Statistical Techniques for Rough Set Data Analysis

Günther Gediga[1] and Ivo Düntsch[2]

[1] FB Psychologie / Methodenlehre, Universität Osnabrück, 49069 Osnabrück, Germany, Guenther@Gediga.de

[2] School of Information and Software Engineering, University of Ulster, Newtownabbey, BT 37 0QB, N. Ireland, I.Duentsch@ulst.ac.uk

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. Traditional "hard" data analysis based on statistical models or are in many cases not equipped to deal with uncertainty, relativity, or non–monotonic processes. Even the recently popular "soft" computing approach with its principal components

> " . . . fuzzy logic, neural network theory, and probabilistic reasoning" [16].

uses quite hard parameters outside the observed phenomena, e.g. representation and distribution assumptions, prior probabilities, beliefs, or membership degrees, the origin of which is not always clear; one should not forget that the results of these methods are only valid up to the – stated or unstated – model assumptions.

The question arises, whether there is a step in the modelling process which is informative for the researcher and, at the same time, does not require additional assumptions about the data.

To make this clearer, we follow [9] in assuming that a data model consists of

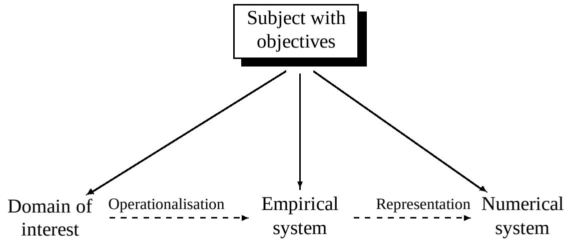
1. A domain $\mathcal{D}$ of interest.
2. An empirical system $\mathcal{E}$, which consists of a body of data and relations among the data, and a mapping $e : \mathcal{D} \rightarrow \mathcal{E}$, called *operationalisation*.
3. A (structural or numerical) model $\mathcal{M}$, and a mapping $m : \mathcal{E} \rightarrow \mathcal{M}$, called *representation*.
4. The agent (literally: the acting subject),

see Figure 1.

The agent with her/his objectives is the central part of the modelling process. Agents choose operationalisation and representation according to their objectives and their view of the world. The numerical models are normally a reduction of the empirical models, and thus of the domain of interest which results in further decontextualisation. We observe that even the soft computing methods reside on the level of the numerical models.

Rough set data analysis (RSDA) which has been developed by Z. Pawlak [10] and his co–workers since the early 1970s is a structural method which stays on the level

Fig. 1. The modelling process



of the empirical model; more formally, the representation mapping is the identity, and thus, there is a one–one relationship between the elements of the empirical model and the representation. In this way, we avoid further reduction, stay closer to the data, and keep the model assumptions to a minimum.

Although designed as a structural – in particular, a non statistical – approach to data analysis, application of RSDA only makes sense, if some basic statistical assumptions are observed. We will show that the application of these assumptions leads quite naturally to

- Statistical testing schemes for the significance of inference rules,
- Entropy measures for model selection, and
- A probabilistic version of RSDA.

1

2 Operationalisation in RSDA

Operationalisation of domain data in RSDA is done via a tabularised OBJECT $\mapsto$ ATTRIBUTE relationship: An *information system* is a tuple $\mathcal{I} = \langle U, \Omega, V_x \rangle_{x \in \Omega}$, where

1. $U = \{a_1, \dots, a_N\}$ is a finite set.
2. $\Omega = \{x_1, \dots, x_T\}$ is a finite set of mappings $x : U \to V_x$.

We interpret U as a set of objects and Ω as a set of attributes or features each of which assigns to an object a its value under the respective attribute. For each nonempty $Q \subseteq \Omega$ we define

$$V_Q = \prod_{a \in Q} V_a. \tag{1}$$

For $a \in U$, we also let

$$Q(a) = \langle x(a) \rangle_{a \in Q}, \tag{2}$$

written as $\boldsymbol{x}^Q(a)$ or just $\boldsymbol{x}^Q$ if a is understood or not relevant in the context. Each $Q(a)$ is called a *Q–granule*; the collection of all Q–granules is denoted by G_Q. A

Q–granule can be understood as a piece of information about objects in U given by the features in Q. The equivalence relation on U induced by Q is denoted by ψ_Q, i.e. for $a_i, a_j \in U$,

$$a_i \equiv_{\psi_Q} a_j \Longleftrightarrow Q(a_i) = Q(a_j). \tag{3}$$

Objects which in this sense belong to the same granule cannot be distinguished with the knowledge of Q. We denote the set of classes of ψ_Q by $\mathcal{P}(Q)$.

Suppose that $\emptyset \neq Q, P \subseteq \Omega$. Our aim is to describe the world according to P with our knowledge according to Q. If, for example, a class M of ψ_Q is contained totally within a class of ψ_P, then $Q(a)$ determines $P(b)$ for all $a, b \in M$. Such an M is called a *P–deterministic class of Q*, and

$$\text{If } Q(a) = \boldsymbol{x}^Q, \text{ then } P(a) = \boldsymbol{x}^P \tag{4}$$

is called a *deterministic Q,P – rule*. Otherwise, M intersects exactly the classes $L_1, \dots, L_k$ of $\mathcal{P}(P)$ with associated $\boldsymbol{x_1}^P, \dots, \boldsymbol{x_k}^P \in G_P$, and we call

$$\text{If } Q(a) = \boldsymbol{x}^Q, \text{ then } P(a) = \boldsymbol{x_1}^P \text{ or } \dots \text{ or } P(a) = \boldsymbol{x_k}^P \tag{5}$$

an *indeterministic Q,P – rule*. The collection of all Q, P – rules is denoted by $Q \to P$, and with some abuse of language, will be sometimes called a rule (of the information system). In writing rules, we will usually identify singleton sets with the element they contain, e.g. we write $Q \to d$ instead of $Q \to \{d\}$.

Note that all constructions above use only the information given by the observed system, and no additional outside parameters.

Throughout this paper, we use $\mathcal{I}$ as above with the given parameters as a generic information system. For further information on RSDA we refer the reader to [11] or [6].

3 Basic statistics

Even though rough set analysis is a structural method, it makes basic statistical assumptions which we briefly want to describe in this section. Suppose that ψ is an equivalence relation on U which may be of the form ψ_Q; the only numerical information we have are the cardinality T of U, and the cardinalities of the classes of ψ.

For $X \subseteq U$, we call

$$\underline{X}_\psi \overset{def}{=} \bigcup\{\psi x : \psi x \subseteq X\} \tag{1}$$

the *lower approximation* or *positive region of X*. These are those elements of U which can be classified with certainty as being in X. The *upper approximation* or *possible region* of X with respect to ψ is defined as

$$\overline{X}^\psi \overset{def}{=} U \setminus (\underline{U \setminus X}_\psi). \tag{2}$$

The lower approximation function leads to the statistic

$$\mu^{\psi}_{*}(X) \overset{def}{=} \frac{|\underline{X}|}{|U|}. \tag{3}$$

We now define $\mu^{\psi *}(X) = 1 - \mu^{\psi}_{*}(-X)$, and it is easy to see that

$$\mu^{\psi *}(X) \overset{def}{=} \frac{|\overline{X}|}{|U|}. \tag{4}$$

We say that a probability measure p on 2^U is *compatible with* ψ, if

$$\mu^{\psi}_{*}(X) \leq p(X) \leq \mu^{\psi *}(X),$$

for all $X \in B_{\psi}$. Compatibility of p expresses the fact that $p(X)$ is within the bounds of uncertainty given by ψ. It is easy to see that the only probability measure on 2^U which is compatible to all functions μ^{ψ}_{*} is given by

$$p(X) = \frac{|X|}{|U|}, \tag{5}$$

so that $p(x) = \frac{1}{|U|}$ for all $x \in U$. In other words, RSDA assumes the *principle of indifference*, where in the absence of further knowledge all basic events are assumed to be equally likely. Unlike statistical models, RSDA does not model the dependency structure of attributes, but assumes that the principle of indifference is the only valid basis for an estimation of probability. If we assume marked dependencies among attributes, there may be better statistics than μ^{ψ}_{*} for the computation of $p(X)$, but even in this situation μ^{ψ}_{*} will remain a reasonable choice for $p(X)$.

The statistics derived from μ^{ψ}_{*} which is normally used in RSDA, the *approximation quality*, is defined as

$$\gamma_{\psi}(X) = \mu_{*}(X) + \mu_{*}(-X). \tag{6}$$

Clearly,

$$\gamma_{\psi}(X) \overset{def}{=} \frac{|\underline{X}_{\psi}| + |\underline{-X}_{\psi}|}{|U|}, \tag{7}$$

so that $\gamma_{\psi}(X)$ is the relative frequency of all elements of U which are correctly classified under the granulation of information by ψ with respect to being an element of X or not.

Generalising (7) to partitions induced by attribute sets, we define the *quality of an approximation* of a an attribute set Q with respect to an attribute set P by

$$\gamma(Q \to P) = \frac{|\bigcup\{X \in \mathcal{P}(Q) : X \text{ is } P\text{–deterministic}\}|}{|U|}. \tag{8}$$

The approximation is *perfect*, if $\gamma(Q \to P) = 1$; in this case, $\psi_Q \subseteq \psi_P$, and all Q, P – rules are deterministic.

If P is fixed, an attribute set which is $\subseteq$ – minimal with respect to $\gamma(Q \to P) = 1$ is called a *reduct* of P.

4 Significance testing

4.1 Rule significance

If we use RSDA for supervised learning, then its results must be controlled by statistical testing procedures; otherwise, we may read more into the results than what is actually in them. For example, if each Q, P – rule is based on a singleton class of ψ_Q – for example, a running number –, then the prediction of P (in fact, of any attribute set) will be perfect, but the rule will usually be rather useless for a different data sample. The underlying assumption on which prediction is based is that the information system $\mathcal{I}$ is a representative sample of the situation.

The assumption of representativeness is a problem of any analysis in most real life data bases. 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).

Table 1. State complexity

# of attr. values	# of attributes		
	10	20	30
	$\log_{10}$ (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

In [4] we have developed two procedures, both based on randomisation techniques, to compute the conditional probability of a rule $Q \to P$, assuming that the null hypothesis

H_0: "Objects are randomly assigned to rules"

is true. Randomisation procedures are particularly suitable to RSDA since they do not require outside information; in particular, it is not assumed that the information system under discussion is a representative sample.

Suppose that $\emptyset \neq Q, P \subseteq \Omega$, and that we want to evaluate the statistical significance of the rule $Q \to P$. Let Σ be the set of all permutations of U, and $\sigma \in \Sigma$. We define new attribute functions x^σ by

$$x^\sigma(a) \overset{def}{=} \begin{cases} x(\sigma(a)), & \text{if } x \in Q, \\ x(a), & \text{otherwise}. \end{cases}$$

The resulting information system $\mathcal{I}_\sigma$ permutes the Q–columns according to σ, while leaving the P–columns constant; we let Q^σ be the result of the permutation in the Q–columns, and $\gamma(Q^\sigma \to P)$ be the approximation quality of the prediction of P by Q^σ in $\mathcal{I}_\sigma$.

The value

$$p(\gamma(Q \to P)|H_0) := \frac{|\{\gamma(Q^\sigma \to P) \geq \gamma(Q \to P) : \sigma \in \Sigma\}|}{|U|!} \quad (1)$$

now measures the significance of the observed approximation quality. If $p(\gamma(Q \to P)|H_0)$ is low, traditionally below 5%, then the rule $Q \to P$ is deemed significant, and the (statistical) hypothesis "$Q \to P$ is due to chance" can be rejected. Otherwise, if $p(\gamma(Q \to P)|H_0) \geq 0.05$, we call $Q \to P$ a *casual rule.*

As an example, consider the following information system [4]:

U	x_1	x_2	d
1	0	0	0
2	0	1	1
3	1	0	2

The rule $\{x_1, x_2\} \to d$ is perfect, since $\gamma(\{x_1, x_2\} \to d) = 1$. Furthermore, $p(\gamma(\{x_1, x_2\} \to d)|H_0) = 1$, because every instance is based on a single observation, and thus, the rule is casual.

Now suppose that we have collected three additional observations:

U	x_1	x_2	d	U	x_1	x_2	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 given in (1) and their approximation qualities. The distribution of the approximation qualities of these 720 rules is given in Table 2, with $\alpha = p(\gamma(\{x_1, x_2\} \to d)|H_0)$. Given the 6–observation example, the probability of obtaining a perfect approximation of d by $\{x_1, x_2\}$ under the assumption of random matching, is 0.067 which is by far smaller than in the 3–observation example, but, using conventional $\alpha = 0.05$, not convincing enough to decide that the rule is sufficiently significant to be not casual.

Table 2. Results of randomisation analysis; 6 observ.

γ	No of cases	α	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	$1, 2, 2', 3, 1', 3'$

A small scale simulation study done in [4] indicates that the randomisation procedure has a reasonable power if the rule structure of the attributes is known.

We have applied the procedures to three well known data sets, and have found that not all claimed results, based on γ alone, can be called significant, and that other significant results were overlooked. Details and more examples can be found in [4].

4.2 Conditional casual attributes

In pure RSDA, the decline of the approximation quality when omitting one attribute is usually used to determine whether an attribute within a perfect rule $Q \to P$ is of high value for the prediction. This interpretation does not take into account that the decline of approximation quality may be due to chance.

As in the preceding section, our approach is to compare the actual $\gamma(Q \to P)$ with the results of a random system; here we randomise the value of a single attribute $t \in Q$ as follows: For each permutation σ of U we obtain a new attribute function $x^{\sigma,t}$ by setting

$$x^{\sigma,t}(a) \overset{def}{=} \begin{cases} x(\sigma(a)) & \text{if } x = t, \\ x(a), & \text{otherwise.} \end{cases}$$

Here, only the values in the t–column are permuted, and we denote the set of the resulting Q–granules by $Q^{\sigma,t}$. Now,

$$p_t(\gamma(Q \to P)|H_0) := \frac{|\{\gamma(Q^{\sigma,t} \to P) \geq \gamma(Q \to P) : \sigma \in \Sigma\}|}{|U|!} \tag{2}$$

measures the significance of attribute t within Q for the prediction of P. If $\alpha = p_t(\gamma(Q \to P)|H_0) \leq 0.05$, the assumption of (random) conditional casualness can be rejected; otherwise we shall call the attribute t *conditional casual within* Q.

The example given in Table 3 shows that, depending on the nature of an attribute, statistical evaluation leads to different expectations of the increase of approximation quality which is not visible under ordinary RSDA methods.

Table 3.

U	x	r_1	r_2	r_3	d	U	x	r_1	r_2	r_3	d
1	0	1	1	1	a	5	1	5	5	3	c
2	0	2	1	1	a	6	1	6	4	3	c
3	0	3	3	3	b	7	2	7	7	3	d
4	0	4	3	3	b	8	2	8	7	3	d

The prediction rule $x \to d$ has the approximation quality $\gamma(x \to d) = 0.5$. Assume that an additional attribute r is conceptualised in three different ways:

- A fine grained measure r_1 using 8 categories,
- A medium grained description r_2 using 4 categories, and
- A coarse description r_3 using 2 categories.

For $1 \leq i \leq 3$ we have $\gamma(\{x, r_i\}) \rightarrow d) = 1$, so that each of these approximations is perfect, and the drop of the approximation quality is 0.5 when r_i is left out. Therefore, we have a situation in which standard RSDA does not distinguish between the different properties of the additional attribute $r_i, 1 \leq i \leq 3$.

If we consider the expectation $E[\gamma(\{x, r_i\}^{\sigma, r_i} \rightarrow p)]$, we observe that

$$E[\gamma(\{x, r_1\}^{\sigma, r_1} \rightarrow p)] = 1.000,$$
$$E[\gamma(\{x, r_2\}^{\sigma, r_2} \rightarrow p)] = 0.880,$$
$$E[\gamma(\{x, r_3\}^{\sigma, r_3} \rightarrow p)] = 0.624.$$

Whereas the statistical evaluation of the additional predictive power differs for each of the three realizations of the new attribute r, the analysis of the decline of the approximation quality tells us nothing about these differences. Therefore, rather than using the decline of approximation quality as a global measure of influence, it is more appropriate to compare the influence of an attribute using the proposed statistical testing procedure.

4.3 Sequential significance testing

One can see that randomisation is a computationally expensive procedure, and it might be said that this fact limits its usefulness in practical applications. We have argued in [4] that, if randomisation is too costly for a data set, RSDA itself will not be applicable in this case, and have suggested several simple criteria to speed up the computations.

Another, fairly simple, tool to shorten the processing time of the randomisation test is the adaptation of a sequential testing scheme to the given situation. Because this sequential testing scheme can be used as a general tool in randomisation analysis, we present the approach in a more general way.

Suppose that θ is a a statistic with realizations θ_i, and a fixed realization θ_c. We can think of θ_c as $\gamma(Q \rightarrow P)$ and θ_i as $\gamma(Q^\sigma \rightarrow P)$. Recall that the statistic θ is called α *– significant*, if the true value $p(\theta \geq \theta_c|H_0)$ is smaller than α. Traditionally, $\alpha = 0.05$, and in this case, one speaks just of *significance*.

An evaluation of the hypothesis $\theta \geq \theta_c$ given the hypothesis H_0 can be done by using a sample of size n from the θ distribution, and counting the number k of θ_i for which $\theta_i \geq \theta_c$. The evaluation of $p(\theta \geq \theta_c|H_0)$ can now be done by the estimator $\hat{p}_n(\theta \geq \theta_c|H_0) = \frac{k}{n}$, and the comparison $\hat{p}_n(\theta \geq \theta_c|H_0) < \alpha$ will be performed to test the significance of the statistic. For this to work we have to assume that the simulation is asymptotically correct, i.e. that

$$lim_{n \rightarrow \infty} \hat{p}_n(\theta \geq \theta_c|H_0) = p(\theta \geq \theta_c|H_0). \tag{3}$$

In order to find a quicker evaluation scheme of the significance, it should be noted that the results of the simulation k out of n can be described by a binomial distribution with parameter $p(\theta \geq \theta_c|H_0)$. The fit of the approximation of $\hat{p}_n(\theta \geq \theta_c|H_0)$ can be determined by the confidence interval of the binomial distribution.

In order to control the fit of the approximation more explicitly, we introduce another procedure within our significance testing scheme. Let

$$H_b : \; p(\theta \geq \theta_c | H_0)) \in [0, \alpha) \tag{4}$$

$$H_a : \; p(\theta \geq \theta_c | H_0)) \in [\alpha, 1] \tag{5}$$

be another pair of statistical hypotheses, which are strongly connected to the original ones: If H_b holds, we can conclude that the test is α–significant, if H_a holds, we conclude that it is not.

Because we want to do a finite approximation of the test procedure, we need to control the precision of the approximation; to this end, we define two additional error components:

1. $r =$ probability that H_a is true, but H_b is the outcome of the approximative test.
2. $s =$ probability that H_b is true, but H_a is the outcome of the approximative test.

The pair (r, s) is called the *precision* of the approximative test. To result in a good approximation, the values r, s should be small (e.g. $r = s = 0.05$); at any rate, we assume that $r + s \lneq 1$, so that $\frac{s}{1-r} \lneq \frac{1-s}{r}$, which will be needed below.

Using the Wald-procedure [15], we define the likelihood ratio

$$LQ(n) = \frac{\sup_{p \in [0,\alpha)} p^k (1-p)^{n-k}}{\sup_{p \in [\alpha,1]} p^k (1-p)^{n-k}}, \tag{6}$$

and we obtain the following approximative sequential testing scheme:

1. If
   $$LQ(n) \lneq \frac{s}{1-r},$$
   then H_a is true with probability at most s.
2. If
   $$LQ(n) \gneq \frac{1-s}{r},$$
   then H_b is true with probability at most r.
3. Otherwise
   $$\frac{s}{1-r} \leq LQ(n) \leq \frac{1-s}{r},$$
   and no decision with precision (r, s) is possible. Hence, the simulation must continue.

With this procedure, which is implemented in our rough set engine GROBIAN[1] [3], the computational effort for the significance test in most cases breaks down dramatically, and a majority of the tests need less than 100 simulations.

[1] `http://www.infj.ulst.ac.uk/~cccz23/grobian/grobian.html`

5 Model selection

In conventional RSDA, the approximation quality $\gamma(Q \rightarrow d)$ is used as a conditional measure to describe prediction success of a dependent decision attribute d from a set Q of independent attributes. However, approximation qualities cannot be compared, if we use different attribute sets Q and R for the prediction of d.

To define an unconditional measure of prediction success, one can use the *minimum description length principle* [13] by 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 [5], we combine the principle of indifference with the maximum entropy principle (where worst possible cases are compared) to arrive at an objective method which combines feature selection with data prediction .

Suppose that R is any nonempty set of attributes and $\mathcal{P}(R) = \{X_i : 1 \leq i \leq k\}$. We define the *entropy of R* by

$$H(R) \overset{def}{=} \sum_{i=1}^{k} \frac{r_i}{n} \cdot \log_2(\frac{n}{r_i}),$$

where $r_i \overset{def}{=} \frac{|X_i|}{n}$.

Now, suppose that the classes of θ_Q are $X_0, \dots, X_m$, and that the probability distribution of the classes is given by $\hat{\pi}_i = \frac{|X_i|}{n}$; let $X_0, \dots X_c$ be the deterministic classes with respect to d, and W be their union.

Since our data is the partition obtained from Q, and we know the world only up to the equivalence classes of θ_Q, an indeterministic observation y is a result of a random process whose characteristics are totally unknown. Given this assumption, no information within our data set will help us to classify the element y, and we conclude that each such y requires a rule (or class) of its own. This results in a new partition of the object set associated with the equivalence relation θ_Q^+ defined by

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

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

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

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

$$H_{\text{rough}}(Q \to d) \overset{def}{=} H(\theta_Q^+) = \sum_i \hat{\psi}_i \cdot \log_2(\frac{1}{\hat{\psi}_i}).$$

To obtain an objective measurement we define the *normalised rough entropy* (NRE) by

$$\mathsf{NRE}(Q \to d) \overset{def}{=} 1 - \frac{H_{\text{rough}}(Q \to d) - H(d)}{\log_2(|U|) - H(d)}. \tag{2}$$

If the NRE has a value near 1, the entropy is low, and the chosen attribute combination is favourable, whereas a value near 0 indicates statistical casualness in the sense of Section 4. The normalisation 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 given a fixed decision attribute makes sense.

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[2], an acronym for Searching Optimal Rough Entropy Sets; SORES is implemented in Grobian [3].

In order to test the procedure, we have used 14 datasets available from the UCI repository[3] from which the appropriate references of origin can be obtained.

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; the mean error value is our measure of prediction success.

In Table 4 we compare the SORES results with the C4.5 performance given in [12]. Column 2 indicates how many attributes were used for prediction out of the total; for example, in the Annealing database, SORES has used 11 out of the 38 attributes.

The results indicate that SORES, even in its present unoptimised 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.

Details can be found in [5].

[2] All material relating to SORES, e.g. datasets, validation data, and a description of the algorithm can be obtained from the SORES website `http://www.psycho.uni-osnabrueck.de/sores/`

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

Table 4. Datasets and SORES validation

Dataset	SORES		C4.5 (Rel. 8)
Name	# attr.	Error	Error
Annealing	11/38	6.26	7.67
Auto	2/25	11.28	17.70
Breast-W	2/9	5.74	5.26
Colic	4/22	21.55	15.00
Credit–A	5/15	18.10	14.70
Credit–G	6/20	32.92	28.40
Diabetes	3/8	31.86	25.40
Glass	3/9	21.79	32.50
Heart–C	2/23	22.51	23.00
Heart–H	5/23	19.43	21.50
Hepatitis	3/19	17.21	20.40
Iris	3/4	4.33	4.80
Sonar	3/60	25.94	25.60
Vehicle	2/18	35.84	27.10
Std. Deviation		10.33	8.77

6 Probabilistic RSDA

RSDA concentrates on finding deterministic rules for the description of dependencies among attributes. Once a (deterministic) rule is found, it is assumed to hold without any error. If a measurement error is assumed to be an immeasurable part of the data – as e.g. statistical procedures do – the pure RSDA approach will not produce acceptable results, because "real" measurement errors cannot be explained by any rule.

In order to formulate a probabilistic version of RSDA, which is able to handle measurement errors as well, we enhance some of the concepts defined before. A *replicated decision system* $\mathcal{D}$ is a structure $\langle U, \Omega, Y, V_x \rangle_{x \in \Omega}$, where

- $\langle U, \Omega, V_x \rangle_{x \in \Omega}$ is an information system,
- $Y = \{y_1, \dots, y_S\}$ is a set of replicated decision attributes, explained more fully below.

We shall use the parameters of the information system $\mathcal{I}$ of p. 2; the set of Ω–granules is $G = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_M}\}$. In the sequel, we shall omit reference to Ω, if no confusion can arise; in particular, we will just speak of *granules* instead of Ω – *granules*.

Although not common in RSDA, the introduction of replicated decision variables offers the opportunity to control the effect of a measurement error: The smaller the agreement among multiple replications of the decision attribute, the more measurement error has to be assumed. This concept of replicated measurements is a way to estimate the reliability of, for example, psychometric tests, using the retest-reliability estimation, which in turn uses a linear model to estimate reliability and error of measurement as well.

With some abuse of notation, we assume that the decision attributes $y_1, \ldots, y_S$ are realizations of an unknown underlying distribution Y considered as a mapping

$$Y : U \times \{r_1, \ldots, r_Y\} \to [0, 1].$$

Y assign to each element a of U and each value r_j of the decision attribute the probability that $Y(a) = r_j$.

We suppose that (each replica of) the decision attribute takes the values $V_Y = \{r_1, r_2, \ldots, r_Y\}$. The classes of ψ_{y_t} are denoted by $M_{t,1}, \ldots, M_{t,r_Y}$; for each granule $\boldsymbol{x}_i$, we let $\xi(i, t, j)$ be the number of objects described by $\boldsymbol{x}_i$ which are in class $M_{t,j}$. In other words,

$$\xi(i, t, j) = |\{a \in U : \Omega(a) = \boldsymbol{x}_i \text{ and } a \in M_{t,j}\}| \tag{1}$$

We also let

$$\nu(\boldsymbol{x}_i) = |\{a \in U : \Omega(a) = \boldsymbol{x}_i\}|. \tag{2}$$

Clearly, $\sum_j \xi(i, t, j) = \nu(\boldsymbol{x}_i)$ for fixed i, and $\sum_{i,j} \xi(i, t, j) = |U|$. Each set $\{\xi(i, t, j) : 1 \leq t \leq s\}$ can be assigned an unknown value $\pi(i, j)$, which is the probability that an element $a \in U$ is assigned to a class $r_j where1 \leq j \leq r_Y$ and $\Omega(a) = \boldsymbol{x}_i$.

An example of the parameters of a decision system with one replica of the decision attribute is shown in Table 5.

Table 5. A decision system

$\boldsymbol{x}_i$	Ω		$Y = r_1$	$Y = r_2$	$\nu(\boldsymbol{x}_i)$
	x_1	x_2	$\xi(i, 1, 1)$	$\xi(i, 1, 2)$	
$\boldsymbol{x}_1$	0	1	5	1	6
$\boldsymbol{x}_2$	1	0	2	8	10
	Σ		7	9	16

The example given in Table 5 shows that indeterministic rules alone do not use the full information given in the database. There is no deterministic rule to predict a value of the decision attribute y_1, given a value of the independent attributes $\langle x_1, x_2\rangle$: Both indeterministic rules will predict both possible outcomes in the decision variable. The pure rough set approach now concludes that no discernible assignment is possible. But if we inspect the table, we see that the error of assigning $\langle 0, 1\rangle$ to 1 is small (1 observation) and that $\langle 1, 0\rangle \mapsto 2$ is true up to 2 observations.

There are several possibilities to reduce the precision of prediction to cope with measurement error. One possibility is the so called *variable precision rough set model* [17], which assumes that rules are only valid within a certain part of the population. The advantage of this approach is that it uses only two parameters (the precision parameter and γ) to describe the quality of a rule system; the disadvantages are that precision and γ are partially exchangeable, and that there is no theoretical background to judge which combination is best suited to the data.

Another approach – based upon standard statistical techniques – is the idea of predicting random variables instead of fixed values. Conceptually, each realization of the distribution Y can be described by a mixture

$$Y = \sum_{1 \leq r \leq R} \omega_r Y_r, \tag{3}$$

with $\sum_r \omega_r = 1$, based on an index R of unknown size, and unknown basic distributions Y_r with unknown weights ω_r.

If we use the granules $\boldsymbol{x}_j$ to predict Y, the maximal number R of basic distributions is bounded by the number M of granules; equality occurs just when each granule $\boldsymbol{x}_j$ determines its own Y_j. In general, this need not to be the case, and it may happen that the same Y_j can be used to predict more than one granule; this can be indicated by an onto function

$$g : \{1, ..., M\} \twoheadrightarrow \{1, ..., R\},$$

mapping the (indices of) the granules to a smaller set of mixture components of Y.

Probabilistic prediction rules are of the form

$$\boldsymbol{x}_j \rightarrow Y_{g(j)},\ 1 \leq j \leq M,$$

where each $Y_{g(j)} : V_Y \rightarrow [0, 1]$ is a random variable. If the probabilities are understood, we shall often just write $\boldsymbol{x} \rightarrow Y$, with Y possibly indexed, for the rule system $\langle \boldsymbol{x}_j \rightarrow Y_{g(j)} \rangle_{1 \leq j \leq M}$.

In the example of Table 5 there are two possibilities for R, and we use maximum likelihood to optimise the binomial distribution, the application of which is straightforward, if we additionally assume that the observations stem from a simple sampling scheme. In case $R = 1$, both granules use the same distribution Y_1. In this case, the likelihood function $L_1 = L(Y_1 | \langle 0, 1 \rangle, \langle 1, 0 \rangle)$ is given by

$$L_1 = \binom{16}{9} \pi^9 (1 - \pi)^7 \tag{4}$$

which, as expected, has a maximum at $\hat{\pi} = \frac{9}{16}$. This leads to the rule system

$$\langle 0, 1 \rangle \text{ or } \langle 1, 0 \rangle \rightarrow \{\langle 1, \tfrac{9}{16} \rangle, \langle 2, \tfrac{7}{16} \rangle\}. \tag{5}$$

If $R = 2$, the samples belonging to $\langle 0, 1 \rangle$ and $\langle 1, 0 \rangle$ are assumed to be different in terms of the structure of the decision attribute, and the likelihood of the sample has to be built from the product of the likelihoods of both subsamples. If we have the rules $\langle 0,1\rangle \to Y_1$ and $\langle 1,0\rangle \to Y_2$, then

$$L_2 = \binom{6}{1}\pi_1^1(1-\pi_1)^5\binom{10}{8}\pi_2^8(1-\pi_2)^2. \tag{6}$$

Using standard calculus, the maximum of L_2 is $(\hat{\pi}_1 = \frac{1}{6},\ \hat{\pi}_2 = \frac{8}{10})$, which gives us the rule system

$$\begin{cases} \langle 0,1\rangle \to \{\langle 1,\frac{5}{6}\rangle, & \langle 2,\frac{1}{6}\rangle\}, \\ \langle 1,0\rangle \to \{\langle 1,\frac{2}{10}\rangle, & \langle 2,\frac{8}{10}\rangle\}. \end{cases} \tag{7}$$

In going from from L_1 to L_2 we change the sampling structure – the estimation of L_2 needs 2 samples, whereas L_1 needs only one sample – and we increase the number of probability parameters π_i by one.

Changing the sampling structure is somewhat problematic, because comparison of likelihoods can only be done within the same sample. A simple solution is to compare the likelihoods based on elements, thus omitting the binomial factors. Because the binomial factors are unnecessary for parameter estimation (and problematic for model comparison) they will be skipped in the sequel. Letting

$$L_1(\max) = \hat{\pi}^9(1-\hat{\pi})^7 \qquad = 0.0000173, \tag{8}$$

$$L_2(\max) = \hat{\pi}_1^1(1-\hat{\pi}_1)^6\hat{\pi}_2^8(1-\hat{\pi}_2)^2 \qquad = 0.0003746, \tag{9}$$

we have to decide which rule offers a better description of the data. Although $L_2(\max)$ is larger than $L_1(\max)$, it is not obvious to conclude that the two rules are really ‘essentially’ different, because the estimation of L_2 depends on more free parameters than L_1.

There are – at least – two standard procedures for model selection, which are based on the likelihood and the number of parameters: The Akaike Information Criterion (AIC) [1] and the Schwarz Information Criterion (SIC) [14]. If $L(\max)$ is the maximum likelihood of the data, P the number of parameters, and K the number of observations, these are defined by

$$AIC = 2(P - \ln(L(\max))) \tag{10}$$

$$SIC = 2\left(\frac{\ln(K)}{2}\cdot P - \ln(L(\max))\right). \tag{11}$$

The lower AIC (and SIC respectively), the better the model. AIC and SIC are rather similar, but the penalty for parameters is higher in SIC then in AIC.

In the example, we have used one parameter π to estimate L_1. Therefore,

$$AIC(L_1(\max)) = 2(1 - \ln(0.0000173)) \qquad = 23.930,$$

$$SIC(L_1(\max)) = 2(\frac{\ln(16)}{2} - \ln(0.0000173)) = 24.702.$$

There are three free parameters to estimate L_2: First, the probabilities π_1, π_2; furthermore, one additional parameter is used, because we need to distinguish between the two granules. Therefore,

$$AIC(L_2(\max)) = 2(3 - \ln(0.0003746)) \qquad =21.779$$
$$SIC(L_2(\max)) = 2(3\ln(16) - \ln(0.0003746))=24.090.$$

and we can conclude that the rule – system (7) is better suited to the data then the simple 1-rule – system (5).

6.1 Finding probabilistic rules

The algorithm of finding probabilistic rules starts by searching for the optimal granule mapping based on a set Ω of (mutually) predicting attributes and a set Y of replicated decision attributes.

```
R=0;
$\Delta(AIC) = +\infty$;
While R $<$ M and $\Delta(AIC) > 0$ do
    R=R+1;
    Compute the best mapping $g : \{1, \dots, M\} \to \{1, \dots, R\}$ in
    terms of the product of the maximum likelihood of the $Y$ replicas;
    Compute number of parameters;
    Compute $AIC_R$;
    if ($R > 1$) then $\Delta(AIC) = AIC_{R-1} - AIC_R$;
    else $\Delta(AIC) = AIC_1$;
```

Finding the best mapping g is a combinatorial optimisation problem, which can be approximated by hill-climbing methods, whereas the computation of the maximum likelihood estimators, given a fixed mapping g, is straightforward: One computes the multinomial parameters $\hat{\pi}_t(i_k)$ of the samples i defined by g for every replication y_t of Y and every value $r_k \in \{r_1, \dots, r_Y\}$, and computes the mean value

$$\hat{\pi}(i_k) = \frac{\sum_{t=1}^{s} \hat{\pi}_t(i_k)}{s}, \tag{12}$$

from which the likelihood can be found. The number of parameters (np) depends on R and r_Y because

$$np = R \times r_Y - 1;$$

the computation of the AIC is now possible.

The result of algorithm offers the most parsimonious description of a probabilistic rule system (in terms of AIC). In order to reduce the number of independent attributes within the rules, a classical RSDA reduct analysis of these attributes can be applied, using the results of the mapping g as a decision attribute.

6.2 Application: Unsupervised learning and nonparametric distribution estimates

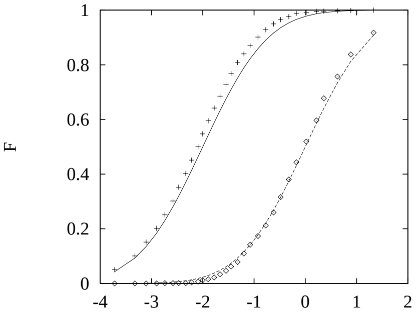
The most interesting feature of the probabilistic granule approach is that the analysis can be used for clustering, i.e. unsupervised learning. In this case the predicting attribute is the identity and any granule consists of one element. If we use more than one replication of the decision attribute, it will be possible to estimate the number of mixture components of Y and the distribution of the mixtures.

The Figures 2 and 3 show the result of the mixture analysis based on granules using the mixture

$$Y = \frac{1}{2}N(-2.0, 1.0) + \frac{1}{2}N(0.0, 1.0). \tag{13}$$

$N(\mu, \sigma)$ is the normal distribution with parameters μ and σ. 1024 observations per replication were simulated; one simulation was done with 2 replications (Figure 2), and another with 5 replications (Figure 3). The simulated data were grouped into 32 intervals with approximately the same frequencies in the replications, and the searching algorithm outlined in section 6.1 was applied.

Fig. 2. Nonparametric estimates of a $\frac{(N(-2,1)+N(0,1))}{2}$ mixture distribution (2 replications; lines denotes theoretical distributions)



The result shows that the underlying distributions can be approximated quite successfully, although

- No parametric distributional assumption was used,
- Y has a inimical shape,
- Only a few replications were considered.

The next numerical experiment was performed with the famous Iris data [7]. These were used by Fisher to demonstrate his discriminant analysis; it consists of 50 specimen of each of the Iris species *Setosa*, *Versicolor*, and *Virginica*, which were measured by Sepal length, Petal length, Sepal width, Petal width. It is well known (e.g. [2]) that Sepal width attribute is not very informative; therefore we shall skip it for the subsequent analysis.

Fig. 3. Nonparametric estimates of a (N(-2,1)+N(0,1))/2 mixture distribution (5 replications; lines denotes theoretical distributions)

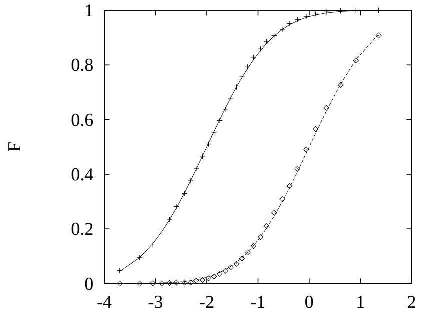
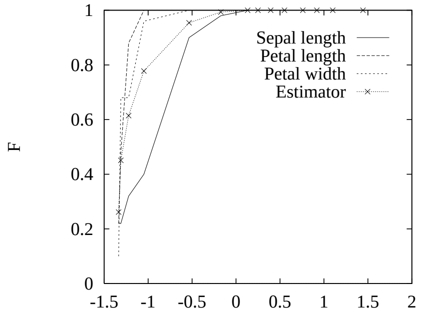


Fig. 4. Setosa distributions of 3 attributes and its estimation



If we assume that the three remaining attributes measure the same variable up to some scaling constants, we can use the z–transformed attributes as a basis for the analysis. The unsupervised AIC search algorithm clearly votes for three classes in the unknown joint dependent attribute. If we use the estimated distribution functions (Figures 4, 5, 6) for the classification of the elements, we find a classification quality of about 85%, which is not too bad for an unsupervised learning procedure.

Table 6. Iris: Classification results

Setosa	50	0	0
Versicolor	7	41	2
Virginica	0	14	36

The procedure does not offer only classification results, but also estimators of the distributions of dependent attributes within the groups without having a prior knowledge about the group structure. The Figures 4, 5, 6 compare three estimated distributions with the respective the distributions of three (normalised) variables within the groups. The results show that the “Sepal length” attribute does not fit very well and that the estimated distributions summarise this aspect of both "Petal" measures.

Fig. 5. Versicolor distributions of 3 attributes and its estimation

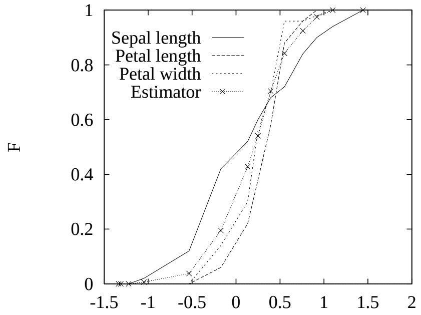
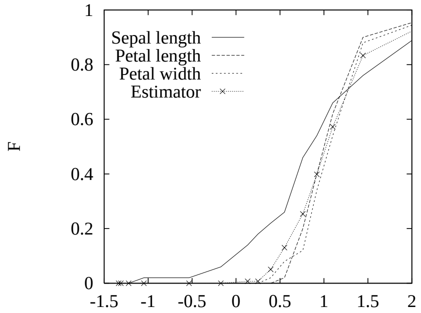


Fig. 6. Virginica distributions of 3 attributes and its estimation



7 Summary and outlook

The paper has introduced a framework for applying statistical tools in concept forming and classification using rule based data analysis. Three different aspects were discussed:

1. Significance testing of rules, rule systems and parts of rules systems can be performed by randomisation procedures, which can be sped up by sequential testing plans.
2. The application of tailored rule based definitions of entropy, which are compatible to the non-numerical philosophy of RSDA, achieves machine learning procedures with excellent reclassification behaviour.

3. The concept of probabilistic granule analysis enables the researcher to use a noise reducing algorithm in advance to standard rule based data analysis. Probabilistic granule analysis itself turns out to be a valuable tool for unsupervised learning and non-parametric distribution estimation.

Whereas significance testing uses the same theoretical assumptions as the classical RSDA, the rough entropy based method SORES and the probabilistic granule analysis (as well as other approaches such as the variable precision model [17]) allow some error within prediction rules. Although all of these are RSDA based, from a strictly modelling point of view, these methods are partially incompatible competitors; their particular strengths and weaknesses need to be determined by further investigation.

Bibliography

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Cáski (Eds.), *Second International Symposium on Information Theory*, 267–281, Budapest. Akademiai Kaidó. Reprinted in *Breakthroughs in Statistics*, eds Kotz, S. & Johnson, N. L. (1992), volume I, pp. 599–624. New York: Springer.

[2] Browne, C., Düntsch, I. & Gediga, G. (1998). IRIS revisited: A comparison of discriminant and enhanced rough set data analysis. In L. Polkowski & A. Skowron (Eds.), *Rough sets in knowledge discovery, Vol. 2*, 345–368, Heidelberg. Physica–Verlag.

[3] Düntsch, I. & Gediga, G. (1997a). The rough set engine GROBIAN. In A. Sydow (Ed.), *Proc. 15th IMACS World Congress, Berlin*, vol. 4, 613–618, Berlin. Wissenschaft und Technik Verlag.

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

[5] Düntsch, I. & Gediga, G. (1998). Uncertainty measures of rough set prediction. *Artificial Intelligence*, **106**, 77–107.

[6] Düntsch, I. & Gediga, G. (2000). Rough set data analysis. In *Encyclopedia of Computer Science and Technology*. Marcel Dekker. To appear.

[7] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179–188.

[8] Gediga, G. & Düntsch, I. (1999). Probabilistic granule analysis. Draft paper.

[9] Gigerenzer, G. (1981). Messung und Modellbildung in der Psychologie. Basel: Birkhäuser.

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

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

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

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

[14] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

[15] Wald, A. (1947). Sequential Analysis. New York: Wiley.

[16] Zadeh, L. A. (1994). What is BISC? `http://http.cs.berkeley.edu/projects/Bisc/bisc.memo.html`, University of California.

[17] Ziarko, W. (1993). Variable precision rough set model. *Journal of Computer and System Sciences*, **46**.