

Rough Set Dependency Analysis in Evaluation Studies – An Application in the Study of Repeated Heart Attacks

Ivo Düntsch School of Information and Software Engineering, University of Ulster
Günther Gediga Fachbereich Psychologie, Universität Osnabrück, Germany

Abstract

One method for modelling uncertain or inaccurate information is *rough set analysis* which was introduced and studied by Pawlak (1982) and his co-workers. Unlike other methods such as fuzzy set theory, Dempster – Shafer theory or statistical methods, rough set analysis requires no external parameters and uses only the information presented in the given data. In the present study we apply rough set analysis to an investigation of the indicators of repeated heart attacks.

1 Introduction

In many cases it is not possible to obtain complete – or certain – information. This may be due to subjective causes such as vagueness or may be due to objective ones such as measuring errors or insufficient knowledge. Faced with vague information, a researcher is limited by

- insufficient distinction of items,
- object sets which can only be approximated,
- insufficient granularity of the knowledge representation,
- decision rules which are not deterministic.

These cause, among others, the following problems:

- Explanation:
 - Which circumstances cause the situation under review?

- To which degree are which factors responsible for the situation?
- Is there a dependence among these factors, and, if so, to what degree?
- Decision: How we should act in a particular set of circumstances.

Several approaches are possible such as fuzzy set methods, Kwakernak (1978a), Kwakernak (1978b), Dempster - Shafer evidence theory, Shafer (1976) or statistical analysis, Pearl (1988). All of these require parameters outside the observed phenomena, or presuppose that the properties are of a quantitative character and are subject to random influences, in order that statistical methods such as variance analysis, regression, or correlation may be applied.

One method which avoids external parameters is *rough set analysis* as introduced by Pawlak (1982). It is a first (and sometimes sufficient) step in analyzing incomplete or uncertain information. It uses only internal information and does not rely on additional model assumptions such as fuzzy set methods or probabilistic models. In other words, instead of using numbers or other additional parameters, rough set analysis utilizes solely the structure of the given data.

Examples of applications of rough set theory to medicine, psychology, conflict analysis and other fields can be found in Pawlak (1991) and Słowiński (1992).

In the present paper, we investigate whether and how rough set analysis can be applied to an investigation into the causes of repeated heart attacks conducted by Rogner & Bartram (1989), Rogner & Bartram (1991), and Rogner et al. (1994). It turns out

that there are substantial differences between the rough set method and more traditional methods of data analysis. Of course, this does not reflect on the quality of either approach. In some instances, however, there are similar results. We try to explain both differences and similarities.

2 The model

Let U be a set and θ an equivalence relation on U . The pair $\langle U, \theta \rangle$ will be called an *approximation space*. In this context, the relation θ is often called an *indiscernibility relation*: Our knowledge of the objects in U extends only up to membership in the classes of θ . The idea now is to approximate our knowledge about a subset X of U modulo the indiscernibility relation θ :

For $X \subseteq U$,

$$\overline{X} = \bigcup \{ \theta x : x \in X \}$$

is the *upper approximation* of X , and

$$\underline{X} = \bigcup \{ \theta x : \theta x \subseteq X \}$$

its *lower approximation*. A *rough subset* of U is a pair $\langle \overline{X}, \underline{X} \rangle$, where $X \subseteq U$.

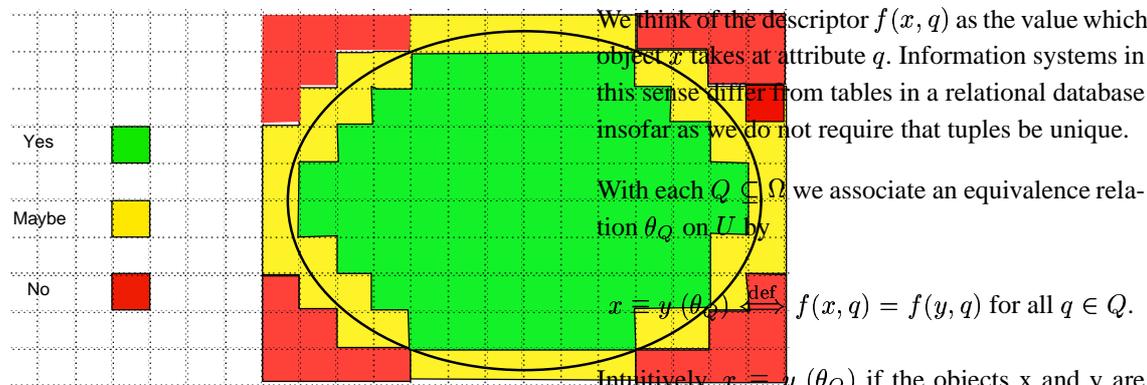
3. $x \in U \setminus \overline{X}$ means that x definitely *does not* have property P .

Thus, the area of uncertainty extends over $\overline{X} \setminus \underline{X}$. In a way, we are dealing with a three-valued logic having the values *yes*, *perhaps*, *no*. Indeed, the collection of all rough sets on U can be made into a regular double Stone algebra; these algebras, in turn, are the algebraic form of three-valued Łukasiewicz logic (see Düntsch (1995) for an overview of the algebraic properties of rough set structures). Alternatively, one can consider the lower approximation as a modal \Box operator (necessity), and the upper approximation as a \Diamond operator (possibility).

Knowledge representation in the rough set model is done via information systems in the following sense: An *information system* $\mathcal{S} = \langle U, \Omega, V_q, f \rangle_{q \in \Omega}$ consists of

1. A set U of objects,
2. A finite set Ω of attributes,
3. For each $q \in \Omega$ a set V_q of attribute values,
4. An information function $f : U \times \Omega \rightarrow V = \bigcup_{q \in \Omega} V_q$ with $f(x, q) \in V_q$ for all $x \in U, q \in \Omega$

Figure 1: Rough approximation



If $X \subseteq U$ is given by a predicate P and $x \in U$, then

1. $x \in \underline{X}$ means that x *certainly* has property P ,
2. $x \in \overline{X}$ means that x *possibly* has property P ,

The finest approximation we can obtain is θ_Ω : The classes of this relation collect all objects which do not differ in any attribute value. θ_Ω is a natural bound

for rough set analysis; within the classes of θ_Ω we cannot make any distinctions with this approach.

Suppose that $P, Q \subseteq \Omega$. We say that P is *dependent on* Q – written as $Q \rightarrow P$ – if $\theta_Q \subseteq \theta_P$. If $Q \rightarrow P$, then every class of θ_P is a union of classes of θ_Q . In other words, the classification of U induced by P can be expressed by the classification induced by Q . If $P = \{p\}$, we usually just write $Q \rightarrow p$.

As a measure of the *quality of an approximation* of a partition \mathcal{P} by a set Q of attributes we define the function $\gamma_Q : \text{Part}(U) \rightarrow [0, 1]$ by

$$(2.1) \quad \gamma_Q(\mathcal{P}) = \frac{\sum_{x \in \mathcal{P}} |X_{\theta_Q}|}{|U|}.$$

Thus, $\gamma_Q(\mathcal{P})$ is the ratio of the number of all elements of U classified with certainty to the total number of elements of U . Note that $P \rightarrow Q$ if and only if $\gamma_P(\theta_Q) = 1$.

If two sets $P, Q \subseteq \Omega$ are mutually dependent, we denote this fact by $\text{dep}(P, Q)$; clearly, dep is an equivalence relation on 2^Ω . If $\text{dep}(P, Q)$, and $P \subseteq Q$, then P and Q generate the same classification on U , and the attributes in $Q \setminus P$ are redundant for this classification. This leads to the following definition:

Let $Q \subseteq \Omega$. A set $P \subseteq Q$ is called a *reduct* of Q , if

1. $\text{dep}(P, Q)$ (i.e. $\theta_P = \theta_Q$),
2. For each $P' \subset P$ we have $\theta_{P'} \neq \theta_Q$.

In other words, $P \subseteq Q$ is a reduct of Q , if P is minimal among all subsets of Q which generate the same classification as Q . It is not hard to see, that each $Q \subseteq \Omega$ has a reduct, though this need not be unique. The intersection of all reducts of Q will be called the *core* of Q .

3 A strategy for modelling

We have now assembled the necessary tools to describe a strategy on how to use rough set methods in investigating uncertain information. The problem which we want to solve is the following:

Suppose that $\langle U, \Omega, V_q, f \rangle_{q \in \Omega}$ is an information system and \mathcal{P} is a partition of U .

1. Can \mathcal{P} be represented by a subset Q of Ω in such a way that $\theta_Q = \theta_{\mathcal{P}}$?
2. Which attributes are most significant for such a representation, if it exists?

The first step is to enlarge the given information system by adding a new attribute p to represent the partition \mathcal{P} . The new attribute set is denoted by Ω' , and the information function is extended by

$$f(x, p) = \theta_{\mathcal{P}} x.$$

The attributes in Ω can be considered to be independent variables, and p a dependent cluster variable.

The next step is to decide whether $\Omega \rightarrow p$. If this is not the case, then we cannot express \mathcal{P} by the attributes in Ω , and we have to be content with an approximation.

Otherwise, suppose that $\Omega \rightarrow p$ which indicates a first connection between \mathcal{P} and the attributes in Ω . To investigate this further, we could try to find all $Q \subseteq \Omega$ for which $Q \rightarrow p$. This "local" search is usually rather demanding on computational resources. Therefore, the search is limited in the first instance to "global" dependencies in the sense that one looks at the reducts of Ω on which p is dependent. If Q is such a reduct, then its approximation quality with respect to p is equal to 1 by Equation (2.1), and every class of θ_p is a union of classes of θ_Q .

Several cases can arise:

1. There is exactly one reduct: in this case, we have found the smallest combination of attributes of Ω which describes \mathcal{P} , and we continue with an analysis of this reduct as described below.
2. There is more than one reduct: in this case, we consider the core:
 - (a) The core is not empty: even though there are combinations of attributes which can be replaced by others and still retain \mathcal{P} , there is a set of attributes which is common to all these combinations.

- (b) The core is empty: every attribute can be replaced by others.

In order to find out which attributes in Q are significant for the representation of p by Q we determine the approximation quality of $Q \setminus \{q\}$ for each $q \in Q$. The lower the resulting value is, the higher the contribution of q .

4 An application: a study of repeated heart attacks

In the investigation Rogner & Bartram (1989), 80 male patients were examined who took part in a rehabilitation programme from March to June 1987 following a heart attack. The criterion was the absence, respectively presence, of certain medical indicators such as myocardiac problems, angina pectoris etc. four years later. The psychological indicators were closely related to the theoretical background:

- A questionnaire on the primary appraisal of the event itself, the *heart attack*.
- A questionnaire on the depressive emotions.
- Because of the central position of the coping mechanism, there were two operationalizations of coping:
 - A questionnaire on the depressive coping (HIKB) which was divided into eight subscales. The subscales – as well as the total – show a substantial correlation with the criterion.
 - A questionnaire on concrete coping mechanisms (HICOP) in specific circumstances which was taken retrospectively for the intensive, acute, and rehabilitation state (Table 1).

Beside these psychological indicators, medical indicators which are part of the standard diagnostics in rehabilitation treatment were considered as data sources. In all cases, ergometry and echocardiometry tests were conducted.

Table 1: HICOP Questionnaire

1.	I was doubtful whether I would survive.
2.	I blamed myself for not having cared until it was too late.
3.	I reflected in detail how life would go on.
4.	I blamed others.
5.	I reacted more calmly than others in this situation.
6.	I have tried to learn as much as possible about my illness.
7.	I have considered what I would do differently in the future.
8.	I decided that I had survived worse things.
9.	I could not understand why of all people I should be affected.
10.	I was worried about my family.
11.	I trusted the doctor to treat me successfully.
12.	I suffered great pain.
13.	I tried to forget that I was ill.
14.	I could not believe that I had had a heart attack.
15.	I was worried about keeping my job.
16.	I took my bad mood out on others.
17.	I felt badly about troubling others.
18.	I did not want to see anyone.
19.	I tried to speak to others about my illness.
20.	I found comfort in faith.
21.	I felt that things were getting better.

4.1 Analysis of the HICOP questionnaire

HICOP interviews using the questionnaire given in Table 1 were conducted in the rehabilitation clinic. For each of the three instances *acute*, *intensive*, and *rehabilitation* the questionnaire results were subjected to a Ward cluster analysis, which resulted in five groups. The problem arose if, and how, these groups could be recovered by a rough set analysis of the variables involved.

For each of the three instances we have

- An information system $\langle U, \Omega, V_q, f \rangle_{q \in \Omega}$, where U is the set of patients, Ω the set of questions, and $V_q = \{0, 1\}$ for all $q \in Q$. The information function f is defined canonically.
- A partition \mathcal{P} of U each class of which is a group of patients obtained by the cluster analysis.

Thus, we have the following situation(s):

Patient	Group	1	2	3	...
William	1	0	0	1	
George	1	0	1	0	
Paul	3	1	1	0	
Peter	5	0	0	1	
...					

The aim now is to find some $Q \subseteq \Omega$ such that $\gamma_Q(\mathcal{P})$ is maximal (see Equation (2.1)). In other words, we want to find out whether the five groups obtained by the cluster analysis could be determined by suitable configurations of attributes.

Following the procedure described in Section 3, we computed the reducts and the core of the information system. In order to determine the influence of a single core attribute on the quality of approximation, we removed each attribute from the core and computed the approximation quality of the resulting set by Equation (2.1). This value is given in column 3 of Table 2. The smaller the number, the greater the influence of the attribute. In Gediga &

Table 2: Result of a rough analysis of the HICOP–cluster results at the last instance

Item	Type	Approx.-quality (worst case)
1	Core	0.750
6	Core	0.833
8	Core	0.800
9	Core	0.767
11	Core	0.850
12	Core	0.833
17	Core	0.867
10,15,18,21	Reduct	0.967

Düntsch (1994) these results were compared with statistical standard methods (analysis of variance, discriminant analysis and other linear modelling approaches). It could be observed that

- Not all variables within the core show a similar effect within a linear modelling approach,
- Linear modelling consistently shows more "significant" results.

The differences between the results can be explained in terms of non–linearity, suppression, interaction, and the value of a chosen, but necessary α probability. It should be noted that none of the points creates a problem for the rough data analysis. On the other hand, rough set analysis offers a configurational description of the data based on the core (or reduct) variables. Because a configurational description is far too complex in most of the cases, a further step of statistical data analysis will usually be necessary.

We conclude that rough set analysis should be used as a first step for data analysis. If the results of the rough set analysis can be expressed in a small number of different configurations, data analysis is complete. If results of the rough set analysis contains more information, the reduced core can be used as the data base for other more restrictive (e.g. linear) modelling approaches. Using such techniques, spurious significant results are obviated, and non–significant results within core variables can be explained either as due to a small effect of the variable or to a misfit of the statistical model.

Because of these differences between linear modelling and rough set analysis some additional indication to the quality of the characterisation would be helpful. For this, we obtained additional information from Rogner & Bartram (1989): Since the 'coping' questionnaire was given retrospectively for the three instances *acute phase*, *intensive phase*, and *rehabilitation phase*, and since it can be assumed that the coping questionnaire is relatively homogeneous during the acute phase, it is reasonable to assume that the characterisation of the groups can be described by the variation of the answers over the three instances. Table 3 compares the results of Rogner & Bartram (1989) with the results of the rough analysis, taken at the final instance.

Table 3: Characterisation of HICOP–items

Item	Var. inf.	Rough inf.	Item	Var. inf.	Rough inf.
1	yes	yes/Core	11	yes	yes/Core
2	yes	no	12	yes	yes/Core
3	no	no	13	no	no
4	no	no	14	no	no
5	($p < 10\%$)	no	15	no	yes/Reduct
6	yes	yes/Core	16	no	no
7	($p < 10\%$)	no	17	yes	yes/Core
8	yes	yes/Core	18	yes	yes/Reduct
9	yes	yes/Core	19	no	no
10	yes	yes/Reduct	20	no	no
			21	no	yes/Reduct

The analysis shows that all core items can be classified as variation–sensitive; hence, the characterisation into five groups by the items in the core can be interpreted as the characterisation of change at the last instance. Summing up, we conclude that

- the reduction of attributes via rough set anal–

ysis is similar to that in linear modelling,

- the model assumptions of rough set analysis are much less stringent,
- there are plausibility reasons that – at least for the chosen example – rough set analysis is the more suitable instrument.

5 Analysis of the HIKB questionnaire

The HIKB questionnaire was divided into three groups:

1. Primary appraisal, specific emotions
2. Depressive coping
3. Medical indicators

Regarding freedom from symptoms after four years as the dependent variable, we first note that the core of the information systems is empty. Table 4 shows some selected reducts. The reduct belonging to Psych1

Table 4: Reducts for the prediction of freedom from symptoms

Item	Psych1	Psych2	Med1	Med2
Primary appraisal	—	0.931	0.868	—
Depressive emotions	—	0.966	—	—
Keep to self	0.966	0.966	—	—
Social supp., fellow patients	0.897	—	0.931	—
Anger	0.914	—	—	0.897
Self blame	0.966	—	—	0.931
Optimism	—	0.931	0.828	0.931
Social support, partner	0.931	—	—	—
Positive reappraisal	0.966	—	—	—
Upward comparison	0.966	0.897	—	—
Ergometry: Wattage <100	—	—	0.931	0.931
Ergometry: RR syst	—	—	—	—
Ergometry: RR diast	—	—	0.931	0.897
Ergometry: Rhythm disturb.	—	—	—	—
Ergometry: Full strain	—	—	—	—
Ergometry: Angina Pectoris	—	—	—	—
Ergometry: Heart rate	0.931	0.966	0.914	0.931
Echocard.: LV enddiast	—	—	0.862	0.828
Echocard.: LV endsyst	—	—	—	—
Echocard.: Shortening coeff.	0.897	0.897	0.966	0.931

contains essentially the items of depressive coping and several necessary medical indicators. Psych2 contains the pre-processing items, *primary appraisal* and *specific emotions*, where we have chosen the smallest reduct with respect to cardinality. The medical reducts contain mainly the medical items and whatever is necessary to obtain a reduct of minimal size.

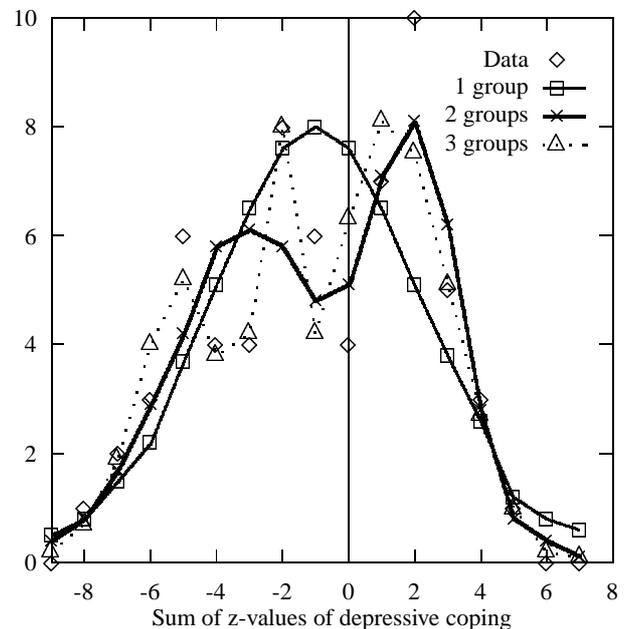
A cautious interpretation of this result could be that the pre-processing variables influence the dependent variable *freedom of symptoms* rather indirectly,

since they can be replaced by the symptoms of depressive coping. Furthermore, it is not unimportant for the validity of the results that after scaling the importance of the medical criteria, the heart rate and the shortening coefficient were recognized as important diagnostic features.

A possible explanation for the empty core is that the reliability of many indicators is low; therefore, random orders may lead to a classification as well. Hence, it may be asked, if there is one reliable psychological indicator whose prediction capacity could be tested in relation to the medical indicators. This has been done in part by Rogner et al. (1994), who replaced the eight psychological indicators with a sum of z-values per patient. This resulted in an extremely high correlation to freedom of symptoms, whereas medical indicators showed consistently a low predictive power.

In order to do a rough set analysis with this aggregated variable, we had to categorise depressive coping. It turned out that a 3-class mixture with normally distributed carrier was adequate (AIC; Bozdogan (1987)). Figure 2 shows that the bounds $z \leq -4$, $-4 < z \leq -1$, $-1 < z$ are adequate for the three groups. The rough set analysis of the aggre-

Figure 2: Mixture analysis of depressive coping



gated psychological variable with the medical variables shows that now there is a non empty core, see Table 5, where the smallest reduct is displayed.

Table 5: Smallest reduct for the prediction of freedom of symptoms with one aggregated psychological variable

Item	Approximation	Core
Depressive coping	0.671	yes
Ergometry: RR diast.	0.899	no
Ergometry: Heart rate	0.873	yes
Ergometry: Rhythm disturbances	0.823	yes
Ergometry: Strain	0.873	no
Ergometry: Angina pectoris	0.949	no
Echocard.: LV enddiast.	0.810	yes
Echocard.: Shortening coeff.	0.747	yes

From the fact that the approximation quality reduces to $\gamma = 0.671$ when removing the psychological variable, we see that depressive coping has a higher information value than any of the medical indicators. Furthermore, the diagnostically important attributes are also present in the core.

Comparing the results of the rough set analysis with the linear modelling approach in Rogner et al. (1994), we summarize:

- The high predictive power of depressive coping remains stable within the changed analysis context. The effect is independent of variation within the medical indicators.
- Rough set analysis, nevertheless, shows that there is also a substantial predictive power of some of the medical indicators. This is a result which was not obtained within the framework of a linear modelling approach.

References

- BOZDOGAN, H. (1987), Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 354–370.
- DÜNTSCH, I. (1995), Rough sets and algebras. In: E. Orłowska (ed.), *Modeling Incomplete Information*, To appear.
- GEDIGA, G. & DÜNTSCH, I. (1994), Grobmen- gen Dependenz Analysen in Evaluationsstudien - eine Anwendung in der Reinfarkt-Forschung. *Forschungsberichte aus dem Fachbereich Psychologie der Universität Osnabrück*, **102**.
- KWAKERNAK, H. (1978a), Fuzzy random variables, Part 1. *Inform. Sci.*, **15**, 1–15.
- KWAKERNAK, H. (1978b), Fuzzy random variables, Part 2. *Inform. Sci.*, **17**, 243–278.
- PAWLAK, Z. (1982), Rough sets. *International Journal of Computer and Information Sciences*, **11**, 341–356.
- PAWLAK, Z. (1991), *Rough sets: Theoretical aspects of reasoning about data*. Kluwer, Dordrecht.
- PEARL, J. (1988), *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.
- ROGNER, J. & BARTRAM, M. (1989), Krankheitsbewältigung und Adaption bei Herzinfarkt-Patienten während der Anschlußheilbehandlung. *Forschungsberichte aus dem Fachbereich Psychologie der Universität Osnabrück*, **70**.
- ROGNER, J. & BARTRAM, M. (1991), Strategien und Muster der Krankheitsbewältigung bei Herzinfarkt-Patienten: Ergebnisse einer Revision der Fragebogen HI-COP und HI-KB. *Forschungsberichte aus dem Fachbereich Psychologie der Universität Osnabrück*, **78**.
- ROGNER, J., BARTRAM, M., HARDINGHAUS, W., LEHR, D. & WIRTH, A. (1994), Depressiv getönte Krankheitsbewältigung bei Herzinfarkt-patienten – Zusammenhänge mit dem längerfristigen Krankheitsverlauf und Veränderbarkeit durch eine Gruppentherapie auf indirekt-suggestiver Grundlage. In: G. Schüßler & E. Leibing (eds.), *Coping. Verlaufs- und Therapiestudien chronischer Krankheit*, pp. 95–109, Hogrefe.
- SHAFER, G. (1976), *A mathematical theory of evidence*. Princeton University Press.

SŁOWINSKI, R. (1992), *Intelligent decision support: Handbook of applications and advances of rough set theory*. Kluwer, Dordrecht.