Uncertainty measures of rough set prediction

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
ggediga@Luce.Psycho.Uni-Osnabrueck.DE

and
Institut für semantische Informationsverarbeitung
Universität Osnabrück

Abstract

The main statistics used in rough set data analysis, the approximation quality, is of limited value when there is a choice of competing models for predicting a decision variable. In keeping within the rough set philosophy of non–invasive data analysis, we present three model selection criteria, using information theoretic entropy in the spirit of the minimum description length principle. Our main procedure is based on the principle of indifference combined with the maximum entropy principle, thus keeping external model assumptions to a minimum. The applicability of the proposed method is demonstrated by a comparison of its error rates with results of C4.5, using 14 published data sets.

Key words: Rough set model, minimum description length principle, attribute prediction

1 Introduction

Most of the commonly used procedures for data prediction 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 data analysis* (RSDA); it has been developed by Z. Pawlak and his co–workers since the early 1970s (Pawlak, 1973, Konrad, Orłowska & Pawlak, 1981a,b, Pawlak, 1982), and has recently received wider attention as a means of data analysis (Pawlak, Grzymała-Busse, Słowiński & Ziarko, 1995). The rationale of the rough set model is the observation that

*Equal authorship implied

> "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" (Pawlak & Słowiński, 1993).

In other words, the original concept behind the model is the realization that sets can only be described "roughly": 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 semantic of a rough set logic corresponds to a fuzzy logic with a three - valued membership function (see Düntsch, 1997, Pagliani, 1997).

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 utilizes solely the granularity structure of the given data, expressed as classes of suitable equivalence relations. Of course, this does not mean that RSDA does not have any model assumptions; for example, we indicate below that the statistical model behind RSDA is the *principle of indifference*. However, model assumptions are such that we admit complete ignorance of what happens within the region of indiscernibility, given by the granularity of information (see Section 2.1).

The results of RSDA must be seen with this background in mind: The rough set model tries to extract as much information as possible from the structural aspects of the data, neglecting, in its pure form, numerical and other contextual information of the attribute domains. This keeps model assumptions to a minimum, and can serve as a valuable indicator of the direction into which possible further analysis can go.

The relationship between RSDA and statistical modeling is quite complementary (see Table 1), and we have discussed it in more detail in Düntsch & Gediga (1997b).

Table 1: RSDA vs statistical modeling

RSDA	Statistical models
Many features/attributes, few data points	Few variables, many data points
Describing redundancy	Reducing uncertainty
Top down, reducing the full attribute set	Bottom up, introducing new variables

Knowledge representation in the rough set model is done via *information systems* which are a tabular form of an OBJECT → ATTRIBUTE VALUE relationship, similar to relational databases (see Section 2.2).

If Q is a set of predictor features and d a decision attribute, then RSDA generates rules of the form

$$\bigwedge_{q \in Q} x^q = m_q \Rightarrow x^d = m_d^0 \vee x^d = m_d^1 \vee \ldots \vee x^d = m_d^k, \tag{1.1}$$

where x^r is the attribute value of object x with respect to attribute r.

We see that in the rough set model rules can be indeterministic in the sense that on the right hand side of (1.1) we can have a proper disjunction. If there is only on term on the right hand side, we call the rule *deterministic*. Whereas RSDA handles deterministic rules in a straightforward manner, the status of the indeterministic rules remains unclear.

If rules are based on a few observations only, the granularity of the system is too high, and the rule may be due to chance. In order to test the significance of rules, one can use randomization methods to compute the conditional probability of the rule, assuming that the null hypothesis

"Objects are randomly assigned to decision classes"

is true. In Düntsch & Gediga (1997c) we have developed two simple procedures, both based on randomization techniques, which evaluate the validity of prediction based on the principle of indifference, which is the underlying statistics of RSDA; this technique is briefly described in Section 2.4.

Although randomization methods are quite useful, they are rather expensive in resources, and are only applicable as a conditional testing scheme:

- Though they tell us when a rule may be due to chance, they do not provide us with a metric for the comparison of two different rules $Q \to d,\ R \to d$, let alone for different models of uncertainty.

Thus, we need a different criterion for model selection: The *minimum description length principle* (MDLP) (see Rissanen, 1978, 1985) states that the best theory to explain a given phenomenon d is one which minimizes the sum of

- The binary length of encoding a hypothesis Q and
- The binary length of encoding the decision data d using the hypothesis Q as a predictor.

In the sequel, we present three different ways of model selection within RSDA, based on three different probability distributions in the spirit of the MDLP. Within each model frame M, the attractiveness of this approach is that information about the uncertainty of rules such as (1.1) is considered in a context where the selection criterion $H^M(Q \to d)$ is the aggregate of the

- Effort of coding a hypothesis Q, expressed by an entropy function $H(Q)$, and
- Uncertainty of "guessing" in terms of the optimal number of decisions to classify a randomly chosen observation given this hypothesis, expressed as a suitable entropy $H^M(d|Q)$.

The paper is organized as follows: In Section 2 we describe the basic tools of RSDA and their main properties, as well as our usage of the entropy functions. Section 3 contains our three approaches to uncertainty, and Section 4 applies our main approach to some well known data sets. Finally, Section 5 consists of a summary and an outlook.

2 Basic tools and constructions

2.1 Approximation spaces

An equivalence θ on a set U is a transitive, reflexive, and symmetric binary relation, and we call the pair $\langle U, \theta \rangle$ an *approximation space*. In our context, we shall sometimes call an equivalence relation an *indiscernibility relation*. Approximation spaces are the core mathematical concept of RSDA, and their usage reflects the idea that granulation of information can be described by classes of an indiscernibility relation.

Recall that a partition $\mathcal{P}$ of a set U is a family of nonempty, pairwise disjoint subsets of U whose union is U. With each equivalence relation θ we associate a partition $\mathcal{P}_\theta$ of U by specifying that $a, b \in U$ are in the same class of $\mathcal{P}_\theta$, if and only if $a \theta b$. The classes of $\mathcal{P}_\theta$ have the form

$$\theta a = \{b \in U : a \theta b\}.$$

By some abuse of language, we also speak of the classes of an equivalence relation when we mean the classes of its associated partition, and call θa *the class of* a *modulo* θ.

The interpretation in rough set theory is that our knowledge of the objects in U extends only up to membership in the classes of θ, and our knowledge about 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

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

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

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

is the *upper approximation* or *possible region* of X.

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

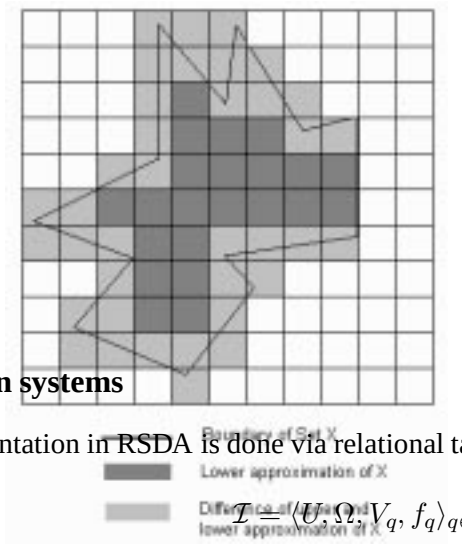
The *area of uncertainty* extends over

$$\overline{X} \setminus \underline{X},$$

and the *area of certainty* is

$$\underline{X} \cup -\overline{X}.$$

Figure 1: Rough approximation



2.2 Information systems

Knowledge representation in RSDA is done via relational tables. 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,
3. For each $q \in \Omega$
   - A set V_q of attribute values,
   - An information function $f_q : U \to V_q$.

In the sequel we shall use $\mathcal{I}$ as above as a generic information system with $|U| = n$ and $P, Q, R \subseteq \Omega$. We also will sometimes write x^q instead of $f_q(x)$ to denote the value of x with respect to attribute q. Furthermore, we suppose that $d \in \Omega$ is a decision attribute which we want to predict with attribute sets $Q, R \subseteq \Omega$.

Example 1. We use the small information system given in Table 2 on the next page as a running example to illustrate the various concepts developed in the sequel. An attribute "Heart Disease" (HD) shall be predicted from two variables "Smoker" (S) and "Body Mass Index" (BMI). □

With each Q we associate an equivalence relation θ_Q on U by defining

$$x \equiv_{\theta_Q} y \overset{\text{def}}{\iff} (\forall q \in Q) f_q(x) = f_q(y),$$

Table 2: An example of an information system

No	S	BMI	HD
1	no	normal	no
2	no	obese	no
3	no	normal	no
4	no	obese	no
5	yes	normal	yes
6	yes	normal	yes
7	yes	obese	no
8	yes	obese	yes
9	no	normal	no

and the partition induced by θ_Q is denoted by $\mathcal{P}(\theta_Q)$ or simply by $\mathcal{P}(Q)$.

The interpretation of θ_Q is the following: If our view of the world U is limited to the attributes given by Q, then we will not be able to distinguish objects within the equivalence classes of θ_Q.

Example 1. cont.
The classes of θ_Q and θ_d are as follows:

Q	Classes of θ
$\{S\}$	$\{1,2,3,4,9\}$, $\{5,6,7,8\}$
$\{BMI\}$	$\{1,3,5,6,9\}$, $\{2,4,7,8\}$
$\{S,BMI\}$	$\{1,3,9\}$,$\{2,4\}$, $\{5,6\}$,$\{7,8\}$
d	
$\{HD\}$	$\{1,2,3,4,7,9\}$, $\{5,6,8\}$.

□

We can now use the definition of upper, resp. lower approximation of sets via θ_Q defined in the previous section. It is not hard to see that for $Y \subseteq U$,

$$\overline{Y}^Q = \{x \in U : \theta_Q x \cap Y \neq \emptyset\} \tag{2.1}$$

is the upper approximation of Y with respect to Q, and

$$\underline{Y}_Q = \{x \in U : \theta_Q x \subseteq Y\} \tag{2.2}$$

is the lower approximation of Y with respect to Q. If Q is understood, we just write $\overline{Y}$ or $\underline{Y}$.

The equivalence relations θ_Q are used to obtain rules in the following way:

Let $Q \to d \subseteq \mathcal{P}(Q) \times \mathcal{P}(d)$ be the relation

$$\langle X, Y \rangle \in Q \to d \stackrel{\text{def}}{\Longleftrightarrow} X \subseteq \overline{Y}^{Q}.$$

Observe that by (2.1),

$$X \subseteq \overline{Y}^{Q} \text{ if and only if } X \cap \overline{Y}^{Q} \neq \emptyset \text{ if and only if } X \cap Y \neq \emptyset,$$

and thus,

$$\langle X, Y \rangle \in Q \to d \Longleftrightarrow X \cap Y \neq \emptyset. \tag{2.3}$$

Observe that we can determine with the knowledge gained from Q whether $X \cap Y = \emptyset$ and also – by (2.2) – whether $X \subseteq Y$

A pair $\langle X, Y \rangle \in Q \to d$ is called a *Q,d – rule* (or just a rule, if Q and d are understood), usually written it as $X \to Y$. By some abuse of language we shall also call $Q \to d$ a rule when there is no danger of confusion, and normally identify singleton sets with the element they contain.

If $\langle X, Y \rangle \in Q \to d$, then X corresponds to the left hand side of the implication (1.1), and Y corresponds to (one of) the disjuncts of the right hand side.

Example 1. cont.
The rule $S \to HD$ consists of the pairs

$$\begin{aligned}
&\langle \{1,2,3,4,9\}, \{1,2,3,4,7,9\} \rangle \\
&\quad \langle \{5,6,7,8\}, \{1,2,3,4,7,9\} \rangle \\
&\quad \langle \{5,6,7,8\}, \{5,6,8\} \rangle,
\end{aligned}$$

$BMI \to HD$ has the pairs

$$\begin{aligned}
&\langle \{1,3,5,6,9\}, \{1,2,3,4,7,9\} \rangle \\
&\langle \{1,3,5,6,9\}, \{5,6,8\} \rangle \\
&\quad \langle \{2,4,7,8\}, \{1,2,3,4,7,9\} \rangle \\
&\quad \langle \{2,4,7,8\}, \{5,6,8\} \rangle,
\end{aligned}$$

and for $\{S, BMI\} \to HD$ we obtain

$$\begin{aligned}
&\langle \{1,3,9\}, \{1,2,3,4,7,9\} \rangle \\
&\quad \langle \{2,4\}, \{1,2,3,4,7,9\} \rangle \\
&\quad \langle \{5,6\}, \{5,6,8\} \rangle \\
&\quad \langle \{7,8\}, \{1,2,3,4,7,9\} \rangle \\
&\quad \langle \{7,8\}, \{5,6,8\} \rangle.
\end{aligned}$$

□

The *deterministic* – or *functional* – part of $Q \to d$, written as $Q \stackrel{det}{\to} d$, is the set

$$\{\langle X, Y\rangle \in Q \to d : X \subseteq Y\}.$$

If $\langle X, Y\rangle \in Q \stackrel{det}{\to} d$, then the class X is called d – *deterministic* or just *deterministic*, if d is understood. In this case, the values of each $x \in U$ on the attributes in Q uniquely determine the values of x with respect to the attribute values of d.

Example 1. cont.
The deterministic classes of $\{S, BMI\} \to HD$ are $\{1, 3, 9\}, \{2, 4\}, \{5, 6\}$; the only deterministic class of $\{S\} \to HD$ is $\{1, 2, 3, 4, 9\}$, and there is no deterministic class of $\{BMI\} \to HD$. □

If $Q \to d = Q \stackrel{det}{\to} d$, i.e. if $Q \to d$ is a function, then we call $Q \to d$ *deterministic* and write $Q \Rightarrow d$; in this case, we say that d *is dependent on* Q. It is not hard to see that

$$Q \Rightarrow d \text{ if and only if } \theta_Q \subseteq \theta_d,$$

so that our terminology is in line with the usual convention in RSDA.

A special role will be played by the deterministic part of $Q \to d$, and we define

$$V_{Q\to d} \stackrel{\text{def}}{=} \bigcup \{X \in \mathcal{P}(Q) : \langle X, Y\rangle \in Q \stackrel{det}{\to} d\}$$

In other words, $V_{Q\to d}$ is the union of all d – deterministic θ_Q classes. If $Q \to d$ is understood, we shall just write V instead of $V_{Q\to d}$. Note that

$$n - |V| = 0 \text{ or } n - |V| \geq 2, \tag{2.4}$$

since every singleton class of θ_Q is deterministic for any d. A class Y of θ_d is called Q – *definable* (or just *definable*, if Q is understood), if $Y \subseteq V$.

Example 1. cont.
The deterministic parts are easily seen to be

$$V_{S\to HD} = \{1, 2, 3, 4, 9\},\ V_{BMI\to HD} = \emptyset,\ V_{\{S,BMI\}\to HD} = \{1, 2, 3, 4, 5, 6, 9\}.\ \square$$

Even though RSDA is a symbolic method, it implicitly makes statistical assumptions which we briefly want to describe, and we start by looking at a single equivalence relation θ on U. The inherent metric of an approximation system $\langle U, \theta\rangle$ is the *approximation quality*

$$\gamma_\theta(X) \stackrel{\text{def}}{=} \frac{|\underline{X}_\theta| + |\underline{-X}_\theta|}{|U|}, \tag{2.5}$$

(Pawlak, 1991, p. 16ff). If θ is understood, we shall usually omit the subscripts.

The value $\gamma(X)$ is the relative frequency of objects of U which can be correctly classified with the knowledge given by θ as being in X or not. The function γ can be generalized for information systems (Pawlak, 1991, p. 22); we choose a different (but equivalent) definition which is more suited

for our purpose. As a measure of the approximation quality of Q with respect to d, we define an *approximation function* by

$$\gamma(Q \to d) = \frac{|\bigcup\{X \in \mathcal{P}(Q) : X \text{ is } d\text{–deterministic}\}|}{|U|}. \tag{2.6}$$

Note that

$$\gamma(Q \to d) = \frac{|V|}{|U|},$$

and

$$Q \Rightarrow d \text{ if and only if } \gamma(Q \to d) = 1.$$

Example 1. cont.
We see that

$$\gamma_{S \to HD} = \frac{5}{9},\ \gamma_{BMI \to HD} = \frac{0}{9},\ \gamma_{\{S,BMI\} \to HD} = \frac{7}{9}. \ \square$$

It is not hard to see that the statistical principle underlying the approximation functions is the *principle of indifference*:

- If one does not have any information about the occurrence of basic events, they are all assumed to be equally likely.

Q is called a *reduct of* d, if it is minimal with respect to the property that $\gamma(Q \to d) = 1$. Reducts are of particular importance in rough set theory as a means of feature reduction.

2.3 Data filtering and discretization

Even though RSDA has no inherent categorization mechanism, it is possible to handle continuous data satisfactorily in several ways. One method which keeps close to the RSDA philosophy of keeping outside assumptions to a minimum is the filtering procedure described in Düntsch & Gediga (1998) which is based only on the information provided by the indiscernibility relations. This technique collects values of a feature into a single value by taking a union of deterministic equivalence classes which are totally contained in a class of the decision attribute; in this way, the underlying statistical basis of the rule may be enlarged, and the significance of the rule is increased (see Section 2.4).

For example, if we have an attribute q and a rule

$$\text{If } q = 2 \text{ or } q = 3 \text{ or } q = 5 \text{ then } d = \text{blue},$$

then we can collect 2,3,5 into a single attribute value of q.

The important feature of this procedure is that the internal dependency structure of the system is kept intact, and that we do not need additional parameters. In other words, this step can be regarded as a

part of the operationalization procedure; it can be implemented as a cheap standard algorithm if the decision attribute is fixed, for example, in our rough set engine GROBIAN (Düntsch & Gediga, 1997a).

Even though the method is simple, it sometimes works surprisingly well as the investigations of Browne, Düntsch & Gediga (1998) and Browne (1997) indicate. Nevertheless, this discretization scheme cannot cope effectively with complex interactions among continuous variable as other, more sophisticated, discretization methods do. For these methods applicable in RSDA (which, however, use external parameters and restrictive modelling assumptions) we invite the reader to consult Bazan (1997) or Nguyen & Nguyen (1998) and the references therein.

The claim that RSDA is not applicable to most real life problems, because it cannot handle continuous variables seems to us to be an open problem, but not a fact. The success of applications of fuzzy controlling, which also requires discretization of continuous data, shows that the distinction of "continuous data" vs. "discrete data" does not necessarily imply that there is a need for different "continuous methods" , respectively, "discrete methods", to handle these different types of data. We also refer the reader to Section 4 below, in which the prediction quality of our RSDA based methods is explored also for data sets which consists of continuous variables.

2.4 Significance testing

Suppose that we want to test the statistical significance of the rule $Q \to d$. Let Σ be the set of all permutations of U. For each $\sigma \in \Sigma$, we define a new set of feature vectors $\overline{x}_\sigma^\Omega$ by

$$x_\sigma^r \stackrel{\text{def}}{=} \begin{cases} \sigma(x)^d, & \text{if } r = d, \\ x^r, & \text{otherwise.} \end{cases} \tag{2.7}$$

In this way, we permute the x^d values according to σ, while leaving everything else constant. The resulting rule system is denoted by $Q \to \sigma(d)$. We now use the permutation distribution $\{\gamma(Q \to \sigma(d)) : \sigma \in \Sigma\}$ to evaluate the strength of the prediction $Q \to d$. The value $p(\gamma(Q \to d)|H_0)$ measures the extremeness of the observed approximation quality and it is defined by

$$p(\gamma(Q \to d)|H_0) := \frac{|\{\sigma \in \Sigma : \gamma(Q \to \sigma(d)) \geq \gamma(Q \to d)\}|}{|U|!} \tag{2.8}$$

If $\alpha = p(\gamma(Q \to d)|H_0)$ is low, traditionally below 5%, we reject the null hypothesis, and call the rule *significant*, otherwise, we call it *casual*. Failure to reject the null hypothesis does not mean that it is true, and thus, such randomization tests are a necessary condition for significance (for a discussion, see Cohen, 1990).

Randomization is a statistical technique which does not require a representative sampling from a population which is a theoretical generalization of the sample under study, because the randomization procedure uses only information within the given sample, well in accord with our stated objective. This aspect is in contrast to most other statistical techniques. Even the bootstrap technique needs some parametric assumptions, because one has to suppose that the percentages of the observed equivalence classes are suitable estimators of the latent probabilities of the equivalence classes in the population.

Example 1. cont.
Table 3 tells us the approximation qualities and the significance of the sets $\{S\}$, $\{BMI\}$, and $\{S, BMI\}$ for the prediction of HD for the example information system of Table 2.

Table 3: Approximation quality and significance of predicting attributes

Attribute Set	γ	Significance	Interpretation
$\{S\}$	0.556	0.047	not casual ($\alpha = 5\%$)
$\{BMI\}$	0.000	1.000	casual ($\alpha = 5\%$)
$\{S, BMI\}$	0.778	0.144	casual ($\alpha = 5\%$)

The best approximation quality is attained by the combination of both predicting attributes S and BMI. However, in terms of statistical significance the set $\{S, BMI\}$ is not a significant predictor for the outcome of HD, because there is no evidence that the prediction success is not due to chance. Therefore, although the approximation quality of $\{S\}$ is smaller than that of $\{S, HD\}$, the set $\{S\}$ should be preferred to predict HD, because it is unlikely that the prediction success is due to chance. □

In most applications one can observe that there are several reducts or attribute sets with an acceptable approximation quality. Significance testing gives some information about their statistical validity, but there are often several sets with comparable good statistical quality. Thus, we need an additional criterion for model selection, the foundations of which will be laid in the next section.

2.5 Partitions and Information Measures

Let $\mathcal{P}$ be a partition of U with classes $X_i, i \leq k$, each having cardinality r_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 X_i is just $\frac{r_i}{n}$. We define the *entropy* of $\mathcal{P}$ by

$$H(\mathcal{P}) \stackrel{\text{def}}{=} \sum_{i=0}^{k} \frac{r_i}{n} \cdot \log_2(\frac{n}{r_i}). \tag{2.9}$$

If θ is an equivalence relation on U and $\mathcal{P}$ its induced partition, we will also write $H(\theta)$ instead of $H(\mathcal{P})$. Furthermore, if Q is a set of attributes, then we usually write $H(Q)$ instead of $H(\theta_Q)$.

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 ϖ, 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 guess 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 maximized.

For two partitions $\mathcal{P}_1, \mathcal{P}_2$ of U with associated equivalence relations θ_1, θ_2, we write $\mathcal{P}_1 \leq \mathcal{P}_2$, if $\theta_1 \subseteq \theta_2$. The following Lemma may be known:

Lemma 2.1. *If $\mathcal{P}_1 \leq \mathcal{P}_2$, then $H(\mathcal{P}_1) \geq H(\mathcal{P}_2)$.*

Proof. Since every class of $\mathcal{P}_2$ is a union of classes of $\mathcal{P}_1$, we can suppose without loss of generality that the probabilities associated with $\mathcal{P}_1$ are $p_1, \ldots, p_m, m \geq 3$, and those associated with $\mathcal{P}_2$ are $p_1 + p_2, p_3, \ldots, p_m$. Now,

$$\begin{aligned} H(\mathcal{P}_1) &= H(p_1, \ldots, p_m) \\ &= H(p_1 + p_2, p_3, \ldots, p_m) + (p_1 + p_2) \cdot H(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}) \\ &= H(\mathcal{P}_2) + (p_1 + p_2) \cdot H(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}) \\ &\geq H(\mathcal{P}_2), \end{aligned}$$

see for example Jumarie (1990), p.21. □

Corollary 2.2. *If $R \subseteq Q \subseteq \Omega$, then $H(R) \leq H(Q)$.*

More classes does not automatically mean higher entropy, and we need a hypothesis such as $\mathcal{P}_1 \leq \mathcal{P}_2$; for example,

$$1.585 \approx H(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) > H(\frac{2}{3}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}) \approx 1.447$$

For later use we mention that the entropy function has the property of strong additivity (see Jumarie, 1990, p 21):

Lemma 2.3. *Suppose that $\{\hat{\pi}_i : i \leq t\}$ and $\{\hat{\eta}_{i,j} : j \leq n_i\}$ are sets of positive parameters such that*

$$\sum_{i \leq t} \hat{\pi}_i = \sum_{j \leq n_i} \hat{\eta}_{i,j} = 1$$

Then,

$$\sum_{i \leq t} \sum_{j \leq n_i} \hat{\pi}_i \hat{\eta}_{i,j} \cdot \log_2(\frac{1}{\hat{\pi}_i \hat{\eta}_{i,j}}) = \sum_{i \leq t} \hat{\pi}_i \log_2(\frac{1}{\hat{\pi}_i}) + \sum_{i \leq t} \hat{\pi}_i \cdot \sum_{j \leq n_i} \hat{\eta}_{i,j} \cdot \log_2(\frac{1}{\hat{\eta}_{i,j}}).$$

3 Rough set prediction

The problem we want to address is a variant of the classical prediction problem:

- Given a decision attribute d, which is the "best" attribute set $Q \subseteq \Omega$ to predict the d – value of an object x, given the values of x under the features contained in Q?

We say "a variant", since the RSDA rules are determined by the equivalence classes of the partitions of U involved – see (1.1) and (2.3) –, and we are combining prediction quality with feature reduction.

The prediction problem raises two questions:

- Which subsets Q of Ω are candidates to be such a "best" attribute set"?
- What should a metric look like to determine and select the "best" attribute set?

In conventional RSDA, the approximation quality γ as defined in 2.6 on page 9 is a measure to describe the prediction success, which is conditional on the choice of attributes and measurement by the researcher. However, approximation qualities cannot be compared, if we use different feature sets Q and R for the prediction of d. To define an unconditional measure of prediction success, one can use the MDLP idea of combining

- Program complexity (i.e. to find a deterministic rule in RSDA) and
- Statistical uncertainty (i.e. a measure of uncertainty when applying an indeterministic rule)

to a global measure of prediction success. In this way, dependent and independent attributes are treated similarly.

In the sequel we discuss three different models M to handle this type of uncertainty, which are based on the information – theoretic entropy functions of Section 2.5. Our model selection criterion will be an entropy value $H^M(Q \to d)$ which aggregates for each set Q of attributes

- The complexity of coding the hypothesis Q, measured by the entropy $H(Q)$ of the partition of its associated equivalence relation θ_Q (see (2.9)), and
- The conditional coding complexity $H^M(d|Q)$ of d, given by the values of attributes in Q,

so that

$$H^M(Q \to d) = H(Q) + H^M(d|Q). \tag{3.1}$$

The estimator $H^M(d|Q)$ measures the uncertainty to predict membership in a class of θ_d given a class of θ_Q; it is important, if we want to gauge the success of a model conditioned to the knowledge given by Q.

The importance of $H^M(Q \to d)$ is due to the fact that it aggregates the uncertainty $H^M(d|Q)$ and the effort $H(Q)$ of coding the hypothesis, i.e. the predicting elements. This enables the researcher to compare different attribute sets Q_i in terms of a common unit of measurement, which cannot be done by a conditional measure of prediction success like γ or $H^M(d|Q)$.

Since all our entropies are defined from probability measures which arise from partitions of an n – element set, we see from the remarks after (2.9) that they have an upper bound of $\log_2(n)$.

In order to be able to compare different entropies within one model M, we define a *normalized entropy measure* – bounded within $[0, 1]$ – as follows: If $H(d) = \log_2(n)$, i.e. if θ_d is the identity, then

$$S^M(Q \to d) \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } \theta_Q = \theta_d, \\ 0, & \text{otherwise.} \end{cases}$$

If $H(d) \lneq \log_2(n)$,

$$S^M(Q \to d) \stackrel{\text{def}}{=} 1 - \frac{H^M(Q \to d) - H(d)}{\log_2(n) - H(d)}. \tag{3.2}$$

The measures $S^M(Q \to d)$ are constructed in such a way that they are comparable to the approximation quality:

- If $S^M(Q \to d) = 1$, the entropy measure is as good as possible, whereas $S^M(Q \to d)$ near 0 shows that the amount of coding information is near the theoretical maximum, which indicates a poor model for predicting the attribute d.

Similarly, based on the bounds $0 \leq H^M(d|Q) \leq \log_2(n)$, we can normalize $H^M(d|Q)$ by

$$S^M(d|Q) = 1 - \frac{H^M(d|Q)}{\log_2(n)}. \tag{3.3}$$

We shall show later that the measures $S^M(d|Q)$ are comparable – and in a special case even identical – to the approximation quality γ.

We assume that prediction requires the specification of a probability distribution; the three models presented below are distinguished by the choice of such distributions and their respective associated parameters.

Throughout, we suppose that the classes of θ_Q are $X_0, \ldots, X_t$ with $r_i \stackrel{\text{def}}{=} |X_i|$, and that the classes of θ_d are $Y_0, \ldots Y_s$. Furthermore, we let $c \leq t$ be such that

$$V = X_0 \cup \ldots \cup X_c,$$

i.e. the $X_i,\ i \leq c$, are exactly the deterministic classes of θ_Q. In accordance with our previous observations, we assume the principle of indifference, and set $\hat{\pi}_i \stackrel{\text{def}}{=} \frac{r_i}{n}$ for $i \leq t$. Also, we shall write γ instead of $\gamma(Q \to d)$, if Q and d are understood.

3.1 Prediction I: Knowing it all

The first approach is based on the assumption that structure and amount of uncertainty can be estimated by the interaction of d and Q. In this case, each class X of θ_Q determines probability distributions based on its intersection with the classes of θ_d. This assumes that we know

1. The classes of θ_d,
2. The classes of θ_Q, and
3. Their interaction, i.e. their intersections.

It follows that, in order to justify any prediction, we have to assume that the data set is a representative sample. This is a general problem of data mining, and we have discussed it within the rough set approach in Düntsch & Gediga (1997c).

Uncertainty in the sense of this model is not predominantly a feature of the predictor set Q (as intended by RSDA) but a local feature of the intersection of equivalence classes $X \in \theta_Q$ and $Y \in \theta_d$. We shall show that the procedure "first code the rules and then apply them" has the same complexity as the simple procedure "guess within $\theta_Q \cap \theta_d$" and can be viewed as identical from this point of view; in other words, we are guided by a purely statistical view. Although this is rather different from the RSDA approach, there is has been some effort to adopt this approach in the RSDA context (Wong, Ziarko & Ye, 1986); we shall discuss some aspects of this work below.

The partition induced by $\theta^{\mathsf{loc}} \stackrel{\text{def}}{=} \theta_Q \cap \theta_d$ are the nonempty sets in $\{X_i \cap Y_j : i \leq t, j \leq s\}$, and its associated parameters are defined by

$$\hat{\nu}_{i,j} = \frac{|X_i \cap Y_j|}{n}. \tag{3.4}$$

Thus,

$$H(\theta^{\mathsf{loc}}) = \sum_{i \leq t} \sum_{j \leq s} \hat{\nu}_{i,j} \cdot \log_2(\frac{1}{\hat{\nu}_{i,j}}) \tag{3.5}$$

Now, we define

$$H^{\mathsf{loc}}(Q \to d) \stackrel{\text{def}}{=} H(\theta^{\mathsf{loc}}).$$

In information theory, $H^{\mathsf{loc}}(Q \to d)$ is usually written as $H(Q, d)$; we use the notation above to emphasize that our view of the world consists of Q and that we want to predict d.

One problem with this approach is the symmetry

$$H^{\mathsf{loc}}(Q \to d) = H^{\mathsf{loc}}(\theta_Q \cap \theta_d) = H^{\mathsf{loc}}(d \to Q).$$

We shall not discuss this problem here, but instead refer the reader to Jumarie (1990), p. 24ff and p. 49ff, and Li & Vitányi (1993), p. 65ff.

The proof of the following proposition is straightforward and is left to the reader.

Proposition 3.1. *Let $d, Q \subseteq \Omega$. Then,*

1. $H^{\mathsf{loc}}(Q \to d) \geq H(d)$,
2. $H^{\mathsf{loc}}(Q \to d) = H(Q)$ *if and only if* $\theta_Q \subseteq \theta_d$.

Applying (3.2), a normalized loc-entropy measure $S^{\mathsf{loc}}(Q \to d)$ is definable and – given $H(d) < \log_2(n)$ – we obtain

$$S^{\mathsf{loc}}(Q \to d) = 1 - \frac{H^{\mathsf{loc}}(Q \to d) - H(d)}{\log_2(n) - H(d)}.$$

For each $i \leq t,\ j \leq s$ let

$$\hat{\eta}_{i,j} \stackrel{\text{def}}{=} \frac{|X_i \cap Y_j|}{r_i}.$$

This is the estimated probability of an element of X_i being in the class $X_i \cap Y_j$. In other words, it is the conditional probability of $x \in Y_j$, given that $x \in X_i$. Observe that

$$\sum_{i \leq t} \hat{\pi}_i = \sum_{j \leq s} \hat{\eta}_{i,j} = 1$$

so that the parameters $\hat{\pi}_i$ and $\hat{\eta}_{i,j}$ satisfy the hypotheses of Lemma 2.3, and that furthermore

$$\hat{\pi}_i \cdot \hat{\eta}_{i,j} = \frac{|X_i \cap Y_j|}{n} = \hat{\nu}_{i,j}. \tag{3.6}$$

Substituting (3.6) into (3.5) and applying Lemma 2.3 we obtain

$$H^{\mathsf{loc}}(Q \to d) = H(Q) + \sum_{i \leq t} \hat{\pi}_i \cdot \sum_{j \leq s} \hat{\eta}_{i,j} \cdot \log_2(\frac{1}{\hat{\eta}_{i,j}}) = H(Q) + \sum_{i=c+1}^{t} \hat{\pi}_i \cdot \sum_{j \leq s} \hat{\eta}_{i,j} \cdot \log_2(\frac{1}{\hat{\eta}_{i,j}}),$$

the latter since $\hat{\eta}_{i,j} = 1$ for $i \leq c$. The conditional entropy of d given Q is now

$$H^{\mathsf{loc}}(d|Q) \stackrel{\text{def}}{=} \sum_{i=c+1}^{t} \hat{\pi}_i \cdot \sum_{j \leq s} \hat{\eta}_{i,j} \cdot \log_2(\frac{1}{\hat{\eta}_{i,j}}). \tag{3.7}$$

This is the usual statistical definition of conditional entropy. Its normalization leads to the expression

$$S^{\mathsf{loc}}(d|Q) = 1 - \frac{H^{\mathsf{loc}}(d|Q)}{\log_2(n)}.$$

Example 1. cont. Table 4 shows the statistical information analysis of prediction quality within the example information system of Table 2 on page 6.

Table 4: Statistical information measures of predicting quality

Attribute Set	$H^{\mathsf{loc}}(Q \to d)$	$S^{\mathsf{loc}}(Q \to d)$	$H^{\mathsf{loc}}(d\|Q)$	$S^{\mathsf{loc}}(d\|Q)$	γ
$\{S\}$	1.352	0.808	0.361	0.835	0.556
$\{BMI\}$	1.891	0.568	0.900	0.587	0.000
$\{S, BMI\}$	2.197	0.432	0.222	0.814	0.778

Although both measures $S^{\mathsf{loc}}(Q \to d)$ and $S^{\mathsf{loc}}(d|Q)$ vote for $\{S\}$ as the best predicting set for the given data – and are in line with the results of the significance test (see Table 3 on page 11), this convergence need not to be true in the general case. □

Example 2. Some simple examples shall demonstrate how the average uncertainty measures H^{loc} and the approximation quality γ work, and how they differ:

Suppose that q_1 and d take the values $0, 1$, and suppose that we observe the probabilities

	$q_1 = 0$	$q_1 = 1$	$\sum$
$d = 0$	1/4	1/4	1/2
$d = 1$	1/4	1/4	1/2
$\sum$	1/2	1/2	1

We calculate $H^{\mathsf{loc}}(q_1 \to d) = 2$, and $H^{\mathsf{loc}}(d|q_1) = 1$.

Now, consider another attribute q_2 with values $0, \ldots 3$, and the observed probabilities

	$q_2 = 0$	$q_2 = 1$	$q_2 = 2$	$q_2 = 3$	$\sum$
$d = 0$	1/4	1/16	1/16	1/8	1/2
$d = 1$	0	3/16	3/16	1/8	1/2
$\sum$	1/4	1/4	1/4	1/4	1

Whereas q_2 enables us to predict 25% of the cases deterministically, namely, by the rule

$$\text{If } q_2 = 0, \text{ then } d = 0,$$

whereas q_1 cannot be used to predict d.

Comparing the entropy measures, we observe that $H^{\mathsf{loc}}(q_2 \to d) = 2.6556 > H^{\mathsf{loc}}(q_1 \to d) = 2$, and $H^{\mathsf{loc}}(d|q_2) = 0.6556 < H^{\mathsf{loc}}(d|q_1) = 1$. Whereas the entropy measure $H^{\mathsf{loc}}(Q \to d)$ favors q_1, the conditional entropy measure $H^{\mathsf{loc}}(d|Q)$ votes for q_2 to be the better predicting attribute. The explanation of this effect is simple: Although in the first example the two large classes predict obviously nothing, the encoding of these small number of classes can be done effectively. The prediction success in the second table is overruled by a large number of small classes with high uncertainty, causing a high coding complexity. If we subtract the coding complexity of the predicting attribute, the effect of the high coding effort is eliminated, and the better prediction success of q_2 results in a smaller conditional entropy measure.

A third table presents an example why $H^{\mathsf{loc}}(d|Q)$ is not optimal for rough set prediction under certain circumstances:

	$q_3 = 0$	$q_3 = 1$	$\sum$
$d = 0$	7/16	1/16	1/2
$d = 1$	1/16	7/16	1/2
$\sum$	1/2	1/2	1

Although q_3 predicts no outcome deterministically, the conditional measure $H^{\mathsf{loc}}(d|q_3) = 0.5436$ is better than $H^{\mathsf{loc}}(d|q_2) = 0.6556$. The essence of the result is that a bet given q_3 is preferable to a bet based on q_2. Having the knowledge $q_3 = 0$ enables us to predict that the outcome $d = 0$ is much more likely than $d = 1$, whereas $q_3 = 1$ predicts $d = 1$ most of the time. With attribute q_2, the bets given the value $q_2 \neq 1$ are comparably bad. Although the betting situation given q_3 is quite satisfactory, for a given observation i, $1 \leq i \leq n$, of the dataset with the knowledge $q_3(i) = 0$ and not knowing anything about d, we cannot find the value $d(i)$ unless we search through the whole set of d-values. In terms of RSDA, the prediction success of q_3 is as bad as that of q_1, and, consequently, $\gamma(q_1 \to d) = \gamma(q_3 \to d) = 0$. □

As the examples show, the statistical entropy measures do not take into account the special layout of the (rough) prediction problem, because the loc – model optimizes guessing outcome of a dependent variable but not necessarily perfect prediction.

In the next sections we will present other entropy measures, which are integrated into the rough set approach and which are more suitable for rough set prediction.

The earliest paper to concern itself with the connection between entropy and rough set analysis was Wong et al. (1986). In their Theorem 2, later restated in Teghem & Benjelloun (1992), Proposition 6, the following strong connection between RSDA and entropy measurement is claimed (translated into our terminology):

Claim *Suppose that for each $c < i \leq t$, $|X_i \cap Y_j| = d_i$ for all $j \leq s$. Then*

$$H^{\mathsf{loc}}(d|Q) = \frac{|\overline{Y_j}^Q \setminus \underline{Y_j}_Q|}{n}$$

for all $j \leq s$. □

Consider the following counterexample:

Suppose that $U = \{0, 1, \ldots, 7\}$, and that the partition given by d has the sets

$$Y_i = \{2 \cdot i, 2 \cdot i + 1\},\ i < 4,$$

and the partition given by Q is

$$X_0 = \{1, 3, 5, 7\},\ X_1 = \{0, 2, 4, 6\}.$$

Now, $\overline{Y_j}^Q = U$ and $\underline{Y_j}_Q = \emptyset$ for all $j < 4$, and thus, $\frac{|\overline{Y_j}^Q \setminus \underline{Y_j}_Q|}{n} = 1$. Furthermore, $|X_i \cap Y_j| = 1$ for all $i < 2,\ j < 4$, so that the hypothesis of the claim is satisfied. We now have

$$\hat{\pi}_i = \frac{1}{2},\ \hat{\eta}_{i,j} = \frac{1}{4},$$

and it follows that

$$\hat{\eta}_{i,j} \cdot \log_2\left(\frac{1}{\eta_{i,j}}\right) = \frac{1}{4} \cdot \log_2(4) = \frac{1}{2},$$
$$\sum_{j<4} \hat{\eta}_{i,j} \cdot \log_2\left(\frac{1}{\eta_{i,j}}\right) = 2.$$

Thus,

$$H^{\mathsf{loc}}(d|Q) = \sum_{i<2} \hat{\pi}_i \cdot 2 = 2,$$

which contradicts the claim.

We can generalize this example to show that under the assumptions of Wong et al. (1986), the value of $H^{\mathsf{loc}}(d|Q)$ does not depend so much on γ as it does on the number of classes of θ_d which are not Q – definable:

Proposition 3.2. *Suppose that no class of θ_Q is deterministic, and that the elements of each X_i are uniformly distributed among the classes Y_j, i.e. for each $i \leq t,\ j \leq s$ we have $|X_i \cap Y_j| = d_i$. Then, $H^{\mathsf{loc}}(d|Q) = \log_2(s+1)$.*

Proof. By the hypothesis we have for all $i \leq t, j \leq s$

$$|X_i \cap Y_j| = d_i,$$

and therefore it follows from $\sum_j \eta_{i,j} = 1$ that

$$\hat{\eta}_{i,j} = \frac{d_i}{r_i} = \frac{1}{s+1}.$$

Thus,

$$\begin{aligned} H^{\mathsf{loc}}(d|Q) &= \sum_i \hat{\pi}_i \cdot \sum_j \hat{\eta}_{i,j} \cdot \log_2\left(\frac{1}{\hat{\eta}_{i,j}}\right) \\ &= \sum_i \hat{\pi}_i \cdot \sum_j \frac{1}{s+1} \cdot \log_2(s+1) \\ &= \sum_i \hat{\pi}_i \cdot \log_2(s+1) \\ &= \log_2(s+1), \end{aligned}$$

which proves our claim. □

3.2 Prediction II: Playing it safe

Whereas the entropy measures in the previous section are good candidates to be measures of optimal guessing strategies, based on the estimated parameters of the distributions of the cross-classification $d \times Q$, a rough set approach should not be based on "guessing" but on "knowing". This means that the observations which can be predicted perfectly are assumed to be the realization of a systematic process, whereas the nature of the indeterministic rules is assumed to be unknown to the researcher.

Based on these arguments, given a class Y of θ_d, any observation y in the region of uncertainty $\overline{Y}^Q \setminus \underline{Y}_Q$ is the result of a random process whose characteristics are unknown; in other words, our given data is the partition obtained from Q, and we know the world only up to the equivalence classes of θ_Q. Given this assumption, no information within our data set will help us to classify an element $y \in U \setminus V$, and we conclude that each such y requires a rule (or class) of its own. In this case, any element of $U \setminus V$ may be viewed as a realization of an unknown probability distribution with its uncertainty $\frac{1}{n}\log_2(n)$. Note that, unlike the previous one, this approach assumes that only the classes of θ_Q are observed within a representative sample, or – in terms of parameters – the approach requires only the probability distribution π_{θ_Q} (and its estimates $\hat{\pi}_{\theta_Q}$) of the classes of θ_Q. Thus, we regard Q (and its associated equivalence relation θ_Q) as the given data, and, in accord with the principles of RSDA, we only know the upper, respectively the lower Q – approximation of any class Y of θ_d.

It follows that we may only apply the deterministic part of $Q \to d$, and ignore whatever might be gained from the indeterministic rules. Thus, we use only those classes of θ_Q which are contained in V, and assume that each $y \in U \setminus V$ is in its own class. In other words, we assume the *maximum entropy principle* as a worst case, and look at the equivalence relation θ^{det} defined by

$$x \equiv_{\theta^{\mathsf{det}}} y \overset{\text{def}}{\iff} x = y \text{ or there exists some } i \leq c \text{ such that } x, y \in X_i.$$

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

$$\hat{\psi}_i \overset{\text{def}}{=} \begin{cases} \hat{\pi}_i, & \text{if } i \leq c, \\ \frac{1}{n}, & \text{otherwise.} \end{cases} \tag{3.8}$$

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

$$H^{\mathsf{det}}(Q \to d) \overset{\text{def}}{=} H(\theta^{\mathsf{det}}) = \sum_i \hat{\psi}_i \cdot \log_2(\frac{1}{\hat{\psi}_i})$$

and have

$$\begin{aligned} H^{\mathsf{det}}(Q \to d) &= \sum_{i \leq c} \hat{\pi}_i \cdot \log_2(\frac{1}{\hat{\pi}_i}) + |U \setminus V| \cdot \frac{1}{n} \cdot \log_2(n) \\ &= \underbrace{\sum_{i \leq c} \hat{\pi}_i \cdot \log_2(\frac{1}{\hat{\pi}_i})}_{\text{Knowledge}} + \underbrace{(1-\gamma) \cdot \log_2(n)}_{\text{Guessing}}. \end{aligned}$$

This gives us

$$\begin{aligned} H^{\mathsf{det}}(d|Q) &\overset{\text{def}}{=} H^{\mathsf{det}}(Q \to d) - H(Q) \\ &= (1-\gamma) \cdot \log_2(n) - \sum_{i > c} \hat{\pi}_i \cdot \log_2(\frac{1}{\hat{\pi}_i}). \end{aligned}$$

Since $\theta_Q^+ \subseteq \theta_Q \cap \theta_d$, we note that $\theta_Q \cap \theta_d$ has no more classes than θ_Q^+, and therefore

$$\begin{aligned} H^{\mathsf{loc}}(Q \to d) &= \sum_{i \leq c} \hat{\pi}_i \cdot \log_2(\frac{1}{\hat{\pi}_i}) + \sum_{i=c+1}^{t} \sum_{j \leq s} \hat{\nu}_{i,j} \cdot \log_2(\frac{1}{\hat{\nu}_{i,j}}) \\ &\leq \sum_{i \leq c} \hat{\pi}_i \cdot \log_2(\frac{1}{\hat{\pi}_i}) + (1-\gamma) \cdot \log_2(n), \\ &= H^{\mathsf{det}}(Q \to d), \end{aligned}$$

which implies $H^{\mathsf{loc}}(d|Q) \leq H^{\mathsf{det}}(d|Q)$.

If we compare the $H^{\mathsf{det}}(Q \to d)$ and $H^{\mathsf{loc}}(Q \to d)$ in terms of necessary parameters, we have to assume for the computation of $H^{\mathsf{loc}}(Q \to d)$ that the deterministic rules as well as the indeterministic rules are representative within the sample of the underlying population. Indeed, the H^{loc}-measures do not distinguish – up to quantitative values – between deterministic and indeterministic rules.

In contrast, $H^{\mathsf{det}}(Q \to d)$ requires a representativeness only for the deterministic rules, and assumes that any indeterministic rule, which is valid for m objects, consists of m unique (individual) rules, gathered from a random world which cannot be replicated.

The proof of the following is straightforward, and is left to the reader:

Proposition 3.3. *Let $d, Q \subseteq \Omega$. Then,*

1. $H^{\mathsf{det}}(Q \to d) \geq H(d)$,
2. $H^{\mathsf{det}}(Q \to d) = H(Q)$ *if and only if* $\theta_Q \subseteq \theta_d$.
3. $H^{\mathsf{det}}(Q \to d) = \log_2(n)$ *if and only if* $V = \emptyset$ *or* V *is a union of singletons of* θ_Q. □

The extremes for $H^{\mathsf{det}}(Q \to d)$ are

- θ_Q is the identity relation, and everything can be explained by Q,
- $\gamma(Q \to d) = 0$, and everything is guessing.

In both cases we have $H^{\mathsf{det}}(Q \to d) = \log_2(n)$.

The following gives the bounds within which $H^{\mathsf{det}}(d|Q)$ varies:

Proposition 3.4. $(1 - \gamma) \leq H^{\mathsf{det}}(d|Q) \leq (1 - \gamma)\log_2(n - |V|)$.

Proof. First, observe that by (2.4) on page 8, $\log_2(n - |V|) \geq \log_2(2) = 1$.

The minimum value of $\sum_{i>c} \hat{\pi}_i \cdot \log_2(\hat{\pi}_i)$ is obtained when $c = t - 1$, and in this case,

$$\begin{aligned}\sum_{i>c} \hat{\pi}_i \cdot \log_2(\hat{\pi}_i) &= \frac{n - |V|}{n} \cdot \log_2(\frac{n}{n - |V|}) \\ &= (1 - \gamma) \cdot \log_2(\frac{1}{1 - \gamma}).\end{aligned}$$

Therefore,

$$\begin{aligned}H^{\mathsf{det}}(d|Q) &= (1 - \gamma) \cdot \log_2(n) - \sum_{i>c} \hat{\pi}_i \cdot \log_2(\frac{1}{\hat{\pi}_i}) \\ &\leq (1 - \gamma) \cdot \log_2(n) - (1 - \gamma) \cdot \log_2(\frac{1}{1 - \gamma}), \\ &= (1 - \gamma) \cdot (\log_2(n) - \log_2(\frac{1}{1 - \gamma}), \\ &= (1 - \gamma) \cdot \log_2(n \cdot (1 - \gamma)) \\ &= (1 - \gamma) \cdot \log_2(n - |V|).\end{aligned}$$

For the other direction, we first note that each nondeterministic class X has at least two elements, and that $\sum_{i>c} \hat{\pi}_i \cdot \log_2(\frac{1}{\hat{\pi}_i})$ has a maximum if either each such class has exactly two elements, or all but

one class have two elements and one class has three elements. Since the value of $\sum_{i>c} \hat{\pi}_i \cdot \log_2(\frac{1}{\hat{\pi}_i})$ is greater in the first case, we assume w.l.o.g. that $n - |V|$ is even, so that

$$\begin{aligned}\sum_{i>c} \hat{\pi}_i \cdot \log_2(\frac{1}{\hat{\pi}_i}) &= \frac{n-|V|}{2} \cdot \frac{2}{n} \cdot \log_2(\frac{n}{2}) \\ &= (1-\gamma) \cdot \log_2(\frac{n}{2}).\end{aligned}$$

Therefore,

$$\begin{aligned}H^{\mathsf{det}}(d|Q) &\geq (1-\gamma) \cdot \log_2(n) - (1-\gamma) \cdot \log_2(\frac{n}{2}) \\ &= (1-\gamma) \cdot (\log_2(n) - \log_2(\frac{n}{2})) \\ &= (1-\gamma) \cdot \log_2(2) \\ &= 1-\gamma,\end{aligned}$$

which proves our claim. □

We see that $H^{\mathsf{det}}(d|Q)$ is independent of the granularity – i.e. the probability distribution – of the deterministic classes of θ_Q, and that it is dependent on the granularity of the classes leading to non-deterministic rules: The higher the granularity of those classes, the lower $H^{\mathsf{det}}(d|Q)$. We use this to show

Proposition 3.5. *If $Q \subseteq R$, then $H^{\mathsf{det}}(d|R) \leq H^{\mathsf{det}}(d|Q)$.*

Proof. By the remark above, we can assume that every deterministic class of θ_Q is a class of θ_R. This implies that $\theta_Q^+ \subseteq \theta_R^+$, and hence,

$$H^{\mathsf{det}}(R \to d) \leq H^{\mathsf{det}}(Q \to d).$$

Since furthermore $H(Q) \leq H(R)$ by Corollary 2.2, the conclusion follows. □

A similar result does not hold for $H^{\mathsf{det}}(Q \to d)$ as the example given in Table 5 shows: There,

$$H^{\mathsf{det}}(\{q_1\} \to \{p\}) = 1.5 < 2 = H^{\mathsf{det}}(\{q_1, q_2\} \to \{p\}) = H^{\mathsf{det}}(\{q_2\} \to \{p\}).$$

Table 5: $H^{\mathsf{det}}(Q \to d)$

U	q_2	q_1	p
1	1	1	1
2	2	1	2
3	3	2	2
4	4	2	2

As in (3.2), we define the normalized relative deterministic prediction success $S^{\mathsf{det}}(Q \to d)$, which we also will call *normalized rough entropy* (NRE): First, let $\theta_d = \varpi$, so that $H(d) = \log_2(n)$. Then

$$S^{\mathsf{det}}(Q \to d) \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } \theta_Q = \varpi, \\ 0, & \text{otherwise.} \end{cases} \tag{3.9}$$

Otherwise, if $H(d) < \log_2(n)$, we set

$$S^{\mathsf{det}}(Q \to d) \stackrel{\text{def}}{=} 1 - \frac{H^{\mathsf{det}}(Q \to d) - H(d)}{\log_2(n) - H(d)}, \tag{3.10}$$

In this way we obtain an measure of prediction success within RSDA, which can be used to compare different rules in terms of the combination of coding complexity and the prediction uncertainty in the sense that a perfect prediction results in $S^{\mathsf{det}}(Q \to d) = 1$, and the worst case is at $S^{\mathsf{det}}(Q \to d) = 0$. S^{det} is an unconditional measure, because both, the complexity of the rules and the uncertainty of the predictions, are merged into one measure.

The question arises, where the approximation function γ is positioned in this model. Proposition 3.4 shows that, for fixed Q,

$$\max\{H^{\mathsf{det}}(d|R) : \gamma(R \to d) = \gamma(Q \to d)\} = (1 - \gamma) \cdot \log_2(n - |V|),$$

and we denote this value by $H^{\mathsf{det}}_{\mathsf{max}}(d|Q)$. The following result tells us that, for fixed d, $H^{\mathsf{det}}_{\mathsf{max}}(d|Q)$ is strictly inversely monotone to $\gamma(Q \to d)$:

Proposition 3.6. $\gamma(Q \to d) < \gamma(R \to d) \iff H^{\mathsf{det}}_{\mathsf{max}}(d|R) < H^{\mathsf{det}}_{\mathsf{max}}(d|Q)$.

Proof. "$\Rightarrow$": The hypothesis $\gamma(Q \to d) < \gamma(R \to d)$ implies that $|V_{Q \to d}| \lneq |V_{R \to d}|$. Thus,

$$\begin{aligned} H^{\mathsf{det}}_{\mathsf{max}}(d|R) &= (1 - \gamma(R \to d)) \cdot \log_2(n - |V_{R \to d}|), \\ &< (1 - \gamma(Q \to d)) \cdot \log_2(n - |V_{Q \to d}|), \\ &= H^{\mathsf{det}}_{\mathsf{max}}(d|Q), \end{aligned}$$

"$\Leftarrow$": First note, that for $k \geq 1$,

$$k \cdot \log_2 k < (k+1) \cdot \log_2(k+1). \tag{3.11}$$

We can also assume that $0 < H^{\mathsf{det}}_{\mathsf{max}}(d|R)$, so that $U \setminus V_{R \to d} \neq \emptyset$. Now,

$$\begin{aligned} & H^{\mathsf{det}}_{\mathsf{max}}(d|R) < H^{\mathsf{det}}_{\mathsf{max}}(d|Q) \\ & \Rightarrow (1 - \gamma(R \to d)) \cdot \log_2(n - |V_{R \to d}|) < (1 - \gamma(Q \to d)) \cdot \log_2(n - |V_{Q \to d}|) \\ & \Rightarrow (n - |V_{R \to d}|) \cdot \log_2(n - |V_{R \to d}|) < (n - |V_{Q \to d}|) \cdot \log_2(n - |V_{Q \to d}|) \\ & \Rightarrow (n - |V_{R \to d}|) < (n - |V_{Q \to d}|) \text{ by (3.11)} \\ & \Rightarrow |V_{Q \to d}| < |V_{R \to d}| \\ & \Rightarrow \gamma(Q \to d) < \gamma(R \to d). \end{aligned}$$

This completes the proof. □

We observe that

- RSDA which tries to maximize γ is a procedure to minimize the maximum of the conditional entropy of deterministic rough prediction.

In terms of conditional uncertainty, we may view $\gamma = \gamma(Q \to d)$ as a crude approximation of a measure of normalized prediction success, because

$$\begin{aligned} S^{\mathsf{det}}_{\mathsf{max}}(d|Q) &= 1 - \frac{H^{\mathsf{det}}_{\mathsf{max}}(d|Q) - \min\{H^{\mathsf{det}}_{\mathsf{max}}(d|R) : R \subseteq \Omega\}}{\max\{H^{\mathsf{det}}_{\mathsf{max}}(d|R) : R \subseteq \Omega\} - \min\{H^{\mathsf{det}}_{\mathsf{max}}(d|R) : R \subseteq \Omega\}} \\ &= 1 - \frac{H^{\mathsf{det}}_{\mathsf{max}}(d|Q) - 0}{\log_2(n) - 0} \\ &= \gamma - (1-\gamma)\frac{\log_2(1-\gamma)}{\log_2(n)} \\ &= \gamma + \mathcal{O}\left(\frac{1}{\log_2(n)}\right). \end{aligned}$$

Proposition 3.5 does not extend to the hypothesis $\gamma(Q \to d) < \gamma(R \to d)$, and thus, a result similar to 3.6 does not hold, as the following example shows: Consider the equivalence relations $\theta_d, \theta_Q, \theta_R$ with the following partitions:

$$\theta_d : \{1,2,3\},\{4,5,6\},\ \theta_Q : \{1,4\},\{2,5\},\{3,6\},\ \theta_R : \{1\},\{2,3,4,5,6\}.$$

Then,

$$\gamma(Q \to d) = 0 < \frac{1}{6} = \gamma(R \to d).$$

On the other hand,

$$H^{\mathsf{det}}(d|Q) = \log_2(6) - \log_2(3) = 1 < \frac{5}{6}\cdot\log_2(5) = \frac{5}{6}\cdot\log_2(6) - \frac{5}{6}\cdot\log_2(\frac{6}{5}) = H^{\mathsf{det}}(d|R).$$

Example 1. cont.
Table 6 presents the rough information analysis for the data of the example given in Table 2. We have skipped the presentation of the $H^{\mathsf{det}}(d|Q)$-measures, because – as shown above – they are identical with γ for the purpose of comparing the prediction success of different attribute sets. The results show

Table 6: Rough information measures of the predicting quality within the example information system

Attribute Set	$H^{\mathsf{det}}(Q \to d)$	$S^{\mathsf{det}}(Q \to d)$	$S^{\mathsf{loc}}(Q \to d)$	γ	Significance
$\{S\}$	1.880	0.573	0.808	0.556	0.047
$\{BMI\}$	3.170	0.000	0.568	0.000	1.000
$\{S, BMI\}$	2.197	0.432	0.432	0.778	0.144

that the NRE $S^{\mathsf{det}}(Q \to d)$ is a good candidate to evaluate the rough prediction quality of attribute set, because it produces the same order of "goodness in predictability" as the significance test, without the limitations of the significance test. Inspecting the results of $\{BMI\}$ in Table 6, one can see that the "defects" of $S^{\mathsf{loc}}(Q \to d)$ have been repaired. □

3.3 Prediction III: Living side by side

In Section 3.2 the prediction $H^{\mathsf{det}}(Q \to d)$ consists of two parts: The absolute correct deterministic part (the union of the lower bound approximations) and the random part. The prediction within the random part is done using an "element – to – class" mapping, because of the assumption that no uncertain observation can be predicted given any available source of data. If we are willing to use the information provided by the indeterministic rules which are offered by RSDA, the uncertainty is restricted by those rules and we need another entropy estimation.

This approach to handle uncertainty recognizes that θ_d induces some structure on $U \setminus V$: If X_i is a class of θ_Q which does not lead to a deterministic rule, there are classes $Y_{i,0}, \ldots, Y_{i,k}$ of $\theta_d,\ k \geq 1$, such that $\langle X_i, Y_{i,j} \rangle \in Q \to d$, i.e. X_i intersects each $Y_{i,j} \setminus V$. Uncertainty given X_i can now be measured by the uncertainty within $\{Y_{i,0} \setminus V, \ldots, Y_{i,k} \setminus V\}$ which also requires knowledge of the probability distribution induced by θ_d. The assumption can be interpreted in the sense that an indeterministic rule produces a certain degree of imprecision in the prediction of θ_d, but that the *amount* of uncertainty is based solely on the uncertainty within d and does not interact with Q. Even though this is not "pure rough set theory", it is certainly consistent with it: The procedure describes the upper bounds of sets defined by θ_d in terms of a suitable probability distribution. As we shall not be using the method in the sequel, we will spare the reader the somewhat involved definitions of the resulting entropy measures $H^*(Q \to d)$ and $H^*(d|Q)$. We shall just mention, that, unlike $H^{\mathsf{det}}(d|Q)$ and $H^{\mathsf{loc}}(d|Q)$, the conditional entropy $H^*(d|Q)$ is not (anti-) monotone. This result is a drawback, because the monotone relationship of $\subseteq$ and a measure of approximation quality seems to be quite natural. As a consequence, within a search process we cannot use $H^*(d|R)$ as stop criterion like the other conditional measures γ, $H^{\mathsf{det}}(d|R)$, or $H^{\mathsf{loc}}(d|R)$. Therefore it seems that the practical value of the H^*–measure is rather limited, although it takes a representativeness assumption which is in between deterministic rough entropy H^{det} and the statistical entropy H^{loc}: The H^* – approach assumes that the probability distributions within the upper bound of any class of θ_d are representative, whereas H^{loc} assumes that any conditional probability distribution is representative, and H^{det} assumes that the probability distribution within the lower bound of any class of θ_d is representative for the population.

We shall investigate this method in more detail in subsequent research.

4 Data analysis and validation

The approach which is closest to the non–invasive philosophy of RSDA is the entropy of deterministic rough prediction $H^{\mathsf{det}}(Q \to d)$ which combines the principle of indifference with the maximum entropy principle in an RSDA context. We advocate this type of entropy because of 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" (Jaynes, 1957).

To obtain an objective measurement we use the normalized rough entropy (NRE) of (3.10) on page 23, where

$$S^{\mathsf{det}}(Q \to d) = 1 - \frac{H^{\mathsf{det}}(Q \to d) - H(d)}{\log_2(|U|) - H(d)}. \tag{4.1}$$

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 (Düntsch & Gediga, 1997a)[1].

4.1 Validation

In order to test the procedure, we have used 14 datasets available from the UCI repository[2] from which the appropriate references of origin can be obtained. These are a subset of the datasets which were used by Quinlan (1996) to test Release 8 of C4.5.

The validation by the training set – testing set method was performed by splitting the full data set randomly into two equal sizes 100 times, assuming a balanced distribution of training and testing data (TT2 method). The mean error value is our measure of prediction success.

We choose only half of the set for training purposes in order to have a basis for testing the predictive power of the resulting attribute sets. Because all data sets contained continuous attributes and most of them missing values as well, a preprocessing step was necessary to apply the SORES algorithm to these data sets. Missing values were replaced by the mean value in case of ordinal attributes, and by the most frequent value (i.e. the mode) otherwise. The preprocessing of the continuous data was done by three different global discretization methods:

Method 1 consists of the global filtering method described in Section 2.3 which influences the NRE, but does not affect γ, and thus has no influence on the dependency structure. This results in minimal granularity of attributes with respect to the decision attribute. The other two discretization methods cluster the values of an attribute into ten, resp. five, classes with approximately the same number of objects. The discretization method can be refined by transforming the H^{loc}–based methods of local

[1] All material relating to SORES, e.g. datasets, a description of the algorithm, as well as GROBIAN, can be obtained from our website `http://www.psycho.uni-osnabrueck.de/sores/`

[2] `http://www.ics.uci.edu/~mlearn/MLRepository.html`

Table 7: 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

discretization of continuous attributes given in Catlett (1991) and Dougherty, Kohavi & Sahami (1995) to the proposed H^{det} – measure. This is a task which still needs to be done, but which is outside the scope of the current introductory article.

In Table 7 we list the basic parameters of the data sets, and compare the SORES results with the C4.5 performance given in Quinlan (1996). This has to be taken with some care, since Quinlan uses 10–fold cross validation (CV10) on data sets optimized by

> " . . . dividing the data into ten blocks of cases that have similar size and class distribution" (Quinlan, 1996, p.81, footnote 3.).

Because TT2 tends to result in smaller prediction success rates than CV10, the comparison of SORES and C4.5 is based on a conservative estimate.

The SORES column "No. of pred. attr." records the number of attributes which are actually used for prediction; this is a prominent feature of RSDA, and in most cases considerably less than the number of all attributes.

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.

5 Summary and outlook

In the first part of the paper we have proposed three approaches to estimate the unconditional prediction success within the context of RSDA using various entropy measures.

The statistical entropy measure is not well suited, because the assumption of a symmetric information exchange of predicting and predicted attributes is not given within the RSDA frame. Two modifications are discussed: The first one, H^{det}, relies only on the information given by the deterministic rules, and assumes an atom–like structure of all other information. The other approach, H^*, additionally uses the knowledge about the distributions within the indeterministic rules, but has the drawback of lacking monotony within the conditional measure $H^*(d|Q)$. The measure $H^{\mathsf{det}}(Q \to d)$ seems to be the most suitable measure to compare attribute sets $Q_1, \ldots, Q_k$ in terms of combined coding complexity and expected prediction uncertainty.

In the second part of the paper, we have applied the method of searching optimal rough entropy sets (SORES) to real life data sets. The method seems to be well applicable, since we show that C4.5 performs better than SORES on only 7 of 14 problems, although C4.5 is used in a fine tuned version (Release 8) and SORES, at present, is still quite "raw".

Fine tuning of the SORES procedure will consist of – at least – the following steps.

- Both types of measures – $H^M(Q \to d)$ and $H^M(d|Q)$ (whatever model M is used) – are to some extent suitable measures for finding optimal sets for prediction, and thus, any weighted sum
$$H^M(Q, d, \omega) = \omega \cdot H^M(Q \to d) + (1 - \omega) \cdot H^M(d|Q),$$
($0 \leq \omega \leq 1$) is a suitable measure as well. If $\omega = 1$, we weight the effort of searching for a rule as high as the effort of reducing uncertainty of the dependent attribute. If $\omega = 0$ is chosen, then the effort of coding the rules is neglected. Finally, any $0 < \omega < 1$ estimates the relative effort of finding a rule with respect to finding an object under uncertainty. The methods in Section 3 are based on an $\omega = 1$ procedure, but it will be worthwhile to compare these results with procedures using $\omega < 1$.

- The proposed method – as a symbolic data analysis procedure – is rather time consuming. In order to enhance the applicability of the procedure to real life data sets, the optimization cannot be performed on big samples, but some kind of subsample optimization must be implemented. The theory of dynamic reducts (Bazan, Skowron & Synak, 1994, Bazan, 1997) is a step towards such an enhancement.

- The discretization of continuous attributes is another problem which has to be solved by any symbolic data analysis technique. Although the global discretization procedures described above work quite well in the presented numerical examples, a local discretization procedure, which optimizes the chosen criterion – e.g. $H^{\mathsf{det}}(Q \to d)$ – directly, can be expected to produce an even better prediction quality.

Finally, we should like to point out that, except for the two numerical global discretization methods, all of the procedures developed in Section 3 do not use any external parameters, and only the representation assumptions stated for each of the three approaches. Thus, model assumptions are kept to a minimum, and the procedures can (at least) serve as a preprocessing mechanism before "harder" computational or statistical methods are applied.

References

Bazan, J., Skowron, A. & Synak, P. (1994). Dynamic reducts as a tool for extracting laws from decision tables. In *Proc. of the Symp. on Methodologies for Intelligent Systems, Charlotte, NC*, Lecture Notes in Artificial Intelligence, 346–355, Berlin. Springer–Verlag.

Bazan, J. G. (1997). A comparison of dynamic and non–dynamic rough set methods for extracting laws from decision tables. Preprint, Institute of Mathematics, University of Rzeszów.

Browne, C. (1997). Enhanced rough set data analysis of the Pima Indian diabetes data. In *Proc. 8th Ireland Conference on Artificial Intelligence, Derry (1997)*, 32–39.

Browne, C., Düntsch, I. & Gediga, G. (1998). IRIS revisited: A comparison of discriminant and enhanced rough set data analysis. In Polkowski & Skowron (1998). To appear.

Catlett, J. (1991). On changing continuous attributes into ordered discrete attributes. In Y. Kodratoff (Ed.), *Proceedings European Working Session on Learning – EWSL-91*, 164–178, Berlin. Springer Verlag.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, **45**, 1304–1312.

Dougherty, J., Kohavi, R. & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Proceedings Twelfth International Conference on Machine Learning*, 194–202, San Francisco. Morgan Kaufmann.

Düntsch, I. (1997). A logic for rough sets. *Theoretical Computer Science*, **179**, 427–436.

Düntsch, I. & Gediga, G. (1997a). The rough set engine GROBIAN. In Sydow (1997), 613–618.

Düntsch, I. & Gediga, G. (1997b). ROUGHIAN – Rough Information Analysis (Extended abstract). In Sydow (1997), 631–636. The full paper is available from `http://www.infj.ulst.ac.uk/~cccz23/papers/roughian.html`.

Düntsch, I. & Gediga, G. (1997c). Statistical evaluation of rough set dependency analysis. *International Journal of Human–Computer Studies*, **46**, 589–604.

Düntsch, I. & Gediga, G. (1998). Simple data filtering in rough set systems. *International Journal of Approximate Reasoning*, **18**, 93–106.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, **106**, 620–630.

Jumarie, G. (1990). Relative Information. Berlin, Heidelberg,New York: Springer–Verlag.

Konrad, E., Orłowska, E. & Pawlak, Z. (1981a). Knowledge representation systems – Definability of informations. ICS Research Report 433, Polish Academy of Sciences.

Konrad, E., Orłowska, E. & Pawlak, Z. (1981b). On approximate concept learning. Tech. Rep. 81–7, Technische Unversität Berlin.

Li, M. & Vitányi, P. (1993). An Introduction to Kolmogorov Complexity and Its Applications. Texts and Monographs in Computer Science. Berlin, Heidelberg, New York: Springer–Verlag.

Nguyen, H. S. & Nguyen, S. H. (1998). Discretization methods in data mining. In Polkowski & Skowron (1998). To appear.

Pagliani, P. (1997). Rough sets theory and logic-algebraic structures. In E. Orłowska (Ed.), *Incomplete Information – Rough Set Analysis*, 109–190. Heidelberg: Physica – Verlag.

Pawlak, Z. (1973). Mathematical foundations of information retrieval. ICS Research Report 101, Polish Academy of Sciences.

Pawlak, Z. (1982). Rough sets. *Internat. J. Comput. Inform. Sci.*, **11**, 341–356.

Pawlak, Z. (1991). Rough sets: Theoretical aspects of reasoning about data, vol. 9 of *System Theory, Knowledge Engineering and Problem Solving*. Dordrecht: Kluwer.

Pawlak, Z., Grzymała-Busse, J. W., Słowiński, R. & Ziarko, W. (1995). Rough sets. *Comm. ACM*, **38**, 89–95.

Pawlak, Z. & Słowiński, R. (1993). Rough set approach to multi–attribute decision analysis. ICS Research Report 36, Warsaw University of Technology.

Polkowski, L. & Skowron, A. (Eds.) (1998). Rough sets in knowledge discovery. Heidelberg: Physica–Verlag. To appear.

Quinlan, R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, **4**, 77–90.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica*, **14**, 465–471.

Rissanen, J. (1985). Minimum – description – length principle. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, 523–527, New York. Wiley.

Sydow, A. (Ed.) (1997). Proc. 15th IMACS World Congress, vol. 4, Berlin. Wissenschaft und Technik Verlag.

Teghem, J. & Benjelloun, M. (1992). Some experiments to compare rough sets theory and ordinal statistical methods. In R. Słowiński (Ed.), *Intelligent decision support: Handbook of applications and advances of rough set theory*, vol. 11 of *System Theory, Knowledge Engineering and Problem Solving*, 267–284. Dordrecht: Kluwer.

Wong, S. K. M., Ziarko, W. & Ye, R. L. (1986). Comparison of rough–set and statistical methods in inductive learning. *Internat. J. Man–Mach. Stud.*, **24**, 53–72.