

ROUGHIAN

Rough Information Analysis

Ivo Düntsch*

School of Information and Software Engineering

University of Ulster

Newtownabbey, BT 37 0QB, N.Ireland

I.Duentsch@ulst.ac.uk

Günther Gediga*

FB Psychologie / Methodenlehre

Universität Osnabrück

49069 Osnabrück, Germany

Guenther@Gediga.de

Summary

Rough set data analysis (RSDA), introduced by Pawlak [46], has become a much researched method of knowledge discovery with over 1200 publications to date [50]. One feature which distinguishes RSDA from other data analysis methods is that, in its original form, it gathers all its information from the given data, and does not make external model assumptions as all statistical and most machine learning methods (including decision tree procedures) do. The price which needs to be paid for the parsimony of this approach, however, is that some statistical backup is required, for example, to deal with random influences to which the observed data may be subjected. In supplementing RSDA by such meta-procedures care has to be taken that the same non-invasive principles are applied.

In a sequence of papers and conference contributions, we have developed the components of a non-invasive method of data analysis, which is based on the RSDA principle, but is not restricted to “classical” RSDA applications. In this article, we present for the first time in a unified way the foundations and tools of such rough information analysis.

1 Introduction

Concept forming and classification in the absence of complete or certain information has been a major concern of artificial intelligence for some time. Faced with such imperfect information, a researcher is limited, for example, by

- Insufficient distinction of items,
- Object sets which can only be approximated,
- Inappropriate granularity of the knowledge representation,

Traditional “hard” data analysis based on statistical models or (binary) rule and reasoning methods are in many cases not equipped to deal with uncertainty, relativity, or non-monotonic processes. As an alternative, a “soft computing” approach has come into fashion, whose main components are fuzzy logic, neural network theory, and probabilistic reasoning [65]. We observe that all of these “soft” methods require “hard” parameters outside the observed phenomena – membership degrees, prior probabilities, parameters for differential equations – the origin of which is not always clear. One should not forget that the results of such methods are valid only up to the given model assumptions which may not always be succinctly stated (or even known).

*The ordering of authors is alphabetical, and equal authorship is implied.

A major problem of any data analysis method is the assumption of representativeness of the observed data for the underlying “true” data structure postulated in a model. The reason for this is the huge state complexity of the space of possible rules, even when there are only a few number of features (Tab. 1). This causes the distribution of the cell frequencies in the observed data to be extremely sparse, and a representative sampling is rather unlikely.

Table 1: State complexity

# of attr. values	# of attributes		
	10	20	30
	$\log_{10}(\text{states})$		
2	3.01	6.02	9.03
3	4.77	9.54	14.31
4	6.02	12.04	18.06
5	6.99	13.98	20.97

One can observe that in most cases, real life problems

- have few data points with respect to state complexity,
- show many attribute dependencies,

and thus, traditional statistical models do not seem optimal tools for data mining. Indeed, it would be advantageous to have an instrumentarium which can

- remove redundant information, and
- discover which features or attributes are relevant for data description and/or prediction, and provide decision support,

without assuming facts which are not contained in the data. Such an instrument seems to be necessary as a pre-processing tool for further analysis, and, as observed in [45], it is sometimes not only the first but also a sufficient step for sensible data analysis.

A rule based approach to achieve these aims is rough set data analysis (RSDA) which has been developed by Z. Pawlak and his co-workers since the early 1970s [34, 35, 45, 46]. Today, rough set theory and its applications in data description and analysis are a prospering research field, and many applications and pointers to recent literature can be found in [37, 41, 43, 50–52, 67].

As we shall see, the rough set model is data driven: Attribute dependencies and decision rules are always extracted (or learned) from existing systems, and they are not part of the design process as is the case, for example, with relational databases. In this respect, RSDA can be considered part of machine learning, or, more concretely, data mining [22, 48], which in turn is a part of “Knowledge discovery in data bases” (KDD). The reader who is interested in the different approaches to imperfect information in data base theory and artificial intelligence is encouraged to consult the survey paper [44].

The original view behind the rough set model is the observation that

The information about a decision is usually vague because of uncertainty and imprecision coming from many sources ... Vagueness may be caused by granularity of representation of the information. Granularity may introduce an ambiguity to explanation or prescription based on vague information [49].

In other words, the original concept behind the model is the realization that sets can be described “roughly”, i.e. there are three regions of knowledge:

An object has a property

■ Certainly, ■ Possibly, ■ Certainly not.

This looks conspicuously like a fuzzy membership function, and indeed, on the algebraic – logical level, we can say that the algebraic semantics of a rough set logic corresponds to a fuzzy logic with a three-valued membership function [see 10, 42]. An analysis of the relation between the rough set approach and fuzzy sets has been done by [8].

Rough set analysis uses only internal knowledge, and does not rely on prior model assumptions as fuzzy set methods or probabilistic models do. In other words, instead of using external numbers or other additional parameters, rough set analysis utilises solely the structure of the given data under the motto

*Let the data speak for themselves*¹.

Of course, this does not mean that RSDA does not have any model assumptions; for example, we show in Section 3.2 that the basic statistical assumption underlying RSDA is the *principle of indifference*. However, model assumptions are such that we admit complete ignorance of what happens within the region of indiscernability given by the granulation of information.

It is important to realise that RSDA does not exist in an isolated world of its own, but needs to be put in the wider context of an overall data model, and, indeed, the KDD process. This includes that questions such as significance of rules, random influences on the data, objectivity of the model selection process etc must be addressed if RSDA wants to be accepted as a serious model for data analysis. In this context, we can strongly recommend the advice given to data miners in [27].

In this paper, we will outline the basic aims of rough set theory, its tools, its limits, and, collecting our earlier work, we show how it can be enhanced by non-invasive methods for significance testing, data filtering, and objective criteria for model selection. In this way, we believe we can present a “rough information analysis” (ROUGHIAN) which is well founded with sufficient epistemological and statistical backup to serve as a fully fledged method of non-invasive data analysis.

The paper is structured as follows: We first present a data model and position RSDA and other methods in this model. This followed by a description of the structural, statistical, and information theoretic tools which are used in RSDA and the proposed additional procedures. These are introduced and discussed in the next Section which describes ROUGHIAN proper: Data filtering, significance testing, and model selection. We conclude with a summary and an outlook.

2 Data models and model assumptions

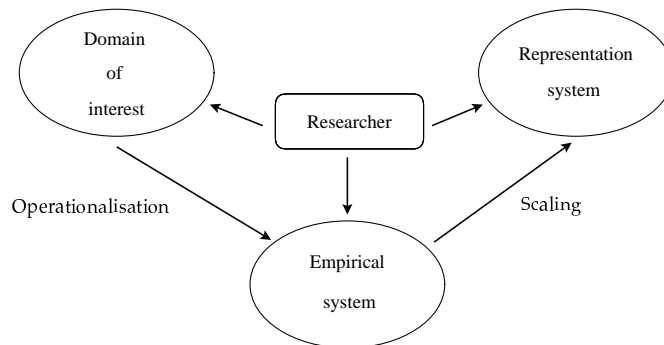
As the basis for our presentation we choose the data model of Gigerenzer [26]: The centre of the modelling process is the researcher who chooses

1. A domain \mathcal{D} of interest.
2. A system \mathcal{E} , which consists of a body of data and relations among the data, called an *empirical system*, and a mapping $e : \mathcal{D} \rightarrow \mathcal{E}$, called *operationalisation*. This mapping is often called *representation* in KDD.
3. A *representation system*, also called *numerical system*² \mathcal{M} , and a mapping $m : \mathcal{E} \rightarrow \mathcal{M}$, called *scaling* which maps the data and the relations among the data to a numerical or graphical scale.

¹Jaynes [32], p. 641, attributes this sentence to R.A. Fisher, but we have not been able to ascertain this. Jaynes goes on stating that “The data *cannot* speak for themselves, and they never have, in any real problem of inference”. We are inclined to agree with him, but only up to the point, that, due to the necessary operationalisation of the domain, there are no “raw data”, and therefore, they cannot tell us anything. We hope that the results of the current paper show that, given the minimal assumptions of an empirical model (which do not include a subjective prior probability), the data tell us much about themselves, and that additional assumptions are not necessary, at least for the first step of data analysis.

²These systems are called “numerical” for historical reasons. Modern scaling theory includes structures such as networks, knowledge spaces and other complex constructs.

Figure 1: Data modelling



The choice of each of the parts of the model is a pragmatic decision by researchers, how they want to model the properties and dependencies of real life criteria in the best possible way, according to their present objectives and their state of knowledge about the world. As a simple example, consider the situation that the knowledge state of individuals in a certain area is to be assessed, which is our domain of interest \mathcal{D} . The empirical system consists of the individuals and problems which they are asked to solve. These problems are given by an expert who assumes that they constitute a true operationalisation of the real knowledge states of the individuals. A numerical system for this domain are the test scores achieved by the students.

The operationalisation is to a large part subjective, and thus the first source of uncertainty (“model selection bias”): One question is whether the elements and relations of the empirical model \mathcal{E} are representative for the objects of the domain \mathcal{D} and the relations among them, another whether the choice of attributes covers the relevant aspects of \mathcal{D} . Operationalisation and an empirical model – along with the assumptions that go with them – are necessary in any type of data analysis, while it is our contention that a numerical system is not.

All statistical and most KDD methods make external model assumptions, and thus reside on the level of the numerical model; a typical example is

“We will consider rectangular datasets whose rows can be modelled as independent, identically distributed (iid) draws from some multivariate probability distribution. ... We will consider three classes of distributions f :

1. the multivariate normal distribution;
2. the multinomial model for cross-classified categorical data, including loglinear models;
and
3. a class of models for mixed model and categorical data ...” [55].

Unlike in the example above, model assumptions are not always spelled out, and thus, it is not clear to an observer on what basis and with which justification a particular method is applied. As we have already pointed out, the assumption of representativeness, for example, is a problem of any analysis in most real life data bases. Furthermore, the influence of the model assumptions on the results is not always taken into account when these are interpreted.

Not clearly separating the operationalisation and scaling processes may result in unstated (or overlooked) model assumptions which may compromise the validity of the result of the analysis. We invite the reader to consult [28] for an indication of what can go wrong when statistical models are applied which are not in concordance with the objectives of the research (if these are known). In particular, all we can hope for is an approximation of the reality that models are supposed to represent, and that there is no panacea for all situations.

Given the difficulties of justifying - or even stating - model assumptions when constructing a numerical system, one may well ask, whether such a system is, indeed, necessary. Just as an empirical system is between the domain of interest and the numerical system, one can argue that one way of avoiding the difficulties of model building is to remain at the level of an empirical system, and investigate what it tells us about the observable and unobservable data. This alleviates to some extent the concerns regarding model uncertainty raised in [5].

RSDA is such a method of data analysis which stays on the level of the empirical system, working only with and from the given operationalisation. Formally, this means that the scale mapping is the identity function. It follows that a question how “noise” is handled within this model is not sensible: The definition of “noise” assumes a numerical model which, at the stage of the empirical model is not yet present. Similarly, there cannot be “outliers”, since they also presuppose a numerical model.

Getting rid of Scylla on the one hand invites Charybdis on the other: For example, not assuming a specific distribution raises the question of rule significance, and, in order to stay consistent with our aim of minimising model assumptions, we need to present tools in which additional restrictions do not creep in through the backdoor. We believe that the methods described below satisfy this condition.

3 The tools of rough set data analysis

3.1 Symbolic tools

Granularity of information can be described by equivalence relations on a set U of objects up to the classes of which objects are discernible; what happens within the classes is not part of our knowledge. Hence, given an equivalence relation θ on a domain U , our knowledge of a subset X of U is limited to the classes of θ and their unions. This leads to the following definition:

For $X \subseteq U$, we say that

$$(3.1) \quad \underline{X}_\theta \stackrel{\text{def}}{=} \bigcup \{ \theta x : \theta x \subseteq X \}$$

is the *lower approximation* or *positive region* of X , and

$$(3.2) \quad \overline{X}^\theta \stackrel{\text{def}}{=} \bigcup \{ \theta x : x \in X \}$$

is the *upper approximation* or *possible region* of X with respect to θ . If θ is understood, we shall usually omit the subscript and the superscript.

A *rough set* is a pair $\langle \underline{X}, \overline{X} \rangle$, and $X \subseteq U$ is called θ -*definable*, if $\underline{X} = \overline{X}$.

Equivalence relations are the major tool of the rough set model, and manipulation of these relations and their classes constitutes its symbolic part.

Knowledge representation in the rough set model is done via information systems which are a tabular form of an OBJECT \rightarrow ATTRIBUTE VALUE relationship as shown in Table 2 [23].

More formally, an *information system*

$$\mathcal{I} = \langle U, \Omega, V_q, f_q \rangle_{q \in \Omega}$$

consists of

1. A finite set U of objects,
2. A finite set Ω of attributes or features,

Figure 2: Rough approximation

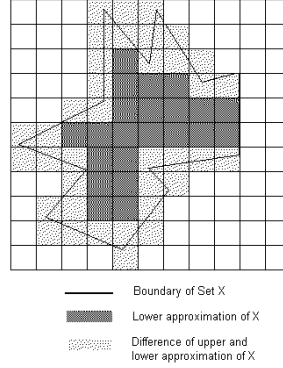


Table 2: Fisher's iris data

Object	Sepal length	Sepal width	Petal length	Petal width	Class
1	50	33	14	2	1
2	46	34	14	3	1
3	65	28	46	15	2
4	62	22	45	15	2
6	67	30	50	17	3
7	64	28	56	22	3
<143 other values>					

3. For each $q \in \Omega$

- A set V_q of attribute values,
- An information function $f_q : U \rightarrow V_q$.

We may think of the descriptor $f_q(x)$ as the value which object x takes at attribute q .

This operationalisation by Object \rightarrow Attribute data tables assumes the “nominal scale restriction” which postulates that each object has exactly one value of each attribute at a given time, and that the observation of this value is without error.

Information systems are a primary source of equivalence relations: With each $Q \subseteq \Omega$ we associate an equivalence relation θ_Q on U by

$$x \equiv y (\theta_Q) \text{ iff } f_q(x) = f_q(y) \text{ for all } q \in Q.$$

Intuitively, $x \equiv y (\theta_Q)$ if the objects x and y are indiscernible with respect to the values of their attributes from Q .

These equivalences form a meet - subsemilattice $\mathbf{E}(\mathcal{I})$ of the lattice $\mathbf{EQ}(U)$ of all equivalence relations on U , and the mapping $\mathbf{E} : \wp(\Omega) \rightarrow \mathbf{EQ}(U)$ defined by

$$Q \mapsto \theta_Q$$

is a homomorphism from $\langle \wp(\Omega), \cup \rangle$ to $\langle \mathbf{EQ}(U), \cap \rangle$ with $\mathbf{E}(\emptyset) = U \times U$. We denote the image of \mathbf{E} by $\mathbf{E}(\mathcal{I})$. The natural partial order on $\mathbf{E}(\mathcal{I})$ is set inclusion \subseteq , and we say for $P, Q \subseteq \Omega$ that

P is dependent on Q if and only if $\theta_Q \subseteq \theta_P$.

If P is dependent on Q we shall usually write this as $Q \Rightarrow P$. In this case, each class of θ_P is the union of classes of θ_Q , and thus, the description of an element by the attributes in P can be locally replaced by a description by the attributes in Q without losing any extensional information.

If $\theta_Q \subseteq \theta_P$, then we say that Q is a *reduct* of P if

$$(3.3) \quad (\forall R \subseteq Q)(R \neq Q \rightarrow \theta_R \not\subseteq \theta_P).$$

The intersection of all reducts of P is called the *core* of P , denoted by $\text{core}(P)$. This is slightly more general than the usual definition. If $P = \Omega$, we just speak of a reduct or the core (of \mathcal{I}).

It may be worth to point out that forming of reducts is a procedure local to the attribute sets involved. In particular, reducts of Ω or its core only describe how the finest partition of U – induced by the whole system – can be obtained by (possibly) fewer features than all of Ω . In algebraic terms, a reduct Q of P corresponds to one concrete inclusion in $\{\theta_R : R \subseteq \Omega\}$, and affects only θ_P . The statement “Attributes in a reduct can replace the whole attribute set” is not globally true, since it only says something about one equation in the whole semilattice $\mathbf{E}(\mathcal{I})$ of induced equivalence relations in U .

The dependency structure of implications $Q \Rightarrow P$ is reflected by the natural order \subseteq of its associated meet-semilattice $\mathbf{E}(\mathcal{I})$, and finding dependencies in \mathcal{I} and reducts is the same as exploring the equational structure of $\mathbf{E}(\mathcal{I})$. We refer the reader to [40] and [12] for details and further references.

Equivalence relations θ_Q, θ_P are used to obtain rules in the following way:

We let $Q \rightarrow P$ be the relation

$$\langle X, Y \rangle \in Q \rightarrow P \iff X \text{ is a class of } \theta_Q, Y \text{ is a class of } \theta_P, \text{ and } X \cap Y \neq \emptyset.$$

A pair $\langle X, Y \rangle \in Q \rightarrow P$ is called a Q, P – *rule* (or just a rule, if Q and P are understood) and usually written it as $X \rightarrow Y$. By some abuse of language we shall also call $Q \rightarrow P$ a rule when there is no danger of confusion.

Each rule has two parts (one of which may be void):

The *deterministic* – or *functional* – part of $Q \rightarrow P$, written as $Q \xrightarrow{\text{det}} P$, is the set

$$(3.4) \quad \{\langle X, Y \rangle \in Q \rightarrow P : X \subseteq Y\}.$$

If $\langle X, Y \rangle \in Q \xrightarrow{\text{det}} P$, then X is called P – *deterministic* or just *deterministic*, if P is understood.

If $Q \rightarrow P = Q \xrightarrow{\text{det}} P$, i.e. if $Q \rightarrow P$ is a function, then we call $Q \rightarrow P$ *deterministic* and write $Q \Rightarrow P$; it is not hard to see that

$$Q \Rightarrow P \text{ if and only if } \theta_Q \subseteq \theta_P,$$

so that the notation is consistent with the definition of dependency given earlier on.

3.2 Statistical tools

Even though rough set analysis is a symbolic method, it uses an inherent metric which is only based on internal information, given by the information system, and not on extraneous parameters which are

required by other methods of data analysis. We first describe it by using one equivalence relation on a set: If $\theta \in \mathbf{EQ}(U)$ and $X \subseteq U$, then the *approximation quality of θ with respect to X* is defined as

$$(3.5) \quad \gamma_\theta(X) \stackrel{\text{def}}{=} \frac{|X_\theta| + |{-X}_\theta|}{|U|},$$

see [47]. This measure is the relative frequency of all elements which are correctly classified under the granulation of information by θ with respect to being an element of X or not. Alternatively, observing that X induces a partition \mathcal{P} of U with the classes X and $-X$, the function γ_θ measures the approximation quality of θ with respect to the partition \mathcal{P} . There are two statistics related to the approximation quality of θ , namely

$$\mu_*(X) \stackrel{\text{def}}{=} \frac{|X|}{|U|}, \quad \mu^*(X) \stackrel{\text{def}}{=} \frac{|\overline{X}|}{|U|}.$$

It is easy to see that $\mu^*(X) = 1 - \mu_*(-X)$, and

$$\gamma(X) = \mu_*(X) + \mu_*(-X),$$

so that we can regard μ_* as the underlying statistics of the rough set model.

For each equivalence θ on U we let B_θ be the subalgebra of $\langle \wp(U), \cap, \cup, -, \emptyset, U \rangle$ whose atoms are the classes of θ . Now, the restriction $\mu_* \upharpoonright B_\theta$ is a probability measure on B whose inner measure is just μ_* , and the measurable sets of $\langle U, B_\theta, \mu_* \upharpoonright B \rangle$ are just the θ -definable sets. We say that a probability measure p on B_θ is *compatible with θ* , if

$$\mu_*(X) \leq p(X) \leq \mu^*(X),$$

for all $X \in B_\theta$. It is easy to see that the only probability measure on $\wp(U)$ which is compatible to all functions μ_* is given by

$$(3.6) \quad p(X) = \frac{|X|}{|U|},$$

so that $p(x) = \frac{1}{|U|}$ for all $x \in U$. In other words, rough set theory assumes the *principle of indifference* [2], where

- In the absence of further knowledge all basic events are assumed to be equally likely.

Thus, the statistical interpretation of the rough set approach is quite simple:

- Rough set analysis neglects the underlying joint distributions of the attributes and the reported statistics μ_* , resp. γ , are sufficient only if the joint distributions of the attributes are constant as in (3.6).

This sounds like a drawback, but one should note that rough set analysis is applied (and applicable!) in a “few – objects – many – attributes” situation which is very different to the situations usually encountered in statistical modelling. In the field of applied regression analysis it was shown that in comparable situations the assumption “simple is better” – e.g. using 0–1 regression weights – results in more stable estimates than using an approach with many parameters [6]. We shall return to this theme in Section 4.4.

Generalising (3.5) to partitions with more than two classes, we define the *quality of an approximation of a an attribute set Q with respect to an attribute set P* by

$$(3.7) \quad \gamma(Q \rightarrow P) = \frac{|\bigcup\{X : X \text{ is a } P\text{-deterministic class of } \theta_Q\}|}{|U|}.$$

Note that $Q \Rightarrow P$ if and only if $\gamma(Q \rightarrow P) = 1$.

3.3 Information theoretic tools

A measure which has not yet been fully exploited in rough set theory is the use of the information theoretic *entropy function* which measures three things [39, p. 16]:

- The amount of information provided by an observation E,
- The uncertainty about E,
- The randomness of E.

The characteristic of this approach is that information of uncertainty described by a probability distribution is mapped into a dimension which has its own meaning in terms of size of a computer program, and which has the consequence that

- Effort of the coding the “knowledge” in terms of optimal coding of given rules and
- Consequences of “guessing” in terms of optimal number of decisions to classify a random chosen observation

can be aggregated in the same dimension.

A comprehensive textbook on these topics is [36], and an introduction as well as pointers to further literature can be found in [1].

Let \mathcal{P} be a partition of U with classes $M_i, i \leq k$, each having cardinality m_i . In compliance with the statistical assumption of the rough set model we assume that the elements of U are randomly distributed within the classes of \mathcal{P} , so that the probability of an element x being in class M_i is just $\frac{m_i}{|U|}$. We now define the *entropy* of \mathcal{P} by

$$H(\mathcal{P}) \stackrel{\text{def}}{=} \sum_{i=0}^k \frac{m_i}{|U|} \cdot \log_2\left(\frac{|U|}{m_i}\right).$$

If θ is an equivalence relation on U and \mathcal{P} its induced partition, we will also write $H(\theta)$ instead of $H(\mathcal{P})$.

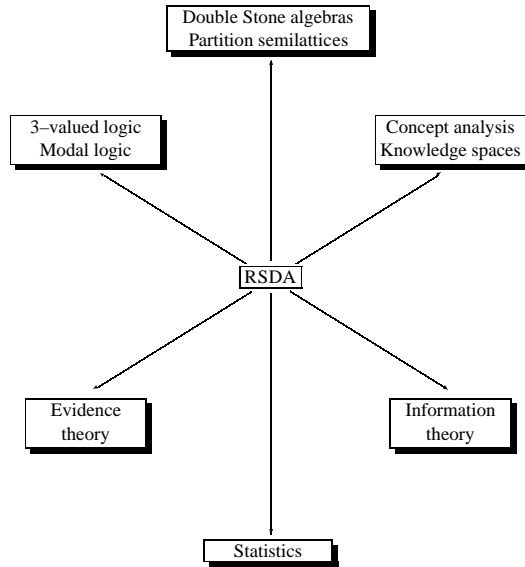
The entropy estimates the mean number of comparisons minimally necessary to retrieve the equivalence class information of a randomly chosen element $x \in U$. We can also think of the entropy of \mathcal{P} as a measure of granularity of the partition: If there is only one class, then $H(\mathcal{P}) = 0$, and if \mathcal{P} corresponds to the identity, then $H(\mathcal{P})$ reaches a maximum (for fixed n). In other words, with the universal relation there is no information gain, since there is only one class and we always choose the correct class of an element; if the partition contains only singletons, the inclusion of an element in a specific class is hardest to predict, and thus the information gain is maximised.

This concludes our presentation of the basic tools of the traditional rough set model. There are strong relations to other fields of Mathematics and Informatics which we shall not go into in this paper; a pictorial pointer is given in Figure 3.

4 ROUGHIAN – Rough information analysis

What we see in Table 2 is often called “raw data”. These are unfiltered measurements of attributes within the domain under investigation. However, it can be argued that there is no such thing as “observed raw data”: Most model building approaches use “features” or “attributes” as well as “measurements” to describe the data. Which attributes are chosen, and which measurements are used, are pragmatic decisions by

Figure 3: Outreach of RSDA



researchers, how they want to represent the dependencies of real life criteria in the best possible way – look again at Fig. 1. The rough set approach is therefore a conditional information analysis strategy, dependent on the choice of attributes and measurement models. It follows that – like other types of information analysis – RSDA needs a pre-processing step such as data filtering which results in data which is suitable for further analysis. This step should be part of the measurement procedure and it is highly desirable in certain situations – for example when a system has an empty core or when the obtained rules are based on a few observations. We shall address this problem in Section 4.2.

We can use the approximation quality defined in (3.7) as an internal index of a rule $Q \rightarrow P$. If $Q \Rightarrow P$, then the prediction is perfect, otherwise, $\gamma(Q \rightarrow P) < 1$. However, a perfect or high approximation quality is not a guarantee that the rule is valid. If, for example, the rough set method discovers a rule $Q \Rightarrow P$ which is based on only a few observations – which we call a *casual rule* – the approximation quality of the rule may be due to chance. Thus, the validity of inference rules for prediction must be validated by statistical techniques – otherwise, application beyond attribute reduction in the concrete situation might as well be done by throwing bones into the air and observing their pattern. We shall present such testing procedures in Section 4.3.

If we have competing rules, say, $Q \rightarrow P$, $R \rightarrow P$, \dots , we need a mechanism to decide which one of these is the best model of the situation and can be used most reliably for prediction. A conditional measure such as the approximation quality is not a good guide, and even the significance testing methods presented in Section 4.3 may be too expensive or not give us clear suggestions. In Section 4.4 we show how one can use the entropy function to describe the amount of uncertainty of deterministic and indeterministic rules.

It is our opinion that with these three additions to the original RSDA, the resulting method – which we call ROUGHIAN³ – can be turned into a useful tool for data analysis and prediction; it can be the first step of a data analysis and, if necessary, may be followed by other, more invasive, methods; applications of ROUGHIAN can be found in [3, 4, 16, 19].

³ROUGH INFORMATION ANALYSIS

4.1 Removing overhead

In this section we shall briefly describe the basic application of ‘pure’ rough set theory, namely the reduction of overhead in a given information system. Here, we are not concerned with learning or generating rules for prediction, but only with the concrete situation under consideration.

Given an information system $\mathcal{I} = \langle U, \Omega, V_q, f_q \rangle_{q \in \Omega}$, a natural question is whether there is a set Δ of attributes which does not change the information which \mathcal{I} provides with regard to the (in-) discernibility of objects. The first step we can take is to remove (or aggregate) attributes which lead to the same partition of U : Two attributes $p, q \in \Omega$ are called *equivalent*, if $\theta_p = \theta_q$. An information system \mathcal{I} over U is called *essential*, if there are no equivalent attributes. In other words,

$$\mathcal{I} \text{ is essential} \iff \mathbf{E} \text{ is one-one on singletons.}$$

If \mathcal{I} is not essential, we can always choose one element of the set $\{q \in \Omega : \mathbf{E}(\{p\}) = \mathbf{E}(\{q\})\}$ for each $p \in \Omega$.

The second step is to remove from \mathcal{I} those attributes which are exactly a combination of others: Since $S \stackrel{\text{def}}{=} \mathbf{E}(\mathcal{I})$ is a finite meet-semilattice, it is generated by the set $\text{lrr}(S)$ of its meet-irreducible elements. $\text{lrr}(S)$ is the smallest generating set for S , and hence it is contained in $\{\theta_q : q \in \Omega\}$. These observations lead to the following definitions:

Let $\Delta \stackrel{\text{def}}{=} \{q \in \Omega : \theta_q \in \text{lrr}(S)\}$. The information system $\mathcal{J} = \langle U, V_q, f_q \rangle_{q \in \Delta}$ is called the *attribute reduct* of \mathcal{I} , denoted by \mathcal{I}^{red} . Note that, once we have chosen an essential system \mathcal{I} from a given system, \mathcal{I}^{red} is uniquely determined, and that Δ is the smallest subset of Ω with respect to the property that $\mathbf{E}(\mathcal{I}) = \mathbf{E}(\mathcal{J})$.

As a method of attribute reduction in a given system, the rough set approach is a safe and useful, if computationally expensive, pre-processing tool which can feed its results into other techniques which are sensitive to redundant information such as statistical methods or neural networks. A good example for this can be found in [7].

4.2 Data filtering

One of the situations in which rough set theory offers no advice is when an information system has an empty core. This indicates a high substitution rate among the attributes which may be due to incomplete pre-processing or operationalisation of the ‘raw data’ which results in an information system where the granularity is still too high.

... nondeterminism is particularly strong if the core knowledge is void. Hence nondeterminism introduces synonymy to the knowledge, which in some cases may be a drawback [47, p. 38].

There is no specific advice within rough set analysis for this situation. In practice, procedures such as the statistical analysis of the appearance of attributes within reducts [e.g. 60] are used to cope with this problem. However, we have shown in [18] that this measure is not reliable.

In [17] we have developed a simple data filtering procedure which is compatible with the rough set approach and which may result in an improved significance of rules. Our approach can be viewed a special case of the more sophisticated methods which were independently developed in [56]. A generalisation of this ‘within attributes’ filtering to several attributes has been proposed in [61, 62].

The main tool are ‘binary information systems’, and we shall outline the procedure with an example from [17].

Consider the information system given in Table 3. We interpret this information system as follows:

Table 3: Heart attack information system \mathcal{I}_1

U	med	psych	H	U	med	psych	H
x_1	1	3	0	x_5	2	4	1
x_2	3	2	0	x_6	4	1	1
x_3	2	1	0	x_7	1	5	1
x_4	3	3	0	x_8	5	4	1

- x_1, \dots, x_8 are persons.
- The attribute *med* is a combined measure of medical indicators for the risk of a heart attack, while *psych* is a combined measure of psychological indicators [see e.g. 11, 54].
- The values of the risk measures are

1 – NO RISK, 2 – SMALL RISK, 3 – MEDIUM RISK, 4 – HIGH RISK, 5 – VERY HIGH RISK.

- The decision variable H is interpreted as the observation of a heart attack within a predefined time span, and we code

1 – HEART ATTACK, 0 – NO HEART ATTACK

One easily sees that $\theta_{\{med, psych\}} = id_U$, and therefore we have the rule

$$\{med, psych\} \Rightarrow H.$$

One the other hand, the statistical rough set analysis of [15] presented in the next section shows that there is no evidence that this dependency is not due to chance. This is surprising, because the dependency

“High medical or high psychological risk leads to heart attack”

is obviously present. However, this statement uses far less information than that given by \mathcal{I}_1 : There are only the two risk values {high, not high}. If we recode the risk values accordingly, we obtain the information system \mathcal{I}_2 :

U	med	psych	H	U	med	psych	H
x_1	0	0	0	x_5	0	1	1
x_2	0	0	0	x_6	1	0	1
x_3	0	0	0	x_7	0	1	1
x_4	0	0	0	x_8	1	1	1

We still have the dependency $\{med, psych\} \Rightarrow H$; the statistical analysis, however, shows that the chance to get the same result by random is about 2.7%. Hence, this dependency can be considered significant. As an example of our method, we outline the procedure how to get from \mathcal{I}_1 to \mathcal{I}_2 .

Let us first discuss binary information systems. These are those systems, in which the range of every information function has exactly two elements. Roughly speaking, we obtain a binary system \mathcal{I}^B from an information system \mathcal{I} by replacing a non-binary attribute q with a set of attributes, each corresponding to an element of V_q ; the associated information functions have value 1 if and only if x has this value under f_q ; binarisation of attributes is not new, see for example [30, 63]. Table 4 shows the binarisation of the example given in Table 3. We see that in the process of binarisation no information is lost; indeed, information is shifted from the columns to the rows.

The procedure now is as follows:

Table 4: The binarised system \mathcal{I}_1^B

U	med					psych					H
	m_1	m_2	m_3	m_4	m_5	p_1	p_2	p_3	p_4	p_5	
x_1	1	0	0	0	0	0	0	1	0	0	0
x_2	0	0	1	0	0	0	1	0	0	0	0
x_3	0	1	0	0	0	1	0	0	0	0	0
x_4	0	0	1	0	0	0	0	1	0	0	0
x_5	0	1	0	0	0	0	0	0	1	0	1
x_6	0	0	0	1	0	1	0	0	0	0	1
x_7	1	0	0	0	0	0	0	0	0	1	1
x_8	0	0	0	0	1	0	0	0	1	0	1

Step 1. Construct the binary extension \mathcal{I}_1^B as shown in Table 4.

Step 2. Find the binary attributes m_i, p_j for which

$$(\forall x \in U)(f_{m_i}(x) = 1 \rightarrow f_H(x) = 1),$$

$$(\forall x \in U)(f_{p_j}(x) = 1 \rightarrow f_H(x) = 1),$$

and build their union within *med*, resp. *psych* in the following sense: If, for example m_{i_0}, \dots, m_{i_k} satisfy **(Step 2)**, then we define a new binary attribute $m_{i_0 \dots i_k}$ by

$$f_{m_{i_0 \dots i_k}}(x) = 1 \stackrel{\text{def}}{\iff} f_{m_j}(x) = 1 \text{ for some } j \in \{i_0, \dots, i_k\},$$

$$\iff \max_{j \in \{i_0, \dots, i_k\}} f_{m_j}(x) = 1,$$

and simultaneously replace m_{i_0}, \dots, m_{i_k} by $m_{i_0 \dots i_k}$.

Because $\{m_4, m_5\}$ as well as $\{p_4, p_5\}$ show this property, we replace the two attributes m_4, m_5 (p_4, p_5) by an aggregate attribute m_{45} (p_{45}).

Similarly, we find the attributes m_i, p_j for which

$$(\forall x \in U)(f_{m_i}(x) = 1 \rightarrow f_H(x) = 0),$$

$$(\forall x \in U)(f_{p_j}(x) = 1 \rightarrow f_H(x) = 0)$$

and build their union within *med*, resp. *psych*. We see that only $\{p_2, p_3\}$ has this property, so that after this step we obtain

U	m_1	m_2	m_3	m_{45}	p_1	p_{23}	p_{45}	H
x_1	1	0	0	0	0	1	0	0
x_2	0	0	1	0	0	1	0	0
x_3	0	1	0	0	1	0	0	0
x_4	0	0	1	0	0	1	0	0
x_5	0	1	0	0	0	0	1	1
x_6	0	0	0	1	1	0	0	1
x_7	1	0	0	0	0	0	1	1
x_8	0	0	0	1	0	0	1	1

Step 3. Perform a rough set dependency analysis with the attributes of this system with respect to the decision attribute H . This results in

1	$\{m_1, m_2, m_3, p_{45}\}$
2	$\{m_2, p_1, p_{45}\}$
3	$\{m_2, p_{23}, p_{45}\}$
4	$\{m_2, p_1, p_{23}\}$
5	$\{m_{45}, p_1, p_{23}\}$
6	$\{m_{45}, p_{45}\}$

Step 4. Choose all reducts with the smallest cardinality⁴ for further processing. In the example there is only one collection of attributes that meets this condition, namely, the set $\{m_{45}, p_{45}\}$.

Because all other binary attributes are superfluous to express the dependency of H from *med* and *psych*, we finally obtain \mathcal{I}_2 on p. 12. \square

The details of the general procedure can be found in [17]. Furthermore, the paper shows that the detection of significant rules is as good or better than without filtering.

It is sometimes said that RSDA can only be applied for nominal data, and that it is unsuitable in case the data is continuous. Experience does not seem to verify this statement. [4] have applied the structural filtering described in this section to the (highly continuous) Iris data. In Table 5 we list the resulting number of classes, and in brackets the number of classes of the unfiltered data. Observe the dramatic fall in the number of classes of the petal attributes.

Table 5: Structural filtering of Iris data

Attribute	Filter	No of classes
Sepal length:	43–48, 53 → 46	22 (35)
	66,70 → 70	
	71–79 → 77	
Sepal width:	35, 37, 39–44 → 35	16 (23)
	20, 24 → 24	
Petal length:	10–19 → 14	8 (43)
	30–44,46,47 → 46	
	50, 52, 54–69 → 50	
Petal width:	1–6 → 2	8 (22)
	10–13 → 11	
	17, 20–25 → 17	

Table 6 shows that, for this data, the prediction quality of the ROUGHIAN method is approximately equal to that of discriminant analysis, even though

- ROUGHIAN does not use the metric information of the data set, except that rules “nearby” have to be evaluated,
- ROUGHIAN does not assume an underlying linear model within the data,
- ROUGHIAN does not make any homogeneity or spatial distributional assumption,

in contrast to discriminant analysis.

4.3 Significance testing

Although rough set theory uses a only few parameters which need simple *statistical estimation procedures*, its results should be controlled using *statistical testing procedures*, in particular, when they are used for modeling and prediction of events. If rules are based on only a few observations, their usefulness as a prediction tool may be rather limited:

Consider a dataset in which there is a nominal attribute that uniquely identifies each example ... Using this attribute one can build a 1 – rule that classifies a given training set 100% correctly: needless to say, the rule will not perform well on an independent test set [29].

⁴Section 4.4 presents a different method of choosing attribute sets.

Table 6: Prediction of Iris data

Discriminant analysis			
Predicted class	Classes in the testing set		
	Setosa	Versicolor	Virginica
Setosa	1.000	0.000	0.000
Versicolor	0.000	0.940	0.069
Virginica	0.000	0.060	0.931
ROUGHIAN			
Setosa	1.000	0.000	0.000
Versicolor	0.000	0.939	0.071
Virginica	0.000	0.061	0.929

In [15] we have developed two procedures, both based on randomization techniques, which evaluate the validity of prediction based on the approximation quality of attributes of rough set dependency analysis. These procedures seem to be particularly suitable to rough information analysis as a data mining tool which is data driven, and does not require outside information. In particular, it is not assumed that the information system under discussion is a representative sample, which, as we have mentioned earlier, is a problem for any method of data analysis (Table 1 on page 2). We invite the reader to consult [21] or [38] for the background and justification of randomization techniques in these situations.

Let σ be a permutation of U , and $P \subseteq \Omega$. We define new information functions $f_r^{\sigma(P)}$ by

$$f_r^{\sigma(P)}(x) \stackrel{\text{def}}{=} \begin{cases} f_r(\sigma(x)), & \text{if } r \in P, \\ f_r(x), & \text{otherwise.} \end{cases}$$

The resulting information system \mathcal{I}_σ permutes the P -columns according to σ , while leaving the Q -columns constant. We let $\gamma(Q \rightarrow \sigma(P))$ be the approximation quality of the prediction of $\sigma(P)$ by Q in \mathcal{I}_σ .

Given a rule $Q \rightarrow P$, we use the permutation distribution $\{\gamma(Q \rightarrow \sigma(P)) : \sigma \in \Sigma\}$ to evaluate the strength of the prediction $Q \rightarrow P$. The value $p(\gamma(Q \rightarrow P)|H_0)$ measures the extremeness of the observed approximation quality, and it is defined by

$$(4.1) \quad p(\gamma(Q \rightarrow P)|H_0) := \frac{|\{\gamma(Q \rightarrow \sigma(P)) \geq \gamma(Q \rightarrow P) : \sigma \in \Sigma\}|}{|U|!}$$

If $p(\gamma(Q \rightarrow P)|H_0)$ is low, traditionally below 5%, then the rule $Q \rightarrow P$ is deemed significant, and the (statistical) hypothesis “ $Q \rightarrow P$ is due to chance” can be rejected. Otherwise, if $p(\gamma(Q \rightarrow P)|H_0) \geq 0.05$, we call $Q \rightarrow P$ a *casual dependency*. A similar idea leads to the definition of *conditional casual attributes* within rough set dependency analysis.

Example 1. [15]

Consider the following information system:

U	p	q	d
1	0	0	0
2	0	1	1
3	1	0	2

The rule $\{p, q\} \rightarrow d$ is perfect, since $\gamma_{\{p,q\}}(d) = 1$. Furthermore, the rule is deterministic casual, because every instance is based on a single observation only.

Now suppose that we have collected three additional observations:

U	p	q	d	U	p	q	d
1	0	0	0	1'	0	0	0
2	0	1	1	2'	0	1	1
3	1	0	2	3'	1	0	2

To decide whether the given rule is casual under the statistical assumption, we have to consider all 720 possible rules as outlined in (4.1) and their approximation qualities. The distribution of the approximation qualities of the 720 possible matching rules is given in Table 7. Given the 6-observation example, the

Table 7: Results of randomization analysis; 6 observ.

$\gamma_{\mathcal{R}}$	No of cases	$p(\gamma(\{p, q\} \rightarrow d) H_0)$	Example of σ
1.00	48	0.067	1, 1', 2, 2', 3, 3'
0.33	288	0.467	1, 1', 2, 3, 2', 3'
0.00	384	1.000	1, 2, 2', 3, 1', 3'

probability of obtaining a perfect approximation of d by $\{p, q\}$ under the assumption of random matching, is 0.067 which is by far smaller than in the 3-observation example, but not convincing enough, using conventional $\alpha = 0.05$, to decide that the rule is sufficiently significant to be not casual. \square

We have applied the procedures to three well known data sets, and have found that not all claimed results can be called significant, and that other significant results were overlooked. Details and more examples can be found in [15].

Although the randomization technique is quite useful, it is rather expensive in resources, and it is not always feasible (or, indeed, possible) to exactly compute α . However, a randomly chosen set of permutations will usually be sufficient; Dwass [20] has shown that the significance level of a randomization test is in a sense exact even when the randomization distribution is only sampled. In [24] we present a sequential randomization test which considerably reduces the computational effort.

Randomization is only applicable as a conditional testing scheme: Though it tells us when a rule may be due to chance, it will not give us any information for the comparison of two different rules. How one can address this problem will be presented in the next section.

4.4 Uncertainty measurement

To compare different rules and/or compare different measures of uncertainty one needs a general framework in which to perform the comparisons.

The traditional measures of RSDA, the approximation quality, as well as lower and upper approximations are conditional constructions, but, from the start of its development, RSDA has been applied in unconditional situations. One problem is that differences of granularity within the set of independent variables is not taken into account by these constructions. Table 1 on page 2 is an illustration how quick granularity can grow, and one cannot rightfully ignore granularity differences. Therefore, approximation qualities cannot in general be compared, if we use different sets Q and R for the prediction of P .

To define an unconditional measure of prediction success one can use the idea of combining program complexity (i.e. to find a certain rule in ROUGHIAN) and statistical uncertainty (i.e. a measure of uncertainty within the indeterministic rules) to a global measure of prediction success. A lack in statistical reliability corresponds to predictions with more complicated rules – an approach that is well known as the *constructive probability approach* or *Kolmogorov complexity* [33, 36].

We have developed three methods related to the ‘Minimum Description Length’ approach [see 36, for a detailed exposition] which enhance RSDA so that it can be used more reliably for prediction and in the situation of competing explanations [18].

Our scenario is as follows:

- We have an information system $\mathcal{I} = \langle U, \Omega, V_q, f_q \rangle_{q \in \Omega}$, where $|U| = n$, and sets Q, P of attributes. The aim is to describe the uncertainty of the prediction of attributes in P by attributes in Q .

A special role will be played by those attributes in Q which induce a deterministic rule; we shall denote by V the union of all θ_Q classes which induce such a rule, i.e. which are a subset of a unique class of θ_P .

Whereas ROUGHIAN handles deterministic rules in a straightforward manner, the status of the indeterministic rules remains somewhat unclear. There are – at least – three different approaches to describe the amount of uncertainty of the indeterministic rules in ROUGHIAN.

The first approach is based on the assumption that, given a class Y of θ_P , any observation y in the set $\bar{Y}^Q \setminus \underline{Y}_Q \subseteq U$ is a result of a random process whose characteristics are totally unknown to the researcher. Given this assumption, no information within our data set will help us to identify the element y , and we conclude that each such y requires a rule – or class – of its own. In this case, any element of $U \setminus V$ is viewed as a realization of a probability distribution with its uncertainty $\frac{1}{n} \log_2(n)$. In other words, we assume the *maximum entropy principle* which minimizes the worst cases, and look at the equivalence relation θ_Q^+ defined by

$$x \equiv_{\theta_Q^+} y \stackrel{\text{def}}{\iff} x = y \text{ or there exists some } i \leq c \text{ such that } x, y \in X_i.$$

Its associated probability distribution is given by $\{\hat{\psi}_i : i \leq c + |U \setminus V|\}$ with

$$(4.2) \quad \hat{\psi}_i \stackrel{\text{def}}{=} \begin{cases} \hat{\pi}_i, & \text{if } i \leq c, \\ \frac{1}{n}, & \text{otherwise.} \end{cases}$$

We now define the *entropy of rough prediction* (with respect to $Q \rightarrow P$) as

$$(4.3) \quad H_{\text{rough}}(Q \rightarrow P) \stackrel{\text{def}}{=} H(\theta_Q^+) = \sum_i \hat{\psi}_i \cdot \log_2\left(\frac{1}{\hat{\psi}_i}\right).$$

This type of entropy seems to be closest to our basic aim to use as few assumptions outside the data as possible:

“Although there may be many measures μ that are consistent with what we know, the *principle of maximum entropy* suggests that we adopt that μ^* which has the largest entropy among all the possibilities. Using the appropriate definitions, it can be shown that there is a sense in which this μ^* incorporates the ‘least’ additional information”. [31].

Note that this approach assumes that classes X of θ_Q have to be observed within a representative sample, or – in terms of parameters – the approach needs only the probability distribution $\hat{\pi}$ of the classes within θ_Q .

To obtain an objective measurement we use the normalized rough entropy (NRE) of (4.3), where

$$(4.4) \quad S_{\text{rough}}(Q \rightarrow d) = 1 - \frac{H_{\text{rough}}(Q \rightarrow d) - H(d)}{\log_2(|U|) - H(d)}.$$

Table 8: Datasets and SORES validation

Dataset					SORES		C4.5(8)
Name	Cases	Classes	Attributes		No. of pred. attr.	Error	Error
			Cont.	Discr.			
Anneal	798	6	9	29	11	6.26	7.67
Auto	205	6	15	10	2	11.28	17.70
Breast-W	683	2	9	-	2	5.74	5.26
Colic	368	2	10	12	4	21.55	15.00
Credit-A	690	2	6	9	5	18.10	14.70
Credit-G	1000	2	7	13	6	32.92	28.40
Diabetes	768	2	8	-	3	31.86	25.40
Glass	214	6	9	-	3	21.79	32.50
Heart-C	303	2	8	15	2	22.51	23.00
Heart-H	294	2	8	15	5	19.43	21.50
Hepatitis	155	2	6	13	3	17.21	20.40
Iris	150	3	4	-	3	4.33	4.80
Sonar	208	2	60	-	3	25.94	25.60
Vehicle	846	4	18	-	2	35.84	27.10
Std. Deviation						10.33	8.77

If the NRE has a value near 1, the entropy is low, and the chosen attribute combination is favorable, whereas a value near 0 indicates casualness. The normalization does not use moving standards as long as we do not change the decision attribute d . Therefore, any comparison of NRE values between different predicting attribute sets makes sense, given a fixed decision attribute.

The implemented procedure searches for attribute sets with a high NRE; since finding the NRE of each feature set is computationally expensive, we use a genetic – like algorithm to determine sets with a high NRE.

We have named the method SORES, an acronym for Searching Optimal Rough Entropy Sets. SORES is implemented in our rough set engine GROBIAN [14]⁵.

In Table 8 we list the basic parameters of several known data sets⁶, and compare the SORES results with the C4.5 performance given in [53]. The column “No. of pred. attr.” records the number of attributes which are actually used for prediction.

The results indicate that SORES in its present version can be viewed as an effective machine learning procedure, because its performance compares well with that of the well established C4.5 method: The odds are 7:7 (given the 14 problems) that C4.5 produces better results. However, since the standard deviation of the error percentages of SORES is higher than that of C4.5, we conclude that C4.5 has a slightly better performance than the current SORES.

The second approach to handle uncertainty recognizes that θ_P induces some structure on $U \setminus V$: If X is a class of θ_Q which does not lead to a deterministic rule, there are classes Y_0, \dots, Y_s of θ_P , $s \geq 2$, such that X intersects each $Y_i \setminus V$. Uncertainty given X can now be measured by the uncertainty within $\{Y_0 \setminus V, \dots, Y_s \setminus V\}$. The assumption can be interpreted in the sense that any rule produces a certain degree of imprecision in the prediction of θ_P , but that the *amount* of uncertainty is based solely on the uncertainty within P and does not interact with Q .

⁵All material relating to SORES, e.g. datasets, a description of the algorithm, as well as GROBIAN, can be obtained from our website <http://www.psychology.uni-osnabrueck.de/sores/>

⁶<http://www.ics.uci.edu/~mllearn/MLRepository.html>

The final approach is based on the assumption that structure and amount of uncertainty can be estimated by the interaction of P and Q . In this case any class $X \in \theta_Q$ determines its own probability distribution based on its intersection with the classes of θ_P . This leads to the traditional notion of conditional entropy which is based solely on combinatorial statistics, and has little connection to the original intention of the rough set model.

Because ROUGHIAN offers several interpretations of uncertainty, we need an objective method to compare different unconditional measures of prediction success. Looking at our three different uncertainty interpretations, we see that they are based on different numbers of parameters. Whereas the first approach only needs the probabilities of the classes within the set θ_Q , the second approach additionally requires the probabilities of the classes of θ_P . The third approach needs even more parameters, since the probabilities of the classes within the set $\theta_Q \times \theta_P$ are required to compute the uncertainty. Only in very rare cases are these probabilities known in advance, and normally, we have to estimate the parameters. However, because estimated parameters and their true values might differ, we encounter another source of uncertainty which depends on the standard error of the estimated parameter. An objective method of comparing the different measures of prediction success will therefore have to take into account the different amounts of uncertainty in the estimation process.

5 Summary and outlook

We have described the tools of basic RSDA and have identified three areas where the traditional rough set method can be enhanced by methods well within the rough set philosophy:

- Data filtering,
- Significance testing of rough set rules,
- Describing an objective uncertainty measure of rough set rules.

For these areas we have proposed tools which supplement the original model, and which can turn it into a fully fledged instrument for information analysis and prediction, complementary to traditional statistical models as indicated in Table 9.

ROUGHIAN	Statistical models
Describing redundancy	Reducing uncertainty
Top down, reducing the full attribute set	Bottom up, introducing new variables
Reducts and Core	Influencing variables

Comparing both approaches, we observe that statistical rules are (usually) short, but far from being deterministic. A statistical model tries to find short rules searching within the set of predictor variables bottom up in order to find the most influential variables (attributes) which reduce the uncertainty as much as possible within the model. Only if we delete a certain amount of the observations we end up with a perfect prediction. Because this can only be done in a training set, statistical rules are interpreted with an amount of uncertainty, which can be estimated and minimised within the learning set. In contrast, the RSDA approach (usually) produces long deterministic rules which can be transported directly from the training situation to the application. The procedures arising in RSDA try to shorten the deterministic rules as good as possible by reducing redundancy among the attributes, which can be done by a top down method.

Obviously, the use of both methods depends on what the researcher had in mind. Nevertheless, the intention of the modeller is not the only criterion to choose RSDA or the statistical approach:

- If there are many observations but only a few attributes with low granularity, RSDA will not have a good chance to produce sound results, and a statistical procedure will be the best way to treat the data description problem.
- If there are few observations but many features, statistical modelling will face a problem, because too many different statistical rules can be used to describe the data. Reporting all the rules will be quite useless in most applications, and any selection of attributes would be subjective to some extent. RSDA tries to reduce the set of attributes to those features which contain the same information as the full attribute set (see Section 4.1). Since reducts are based on the conditional approximation quality, care has to be taken when comparing the prediction quality of different reducts. An alternative to reduct search is the search for attribute sets with optimal entropy, which constitutes an unconditional measure, thus allowing comparison of predictor sets.

The presentation of rough set data analysis and its enhancements given in this paper is mainly based on our own work. RSDA being a very active and fruitful research field has, of course, many other facets which we could not go into in the available space (short of writing a book). While we have tried to overcome the inherent limitations of RSDA by keeping the original model and adding procedures which are in the original spirit of RSDA, there are other directions, for example, making the crisp bounds of equivalence classes soft by looking at variable precision, or investigating other types of indiscernibility relations and neighbourhood systems, [see e.g. 50, 57, 66] and the references therein.

In most applications, the objects of an information system are themselves unstructured, while the attributes have, say, an ordinal domain. It would be advantageous to have a rough approach to the modelling of ordinal relations [13]. It is also of interest to develop procedures which take into account that the objects themselves have a structure, for example, if they are binary relations [9, 64].

Finally, since RSDA assumes implicitly that the collected information is without error, a non-invasive probabilistic error theory would further improve the quality of RSDA. We have started such investigations in [25].

References

- [1] Allison, L. (1994). Minimum message length encoding (MML). WWW page, <http://www.cs.monash.edu.au/~lloyd/tildeMML/index.html>.
- [2] Bacchus, F., Grove, A. J., Halpern, J. Y., and Koller, D. (1994). From statistical knowledge bases to degrees of belief. Technical report 9855, IBM.
- [3] Browne, C. (1997). Enhanced rough set data analysis of the Pima Indian diabetes data. In *Proc. 8th Ireland Conference on Artificial Intelligence, Derry (1997)*, pages 32–39.
- [4] Browne, C., Düntsch, I., and Gediga, G. (1998). IRIS revisited: A comparison of discriminant and enhanced rough set data analysis. In [51], pages 345–368.
- [5] Chatfield, C. (1995). Model uncertainty, data mining, and statistical inference. *J. Roy. Statist. Soc. Ser. A*, 158:419–466.
- [6] Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45:1304–1312.
- [7] Czyżewski, A. and Kaczmarek, A. (1994). Speech recognition system based on rough sets and neural networks. ICS Research Report 30, Warsaw University of Technology.
- [8] Dubois, D. and Prade, H. (1992). Putting rough sets and fuzzy sets together. In [58], pages 203–232.
- [9] Düntsch, I. (1994). Rough relation algebras. *Fundamenta Informaticae*, 21:321–331.
- [10] Düntsch, I. (1997). A logic for rough sets. *Theoretical Computer Science*, 179(1-2):427–436.

- [11] Düntsch, I. and Gediga, G. (1995). Rough set dependency analysis in evaluation studies: An application in the study of repeated heart attacks. *Informatics Research Reports*, 10:25–30.
- [12] Düntsch, I. and Gediga, G. (1997a). Algebraic aspects of attribute dependencies in information systems. *Fundamenta Informaticae*, 29:119–133.
- [13] Düntsch, I. and Gediga, G. (1997b). Relation restricted prediction analysis. In [59], pages 619–624.
- [14] Düntsch, I. and Gediga, G. (1997c). The rough set engine GROBIAN. In [59], pages 613–618.
- [15] Düntsch, I. and Gediga, G. (1997d). Statistical evaluation of rough set dependency analysis. *International Journal of Human–Computer Studies*, 46:589–604.
- [16] Düntsch, I. and Gediga, G. (1998a). Feature selection and data prediction by rough entropy. In *Fifth European Congress on Intelligent Techniques and Soft Computing, (EUFIT'98)*.
- [17] Düntsch, I. and Gediga, G. (1998b). Simple data filtering in rough set systems. *International Journal of Approximate Reasoning*, 18(1–2):93–106.
- [18] Düntsch, I. and Gediga, G. (1998c). Uncertainty measures of rough set prediction. *Artificial Intelligence*, 106(1):77–107.
- [19] Düntsch, I., Gediga, G., and Rogner, J. (2000). Archetypal psychiatric patients: An application of rough information analysis. Submitted for publication, Fachbereich Psychologie, Universität Osnabrück.
- [20] Dwass, M. (1957). Modified randomization tests for non-parametric hypothesis. *Annals of Mathematical Statistics*, 28:181–187.
- [21] Edgington, E. S. (1987). *Randomization Tests*, volume 31 of *Statistics: Textbooks and Monographs*. Marcel Dekker, New York and Basel.
- [22] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *Artificial Intelligence Magazine*, 17:37–54.
- [23] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7:179–188.
- [24] Gediga, G. and Düntsch, I. (2000a). A fast randomisation test for rule significance. Submitted for publication.
- [25] Gediga, G. and Düntsch, I. (2000b). Statistical techniques for rough set data analysis. In [52]. To appear.
- [26] Gigerenzer, G. (1981). *Messung und Modellbildung in der Psychologie*. Birkhäuser, Basel.
- [27] Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. (1997). Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery*, 1:11–28.
- [28] Hand, D. J. (1994). Deconstructing statistical questions. *J. Roy. Statist. Soc. Ser. A*, 157:317–356.
- [29] Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63–91.
- [30] Iwinski, T. B. (1988). Contraction of attributes. *Bull. Polish Acad. Sci. Math.*, 36:623–632.
- [31] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106:620–630.
- [32] Jaynes, E. T. (1996). Probability Theory: The Logic of Science. Fragmentary edition of March 1996, <http://www.math.albany.edu:8008/JaynesBook.html>.

- [33] Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Prob. Inf. Trans.*, 1:1–11.
- [34] Konrad, E., Orłowska, E., and Pawlak, Z. (1981a). Knowledge representation systems – Definability of informations. ICS Research Report 433, Polish Academy of Sciences.
- [35] Konrad, E., Orłowska, E., and Pawlak, Z. (1981b). On approximate concept learning. Technical Report 81–7, Technische Universität Berlin.
- [36] Li, M. and Vitányi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications*. Graduate Texts in Computer Science. Springer-Verlag, New York, 2 edition.
- [37] Lin, T. Y. and Cercone, N., editors (1997). *Rough sets and data mining*, Dordrecht. Kluwer.
- [38] Manly, B. F. J. (1997). *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. Chapman and Hall, London.
- [39] McEliece, R. J. (1977). *The Theory of Information and Coding*, volume 3 of *Encyclopedia of Mathematics and its Applications*. Addison–Wesley, Reading.
- [40] Novotný, M. (1997). Dependence spaces of information systems. In [41], pages 193–246.
- [41] Orłowska, E., editor (1997). *Incomplete Information – Rough Set Analysis*. Physica – Verlag, Heidelberg.
- [42] Pagliani, P. (1997). Rough sets theory and logic-algebraic structures. In [41], pages 109–190.
- [43] Pal, S. and Skowron, A., editors (1999). *Rough Fuzzy Hybridization*. Springer–Verlag.
- [44] Parsons, S. (1996). Current approaches to handling imperfect information in data and knowledge bases. *IEEE Transactions on Knowledge and Data Engineering*, 8(3):353–372.
- [45] Pawlak, Z. (1973). Mathematical foundations of information retrieval. ICS Research Report 101, Polish Academy of Sciences.
- [46] Pawlak, Z. (1982). Rough sets. *Internat. J. Comput. Inform. Sci.*, 11:341–356.
- [47] Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data*, volume 9 of *System Theory, Knowledge Engineering and Problem Solving*. Kluwer, Dordrecht.
- [48] Pawlak, Z., Grzymała-Busse, J. W., Słowiński, R., and Ziarko, W. (1995). Rough sets. *Comm. ACM*, 38:89–95.
- [49] Pawlak, Z. and Słowiński, R. (1993). Rough set approach to multi–attribute decision analysis. ICS Research Report 36, Warsaw University of Technology.
- [50] Polkowski, L. and Skowron, A., editors (1998a). *Rough sets in knowledge discovery, Vol. 1*. Physica–Verlag, Heidelberg.
- [51] Polkowski, L. and Skowron, A., editors (1998b). *Rough sets in knowledge discovery, Vol. 2*. Physica–Verlag, Heidelberg.
- [52] Polkowski, L., Tsumoto, S., and Lin, T. Y., editors (2000). *Rough Set Theory and Applications: New Developments*. Physica Verlag, Heidelberg. To appear.
- [53] Quinlan, R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77–90.

- [54] Rogner, J., Bartram, M., Hardinghaus, W., Lehr, D., and Wirth, A. (1994). Depressiv getönte Krankheitsbewältigung bei Herzinfarktpatienten – Zusammenhänge mit dem längerfristigen Krankheitsverlauf und Veränderbarkeit durch eine Gruppentherapie auf indirekt–suggestiver Grundlage. In Schüßler, G. and Leibing, E., editors, *Coping. Verlaufs– und Therapiestudien chronischer Krankheit*, pages 95–109. Hogrefe, Göttingen.
- [55] Schafer, J. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall.
- [56] Skowron, A. and Polkowski, L. (1996). Analytic morphology: Mathematical morphology of decision tables. *Fundamenta Informaticae*, 27(7):255–271.
- [57] Skowron, A. and Stepaniuk, J. (1997). Constructive information granules. In [59], pages 625–630.
- [58] Słowiński, R., editor (1992). *Intelligent decision support: Handbook of applications and advances of rough set theory*, volume 11 of *System Theory, Knowledge Engineering and Problem Solving*. Kluwer, Dordrecht.
- [59] Sydow, A., editor (1997). *Proc. 15th IMACS World Congress*, volume 4, Berlin. Wissenschaft und Technik Verlag.
- [60] Teghem, J. and Charlet, J.-M. (1992). Use of “rough sets” method to draw premonitory factors for earthquakes by emphasizing gas geochemistry: the case of a low seismic activity context in Belgium. In [58], pages 165–179.
- [61] Wang, H., Düntsch, I., and Bell, D. (1998). Data reduction based on hyper relations. In Agrawal, R., Stolorz, P., and Piatetsky-Shapiro, G., editors, *Proceedings of KDD’98*, pages 349–353, New York.
- [62] Wang, H., Düntsch, I., and Gediga, G. (2000). Classificatory filtering in decision systems. *International Journal of Approximate Reasoning*, pages 111–136.
- [63] Wille, R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In Rival, I., editor, *Ordered sets*, volume 83 of *NATO Advanced Studies Institute*, pages 445–470. Reidel, Dordrecht.
- [64] Yao, Y. and Wang, T. (1999). On rough relations: An alternative formulation. In [67], pages 82–90.
- [65] Zadeh, L. A. (1994). What is BISC? <http://http.cs.berkeley.edu/projects/Bisc/bisc.memo.html>, University of California.
- [66] Ziarko, W. (1993). Variable precision rough set model. *Journal of Computer and System Sciences*, 46.
- [67] Zong, N., Skowron, A., and Ohsuga, S., editors (1999). *New directions in rough sets, data mining, and granular soft computing*, volume 1711 of *Lecture Notes in Artificial Intelligence*, Berlin. Springer–Verlag.