and the specificity is generally very good. These two performance measurements estimate probabilities from the confusion matrix *C* by dividing an entry by the sum of the row or column the entry appears in. Decision class 2 is, however, an exception in that as many as 6 out of 16 objects are unclassified in split 1. The domain experts also find it difficult to produce general rules to describe the situations in class 2 because 26 of the objects belonging to class 2 also belong to three other classes. Hence, the discussion is then whether this class should have been split into several subclasses.

ANALYSIS

The question is, however, what should rule generation for classification be used for in process plants. Classification of situations can be used in two ways; to detect and classify situations automatically and to support the operator in diagnosing a situation. In the first approach, an operator support system directs the operator's attention to new problem situations which the operator may have overlooked, or may not have recognised yet. Furthermore, the system

automatically suggests which class the situation belongs to. In the latter approach, the operator support system aids the operator when he has observed a new abnormal situation that he is unable to recognise. The operator support system then classifies the new situation regarding generated rules for existing situation classes.

LEGIBILITY

The generated rule sets are typically small and consist of short, readable rules. The rules have a maximum of two premises while univariate rules are common. Thus, the rules become easy to read and understand for humans. Figure 1 illustrates an example of rules generated for class 3, split 1.

Power(trend, Increasing) Þ Category(not 3)
Power(trend, Constant) Þ Category(not 3)
Power(MW, 750) Þ Category(not 3)
Power(MW, 300) Þ Category(not 3)
FlowFWpump2(trend, Increasing) Þ Category(3)
FlowFWpump2(trend, Strongly increasing) Þ Category(3)

Figure 1: Example of a rule set (class 3, split 1)

In addition, the rule sets contain between 4 and 24 rules which makes it possible for humans to quickly get an overview of the generated rules. The rules furthermore represent the time dimension as gradients of the trends of process variables in agreement with operators' perception and interpretation of such variables. The operator is then already familiar with this representation.

USEFULNESS OF RULES

Knowledge extracted from process data in form of reducts and rules also extends and improves operator's mental models, that is, process knowledge. First, even for a small number of variables, human operators have difficulty in judging which of the process variables are redundant. An analysis of the generated reducts shows that the domain expert became aware of dependencies consisting of a smaller number of variables than noticed in advance. For example, in classifying situations to class 3, the operator monitors especially the flow of feed water through feed water pump 1, 2 and 3 whereas the reducts allow the operator to focus his attention only on the flow of feed water through feed water pump 2. Second, the generated rules incorporate knowledge about exact relationships between process variables. Although humans are able to recognise simple patterns, they find it hard to deduce more precise relationships between process variables. Thus, the generated rules can be used to teach operators simpler and more exact rules-of-thumb to recognise and classify new situations. For example, operators as well as domain experts find it difficult to produce general rules to classify situations in class 4. The objects of this class differ in that the levelstands of high-pressure preheater 61/62 and of high-pressure preheater 71/72, for example, do not have the same influence on the rest of the process. The generated rules, however, show that there exists a set of simple relationships for the classification of this class as well. The domain expert was not able to extract equal knowledge in advance. The three splits in class 4 have nine rules in common.
These results prove that process plant data actually includes hidden knowledge that is significant and valuable for operators in plant operation.

CONCLUSION

The work presented in this paper is a pilot study using a data set of 49 objects. Based on this data set, the classification performance of the generated rules has proven to be high. Furthermore, the legibility and the usefulness of the rules strongly support further studies on knowledge extraction using the rough sets approach on process plant data. Knowledge extracted from process data can mainly be used in two ways; to classify new situations and to extend the operator's process knowledge. The latter is supported by presenting reducts which the operator was unaware of and by presenting unrecognised relationships between process variables. The successful results have verified that data mining techniques can be applied to time-dependent process data mainly represented by parameters describing the gradients of the process variables.

ACKNOWLEDGEMENT

Parts of this research is supported by the Norwegian Research Council and is a part of a research programme taken place at Department of Engineering Cybernetics, NTNU. The authors thank the Training and Mobility Programme of the European Commission for the financial support which enabled one of the authors to spend time at Institute for Process and Production Control Technology of the Technical University of Clausthal under the COPES project. The authors also thank M.Sc. Grethe Røed for her work on her master thesis which has contributed to the work presented in this paper.

REFERENCE

Berndt, Donald J. and Clifford, James, 1996, "Finding patterns in time series: A dynamic programming approach", in: "Advances in Knowledge Discovery and Data Mining", U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusam (eds.), MIT Press, pp. 229 - 248.

Komorowski, Jan; Pawlak, Zdzislaw; Polkowski, Lech and Skowron, Andrzej, 1999a, "A Rough Set Perspective on Data and Knowledge", to appear in: "Handbook of Data Mining and Knowledge Discovery", W. Kloesgen and J. Zytkow (eds.), Oxford University Press.

Komorowski, Jan; Skowron, Andrzej and Øhrn, Aleksander, 1999b, "Rosetta and Rough Set Software Systems for Data Mining and Knowledge Discovery", to appear in: "Handbook of Data Mining and Knowledge Discovery", W. Kloesgen and J. Zytkow (eds.), Oxford University Press.

Mrozek, Adam, 1992, "Rough sets in computer implementation of rule-based control of industrial processes", in: "Intelligent Decision Support: Handbook of Applications and Advances of the Rough Set Theory", S.-Y. Huang (ed.), Kluwer Academic Press, pp. 19 - 31.

Pawlak, Zdzislaw, 1991, "Rough Sets - Theoretical Aspects of Reasoning about Data", Kluwer Academic Publishers, the Netherlands.

Szladow, Adam J. and Ziarko, Wojciech P., 1992, "Knowledge-based process control using rough sets". In: "Intelligent Decision Support: Handbook of Applications and Advances of the Rough Set Theory", S.-Y. Huang (ed.), Kluwer Academic Press, pp. 49 - 60.

Øhrn, Aleksander; Komorowski, Jan; Skowron, Andrzej and Synak, Piotr, 1998, "The Design and Implementation of a Knowledge Discovery Toolkit Based on Rough Sets - The Rosetta System", in: "Rough Sets in Knowledge Discovery", A. Skowron and L. Polkowski (eds.), Springer Verlag.

Øhrn, Aleksander, 1999, "Rosetta: Technical Reference Manual", The Norwegian University of Science and Technology, Trondheim, Norway. The Rosetta Homepage, URL: http://www.idi.ntnu.no/~aleks/rosetta/