# Rough Set Theory for Data Mining in an On-line Cable Condition Monitoring System

Chengke Zhou Heriot-Watt University (UK) C.Zhou@hw.ac.uk

Matthieu Michel EDF Energy (UK) Matthieu.Michel@edfenergy.com

D.M.Hepburn Glasgow Caledonian University (UK) D.M.Hepburn@gcal.ac.uk

Guobin Zhang Heriot-Watt University (UK) G.Zhang@hw.ac.uk

## ABSTRACT

This paper presents the application of a new methodology, the Rough Set (RS) theory, for data mining or knowledge acquisition in an on-line cable condition monitoring system. The attractiveness of the RS theory is that it allows automated generation of knowledge models of clear semantics which offer explanations of the inferences performed when used in diagnostics.

Following an introduction to a cable on-line condition monitoring system, where High Frequency Current Transformers (HFCT) are used for continuous monitoring of Partial Discharge (PD) activity, and the desired data mining or knowledge acquisition technology for such a system, the paper briefly described the concept of RS and the procedure of its implementation in practical applications. It then presents an example of the application of the method to knowledge deduction which can be easily adopted in the on-line 11-kV cable condition monitoring system. In such a system effective data mining proved to be the bottleneck due to the large volume of data generated by the system (sampling rate at 100 Mega Samples/second). The paper demonstrates that the RS theory is effective in mining knowledge rules from large volume of data. It requires little computing time and storage space, when compared with other data mining algorithms, as it simply removes redundant data and those data containing no information. It has the potential to be applied to automatically analyse utilities' condition monitoring database in data mining and knowledge acquisition.

## INTRODUCTION

It is well recognized that condition monitoring and early diagnosis of defects in HV/MV power plant items is key to the provision of a reliable power supply. Following the global privatization of, and the resultant competition in, the electricity market, financial pressures among utility companies limit the scope for investment in new plant and infrastructure. Given that power plant items are being used up to, and beyond, their design life and that demand for electricity is growing in a modern society which is reliant on a secure electricity supply, the importance of properly maintained equipment cannot be over-emphasized.

In the initial monitoring programmes for distribution plant, partial discharge activity and degradation processes were monitored off-line using electrical and acoustic methods and by Dissolved Gas Analysis at remote laboratories. Although many other monitoring schemes for the broad range of plant items have been developed, these methods do not allow monitoring of the plant under in-service conditions and delay can exist between data collection and analysis. Excellent progress has, however, been made in recent years in relation to on-line condition monitoring instrumentation, e.g. Partial Discharge (PD) monitors [1]. In EDF Energy, over 1100 on-line PD condition monitors have been installed to monitor PD activity in cables and switchgears. Data obtained from these systems have been used to identify the imminent failure of distribution plant items and to prioritize the replacement of network components. At a sampling rate of 100 Mega Samples/second, the amount of data produced from these systems presents a tremendous challenge, in terms of storage and data processing. Improvements in data processing techniques, such as the use of Wavelet Transforms, have allowed effective signal denoising by providing simultaneous characterization and analysis of time-series data in both time and frequency domains [2]. A shortage of manpower with the skills and experience of interpreting the data acquired by systems has been a major bottleneck in allowing utility companies to fully utilize their condition monitoring data and moving their asset management from time-based to condition-based maintenance.

The data created in monitoring electrical systems may contain underlying trends or features within the data that are not evident from the usual analysis techniques or to diagnostics experts [3]. In addition, the data also often contains superfluous or redundant information, e.g. where no fault exists. This complexity, in conjunction with the relatively short history of electrical plant condition monitoring, has resulted in the known deficiencies in electrical plant diagnostics systems. To improve the reliability of the electrical network, the full potential of the enormous amount of condition monitoring data should be explored using new and comprehensive data mining techniques. Ideally these should be computer based techniques with the ability to extract new information or knowledge and to develop diagnosis rules from raw data gathered during condition monitoring of power plant. Computers offer the opportunity to operate continuously and to continue to investigate data for new correlations, whilst being less expensive than human experts. The desirable features of a data mining system are: the capability of interrogating and learning from any existing database; the ability to learn and evolve continuously using newly acquired data; and the ability to remove redundant data from the system.

Despite there being a great deal of research into the application of Artificial Neural Networks (ANN), Genetic Algorithms (GA) and Fuzzy Rule Induction Algorithms (FRIA), none of these, unfortunately, meets the requirements of the desired system outlined above. This paper presents the use of a relatively new method, the Rough Set (RS) theory for knowledge acquisition in power plant condition monitoring. Apart from satisfying the requirements, as described in the preceding paragraph, an additional attraction of the RS theory is that it allows automated generation of knowledge models, offering clear explanations to the inferences performed in diagnosis.

## THE ON-LINE, CABLE CONDITION MONITORING SYSTEM

The system shown in Figure 1 is a typical on-line cable condition monitoring system deployed in EDF Energy [2]. The system uses an HFCT, with a frequency bandwidth of 13kHz – 15MHz, attached to the earth strap of an 11kV feeder: the signal is sampled at 100MS/s. Every data set acquired from each sensor consists of 2 million data points which, at a sampling rate of 100 MS/s, covers 20ms or a full AC cycle's time. The main purpose of the system is to monitor the PD activity in the cable insulation, a key indicator of the remaining life of cables in service. Analysis carried out on the data generated from the system includes denoising [2], PD event recognition and diagnosis [1].

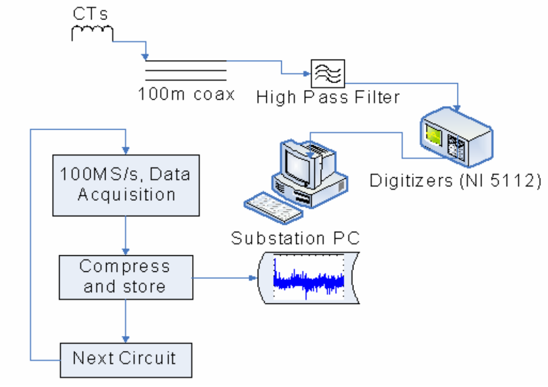

Figure 1: The On-line Cable Condition Monitoring System Under Study.

The monitoring system generates large volumes of data which require to be analysed to extract features for use in determining the insulation condition. In some cases severe noise is present and this can lead to acquisition of inaccurate or misleading information. It should be noted that not all the extractable attributes will be essential in determining the operational status of the cable and, in addition, various insulation defects and problems may not have been encountered by engineers, because most of the cables in UK utility companies were installed in 1950s or 1960s, these plant items are just approaching the end of their design life.. As a result, it is highly desirable that a technique to automatically analyse both the historical and newly reported cable problems be developed and that knowledge rules correlating all the defects or problems with the attributes that can be extracted from the condition monitoring data be produced, as discussed in the previously.

## ROUGH SET METHOD

### Information system and Rough Set

Rough Set is a mathematical tool that extracts information automatically from data by using matrix algebra and Set Theory to investigate correlations amongst data. In comparison with other data mining techniques, such as Genetic Algorithms and Artificial Neural Network, RS is relatively new [4]. Few applications in the area of power industry have been reported as far as the present authors are aware, e.g. [5].

When the RS method is applied to data mining and knowledge acquisition, it starts with an information system S which contains a pair (U, A), where U is a non-empty, finite set of objects, called the universe or search space, and A is a non-empty finite set of attributes. For example, Table 1 lists an information system, summarized by the present authors from a CIGRE report [6], consisting of possible insulation problems in relation to the observed phenomena in PD measurements. For the recently installed on-line cable condition monitoring systems, these attributes can be extracted from the condition monitoring data.

Table 1: Information System representing cable insulation problems.

| Case | A1 | A2 | A3 | A4 | A5 | D |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | PD-1 |
| 2 | 1 | 1 | 1 | 1 | 2 | PD-2 |
| 3 | 1 | 1 | 1 | 2 | 1 | PD-3 |
| 4 | 1 | 1 | 1 | 2 | 2 | PD-4 |
| 5 | 1 | 1 | 1 | 2 | 3 | PD-5 |
| 6 | 1 | 1 | 2 | 1 | 1 | PD-6 |
| 7 | 1 | 1 | 2 | 2 | 1 | PD-7 |
| 8 | 1 | 2 | 1 | 1 | 1 | Disturbance |
| 9 | 2 | 3 | 3 | 1 | 1 | Transients |
| 10 | 2 | 4 | 3 | 1 | 1 | Corona |
| 11 | 2 | 2 | 2 | 3 | 1 | Corona |
| 12 | 3 | 1 | 1 | 2 | 1 | Contact noise |
| 13 | 4 | 2 | 2 | 2 | 1 | Harmonics |

The attributes for the Information System presented in Table 1 are:

A1 - location of discharge on voltage waveform

1 - most pulses in advance of the voltage peaks
2 - on both sides of voltage peaks
3 - on both sides of voltage zeros
4 - proportional to voltage magnitude

A2 - Variability of response
1 - random movement
2 - steady or repeated motion
3 - stationary

A3 - Magnitude of discharge on +ve and –ve half cycles
1 - similar magnitude on both half cycles
2 - different magnitude on two half cycles
3 - on positive half cycle only
4 - on negative half cycle only

A4 - Variation of discharge magnitude with voltage
1 - constant with voltage
2 - rising with voltage
3 - constant on one half and rising on the other half cycle

A5 - Variation with time
1 - constant with time
2 - falls slowly with time
3 - rises rapidly with time

and

PD-1 - Internal discharge in a dielectric bounded cavity
PD-2 - Elastomeric insulation containing cavity in the form of a fissure or inhibitor
PD-3 - Internal discharge in a number of dielectric bounded cavities of various size, discharges on external dielectric surfaces between two touching insulated conductors, or discharges on external dielectric surfaces at areas of high tangential stress.
PD-4 - Internal discharges in a number of dielectric-bounded cavities of various size, mainly found in cast-resin insulation.
PD-5 - Internal discharges in gas bubbles in an insulating liquid in contact with moist cellulose.
PD-6 - Internal discharges between metal or carbon and dielectric in a cavity.
PD-7 - Internal discharges between metal or carbon and dielectric in a number of cavities of various sizes or surface discharges taking place between external metal or carbon and dielectric surfaces.

In this instance, the search space U includes all the cases, or all the examples shown in the table. The "A" attributes described above are all believed to be related to cable conditions. "D" the rightmost column of the table, also known as the decision attribute, gives the known cable fault conditions.

Using the given data in Table 1, the Rough Set theory will, by using combinations and permutations of the data sets and comparing these with the decision attribute, determine which groupings of data are "responsible" for the given outcomes. Although a summary of procedures for applying RS to data are given in the following section, more information on the application of the method can be found in [4,5].

### The procedures of applying Rough Set theory

Let $S=(U, A)$ is an information system, and $U=\{x_1, x_2, \ldots, x_n\}$. A is condition attributes set, D is decision attribute, and $a(x_i)$ is the value of record $x_i$ on attribute a.
Then the discernibility matrix is defined as:
$(C_{ij})=0,\quad D(x_i)=D(x_j),\ i,j=1,2,\ldots,n.$
$(C_{ij})=-1,\quad a(x_i)=a(x_j),\ D(x_i)\neq D(x_j),\ i,j=1,2,\ldots,n.$
$(C_{ij})=\{a\in A/a(x_i)\neq a(x_j)\},\quad D(x_i)\neq D(x_j),\ i,j=1,2,\ldots,n.$

The arithmetic of data reduction based on rough set is:
**If** Redu is the attribute set after reduction.
(1) let us arrange the nucleus attribute into the attribute set, Redu= C0;
(2) find the combination S which does not include the nucleus attribute in discernible matrix, $Q=\{ B_i: B_i\cap Redu \neq \varphi, i=1, 2, \ldots, s \}$, $S=S-Q$;
(3) change the combination S into the following pattern, $P=\wedge\ \{ \vee\ b_{i,k}: i=1,2,\ldots,s; k=1,2,\ldots,m\}$;
(4) change P into the extract pattern;
(5) select the satisfying result according to the demand.

According to the above arithmetic of data reduction based on rough set theory, an algorithm for rule induction, after removing repetitive rows and columns, is given as follows:

(1) Observe each condition attribute in information table. If there are conflict records after delete the row and the column the attribute situates, the condition attributes value is reserved. Otherwise if there are not conflict records, but there are repeated records, the attribute value which causes the conflict will be mark "*".
(2) Delete instances of repeated records.
(3) Delete records in which all condition attributes are marked "*".
(4) In a situation where there are two records in which only one condition attribute is different and one record's attribute is marked "*", when the decision record can be judged by the attribute which has the attribute marked "*" then the record will be deleted, otherwise the record which contains a value will be deleted.
(5) Repeat the first 4 steps until no further reduction can be made.
(6) Knowledge rules are derived from the matrix which is the result of applying the steps above.

### Practical application of RS theory to Table 1

Table 2, below, is generated from Table 1 by application of the above procedures. Although at first inspection there does not appear to be a major reduction in the table contents, Table 2 does in fact contain a lower number of data elements, as the elements marked with "*" can be ignored when knowledge rules are being deduced from the resultant table.

Table 2: Reduced data set

| Case | A1 | A2 | A3 | A4 | A5 | D |
|---|---|---|---|---|---|---|
| 1 | * | 1 | 1 | 1 | 1 | PD-1 |
| 2 | * | * | 1 | 1 | 2 | PD-2 |
| 3 | 1 | * | 1 | 2 | 1 | PD-3 |
| 4 | * | * | 1 | 2 | 2 | PD-4 |
| 5 | * | * | 1 | 2 | 3 | PD-5 |
| 6 | * | * | 2 | 1 | 1 | PD-6 |
| 7 | * | * | 2 | 2 | 1 | PD-7 |
| 8 | * | 2 | 1 | 1 | 1 | Disturbance |
| 9 | * | 3 | 3 | 1 | 1 | Transients |
| 10 | * | 4 | 3 | 1 | 1 | Corona |
| 11 | * | * | 2 | 3 | 1 | Corona |
| 12 | 3 | * | 1 | 2 | 1 | Contact noise |
| 13 | 4 | * | * | * | * | Harmonics |

From Table 2 the following rules can be deduced:

- If (A2=1, A3=1, A4=1 and A5=1), fault =PD-1
- If (A3=1, A4=1, and A5=2), fault =PD-2
- If (A1=1, A3=1, A4=2, and A5=1) fault =PD-3
- If (A3=1, A4=2 and A5=2) fault =PD-4
- If (A3=1, A4=2 and A5=3) fault =PD-5
- If (A3=2, A4=1 and A5=1) fault = PD-6
- If (A3=2, A4=2, and A5=1) fault =PD-7

A further comparison with the original table of data shows that the algorithm helped to remove redundant information from the information set and has produced valid linguistically expressed knowledge rules.

## CONCLUSIONS AND DISCUSSION

The paper has demonstrated, using a simple example, that the Rough Set theory is effective in mining useful, linguistically expressed knowledge rules from large volumes of data. Figure 2, below, shows how the method can be applied to practical systems. The left hand side of Figure 2 shows the procedure for diagnosis whilst right hand side illustrates the data mining process which runs every time a new type of fault or new sets of data are reported.

The RS method requires negligible computing time and small storage space when compared with other data mining algorithms as it involves no iterative algorithm and it simply removes all the repetitive and redundant data and those data which are determined as containing no information. The RS methodology has the potential to be applied to automatically analyse utilities' condition monitoring databases for data mining and knowledge acquisition purposes.

It should be noted, however, that the system demonstrated in the paper is small in size. Further investigation is required to include additional features such as the description of individual PD pulse shape and fault types.

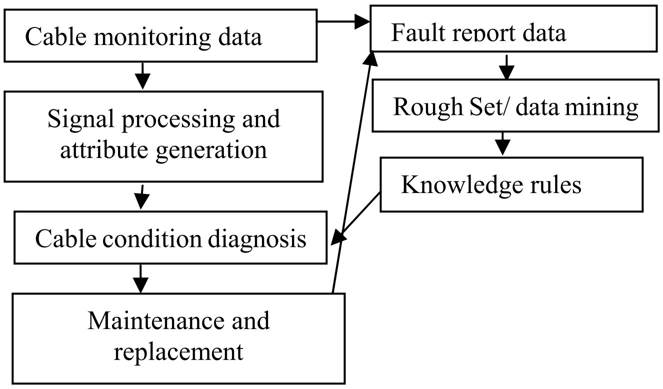
Cable monitoring data
Fault report data
Signal processing and attribute generation
Rough Set/ data mining
Knowledge rules
Cable condition diagnosis
Maintenance and replacement

Figure 2: Application of the proposed Rough Set method for analysing an on-line cable condition monitoring system.

## REFERENCES

[1] L. Renforth, R. MacKinlay and M. Michel: "MV Cable Diagnostics – Applying Online PD Testing and Monitoring", Asia Pacific Conference on MV Power Cable Technologies, Hongkong, 6-8 September 2005.

[2] C. Zhou et al: "Comparisons of Digital Filter, Matched Filter and Wavelet Transform in PD Measurement Denoising", CIGRE paper D1-111, Paris, August, session 2006.

[3] A.J. McGrail et al: "Data Mining Techniques to Assess the Condition of High Voltage Electrical Plant", WG 15.11 report, CIGRE session 2002, Paris, August 2002.

[4] Z. Pawlak: "Rough Set classification", International Journal on Man-machine Studies, Vol.20, 1984.

[5] Z. Wang etal: "A Fault Diagnosis Method for Transformer Integrating Rough Set with Fuzzy Rules", Transaction of Institute of Measurement and Control, 28, 3, 2006.

[6] F.H. Kreuger, 1992, "Recognition of Discharges", *Electra.* No. 11, pp61-98.