Analysis of an Incomplete Information System using the Rough Set Theory

C. I. Faustino Agreira[1], M. M. Travassos Valdez[1], C. M. Machado Ferreira[1] and F. P. Maciel Barbosa[2]

[1]Instituto Superior de Engenharia de Coimbra,
Instituto Politécnico de Coimbra, Coimbra, Portugal
cif@isec.pt

[2]Inesc Tec and Faculdade de Engenharia,
University of Porto, Porto, Portugal
fmb@fe.up.pt

Abstract *— In this paper it is applied a Rough Set approach that takes into account an incomplete information system to study the steady–state security of an electric power system. The Rough Set Theory has been conceived as a tool to conceptualize, organize and analyze various types of data, in particular, to deal with inexact, uncertain or vague knowledge. The knowledge acquisition process is a complex task, since the experts have difficulty to explain how to solve a specified problem. So, an incomplete set of relevant information may arise. The study presents a systematic approach to transform examples in a reduced set of rules. These rules can be used successfully to avoid security problems and provides a deeper insight into the influence of parameters on the steady-state system performance.*

Keywords *— Incomplete information systems, Rough Set Theory*

1 Introduction

Recently, the Rough Sets theory (RST) has been used successfully to handle efficiently problems where large amounts of data are produced [1]. RST constitutes a framework for inducing minimal decision rules. These rules can be used in turn to perform a classification task. Important concepts include the elimination of redundant criteria to give more compact rules. The strength of a rule can be quantified using rough membership. The main goal of the rough set analysis is to search large databases for meaningful decision rules and, finally, acquire new knowledge. This approach is based in four main topics: indiscernibility, approximation, reducts and decision rules [1]. A reduct is a minimal set of attributes, from the whole attributes set, that preserves the partitioning of the finite set of objects and, therefore, the original classes. It means that the redundant attributes are eliminated. When the reducts are evaluated, the task of creating definite

rules for the value of the decision attribute of the information system is practically performed. Decision rules are generated combining the attributes of the reducts with the values. Decision rules extracte knowledge, that can be used when classifying new objects not in the original information system.

The RST has been conceived as a tool to conceptualize, organize and analyze various types of data, in particular, to deal with inexact, uncertain or vague knowledge. In the Rough Sets analysis the concept of an information system is used to construct the approximation space. It enables representation of data in a useful form of a table. The information system is, in fact, a finite data table where columns are labelled by attributes and rows are labelled by objects [2]. Attributes are generally classified into conditions and decisions. Usually, a number of condition attributes and a single decision attribute are presented. In an incomplete information system the attribute values for objects may be unknown (missing or null) [3].

The knowledge acquisition process is a complex task, since the experts have difficulty to explain how to solve a specified problem. So, an incomplete set of relevant information may arise. This study presents a systematic approach to transform examples in a reduced set of rules [4]. These rules can be used successfully to avoid security problems and provides a deeper insight into the influence of parameters on the system performance.

This paper is organised as follows. Section 1 presents an introduction to the problem. Section 2 is devoted to the Rough Set Theory considering an incomplete information system. In section 3 is presented the test power network that was analysed and shows the results obtained using the proposed approach [5]. Finally, in section 4, some conclusions that provide a valuable contribution to the understanding the RST applied to the security analysis of the electric power system are presented.

2 Rough Set Theory

Rough Set Theory can be considered as an extension of the Classical Set Theory, for use when representing incomplete knowledge. Rough sets can be considered sets with fuzzy boundaries − sets that cannot be precisely characterized using the available set of attributes. Many different problems can be addressed by RST. During the last few years this formalism has been approached as a tool used in connection with many different areas of research. It has also been used for, among many others, knowledge representation, data mining, dealing with imperfect data, reducing knowledge representation and for analysing attribute dependencies.

2.1 Information system

Information System (IS) can be defined as a $K = (U,R,V,\rho)$, where U is a finite set of objects, R is a finite set of attributes, V is the domain of each attribute of R, and ρ is a total function that defines the following application: ρ. $U \times R \rightarrow V$, i.e, the examples. Va is called the value set of a.

2.2 Approximations sets

The Rough Set Theory (RST) is a new mathematical tool presented to dispose incomplete and uncertainty problem [1]. It works with lower and upper approximation of a set as it is shown in Fig. 1. The discernibility relation is used for two basic operations in rough set

theory i.e. upper $\overline{R}X$ and lower $\underline{R}X$ approximations, which defines crisp and vague manner in the sets. If any concept of the universe can be formed as a union of some elementary sets, it is referred as crisp (precise). On the contrary, if the concept cannot be presented in such a way, it is referred as vague (imprecise, rough). $\underline{R}X$ is defined as the collection of cases whose equivalence classes are fully contained in the set of cases to approximate. $\overline{R}X$ is defined as the collection of cases whose equivalence classes are at least partially contained in (i.e. overlap with) the set of cases to approximate [6].
There are five regions of interesting: $\overline{R}X$ and $\underline{R}X$, and $POS_R(X)$, $BN_R(X)$ and $NEG_R(X)$. These sets are defined as shown below.

Let a set $X \subseteq U$, R be an equivalence relation and knowledge. Two subsets base can be associated:
i) R - Lower: $\underline{R}X = U\{Y \in U/R : Y \subseteq X\}$
ii) R - Upper: $\overline{R}X = U\{Y \in U/R : Y \cap X \neq \phi\}$

It means that the elements belonging to $\underline{R}X$ set can be with certainty classified as elements of X; while the elements belong to $\overline{R}X$ set can be possibly classified as elements of X. In the same way, $POS_R(X)$, $BN_R(X)$ and $NEG_R(X)$ are defined below [1].

iii) $POS_R(X) = \underline{R}X \Rightarrow$ certainly member of X
iv) $NEG_R(X) = U - \overline{R}X \Rightarrow$ certainly non member of X
v) $BN_R(X) = \overline{R}X - \underline{R}X \Rightarrow$ possibly member of X

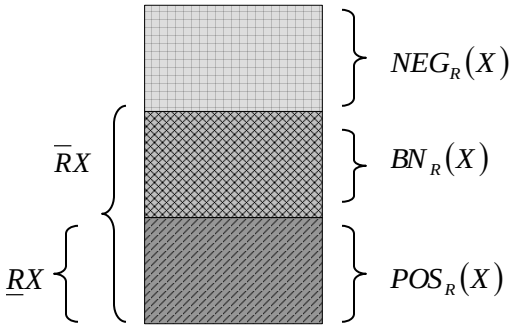

Figure 1: Definition of *R*-approximation sets and *R*-regions

Before the presentation of the algorithm, it is necessary to define two major concepts in Rough Set Theory, reduct and core. These concepts are important in the knowledge base reduction. Let R be a family of equivalence relations. The reduct of R, RED(R), is defined as a reduced set of relations that conserves the same inductive classification of set R. The core of R, CORE(R), is the set of relations that appears in all reduct of R, i.e., the set of all indispensable relations to characterize the relation R. As the core is the intersection of all reducts, it is included in every reduct, i.e., each element of the core belongs to some reduct. Therefore, in a sense, the core is the most important subset of

attributes, since none of its elements can be removed without affecting the classification strength of attributes.

The approximation of classification is a simple extension of the definition of approximation of sets. Namely if $F=\{X_1, X_2,...,X_N\}$ is a family of non empty sets, then $\underline{R}F=\{\underline{R}X_1, \underline{R}X_2,...,\underline{R}X_n\}$ and $\bar{R}F=\{\bar{R}X_1, \bar{R}X_2,..., \bar{R}X_n\}$, are called the RF – lower and the $\bar{R}F$ – upper approximation of the family F [3].

Two measures can be defined to describe inexactness of approximate classification. The first one is the extension of the measure defined to describe accuracy of approximation sets.

The accuracy of approximation of F by R is defined as [1]:

$$\alpha R(F) = \frac{\sum \text{card} \underline{R}X_i}{\sum \text{card} \bar{R}X_i} \tag{1}$$

where card(X) denotes the cardinality of X = ϕ.

The accuracy of approximation can be used to measure the quality of approximation of decision classes on the universe U. It is possible to use another measure of accuracy defined by 1 − αR(X). Some other measures of approximation accuracy are also used based on entropy or some more specific properties of boundary regions. The choice of a relevant accuracy of approximation depends on a particular data set. The accuracy of approximation of X can be tuned by R.

The second measure, called the quality of approximation of F by R, is the following [1]:

$$\gamma R(F) = \frac{\sum \text{card} \underline{R}X_i}{\text{card } U} \tag{2}$$

The accuracy of classification expresses the percentage of possible correct decision, when classifying objects, employing the knowledge R. The quality of classification expresses the percentage of objects that can be correctly classified as belonging to classe F employing knowledge R. By selecting a proper balance between the accuracy of classification and the description size it is expected to define the classifier with the high quality of classification also on unseen objects.

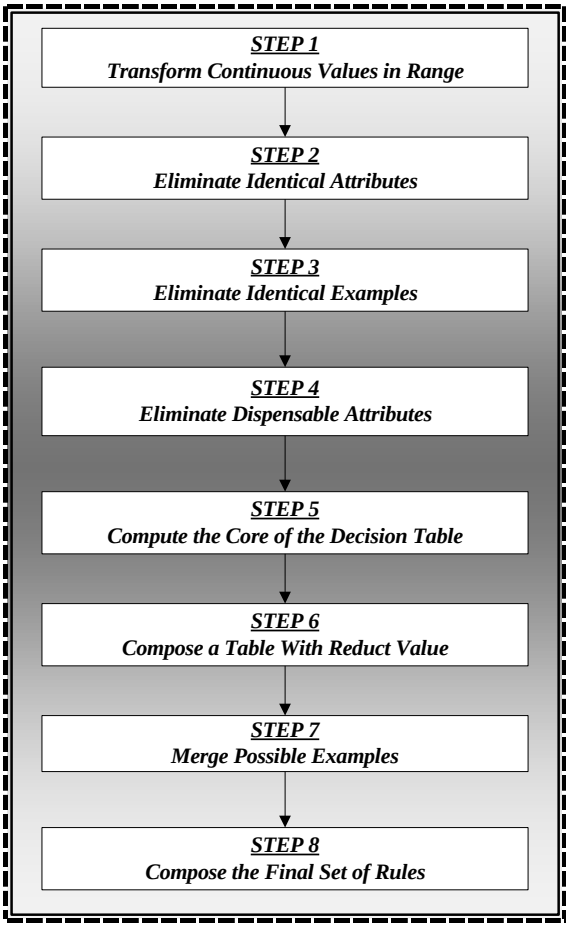

Figure 2: Reduction Algorithm

One of the most important applications of RST is the generation of decision rules for a given information system for the prediction of classes for new objects which are beyond observation. The rules are presented in an "If condition(s) then decision(s)" format.

2.3 Incomplete Information System

It may happen that some attribute values for an object are missing. To indicate such a situation a distinguished value, so-called null value, is usually assigned to those attributes [2]. If *Va* contains null value for at least one attribute $a \in U$ then *K* is called an incomplete information system, otherwise it is complete. Further on, we will denote null value by * [2]. The algorithm of the reduction of a decision table is shown in Figure 2 [7].

3 Application Examples

In Figure 3 it is shown the 118IEEE Test Power Network that was used in this study [8]. The input numerical values for the Rough Set approach considering an incomplete information system were obtained using the software package *SecurMining 1.0*, developed by the authors [5]. The ROSE software package was used to perform the RST analysis [9].

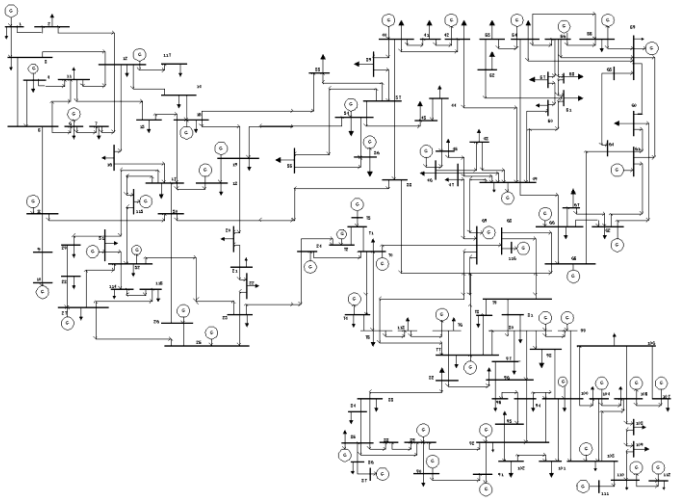

Figure 3: IEEE 118 Test Power Network

In this section it is presented the final results using the Rough Set Theory. A first order contingency study was carried out and it was obtained a list of 231 contingencies that allows the construction of a contingency control database. The specified attributes are as follows:

A – overloads in the transmission lines;
B – number of overloaded transmission lines;
C – voltage levels;
D – number of busbars with voltage violation;
E – severity indices related to the power and the voltage
F – severity indices related to the power losses.

Table 1 presents a set of information related to a contingency control database. Table 2 shows the chosen range for the coded qualitative attributes. The condition attributes are coded into three qualitative terms: Low, Medium and High. The decision attribute is coded into four qualitative terms: Normal (N), Alert (A), Emergency 1 (E1) and Emergency 2 (E2).

Attributs	*Codes*			
	0	1	2	3
A		90 % <	90 % ≤ a ≤ 110 %	> 110 %
B		2 ≤	3 ≤ b ≤ 4	> 4
C		0.85 <	0.85 ≤ c ≤ 1.05	> 1.05
D		2 ≤	3 ≤ d ≤ 5	> 5
E		0.800 <	800 ≤ e ≤ 0.900	> 0.900
F		0.800 <	0.800 ≤ e ≤ 0.900	> 0.900
S	N	A	E_1	E_2

Table 2: Definition of Range Attributes Coding

Cont Nº	Attributes A	B	C	D	E	F	S
1	2	1	1	1	3	3	A
2	0	1	1	1	3	3	A
3	3	1	1	1	3	3	E_2
4	2	1	1	1	3	3	E_1
5	2	1	1	1	3	3	E_1
6	2	1	1	1	3	3	A
7	3	3	0	1	3	3	E_2
8	2	1	1	1	3	3	E_1
9	2	1	1	1	3	3	E_1
10	2	1	1	1	3	3	E_1
11	2	1	2	1	3	3	A
12	2	1	1	1	3	3	A
13	2	0	1	1	3	3	A
14	2	1	2	1	3	3	A
15	2	1	1	1	3	3	A
16	2	1	1	1	3	3	A
17	2	1	1	1	3	3	A
18	2	1	1	1	3	3	A
19	3	1	1	1	3	3	E_2
20	2	1	1	1	3	0	A
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
222	3	1	1	1	3	1	E_2
223	2	1	1	1	3	1	A
224	2	1	1	1	3	1	A
225	2	1	1	1	3	1	A
226	2	1	1	1	3	1	A
227	2	1	1	1	3	1	A
228	2	1	1	1	3	1	A
229	2	1	1	1	3	0	A
230	2	1	1	1	3	1	A
231	2	1	1	1	3	1	A

Table 1: The Attributes Represented by the Set

***Step 1*:**
The first step of the algorithm is to redefine the value of each attribute according to a certain metric that was described above. Using these redefinitions for each contingency of Table 1, Table 3 arises.

Step 2 and 3:
The next step of the algorithm is to verify if any attribute can be eliminated by repetition. It can be verified that the attributes are different for all examples. Some examples are identical (for instance, contingencies 5 - 6, 9 and 10). The similar examples are also merged.

Step 4:
The next step is to verify if the decision table contains only indispensable attributes. This task can be accomplished eliminating each attribute step–by–step and verifying if the table gives the correct classification. For example, if the attribute E is eliminated, the table continues to give a correct classification. So, it can be said that E is a dispensable attribute for this decision table. However, when the attribute C is eliminated it can be verify

that the contingencies 1 and 4 have the same set of attributes but they give different classification. In this case, we say that the attribute C is indispensable for all attributes, so we can realize that the attributes A, B, C, D and F are indispensable, and E is dispensable for this decision table.

Cont Nº	Attributes						
	A	**B**	**C**	**D**	**E**	**F**	**S**
1	M	L	L	L	H	H	A
2	*	L	L	L	H	H	A
3	H	L	L	L	H	H	E_2
4	M	L	L	L	H	H	E_1
5	M	L	L	L	H	H	E_1
6	M	L	L	L	H	H	E_1
7	H	H	*	L	H	H	E_2
8	M	L	L	L	H	H	E_1
9	M	L	L	L	H	H	E_1
10	M	L	L	L	H	H	E_1
11	M	L	M	L	H	H	A
12	M	L	L	L	H	H	A
13	M	*	L	L	H	H	A
14	M	L	M	L	H	H	A
15	M	L	L	L	H	H	A
16	M	L	L	L	H	H	A
17	M	L	L	L	H	H	A
18	M	L	L	L	H	H	A
19	H	L	L	L	H	H	E_2
20	M	L	L	L	H	*	A
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
223	M	L	L	L	H	L	A
224	M	L	L	L	H	L	A
225	M	L	L	L	H	L	A
226	M	L	L	L	H	L	A
227	M	L	L	L	H	L	A
228	M	L	L	L	H	L	A
229	M	L	L	L	H	*	A
230	M	L	L	L	H	L	A
231	M	L	L	L	H	L	A

Table 3: Database with Range Values

Step 5 and 6:
Using the last information, can be computed the core of the set of contingencies. This computation can be done eliminating each attribute, step-by-step, and verifying if the decision table continues consistent. Using the compute package [9] it can be verified that the attributes A, B, C, D and F are the Core and the Reduct of the Problem.

Step 7 and 8:
According to the step 5 and 6, and using logical arithmetic, we can compose the set of rules. Incorporating the range values the final set of rules and approximate rules that contains the knowledge of Table 1, can be expressed the quality of classification for all conditions and the attributes in the core is 0.1385.
The Table 4 shows the approximation of the objects in the Decision levels.

Decision Level	Number of objects	Approximation Upper	Approximation Lower	Precision the approximation. of classification
1 – Alert	167	2	201	0.0100
2 – Emerg. I	34	0	199	0.0000
3 – Emerg. II	30	30	30	1.0000

Table 4: Approximation of the objects

According to the algorithm described, and using logical arithmetic, it is possible to compose the set of rules. Also, incorporating the range values the final set of rules and approximate rules that contains the knowledge of a initial database range values were obtained with the software package *SecurMining 1.0* and the ROSE computer programme [5][9].

Exact Rules:

1 – If (A is M and D is M) then S = A.

2 – If (C is M and E is L) then S = A.

3 – If (A is H) then S = E_2.

Approximate Rules:

4 – If (A is M and C is L) then S = A or S = E_1.

5 – If (A is M and D is L) then S = A or S = E_2.

The above rules can be written in a more clear way:

Exact Rules:

*1 – If the overloads in the transmission lines present a medium value and the number of busbars with voltage violation are medium then the Power System is in **Alert state**.*

*2 – If the voltage levels assume a medium values and the severity indices related to the power and the voltage are lower values then the Power System is also in **Alert state**.*

*3 – If the overloads in the transmission lines present a high values then the Power System is in **Emergency state II**.*

Approximate Rules:

*4 - If the voltage levels assume a medium values and the voltage levels present a Lower values then the Power System is in **Alert** or in **Emergency state I**.*

*5 - If the voltage levels present a medium values and the number of busbars with voltage violation assume lower values then the Power System is in **Alert** or in **Emergency state II**.*

4 Conclusions

In this paper it was presented the RST applied to an Electric Power System considering an incomplete information system. The Knowledge acquisition process is a complex task, since the experts have difficulty to explain how to solve a specified problem. The proper definitions of reducts allow to define knowledge reduction that does not diminish the original system's abilities to classify objects or to make decisions. Both reduction of dispensable knowledge and finding of optimal decision rules are transformable to the problem of computing prime implicates discernibility functions. It was also shown that discernibility functions for incomplete information systems may be constructed in conjunctive normal form. Consequently, an incomplete set of relevant information may arise. In order to overcome this problem it is proposed a new methodology to study and analyse the steady-state contingency classification using the RST. The study presents a systematic approach to transform examples in a reduced set of rules.

Acknowledgments

The first author would like to thank Fundação para a Ciência e Tecnologia, FCT, that partially funded this research work through the PhD grant n º SFRH/BD/38152/2007.

References

[1] Z. Pawlak, *Rough Sets – Theoretical Aspects of Reasoning about Data*, Kluwer, 1991

[2] M. Kryszkiewicz. Rought set approach to incomplete information systems. *Elsevier, Information Sciences*, 112: 39-49, 1998.

[3] M. Kryszkiewicz, "Rules in incomplete information systems", Elsevier, Information Sciences, 113, 1999, pp. 271-292.

[4] Maurílio Pereira Coutinho, Germano Lambert-Torres, Luiz Eduardo Borges da Silva, Edison F. Fonseca and Horst Lazarek. A methodology to extract rules to identify attacks in power system critical infrastructure: New results. *In Proceedings. IEEE Power Engineering Society* General Meeting, IEEE PES GM 2007, Tampa.

[5] C. I. Faustino Agreira: "Data Mining Techniques for Security Study and analysis to the Electrical Power Systems", Ph.d – dissertation, University of Porto, 2010.

[6] Ching-Lai, Crossley, P., MIEEE. and Franck Dunand. Knowledge extraction within distribution substation using rough set approach. *Power Engineering Society Winter Meeting*, Vol. 1: 654-659, 2002.

[7] Lambert-Torres, Alves da Silva, A. P., et al. Classification of power system operating point using rough set techniques. *IEEE International. Conference on Systems, Man and Cybernetics*, 1996.

[8] "Power Systems test Case Archive: 118 Bus Power Flow Test Case" Department of Electrical Engineering, University of Washington, [Online]. Available: http://www.ee.washington.edu/research/pstca/

[9] ROSE2 – Rough sets data explorer. Laboratory of Intelligent Decision Support Systems of the Institute of Computing Science, Poznan [Online] Available: http://www.idss.cs.put.poznan.pl/software/rose/