

# IRIS revisited: A comparison of discriminant and enhanced rough set data analysis

Ciarán Browne<sup>1\*</sup>, Ivo Düntsch<sup>1\*</sup>, Günther Gediga<sup>2\*</sup>

<sup>1</sup> School of Information and Software Engineering, University of Ulster, Newtownabbey, BT 37 0QB, N.Ireland. E-mail: {C. Browne, I. Düntsch}@ulst.ac.uk

<sup>2</sup> FB Psychologie / Methodenlehre, Universität Osnabrück, 49069 Osnabrück, Germany. E-mail: ggediga@luce.psych.uni-osnabrueck.de

## 1 Introduction

Rough set data analysis (RSDA) was introduced to Computer Science in the early 1980s by Z. Pawlak [Pawlak(1982)] and has since come into focus as an alternative to the more widely used methods of machine learning and statistical data analysis. A good overview of the state of the art are *Fundamenta Informaticae*, Vol. 27 (1996), and [Lin & Cercone(1997)].

Just like other new approaches, RSDA needs to show that its methods are as good as or even superior to commonly used – mainly statistical – data analysis strategies. Even though singular attempts have been made to compare RSDA to other approaches, e.g. [Wong et al.(1986)Wong, Ziarko & Ye, Teghem & Benjelloun(1992), Krusińska et al.(1992a)Krusińska, Babic, Słowiński & Słowiński, Krusińska et al.(1992b)Krusińska, Słowiński & Stefanowski], a systematic comparison is as yet missing.

In this paper, we use the IRIS data set to compare the ROUGHIAN extension of RSDA developed by two of the authors [Düntsch & Gediga(1997c)] with Fisher’s discriminant analysis method, and exhibit some general principles regarding the power of the two approaches. This need of further comparison arises, because the methods of [Düntsch & Gediga(1997c)] enable the researcher to treat quantitative attributes in an “RSDA compatible” way, which could not be done in the previous comparison studies.

The structure of this paper is as follows: To make the paper self contained, we first briefly describe Fisher’s discriminant analysis and its application to the IRIS data set, and then proceed to highlight the main points of the ROUGHIAN method.

Section 4 gives an RSDA analysis of the IRIS data set, and comments on earlier comparisons. In Sect. 5 we present the ROUGHIAN analysis of the IRIS data, as well as validation and testing procedures of prediction.

## 2 IRIS Data: The Historical Perspective

### 2.1 Fisher’s Discriminant Analysis

Suppose that we have a situation where we have cases or subjects divided into groups, and quantitative attributes which should predict the group membership; it was Fisher [Fisher(1936)] who discovered *discriminant analysis* (DA), which enables the researcher to find linear combinations of the predicting attributes – called *canonical discriminant functions* (CDF) – which best characterize the differences between the groups. Once one has found these characterizations, one can compute the differences between any object and every centroid of the group means in terms of the canonical classification functions; one is able to assign every object – and even new objects – to the group whose centroid is nearest to the

---

\* Equal authorship implied

coordinates of the object. Other assignment procedures are also possible: For example, if we assume that the prediction attributes can be described by the same multivariate normal density within any class of the decision attribute, the assignment can be done by choosing the group which maximizes the likelihood  $f(object = i | group = j)$ .

Nowadays – more than 60 years later – discriminant analysis has turned into a class of methods using the same spirit and some ideas of the original Fisherian one. In order to compare RSDA results with results of “the” discriminant analysis, we have to specify what kind of discriminant analysis we shall use. Our idea is that we take a *modal type analysis*, which uses the following underpinnings:

- The predicting attributes are quantitative variables and no data transformation is done before entering the discriminant analysis process.
- It is assumed that the within-group covariance matrices of the predicting attributes are identical. Although this assumption puts severe restrictions on the data, the “modal” DA is run with this restriction.
- We use the simple centroid classification rule to (re)classify objects to classes.

## 2.2 Fisher’s IRIS Data

The data used by Fisher to demonstrate his discriminant analysis consists of 50 specimen of each of the iris species *Setosa*, *Versicolor*, and *Virginica*, measured by the features given in Tab. 1.

**Table 1.** IRIS Data

No	Attribute	Range in mm	No	Attribute	Range in mm
<b>A1</b>	Sepal length	$43 \leq x \leq 79$	<b>A3</b>	Petal length	$10 \leq x \leq 69$
<b>A2</b>	Sepal width	$22 \leq x \leq 44$	<b>A4</b>	Petal width	$1 \leq x \leq 25$

Applying DA to the data, it turns out that there are two canonical discriminant functions necessary to describe the differences between the groups. It is well known that petal length is the most prominent variable to constitute the first CDF, which is indicated by the highest pooled-within-groups correlations between petal length and the first CDF (Tab. 2). Petal width and sepal width turned out to be equally important, whereas the sepal length has no remarkable impact on the CDF.

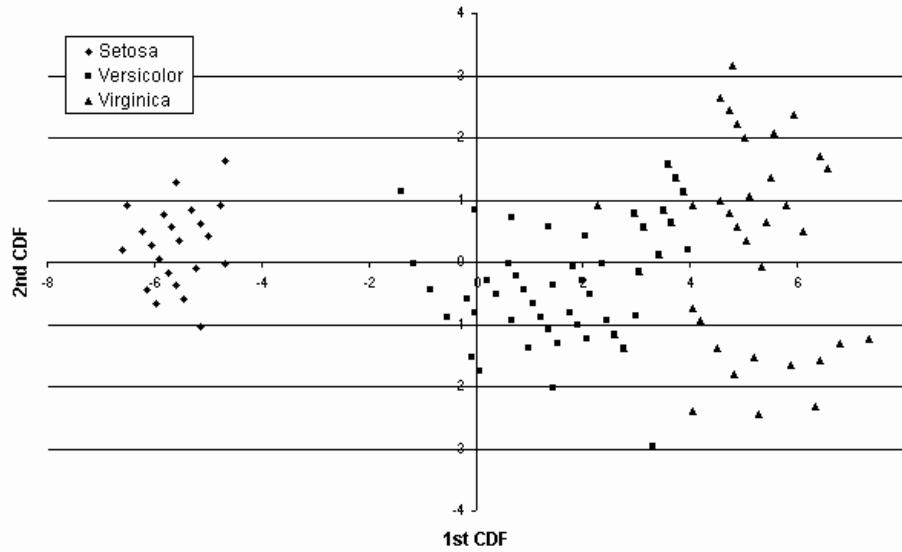
**Table 2.** Pooled-within-groups Correlations between Discriminating Variables and Canonical Discriminant Functions

	CDF 1	CDF 2
Petal length	.73 (.91)	.19 (-.41)
Petal width	.65 (.81)	.72 (.58)
Sepal length	.24	.34
Sepal width	-.13	.87

In order to be compatible with the results of the RSDA, we choose the pair (*petal length*, *petal width*) as attributes for further analysis. Reclassification using these two variables works very well, as the results of Tab. 3 indicate. A geometrical interpretation of the DA can be given by plotting each case as a point in the space built by the axes of the two CDFs. Figure 1 shows the data projected into the space of the two CDFs based on petal length and petal width.

**Table 3.** Classification Results Using Bayesian Reclassifier

Predicted classes	IRIS species		
	Setosa	Versicolor	Virginica
Setosa	1.000	0.000	0.000
Versicolor	0.000	0.960	0.080
Virginica	0.000	0.040	0.920

**Fig. 1.** The Space of Two CDFs

### 3 ROUGHIAN – Rough Information Analysis

[Dütsch & Gediga(1997c)] have developed a *rough information analysis* (ROUGHIAN) which enhances traditional rough set data analysis by three additional procedures, namely,

- Significance testing,
- Data filtering
- Uncertainty measuring.

In this section we shall describe these features as well as our notation of RSDA. To make this chapter more self contained, some of the material of this section was taken from [Dütsch & Gediga(1996)] and [Dütsch & Gediga(1997c)]. Further applications of the ROUGHIAN model can be found in [Browne(1997)] and [Dütsch et al.(1997)Dütsch, Gediga & Rogner]. All computations were done using the rough set engine Grobian [Dütsch & Gediga(1997b)].

#### 3.1 Basics

We assume that the reader is familiar with the philosophy and the basic terms of RSDA, so that we will just outline our notation and definitions.

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  $\Omega$  of attributes,
3. For each  $q \in \Omega$ 
  - (a) A set  $V_q$  of attribute values,
  - (b) An information function  $f_q : U \rightarrow V_q$ ,

We extend the information functions  $f_q$  to functions  $f_Q$ ,  $Q \subseteq \Omega$  in the canonical way.

The subsets  $Q$  of  $\Omega$  can be used to define the indiscernibility relations  $\theta_Q$  by

$$x\theta_Q y \stackrel{\text{def}}{\iff} (\forall q \in Q)(f_q(x) = f_q(y)).$$

If  $x \in X$ , then the class of  $x$  with respect to  $\theta_Q$  is written as  $Q[x]$ , and the set of classes of  $\theta_Q$  is denoted by  $K_Q$ .

If  $P, Q \subseteq \Omega$  we call a class  $X$  of  $\theta_Q$  *P-deterministic* (or just *deterministic* if  $P$  is understood), if it is contained in a class of  $\theta_P$ . Each such class induces a rule of the form

$$f_Q(x) = \bar{t} \Rightarrow f_P(x) = \bar{s},$$

where  $\bar{t}$  and  $\bar{s}$  are the feature vectors of  $x$  determined by  $Q$ , resp.  $P$ .

If  $X \in K_Q$  intersects  $Y_0, \dots, Y_k \in K_P$ ,  $k > 0$ , then we obtain an indeterministic rule

$$f_Q(x) = \bar{t} \Rightarrow f_P(x) = \bar{s}_0 \vee \dots \vee f_P(x) = \bar{s}_k.$$

We write  $Q \rightarrow P$  for the conjunction of all deterministic and indeterministic rules obtained this way, and – with some abuse of language – call  $Q \rightarrow P$  a rule as well. Strictly speaking, we should consider equivalence classes of rules, but we will not do this, as it is clear what we mean.

### 3.2 Rough Sets and Statistics

While RSDA is a non-numeric method of data analysis, it implicitly makes statistical assumptions which we want to explore in this section. We first review briefly some properties of finite general statistics. A *probability space* is a triple  $\langle U, B, p \rangle$ , where  $U$  is a finite non-empty set,  $B$  a Boolean subalgebra of  $\langle \mathfrak{P}(U), \cap, \cup, -, \emptyset, U \rangle$ , and  $p$  a probability measure on  $B$ , i.e. a function  $p : B \rightarrow [0, 1]$  which satisfies the Kolmogorov axioms

1.  $p(\emptyset) = 0$ ,
2.  $p(U) = 1$ ,
3.  $p(\bigcup_{i \in I} X_i) = \sum_{i \in I} p(X_i)$ , if each  $X_i \in B$ , and the sets  $X_i$  are pairwise disjoint.

If  $B$  is a proper subalgebra of  $\mathfrak{P}(U)$ , then the function  $p$  is not defined on all of  $\mathfrak{P}(U)$ ; there are two standard ways to extend  $p$  over all of  $\mathfrak{P}(U)$ , see e.g. [Halpern & Fagin(1992)]:

$$p_*(Y) = \sup \{p(X) : X \subseteq Y, X \in \mathcal{P}\}, \quad (\text{Inner measure}) \quad (1)$$

$$p^*(Y) = \inf \{p(X) : X \supseteq Y, X \in \mathcal{P}\}. \quad (\text{Outer measure}) \quad (2)$$

Suppose that  $\langle U, \theta \rangle$  is an approximation space (i.e.  $U$  is a nonempty set and  $\theta$  an equivalence relation on  $U$ ), and that  $\mathcal{P}$  is the partition associated with  $\theta$ . The *lower*, resp. *upper approximation* of  $X$  by  $\theta$  is defined by

$$\underline{X}_\theta \stackrel{\text{def}}{=} \bigcup \{Y \in \mathcal{P} : Y \subseteq X\},$$

resp.

$$\overline{X}^\theta \stackrel{\text{def}}{=} \bigcup \{Y \in \mathcal{P} : Y \cap X \neq \emptyset\}.$$

The metrics of  $\langle U, \theta \rangle$  are the two “approximation functions”

$$\gamma_\theta(X) \stackrel{def}{=} \frac{|X_\theta| + |-X_\theta|}{|U|}, \quad (3)$$

$$\alpha_\theta(X) \stackrel{def}{=} \frac{|X_\theta|}{|X^\theta|} \text{ (for } X \neq \emptyset\text{)}, \quad (4)$$

see [Pawlak(1991)], p. 16ff. If  $\theta$  is understood, we shall usually omit the subscripts.

Usually, one interprets  $\gamma(X)$  as the percentage of objects of  $U$  which can be correctly classified with the knowledge given by  $\theta$  as being in  $X$  or not, while  $\alpha(X)$  expresses the degree of completeness of our knowledge of  $X$ .

We define two associated statistics for  $\langle U, \theta \rangle$  by

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

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

$$\gamma(X) = \mu_*(X) + \mu_*(-X), \quad \alpha(X) = \frac{\mu_*(X)}{\mu^*(X)},$$

so that we can regard  $\mu_*$  as the basic measure of RSDA.

For each equivalence  $\theta$  on  $U$  we let  $B_\theta$  be the subalgebra of  $\langle \mathfrak{P}(U), \cap, \cup, -, \emptyset, U \rangle$  whose atoms are the classes of  $\theta$ . Now, the restriction  $\mu_* \upharpoonright B_\theta$  is a probability measure on  $B$  whose inner measure is just  $\mu_*$ , and the measurable sets of  $\langle U, B_\theta, \mu_* \upharpoonright B \rangle$  are just the  $\theta$ -definable sets. Following [Halpern & Fagin(1992)], we say that a probability measure  $p$  on  $B_\theta$  is *compatible with  $\theta$* , if

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

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

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

so that  $p(x) = \frac{1}{|U|}$  for all  $x \in U$ . In other words, rough set theory assumes the *random world model* described in [Bacchus et al.(1994)Bacchus, Grove, Halpern & Koller], also called the *principle of indifference*, where in the absence of further knowledge all basic events are assumed to be equally likely.

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

- *Rough set analysis neglects the underlying joint distributions of the attributes and the reported statistics  $\mu_*$ , resp.  $\gamma$ , are sufficient only if the joint distributions of the attributes are constant as in (5).*

This sounds like a drawback, but one should note that rough set analysis is applied (and applicable!) in a “few – objects – many – attributes” situation which is very different to the situations usually encountered in statistical modeling. In the field of applied regression analysis it was shown that in comparable situations the assumption “simple is better” – e.g. using 0–1 regression weights – results in more stable estimates than using an approach with many parameters [Cohen(1990)].

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

$$\gamma(Q, \mathcal{P}) = \frac{\sum_{X \in \mathcal{P}} |X_{\theta_Q}|}{|U|}, \quad (6)$$

thus generalizing  $\gamma_\theta$  of (3) to include partitions with more than two classes. In case  $\mathcal{P}$  is induced by  $\theta_P$  for some  $P \subseteq \Omega$ , we will write  $\gamma(Q \rightarrow P)$  instead of  $\gamma(Q, \mathcal{P})$  to indicate that  $\gamma$  measures the approximation quality of the rule  $Q \rightarrow P$ . It is not hard to see that

$$\gamma(Q \rightarrow P) = \frac{|\bigcup\{M \in K_Q : M \text{ is } P\text{-deterministic}\}|}{|U|}$$

If  $\gamma(Q \rightarrow P) = 1$ , we call  $P$  *dependent on*  $Q$ , and write  $Q \Rightarrow P$ . This is the case exactly when  $\theta_Q \subseteq \theta_P$ .

### 3.3 Significance Testing

We can use the approximation quality defined in (6) as an internal index of a rule  $Q \rightarrow P$ . If  $Q \Rightarrow P$ , then the prediction is perfect, otherwise,  $\gamma(Q \rightarrow P) < 1$ . However, a perfect or high approximation quality is not a guarantee that the rule is valid. If, for example, the rough set method discovers a rule  $Q \rightarrow P$  which is based on only a few observations – which one might call a *casual rule* – the approximation quality of the rule may be due to chance. Thus, the validity of inference rules for prediction must be validated by statistical techniques – otherwise, application beyond attribute reduction in the concrete situation might as well be done by throwing bones into the air and observing their pattern. We are certainly not the first to observe this phenomenon:

“Consider a dataset in which there is a nominal attribute that uniquely identifies each example . . . Using this attribute one can build a 1 – rule that classifies a given training set 100% correctly: needless to say, the rule will not perform well on an independent test set” [Holte(1993)].

Thus, although rough set theory uses a only few parameters which need simple statistical estimation procedures (e.g. the cardinalities of equivalence classes and the associated probability function on its partition), the validity of obtained rules should be controlled using statistical testing procedures, in particular, when they are used for modeling and prediction of events.

[Düntsch & Gediga(1997e)] have developed two simple procedures, both based on randomization techniques, which evaluate the validity of a rule based on the approximation quality of attributes. These procedures seem to be particularly suitable for the soft computing approach of RSDA since they do not require information from outside the data under consideration; in particular, it is not assumed that the information system under discussion is a representative sample. The reader is invited to consult [Edgington(1987)] or [Manly(1991)] for the background and justification of randomization techniques in these situations.

Let  $\Sigma$  be the set of all permutations of  $U$ ,  $\sigma \in \Sigma$ , and suppose that we want to test the significance of  $Q \rightarrow P$ . We define new information functions  $f_r^{\sigma(P)}$  by

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

The resulting information system  $\mathcal{I}_\sigma$  permutes the  $P$ -columns according to  $\sigma$ , while leaving the  $Q$ -columns constant. We now use the permutation distribution  $\{\gamma(Q \rightarrow \sigma(P)) : \sigma \in \Sigma\}$

to evaluate the strength of the prediction  $Q \rightarrow P$ . The value  $p(\gamma(Q \rightarrow P)|H_0)$  measures the extremeness of the observed approximation quality and it is defined by

$$p(\gamma(Q \rightarrow P)|H_0) := \frac{|\{\sigma \in \Sigma : \gamma(Q \rightarrow \sigma(P)) \geq \gamma(Q \rightarrow P)\}|}{|U|!} \quad (7)$$

If  $\alpha = p(\gamma(Q \rightarrow P)|H_0)$  is low, traditionally below 5%, then the rule  $Q \rightarrow P$  is deemed significant, and the (statistical) hypothesis “ $Q \rightarrow P$  is due to chance” can be rejected.

One can see that the procedure is computationally expensive, and that it is not always feasible (or, indeed, possible) to exactly compute  $\alpha$ . However, a randomly chosen set of permutations will usually be sufficient: It is known [Dwass(1957)] that the significance level of a randomization test is in a sense exact even when the randomization distribution is only sampled.

In rough set analysis, the decline of the approximation quality when omitting one attribute is normally used to determine whether an attribute within a reduct is of high value for the prediction. However, this view does not take into account that the decline of approximation quality may be due to chance. This observation leads to the following definition: We call an attribute  $q \in Q$  *conditional casual*, if there are only a few observations in which the attribute  $q$  is needed to predict  $P$ . More precisely, the statistical approach is to compare the actual  $\gamma(Q \rightarrow P)$  with the results of a random system: For each permutation  $\sigma$  of  $U$  and each  $q \in Q$  we obtain a new information function  $f^{\sigma,q}$  by setting

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

The resulting approximation quality of  $P$  by  $Q$  is denoted by  $\gamma(Q, \sigma(q) \rightarrow P)$ , and we define  $p(\gamma(Q, q \rightarrow P)|H_0)$  in analogy to (7) and call it the *relative significance* of  $q$  within  $Q$ .

As above, if  $p(\gamma(Q, q \rightarrow P)|H_0)$  is below 5%, the assumption of (random) conditional casualness can be rejected, otherwise we will call the attribute *conditional casual within  $Q$* , or just *conditional casual*, if  $Q$  is understood.

### 3.4 Data Filtering

As we seen in the previous section, if the granularity of an information system is high, it may lead to rules which are based on a few observations only, and thus, their validity is doubtful. In this case, the  $\alpha$  value will be high, and the rule may be due to chance. Thus, rough set analysis as a conditional method needs a preprocessing step in which unnecessary granularity is removed, but in which no essential (dependency) information is lost. One way to increase the significance is to reduce the granularity of information by using appropriate data filters on the sets  $V_q$ , which may reduce the number of classes of  $\theta_Q$  while at the same time keeping the dependency information.

[Düntsch & Gediga(1997d)] develop a simple data filtering procedure which is compatible with the rough set approach and which may result in an improved significance of rules.

The main tool are ‘binary information systems’. These are those systems, in which every attribute has exactly two values. Roughly speaking, we obtain a binary system  $\mathcal{I}^B$  from an information system  $\mathcal{I}$  by replacing a non–binary attribute  $q$  with a set of attributes, each corresponding to an attribute value of  $q$ ; the associated information functions have value 1 if and only if  $x$  has this value under  $f_q$ . In the process of binarization no information is lost; indeed, information is shifted from the columns to the rows.

Strictly speaking, we should distinguish between “symmetric” and “asymmetric” binary attributes, but we shall omit this here for reasons of brevity.

Let us consider  $Q \rightarrow d$ , and choose some  $m \in Q$ ; suppose that  $m$  leads to the binary attributes  $m_0, \dots, m_r$ . For each  $t \in \{f_d(x) : x \in U\}$  do the following:

1. Find the binary attributes  $m_i$  for which

$$(\forall x \in U)(f_{m_i}(x) = 1 \rightarrow f_d(x) = t).$$

If there is no such  $m_i$ , go to step 3.

2. Build their union within  $m$  in the following sense: If, for example  $m_{i_0}, \dots, m_{i_k}$  satisfy the condition above, then we define a new binary attribute  $m_{i_0 \dots i_k}$  by

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

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

3. Collect the resulting binary attributes in  $m$  to arrive at the filtered attribute.

Step 3 aggregates all classes of  $\theta_m$  (i.e. attribute values) which are totally contained in a class of  $\theta_d$ .

The main result shows that filtering preserves the dependency structure and may improve the statistical significance of the rule:

**Proposition 1.** *Let  $Q \rightarrow P$  be a rule of  $\mathcal{I}$  and  $Q' \rightarrow P$  its filtered version. Then,*

1.  $\gamma(Q \rightarrow P) = \gamma(Q' \rightarrow P)$ .
2.  $p(\gamma(Q \rightarrow P)|H_0) \geq p(\gamma(Q' \rightarrow P)|H_0)$ .

Details and applications, as well as a proof of Proposition 1, can be found in [Düntsch & Gediga(1997d)].

It may be worth to point out that this type of filtering is applicable to any type of attribute, and that it does not use any metric information from within the attribute domains. If one is willing to take these into account and also use e.g. genetic algorithms, there are more sophisticated methods available, for example, [Skowron & Nguyen(1996)], [Nguyen et al.(1996)Nguyen, Nguyen & Skowron], [Skowron & Polkowski(1996)], or [Düntsch & Gediga(1997a)] for a purely data driven approach.

### 3.5 Uncertainty Measures

To compare different rules and/or compare different measures of uncertainty one needs a general framework in which to perform the comparisons. To define an unconditional measure of prediction success one can use the idea of combining program complexity (i.e. to find a deterministic rule) and statistical uncertainty (i.e. a measure of uncertainty within the indeterministic rules) to a global measure of prediction success. The broad idea behind this is the well known approach of *constructive probability* or *Kolmogorov complexity*; we invite the reader to consult [Li & Vitányi(1993)] for a detailed exposition of the theory.

The tool which we use is (information theoretic) entropy: If  $\{p_i : i \leq n\}$  is a probability distribution, then its entropy is given by

$$H(p_0, \dots, p_n) = \sum_{i \leq n} p_i \cdot \log_2 \frac{1}{p_i}.$$

The entropy measures three things [McEliece(1977)]:

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



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

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

can be aggregated in the same dimension.

There are several possibilities to describe what is meant by “quality of non-deterministic prediction” in RSDA, and [Dütsch & Gediga(1997f)] present three different approaches to handle uncertainty of a rule  $Q \rightarrow P$ . Within each approach it has to be made explicit how deterministic rules and guessing should work together to predict a class of  $\theta_P$ ; different models  $M$  how to predict such a class, given  $\theta_Q$ , are then mapped to an entropy value  $H_M(Q \rightarrow P)$ .

Entropy has been discussed in the RSDA context before, e.g. by [Wong et al.(1986)Wong, Ziarko & Ye] or [Teghem & Benjelloun(1992)]. However, the class of models studied there is very narrow, which prohibits its use as a general method; furthermore, [Dütsch & Gediga(1997e)] have shown that the main theoretical result of [Wong et al.(1986)Wong, Ziarko & Ye] is incorrect.

In this paper we shall concentrate on the approach closest to the philosophy of RSDA. Let us suppose that  $U$  is our set of objects with cardinality  $n$ , and 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 (see Sect. 3.2) we assume that the elements of  $U$  are uniformly 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 now define the *entropy* of  $\mathcal{P}$  by

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

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

The entropy estimates the mean number of comparisons minimally necessary to retrieve the equivalence class information of a randomly chosen element  $x \in U$ ; we can also think of the entropy of  $\mathcal{P}$  as a measure of granularity of the partition.

Suppose that the classes of  $\theta_Q$  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  $P$ , and  $V$  be their union.

The approach is based on the pure rough set assumption that we know the world only up to the equivalence classes of  $\theta_Q$ , and that we admit complete ignorance about what happens “inside” these classes.

Consequently, given a class  $Y$  of  $\theta_P$ , any observation  $y$  in the set  $Y \setminus V$  is the result of a random process whose characteristics are totally unknown to the researcher; according to the principle of indifference, any element of  $U \setminus V$  must be viewed as a realization of a probability distribution with uncertainty  $\frac{1}{n} \log_2(n)$ . Hence, we use only those classes of  $\theta_Q$  which are contained in  $V$ , and put each  $x \in U \setminus V$  in its own class. In other words, we assume the maximum entropy principle, and look at the equivalence relation  $\theta_Q^+$  defined by

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

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

$$\hat{\psi}_i \stackrel{def}{=} \begin{cases} \hat{\pi}_i, & \text{if } i \leq c, \\ \frac{1}{n}, & \text{otherwise.} \end{cases} \quad (8)$$

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

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

We choose 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  $\mu$  that are consistent with what we know, the *principle of maximum entropy* suggests that we adopt that  $\mu^*$  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  $\mu^*$  incorporates the ‘least’ additional information [Jaynes(1957)]”.

We invite the reader to consult [Grove et al.(1994)Grove, Halpern & Koller] (from which the quote above is taken) for more details of the interplay of the principle of indifference and the maximum entropy principle.

There are other possibilities, for example, taking into account the distribution of elements in  $Y \setminus V$ . It would be outside the scope of this paper to discuss these approaches in detail, and we refer the interested reader to [Düntsch & Gediga(1997f)].

The entropy of the combined information  $Q \cup P$

$$H_{\text{total}}(Q \rightarrow P) \stackrel{\text{def}}{=} H(Q \cup P).$$

– more traditionally written as  $H(Q, P)$  – measures the uncertainty of the overall system. The boundary of both entropy measures is given by

$$H(P) \leq H_{\text{rough}}(Q \rightarrow P), H_{\text{total}}(Q \rightarrow P) \leq \log_2(|U|).$$

A measure  $H_{\text{rough}}(Q \rightarrow P)$  near  $H(P)$  is favourable, since little or no additional information is needed to code the prediction attributes  $Q$ . If  $H_{\text{rough}}(Q \rightarrow P)$  is close to  $\log_2(|U|)$ , the worst case in terms of entropy is met.

In order to normalize the outcome of the uncertainty estimation we transform the measures to *normalized overall information* (NOI) and *normalized rough information* (NRI) by the functions

$$\begin{aligned} \text{NOI}(Q \rightarrow P) &\stackrel{\text{def}}{=} 1 - \frac{H_{\text{total}}(Q \rightarrow P) - H(P)}{\log_2(|U|) - H(P)}, \\ \text{NRI}(Q \rightarrow P) &\stackrel{\text{def}}{=} 1 - \frac{H_{\text{rough}}(Q \rightarrow P) - H(P)}{\log_2(|U|) - H(P)}. \end{aligned}$$

If both normalized measures have a value near 1, the chosen attribute combination is favourable, whereas a value near 0 indicates casualness. Note, that the normalization does not use moving standards as long as we do not change the decision attribute  $P$ . Therefore, any comparison of NOI or NRI values between different predicting attribute sets given a fixed set of decision attributes is feasible. The normalized rough information is always smaller than the normalized overall information. Big differences between both indicate that the local structure within  $Q$  determines indeterministically much of the local structure within  $P$ .

A discussion of where the approximation quality  $\gamma$  can be located within this context can be found in [Düntsch & Gediga(1997f)].

## 4 Rough Set Analysis of IRIS Data

Several earlier studies compare statistical techniques such as discriminant analysis with RSDA [Krusińska et al.(1992a)Krusińska, Babic, Słowiński & Stefanowski, Krusińska et al.(1992b)Krusińska, Słowiński & Teghem & Benjelloun(1992)]. Their result can be summarized to the claim that RSDA and statistical techniques offer similar approaches. If so, RSDA would be the method of choice, because RSDA is a “soft” data analysis method, which does not assume structural information outside the data.

One may have reservations about this claim:

- An attribute with continuous values usually cannot be used by RSDA in its pure original form, whereas discriminant analysis is based on the interpretation of metric information within the data. Therefore, discriminant analysis uses more details within the data for the price of using a “hard” dimensional data representation as an underpinning.
- RSDA needs a fixed number of equivalence classes within any attribute. If we use data with continuous metric information, the number of equivalence classes of the raw data may be as high as the number of objects under study. Hence, if we like to result in statistically stable rules, a preprocessing stage (which we call *filtering*) has to be performed before data can be analysed. Although a filter procedure is a precondition to perform a reliable RSDA using continuous metric attributes, a “dependency preserving” filtering procedure was not included in the previous studies.

In the next subsections we will show how the IRIS data are processed by the traditional RSDA approach, and discuss the filtering of the IRIS data used in [Teghem & Benjelloun(1992)].

### 4.1 Pure RSDA Description of IRIS Data

RSDA starts by finding (global) dependency information, i.e. computation of reducts and core, as well as the rules of the information system under review. The ranges and the number of classes of each attribute are given in Tab. 4

**Table 4.** IRIS - Unfiltered Data

Attribute	Interval	No of classes
Sepal length:	[43,79]	35
Sepal width:	[20,44]	23
Petal length:	[10,69]	43
Petal width:	[1,25]	22

The full IRIS data set has each three element set of attributes as a reduct, and thus, it has an empty core. This indicates a high substitution rate among the attributes. The approximation qualities of the nonempty attribute sets are given in Tab. 5. We see that petal length (A3) has a high classification quality, followed by petal width (A4). Together, they can account for 98% of all cases.

Using all four dependent attributes, Grobian has found a total of 243 rules. We give the 58 deterministic rules for single petal attributes in Tab. 6.

### 4.2 A Previous RSDA Analysis of IRIS Data

In the rough set context, the IRIS data have been explored by [Teghem & Benjelloun(1992)] with a data filtering displayed in Tab. 7. The resulting system does not explain the data, since  $\gamma(\{A1, A2, A3, A4\} \rightarrow D) = 0.77$ . If we compare this result with the 96% reclassification success of discriminant analysis using two attributes only (Tab. 3), the result does

**Table 5.** Approximation Qualities

Attributes	$\gamma$	Attributes	$\gamma$
A1, A2, A3	1.00	A2, A3	0.97
A1, A2, A4	1.00	A2, A4	0.94
A1, A3, A4	1.00	A3, A4	0.98
A2, A3, A4	1.00	A1	0.21
A1, A2	0.85	A2	0.13
A1, A3	0.97	A3	0.82
A1, A4	0.94	A4	0.73

**Table 6.** IRIS Rules, Petal Attributes (Unfiltered Full Set)

Rule	Instances	Rule	Instances	Rule	Instances
A3=14 $\Rightarrow$ D=1	13	A3=37 $\Rightarrow$ D=2	1	A3=69 $\Rightarrow$ D=3	1
A3=10 $\Rightarrow$ D=1	1	A3=43 $\Rightarrow$ D=2	2	A3=63 $\Rightarrow$ D=3	1
A3=17 $\Rightarrow$ D=1	4	A3=30 $\Rightarrow$ D=2	1	A4=2 $\Rightarrow$ D=1	29
A3=13 $\Rightarrow$ D=1	7	A3=36 $\Rightarrow$ D=2	1	A4=3 $\Rightarrow$ D=1	7
A3=16 $\Rightarrow$ D=1	7	A3=50 $\Rightarrow$ D=3	4	A4=5 $\Rightarrow$ D=1	1
A3=19 $\Rightarrow$ D=1	2	A3=56 $\Rightarrow$ D=3	6	A4=1 $\Rightarrow$ D=1	5
A3=12 $\Rightarrow$ D=1	2	A3=52 $\Rightarrow$ D=3	2	A4=6 $\Rightarrow$ D=1	1
A3=11 $\Rightarrow$ D=1	1	A3=55 $\Rightarrow$ D=3	3	A4=4 $\Rightarrow$ D=1	7
A3=15 $\Rightarrow$ D=1	13	A3=59 $\Rightarrow$ D=3	2	A4=11 $\Rightarrow$ D=2	3
A3=46 $\Rightarrow$ D=2	3	A3=54 $\Rightarrow$ D=3	2	A4=13 $\Rightarrow$ D=2	13
A3=48 $\Rightarrow$ D=2	2	A3=67 $\Rightarrow$ D=3	2	A4=12 $\Rightarrow$ D=2	5
A3=39 $\Rightarrow$ D=2	3	A3=57 $\Rightarrow$ D=3	3	A4=10 $\Rightarrow$ D=2	7
A3=47 $\Rightarrow$ D=2	5	A3=66 $\Rightarrow$ D=3	1	A4=17 $\Rightarrow$ D=3	2
A3=40 $\Rightarrow$ D=2	5	A3=53 $\Rightarrow$ D=3	1	A4=22 $\Rightarrow$ D=3	3
A3=38 $\Rightarrow$ D=2	1	A3=64 $\Rightarrow$ D=3	1	A4=24 $\Rightarrow$ D=3	3
A3=44 $\Rightarrow$ D=2	4	A3=60 $\Rightarrow$ D=3	2	A4=23 $\Rightarrow$ D=3	8
A3=33 $\Rightarrow$ D=2	2	A3=48 $\Rightarrow$ D=3	2	A4=20 $\Rightarrow$ D=3	6
A3=41 $\Rightarrow$ D=2	3	A3=61 $\Rightarrow$ D=3	3	A4=25 $\Rightarrow$ D=3	3
A3=35 $\Rightarrow$ D=2	2	A3=58 $\Rightarrow$ D=3	3	A4=21 $\Rightarrow$ D=3	6
A3=42 $\Rightarrow$ D=2	4				

not look favorable for RSDA, if this is the best such data analysis can offer. The original unfiltered data show that  $\gamma(\{A3, A4\} \rightarrow D) = 0.98$ , so that the low approximation quality is only due to the filtering.

The attribute sets  $\{A3, A4\}$  and  $\{A1, A2, A4\}$  have an approximation quality of  $\gamma = 0.75$ , resp.  $\gamma = 0.72$ ; thus, it seems that these sets should have been considered in the data analysis as well. If one is prepared to accept an approximation quality of  $\gamma = 0.77$  with four features, it is surely acceptable to eliminate two of these in return for a drop in the approximation quality of only 0.02.

In order to show that the attribute set  $\{A3, A4\}$  is the optimal combination, we can

**Table 7.** Data Filtering of [Teghem & Benjelloun(1992)]

	Very small (1)	Small (2)	Large (3)	Very large (4)
Sepal length	$x < 50$	$50 \leq x < 60$	$60 \leq x < 70$	$70 \leq x$
Sepal width	$x < 24$	$24 \leq x < 31$	$31 \leq x < 38$	$38 \leq x$
Petal length	$x < 30$	$30 \leq x < 40$	$40 \leq x < 55$	$55 \leq x$
Petal width	$x < 10$	$10 \leq x < 14$	$14 \leq x < 21$	$21 \leq x$

compare the uncertainty measures of the attribute sets

$$\{A1, A2, A3, A4\}, \{A1, A2, A4\}, \{A3, A4\},$$

see Tab. 8. The results show that the petal attributes are by far preferred, and that in terms of uncertainty measure the complete set of attributes is the worst.

**Table 8.** Entropy Values

$Q$	$H(Q)$	$H(Q, D)$	NOI	NRI
$\{A1, A2, A3, A4\}$	4.416	4.621	0.462	0.362
$\{A1, A2, A4\}$	4.023	4.290	0.520	0.392
$\{A3, A4\}$	2.520	2.763	0.791	0.607

[Teghem & Benjelloun(1992)] offer, among others, the following conclusions to their work:

“The three main advantages of rough sets theory are

- its very clear interpretation for the user,
- its independence to any statistical assumptions,
- its efficiency and its rapidity.”

We are somewhat more skeptical. Even though the basis of RSDA consists of a very simple mathematical model which is clearly understandable, the interpretation of results is not always all that clear; we believe that e.g. the considerations of Sect. 3.3 regarding the statistical validation of rough set rules show that care has to be taken when interpreting the results of a rough set data analysis, and that the results are by no means always clear and straightforward.

The rough set model is not independent of any statistical assumptions. Even though it requires no (exterior) prior probabilities it has an underlying statistical model as shown in Sect. 3.2.

In view of the fact that for example minimal reduct search is NP hard, it seems hardly justified to claim efficiency and rapidity for RSDA except for very small databases. Having said this, one should mention that heuristic tools for reduct finding have been developed, for example [Wroblewski(1995)] or [Bjorvand & Komorowski(1997)]. On the other hand, these methods need assumptions outside the data at hand, which we are trying to avoid.

## 5 Rough Information Analysis of IRIS Data

### 5.1 Data Filtering

We have used the procedure outlined in Sect. 3.4 to obtain the data conversions given in Tab. 9; there, the choice of a value is irrelevant. We also list the resulting number of classes, and in brackets as a reminder the number of classes of the unfiltered data. Observe the dramatic fall in the number of classes of the petal attributes. We shall use this filtering for all subsequent computations, unless indicated otherwise.

### 5.2 Significance

We found that no attribute set  $\emptyset \neq Q \subseteq \{A1, A2, A3, A4\}$  was casual with respect to  $D$ . The values of relative significances is given in Tab. 10. The results clearly indicate that the combination (A3, A4) is the best choice to describe the IRIS data in terms of rules. Whereas any other combination of attributes contains at most one attribute which is not conditional casual, A3 and A4 show significant (0.004), resp. marginally significant (0.063) test results of the hypothesis of conditional casualness.

**Table 9.** Rough Filtering

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

**Table 10.** Relative Significance

Attributes	A1	A2	A3	A4
A1, A2, A3, A4	1.00	1.00	1.00	1.00
A1, A2, A3	0.901	0.862	0.604	
A1, A2, A4	0.843	0.710		0.336
A1, A3, A4	0.944		0.857	0.857
A2, A3, A4		0.860	0.843	0.790
A1, A2	0.127	0.213		
A1, A3	0.874		0.001	
A1, A4	0.727			0.213
A2,A3		0.814	0.001	
A2,A4		0.884		0.001
A3, A4			0.004	0.063

### 5.3 Uncertainty Measures

The entropy values are given in Tab. 11. Observe that in case  $Q \Rightarrow D$ , we have of course  $H(Q) = H(Q, D)$ , since there is no unexplained information, and also  $\text{NOI} = \text{NRI}$ .

The remarks above concerning significance are reflected in the results of the entropy values of Tab. 11. Large values for NOI and NRI are recorded for attribute sets  $\{A3\}$ ,  $\{A4\}$ , and  $\{A3, A4\}$ , with corresponding low values for  $H(Q)$  and  $H(Q,D)$ .

Attribute A2 records a high value for NOI, with corresponding values for  $H(Q)$  and  $H(Q,D)$  being about average. Perhaps this indicates A2's rank in ability to distinguish between the species, i.e., not as good as A3 or A4 but better than A1. It is interesting to note that the reducts of the full information system perform badly on both entropy results and relative significance.

To summarize the NOI/NRI analysis, both unconditional measures vote for A3 – and if we want to predict more objects for the price of a small decrease in information – the prediction set  $\{A3, A4\}$ . Any other combination of two or more attributes is by far too crude in terms of NOI/NRI.

### 5.4 Rules

Rough filtering not only reduces the number of classes and increases the significance of rules, it also reduces the number of rules. Using all four independent attributes, Grobian has found 118 deterministic rules. The 6 deterministic rules for the petal attributes alone are given in Tab. 12, as well as some other rules which make up the whole data set.

**Table 11.** Entropy Values (Filtered)

$Q$	$H(Q)$	$H(Q, D)$	NOI	NRI
Reducts				
{A2, A3, A4}	5.683	5.683	0.274	0.274
{A1, A3, A4}	5.657	5.657	0.279	0.279
{A1, A2, A3}	6.683	6.683	0.097	0.097
{A1, A2, A4}	6.724	6.724	0.090	0.090
Non – reducts				
{A1, A2}	6.500	6.644	0.104	0.094
{A1, A3}	5.310	5.337	0.335	0.335
{A1, A4}	5.492	5.550	0.297	0.294
{A2, A3}	5.340	5.371	0.329	0.326
{A2, A4}	5.400	5.459	0.314	0.311
{A3, A4}	3.285	3.303	0.696	0.693
{A1}	4.314	5.020	0.391	0.139
{A2}	3.759	4.818	0.427	0.694
{A3}	2.358	2.488	0.840	0.780
{A4}	2.562	2.722	0.799	0.674

**Table 12.** Some IRIS Rules (Filtered Full Set)

Rule	Instances
A3=14 $\Rightarrow$ D=1	50
A4=2 $\Rightarrow$ D=1	50
A3=46 $\Rightarrow$ D=2	37
A4=11 $\Rightarrow$ D=2	28
A3=45, A4=15 $\Rightarrow$ D=2	5
A1=60, A4=16 $\Rightarrow$ D=2	2
A2=32, A3=48 $\Rightarrow$ D=2	1
A3=49, A4=15 $\Rightarrow$ D=2	2
A3=53, A4=19 $\Rightarrow$ D=2	1
A1=68, A4=14 $\Rightarrow$ D=2	1
A3=50 $\Rightarrow$ D=3	36
A4=17 $\Rightarrow$ D=3	31
A3=51, A4=15 $\Rightarrow$ D=3	1
A3=51, A4=19 $\Rightarrow$ D=3	2
A2=30, A4=18 $\Rightarrow$ D=3	4
A3=49, A4=18 $\Rightarrow$ D=3	2
A1=62, A4=18 $\Rightarrow$ D=3	2

The class *Setosa* needs only one prediction rule ( $A3 = 14$  or  $A4 = 2$ ), and it is obvious that it is rather different from the other two. The class *Virginica* is fairly well explained: The (filtered) values  $A3 = 50$  or  $A4 = 17$  explain 42 instances of *Virginica*. There is only one object which needs an A1-based rule, and there are only four objects that need an A2-based rule.

The class *Versicolor* causes difficulties. The rules “( $A3 = 53, A4 = 19$ )  $\Rightarrow$  D = 2”, and “( $A3 = 51, A4 = 19$ )  $\Rightarrow$  D = 3” have no frame of interpretation because of their “closeness”. Observe that none of the occurring values is filtered.

### 5.5 ROUGHIAN Prediction of IRIS Data

In machine learning, a common procedure to test the accuracy of the prediction value of a rule set is done in the following way:

1. Split the data into a *training set* and a *testing set*.

2. Learn a rule set in the training set.
  3. Measure the accuracy of the rule set (in the given theoretical system) in the testing set.
- Repeat this procedure about twenty times and find the mean and standard deviation of the obtained values.

We have followed this procedure to find out the prediction quality of the filtered data set for certain attribute sets. To this end, we have generated twenty random partitions of the whole data set into two equally sized classes of 75 specimen each. We then have filtered the training set and, with this filtering, computed the approximation quality of the attribute sets

$$\{A3, A4\}, \{A1, A3, A4\}, \{A2, A3, A4\}$$

on the testing set, and the  $\alpha$ -value of (4) for each species. We have also tested on the whole data set whether the random partition is conditional casual for the attribute set under consideration, i.e. whether we can assume that the partition is really random. The mean and the standard deviation of the resulting sequences of values can be found in Tab. 13. The

**Table 13.** Means and Standard Deviations of the 20 IRIS Files for  $\gamma$ ,  $\alpha$  for Each Species, and the Relative Significance of the Random Variable

Attributes	$\gamma$		$\alpha$ , Setosa		$\alpha$ , Versicolor		$\alpha$ , Virginica		Rel. signif.	
	Mean	Std.dev.	Mean	Std.dev.	Mean	Std.dev.	Mean	Std.dev.	Mean	Std.dev.
A1,A3,A4	0.972	0.028	1	0	0.791	0.130	0.797	0.157	0.685	0.347
A2,A3,A4	0.992	0.014	1	0	0.818	0.119	0.780	0.151	0.863	0.275
A3,A4	0.896	0.088	1	0	0.613	0.138	0.602	0.167	0.329	0.358

results show that the filter derived from the first half of the data can be used in principle in the second half of the data set. The prediction quality is about as high as in the overall data analysis (see Table 5). A further check of the admissibility of the filter procedure is presented in the last two columns of Table 13: The training set / testing set coding should have no influence on the prediction in the overall system given the filter of the training set; hence, the random variable should be conditional casual. As the last two columns of Table 13 indicate, we cannot observe a significant influence of the random variable to the overall prediction success.

RSDA faces a problem which is common to every structural data analysis: There may be observations within the testing set which cannot be expressed by a rule extracted from the training set, simply because this rule does not occur in the training situation. To solve this problem, [Słowiński & Stefanowski(1992)] adopt metric information about the data in their 'ROUGHCLASS' approach to compute the best 'nearby' rule(s) which can be used for prediction. Although this approach is not in line with the original soft computing aims of the RSDA (since one has to enter external assumptions about the data), we are forced to use a similar approach, because – up to now – no non-invasive data analysis counterpart for the classification problem is at hand.

Our validation using the “training - testing - set” paradigm runs as follows:

- Initialize 9 counters  $N(P|D)$ , where  $D$  is true value of the specimen in the testing set, and  $P$  is the predicted value of the specimen from rules of the training set.
- For each of the 20 training sets do
  - Compute the filter rules based on the 4 predicting attributes and the specimen decision attribute.
  - Compute the rules based on petal length and petal width within the training set based on the respective filter.



- Apply the filter derived from the training set to the testing set objects.
- For every object of the testing set do
  - \* Compute those rules which have the same *minimal* Euclidian distance to the current (petal length, petal width) combination.
  - \* Choose one of these rules with minimal distance randomly, and use the value of the decision attribute of the chosen rule as the prediction  $P$  for the testing object under study.
  - \* Increase the counter  $N(P|D)$  by 1.
- Normalize each counter  $N(P|D)$  by  $N(\cdot|D) = N(1|D)+N(2|D)+N(3|D)$  resulting in  $\hat{p}(P|D) = \frac{N(P|D)}{N(\cdot|D)}$ .

Table 14 shows the result of the RSDA validation procedure. Whereas Setosa is captured perfectly, the error rates of the classification of Versicolor and Virginica in the testing set are between are 8% and 9%.

**Table 14.** Mean Prediction Quality Using Rough Analysis (Half Sample Prediction)

Predicted class	Classes in the testing set		
	Setosa	Versicolor	Virginica
Setosa	1.000	0.000	0.000
Versicolor	0.000	0.914	0.080
Virginica	0.000	0.086	0.920

With the same simulation procedure, but using discriminant analysis instead, we see in Table 15 that discriminant analysis outperforms RSDA by about additional 2% correct predictions.

**Table 15.** Mean Prediction Quality Using Discriminant Analysis

Predicted class	Classes in the testing set		
	Setosa	Versicolor	Virginica
Setosa	1.000	0.000	0.000
Versicolor	0.000	0.940	0.069
Virginica	0.000	0.060	0.931

It should be noted that the jack–knife (“leave-one-out”) validation results in more optimistic estimations of prediction success (Tab. 16), and these probabilities are comparable to those of the discriminant analysis. [Krusińska et al.(1992a)Krusińska, Babic, Słowiński & Stefanowski]

**Table 16.** Mean Prediction Quality Using Rough Analysis (Jack-knife Validation)

Predicted class	Classes given in data		
	Setosa	Versicolor	Virginica
Setosa	1.000	0.000	0.000
Versicolor	0.000	0.939	0.071
Virginica	0.000	0.061	0.929

use a jack–knife procedure for validation, and thus, the goodness of the RSDA classification compared to DA may be overestimated and should be taken with care. Furthermore, all

their attributes are conditional casual, which indicates a large inhomogeneity of the data, and one could conclude that the sample size is too small to allow a reliable prediction.

The dependency of the prediction success on the chosen validation method seems to be a disadvantage of the RSDA. If we compare the jack-knife approach with the half sample prediction, we observe that the number of rules generated by RSDA tends to be smaller in case of the half sample prediction. Because sometimes essential rules may be missing, a misclassification will occur, and the prediction quality will decrease. In case of discriminant analysis, a smaller number of subjects will only decrease the precision of estimators, and will not decrease the number of structural parameters as in case of RSDA.

## 6 Conclusion

We have performed a traditional RSDA analysis of Fisher’s IRIS data, and supplemented it with the ROUGHIAN procedures *data filtering*, *significance testing*, and *uncertainty measures*.

Given a measurement situation like in the case of the IRIS data, which is a standard one for performing discriminant analysis, there is a need for reducing the granularity of the predicting attributes. Categorization by researchers may be suboptimal in terms of approximation quality of the filtered attributes, as e.g. the results of [Teghem & Benjelloun(1992)] indicate. The data filtering procedure offers a method to reduce the granularity of the data, which does not change the prediction success of the attribute under study.

Significance testing is necessary for a decision whether the rules derived from an information system are more than just rules generated by a random process (casualness) or whether part of the rules can be explained by chance (conditional casualness) respectively. Classical approaches of RSDA usually overfit the data by using either casual systems or an abundance of conditional casual attributes. It turns out that the increase in rule significance obtained by our filtering procedure makes ROUGHIAN a viable data description method even in those cases where DA seems to be the method of choice. We argue that this result could not have been obtained by RSDA alone since the  $\gamma$  statistics as a measure of “determinacy” does not generally suffice to evaluate the quality of rules. For example, in the unfiltered system, the relative significance of  $A_3$  and  $A_4$  in  $\{A_3, A_4\}$  with 1000 simulations is 0.621, resp. 0.516 – thus, both attributes are conditional casual – , but in the filtered system only 0.004, resp. 0.063.

We have shown that the combination of filtering and significance testing achieved the same combination of variables in which the DA resulted, with about the same coverage in terms of posterior probabilities.

Using the significance and entropy procedures as additional information providers significantly simplifies model selection within RSDA, and justifies the appropriate choice.

Using the IRIS data set, we have shown that prediction using the ROUGHIAN model is nearly as good as that of discriminant analysis, even though

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

in contrast to the discriminant analysis.

However, the problem of the dependency of the prediction success on the choice of the validation method is a problem which should not be underestimated, and should be investigated further.

## References

- [Bacchus et al.(1994)Bacchus, Grove, Halpern & Koller] Bacchus, F., Grove, A. J., Halpern, J. Y. & Koller, D. (1994). From statistical knowledge bases to degrees of belief. Technical report 9855, IBM.
- [Bjorvand & Komorowski(1997)] Bjorvand, A. T. & Komorowski, J. (1997). Practical applications of genetic algorithms for efficient reduct computation. In [Sydow(1997)], 601–606.
- [Browne(1997)] Browne, C. (1997). Enhanced rough set data analysis of the Pima Indian diabetes data. In *Proc. 8<sup>th</sup> Ireland Conference on Artificial Intelligence, Derry (1997)*, 32–39.
- [Cohen(1990)] Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, **45**, 1304–1312.
- [Dütsch & Gediga(1996)] Dütsch, I. & Gediga, G. (1996). The rough set model for data analysis – introduction and overview. Preprint, <http://www.infj.ulst.ac.uk/~ccc23/papers/roughmod.html>.
- [Dütsch & Gediga(1997a)] Dütsch, I. & Gediga, G. (1997a). Relation restricted prediction analysis. In [Sydow(1997)], 619–624. Abstract: <http://www.infj.ulst.ac.uk/~ccc23/papers/ordg.html>.
- [Dütsch & Gediga(1997b)] Dütsch, I. & Gediga, G. (1997b). The rough set engine GROBIAN. In [Sydow(1997)], 613–618. Abstract: <http://www.infj.ulst.ac.uk/~ccc23/papers/grobrian.html>.
- [Dütsch & Gediga(1997c)] Dütsch, I. & Gediga, G. (1997c). ROUGHIAN – Rough information analysis, an introduction. Technical report, University of Ulster, <http://www.infj.ulst.ac.uk/~ccc23/papers/roughian.html>. Extended abstract in *Proc. 15th IMACS World Congress*, Vol 4., 631–636.
- [Dütsch & Gediga(1997d)] Dütsch, I. & Gediga, G. (1997d). Simple data filtering in rough set systems. *International Journal of Approximate Reasoning*. To appear. Abstract: <http://www.infj.ulst.ac.uk/~ccc23/papers/bininf.html>.
- [Dütsch & Gediga(1997e)] Dütsch, I. & Gediga, G. (1997e). Statistical evaluation of rough set dependency analysis. *International Journal of Human–Computer Studies*, **46**, 589–604. Abstract: <http://www.infj.ulst.ac.uk/~ccc23/papers/rougheva.html>.
- [Dütsch & Gediga(1997f)] Dütsch, I. & Gediga, G. (1997f). Uncertainty measures of rough set prediction. Submitted for publication, University of Ulster. <http://www.infj.ulst.ac.uk/~ccc23/papers/rmml.html>.
- [Dütsch et al.(1997)Dütsch, Gediga & Rogner] Dütsch, I., Gediga, G. & Rogner, J. (1997). Archetypal psychiatric patients: An application of rough information analysis. Preprint, Fachbereich Psychologie, Universität Osnabrück.
- [Dwass(1957)] Dwass, M. (1957). Modified randomization tests for non-parametric hypothesis. *Annals of Mathematical Statistics*, **28**, 181–187.
- [Edgington(1987)] Edgington, E. S. (1987). Randomization Tests, vol. 31 of *Statistics: Textbooks and Monographs*. New York and Basel: Marcel Dekker.
- [Fisher(1936)] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179–188.
- [Grove et al.(1994)Grove, Halpern & Koller] Grove, A. J., Halpern, J. Y. & Koller, D. (1994). Random worlds and maximum entropy. *Journal of AI Research*, **2**, 33–88.
- [Halpern & Fagin(1992)] Halpern, J. Y. & Fagin, R. (1992). Two views of belief: belief as generalized probability and belief as evidence. *Artificial Intelligence*, **54**, 275–317.
- [Holte(1993)] Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, **11**, 63–91.
- [Jaynes(1957)] Jaynes, E. (1957). Information theory and statistical mechanics. *Physical Review*, **106**, 620–630.
- [Krusińska et al.(1992a)Krusińska, Babic, Słowiński & Stefanowski] Krusińska, E., Babic, A., Słowiński, R. & Stefanowski, J. (1992a). Comparison of the rough sets approach and probabilistic data analysis techniques on a common set of medical data. In [Słowiński(1992)], 251–265.
- [Krusińska et al.(1992b)Krusińska, Słowiński & Stefanowski] Krusińska, E., Słowiński, R. & Stefanowski, J. (1992b). Discriminant versus rough set approach to vague data. *Appl. Stochastic Models and Data Analysis*, **8**, 43–56.

- [Li & Vitányi(1993)] 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.
- [Lin & Cercone(1997)] Lin, T. Y. & Cercone, N. (Eds.) (1997). Rough sets and data mining, Dordrecht. Kluwer.
- [Manly(1991)] Manly, B. F. J. (1991). Randomization and Monte Carlo Methods in Biology. London: Chapman and Hall.
- [McEliece(1977)] McEliece, R. J. (1977). The Theory of Information and Coding, vol. 3 of *Encyclopedia of Mathematics and its Applications*. Reading: Addison-Wesley.
- [Nguyen et al.(1996)Nguyen, Nguyen & Skowron] Nguyen, H. S., Nguyen, S. H. & Skowron, A. (1996). Searching for features defined by hyperplanes. In Z. Ras & M. Michalewicz (Eds.), *ISMIS-96, Ninth International Symposium on Methodologies for Intelligent Systems*, vol. 1079 of *Lecture Notes in Artificial Intelligence*, 366–375, Berlin. Springer-Verlag.
- [Pawlak(1982)] Pawlak, Z. (1982). Rough sets. *Internat. J. Comput. Inform. Sci.*, **11**, 341–356.
- [Pawlak(1991)] Pawlak, Z. (1991). Rough sets: Theoretical aspects of reasoning about data. Dordrecht: Kluwer.
- [Skowron & Nguyen(1996)] Skowron, A. & Nguyen, H. S. (1996). Quantization of real value attributes: Rough set and Boolean reasoning approach. *Bulletin of International Rough Set Society*, **1**, 5–16.
- [Skowron & Polkowski(1996)] Skowron, A. & Polkowski, L. (1996). Analytic morphology: Mathematical morphology of decision tables. *Fundamenta Informaticae*, **27**, 255–271.
- [Słowiński(1992)] Słowiński, R. (1992). Intelligent decision support: Handbook of applications and advances of rough set theory, vol. 11 of *System Theory, Knowledge Engineering and Problem Solving*. Dordrecht: Kluwer.
- [Słowiński & Stefanowski(1992)] Słowiński, R. & Stefanowski, J. (1992). ‘ROUGHIDAS’ and ‘ROUGHCLASS’ software implementations of the rough sets approach. In [Słowiński(1992)], 445–456.
- [Sydow(1997)] Sydow, A. (Ed.) (1997). Proc. 15th IMACS World Congress, vol. 4, Berlin. Wissenschaft und Technik Verlag.
- [Teghem & Benjelloun(1992)] Teghem, J. & Benjelloun, M. (1992). Some experiments to compare rough sets theory and ordinal statistical methods. In [Słowiński(1992)], 267–284.
- [Wong et al.(1986)Wong, Ziarko & Ye] 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.
- [Wroblewski(1995)] Wroblewski, J. (1995). Finding minimal reducts using genetic algorithms. ICS Research Report 16, Warsaw University of Technology.