

Statistical Evaluation of Rough Set Dependency Analysis

Ivo Düntsch¹

School of Information and Software Engineering

University of Ulster

Newtownabbey, BT 37 0QB, N.Ireland

I.Duentsch@ulst.ac.uk

Günther Gediga¹

FB Psychologie / Methodenlehre

Universität Osnabrück

49069 Osnabrück, Germany

gg@Luce.Psycho.Uni-Osnabrueck.DE

and

Institut für semantische Informationsverarbeitung

Universität Osnabrück

December 12, 1996

¹Equal authorship implied

Summary

Rough set data analysis (RSDA) has recently become a frequently studied symbolic method in data mining. Among other things, it is being used for the extraction of rules from databases; it is, however, not clear from within the methods of rough set analysis, whether the extracted rules are valid.

In this paper, we suggest to enhance RSDA by two simple statistical procedures, both based on randomization techniques, to evaluate the validity of prediction based on the approximation quality of attributes of rough set dependency analysis. The first procedure tests the casualness of a prediction to ensure that the prediction is not based on only a few (casual) observations. The second procedure tests the conditional casualness of an attribute within a prediction rule.

The procedures are applied to three data sets, originally published in the context of rough set analysis. We argue that several claims of these analyses need to be modified because of lacking validity, and that other possibly significant results were overlooked.

Keywords: Rough sets, dependency analysis, statistical evaluation, validation, randomization test

1 Introduction

Rough set analysis, an emerging technology in artificial intelligence (Pawlak et al. (1995)), has been compared with statistical models, see for example Wong et al. (1986), Krusińska et al. (1992a) or Krusińska et al. (1992b). One area of application of rough set theory is the extraction of rules from databases; these rules then are sometimes claimed to be useful for future decision making or prediction of events. However, if such a rule is based on only a few observations, its usefulness for prediction is arguable (see also Krusińska et al. (1992a), p 253 in this context).

The aim of this paper is to employ statistical methods which are compatible with the rough set philosophy to evaluate the “prediction quality” of rough set dependency analysis. The methods will be applied to three different data sets:

- The first set was published in Pawlak et al. (1986) and Słowiński & Słowiński (1990). It utilizes rough set analysis to describe patients after highly selective vagotomy (HSV) for duodenal ulcer. The statistical validity of the conclusions will be discussed.
- The second example is the discussion of earthquake data published by Teghem & Charlet (1992). The main reason why we use this example is that it demonstrates the applicability of our approach in the situation when the prediction success is perfect in terms of rough analysis.
- The third example is used by Teghem & Benjelloun (1992) to compare statistical and rough set methods. We show how statistical methods *within* rough set analysis highlight some of their results in a different way.

2 Rough set data analysis

A major area of application of rough set theory is the study of dependencies among attributes of information systems. An *information system* $\mathcal{S} = \langle U, \Omega, V_q, f \rangle_{q \in \Omega}$ consists of

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

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

With each $Q \subseteq \Omega$ we associate an equivalence relation θ_Q on U by

$$x \equiv y (\theta_Q) \stackrel{\text{def}}{\iff} f(x, q) = f(y, q) \text{ for all } q \in Q.$$

If $x \in U$, then $\theta_Q x$ is the equivalence class of θ_Q containing x .

Intuitively, $x \equiv y (\theta_Q)$ if the objects x and y are indiscernible with respect to the values of their attributes from Q . If $X \subseteq U$, then *the lower approximation of X by Q*

$$\underline{X}_{\theta_Q} = \bigcup \{ \theta_Q x : \theta_Q x \subseteq X \}$$

is the set of all correctly classified elements of X with respect to θ_Q , i.e. with the information available from the attributes given in Q .

Suppose that $P, Q \subseteq \Omega$. We say that P is *dependent on Q* – written as $Q \rightarrow P$ – if every class of θ_P is a union of classes of θ_Q . In other words, the classification of U induced by θ_P can be expressed by the classification induced by θ_Q .

In order to simplify notation we shall in the sequel usually write $Q \rightarrow p$ instead of $Q \rightarrow \{p\}$ and θ_p instead of $\theta_{\{p\}}$.

Each dependency $Q \rightarrow P$ leads to a set of rules as follows: Suppose that $Q \stackrel{\text{def}}{=} \{q_0, \dots, q_n\}$, and $P \stackrel{\text{def}}{=} \{p_0, \dots, p_k\}$. For each set $\{t_0, \dots, t_n\}$ where $t_i \in V_{q_i}$ there is a uniquely determined set $\{s_0, \dots, s_k\}$ with $s_i \in V_{p_i}$ such that

$$(2.1) \quad (\forall x \in U)[f(x, q_0) = t_0 \wedge \dots \wedge f(x, q_n) = t_n \Rightarrow (f(x, p_0) = s_0 \wedge \dots \wedge f(x, p_k) = s_k)].$$

Of particular interest in rough set dependency theory are those sets Q which use the least number of attributes, and still have $Q \rightarrow P$. A set with this property called a *minimal determining set for P* . In other words, a set Q is minimal determining for P , if $Q \rightarrow P$, and $R \not\rightarrow P$ for all $R \subsetneq Q$.

If such Q is a subset of P we call Q a *reduct of P* . It is not hard to see, that each $P \subseteq \Omega$ has a reduct, though this need not be unique. The intersection of all reducts of P is called the *core of P* . Unless P has only one reduct, the core of P is not itself a reduct.

For each $R \subseteq \Omega$ let \mathcal{P}_R be the partition of U induced by θ_R . Define

$$(2.2) \quad \gamma_Q(P) = \frac{\sum_{X \in \mathcal{P}_P} |\underline{X}_{\theta_Q}|}{|U|}.$$

$\gamma_Q(P)$ is the relative frequency of the number of correctly Q -classified elements with respect to the partition induced by P . It is usually interpreted in rough set analysis as a measurement of the prediction success of a set of inference rules based on value combinations of Q and value combinations of P of the form given in (2.1). The prediction success is perfect, if $\gamma_Q(P) = 1$; in this case, $Q \rightarrow P$.

Suppose that Q is a reduct of P , so that $Q \rightarrow P$, and $Q \setminus \{q\} \not\rightarrow P$ for any $q \in Q$. In rough set theory, the impact of attribute q on the fact that $Q \rightarrow P$ is usually measured by the drop of the approximation function γ from 1 to $\gamma_{Q \setminus \{q\}}(P)$: The larger the difference, the more important one regards the contribution of q . We shall show below that this interpretation needs to be taken with care in some cases, and additional statistical evidence may be needed.

3 Casual rules and randomization analysis

3.1 Casual dependencies

In the sequel we consider the case that a rule $Q \rightarrow P$ was given *before* performing the data analysis, and not obtained by optimizing the quality of approximation. The latter needs additional treatment and will be discussed briefly in Section 3.5.

Suppose that θ_Q is the identity relation id_U on U . Then, $\theta_Q \subseteq \theta_P$ for all $P \subseteq \Omega$, i.e. $Q \rightarrow P$ for all $P \subseteq \Omega$. Furthermore, each class of θ_Q consists of exactly one element, and therefore, any rule $Q \rightarrow P$ is based on exactly one observation. We call such a rule *deterministic casual*.

If a rule is not deterministic casual, it nevertheless may be based on a few observations only, and thus, its prediction quality could be limited; such rules may be called *casual*. Therefore, the need arises for a statistical procedure which tests the casualness of a rule based on mechanisms of rough set analysis.

Assume that the information system is the realization of a random process in which the attribute values of Q and P are realized independently of each other. If no additional information is present, it may be assumed that the attribute value combinations within Q and P are fixed and the matching of the Q, P – combinations is drawn at random.

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

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

and let $\gamma_{\sigma(Q)}(P)$ be the approximation of the prediction of P by Q in the new information system. Note that the structure of the equivalence relation $\theta_{\sigma(Q)}$ determined by Q in the revised system is the same as that of the original θ_Q . In other words, there is a bijective mapping

$$\tau : \{\theta_{\sigma(Q)}x : x \in U\} \rightarrow \{\theta_Qx : x \in U\}$$

which preserves the cardinality of the classes. In particular, if θ_Q is the identity on U , so is $\theta_{\sigma(Q)}$. It follows that for a rule $Q \rightarrow p$ with $\theta_Q = id_U$, we have $\gamma_{\sigma(Q)}(p) = 1$ as well for all permutations σ of U .

The distribution of the prediction success is given by the set

$$\mathcal{R}_{P,Q} \stackrel{\text{def}}{=} \{\gamma_{\sigma(Q)}(P) : \sigma \text{ a permutation of } U\}.$$

Let H be the null hypothesis; we have to estimate the position of the observed approximation quality $\gamma_{\text{obs}} \stackrel{\text{def}}{=} \gamma_Q(P)$ in the set $\mathcal{R}_{P,Q}$, i.e. to estimate the probability $p(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}}|H)$. Standard randomization techniques – for example Manly (1991), Chapter 1 – can now be applied to estimate this probability.

If $p(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}}|H)$ is low – conventionally in the upper 5% region –, the assumption of randomness can be rejected, otherwise, if

$$p(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}}|H) > 0.05,$$

we call the rule (random) *casual*.

Example 1. Consider the following information system:

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

The rule $\{p, q\} \rightarrow d$ is perfect, since $\gamma_{\{p,q\}}(d) = 1.0$. Furthermore, the rule is deterministic casual, because every instance is based on a single observation only.

Now suppose that we have collected three additional observations:

U	p	q	d
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 $\{\sigma(p), \sigma(q)\} \rightarrow d$ and their approximation qualities. The distribution of the approximation qualities of the 720 possible matching rules is given in Table 1.

Table 1: RESULTS OF RANDOMIZATION ANALYSIS; 6 OBSERV.

$\gamma_{\mathcal{R}}$	Number of cases	$\hat{p}(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}} H)$	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'

Given the 6-observations example, the probability of obtaining a perfect approximation of d by $\{p, q\}$ under the assumption of random matching, is 0.067 which is by far smaller than in the 3-observations example, but not convincing enough, using conventional $\alpha = 0.05$, to decide that the rule is sufficiently significant to be not casual. \square

A problem similar to casualness of attributes, namely, the *reliability of rules*, was addressed by Krusińska et al. (1992a). The authors define an index which they call *strength* of a rule by counting the number of objects a rule refers to. It was argued that strength of a rule is connected with the possibility that such a rule will be observable in a population (Krusińska et al. (1992a), p. 253). Applying the randomization arguments above, it is easy to see that there are situations in which the relation "lower strength is monotone related to higher randomness" does not hold, as the following example demonstrates:

U	p	q
1	0	a
2	1	b
3	1	b
4	2	c
5	2	c

The strength of the rule

$$(3.1) \quad (\forall x \in U)[f(x, p) = 0 \Rightarrow f(x, q) = a]$$

is smaller than the strength of any other rule. Looking at all possible randomized predictions $p \rightarrow q$, we observe that there is only 1 (out of 120) possibilities in which $f(x, q) = a$ will be predicted by an element of p . Considering the other rules, we observe that there are, for example, 4 possibilities in which b will be predicted perfectly by a rule at random. Therefore, although the rule (3.1) has a lower strength than

$$(3.2) \quad (\forall x \in U)[f(x, p) = 1 \Rightarrow f(x, q) = b],$$

in this situation, rule (3.1) is not as likely to be produced at random as rule (3.2)

3.2 How the randomization procedure works

The proposed randomization test procedure is one way to model errors in terms of a statistical approach. We neither want to reiterate a general discussion of choosing randomization over other techniques (or vice versa), nor look at the different views of the world held by Fisherian and Neyman-Pearson statisticians (see e.g. Edgington, 1987, Manly, 1991, Efron & Tibshirani, 1993), but we should like to put forward several arguments which provide justification for the proposed randomization procedure of testing the casualness of a rough set rule system.

Randomization is a statistical technique which does not require a representative sampling from a population which is a theoretical generalization of the sample under study, because the randomization uses only information within the given sample. The method is well in accord with the philosophy behind RSDA and, indeed, soft computing, whose motto is

LET THE DATA SPEAK FOR THEMSELVES.

This aspect is in contrast to most other statistical techniques. Even the bootstrap technique (discussed in the rough set context in Tsumoto & Tanaka, 1996) needs some parametric assumptions, because one has to suppose that the percentages of the observed equivalence classes are suitable estimators of the latent probabilities of the equivalence classes in the population.

Because our approach is aimed to test the casualness of a rule system – and assume for a moment that this assumption really holds –, 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,

**Table 2: STATE COMPLEXITY OF INFORMATION SYSTEMS
WITH A MODERATE NUMBER OF ATTRIBUTES**

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

even when there are only a few number of attributes (Table 2). We observe that any real life data base contains only few data with respect to the state complexity. Suppose that we sample 100 observations, and use 10 attributes with four different values each. We observe empirical casualness with 100 different equivalence classes with $\hat{\pi} = 0.01$ per class. If there is no structure at all within the data, the probability of observing any class is $\pi = 0.000001$. Given the small empirical basis, we cannot decide whether $\hat{\pi} = 0.01$ is near the true value π or not. We need additional modeling assumption to narrow the huge uncertainty interval $[0.000001, 0.01]$. Because randomization techniques do not need the assumption of representativeness, we do not have the problem of modeling the sampling process and restrictions within the data.

In order to show that the randomization procedure really works – and has a reasonable power, if we know the dependency structure of the attributes –, we have done a small scale simulation study. We assume nine equivalence classes in θ_Q and three equivalence classes in θ_P . There are three rules $q_1 \rightarrow p_1, q_2 \rightarrow p_2, q_3 \rightarrow p_3$, which are assumed to hold without any error. Any observation within the other six classes of θ_Q was randomly assigned to one of the three classes of θ_P . The percentage of the three rules – which is the true value of the approximation quality γ – is varied by

γ			
0.0	0.1	0.2	0.3

We have performed 100 simulations using $N = 10, 20, \dots, 70$ observations, and 1000 simulated randomizations within each simulated trial.

Figure 1 shows the problem of granularity: Given $N = 10$ observations and a true value of $\gamma = 0.0$, the expectation of $\hat{\gamma}$ is about 0.32; the granularity overshoot vanishes at about $N = 40$.

Figure 2 presents the test characteristic of the randomization tests using the conventional α -risk of 5%. Given no effect ($\gamma = 0.0$), we see that the recovered α of the randomization procedure has its maximum at $N = 30$ with $\alpha = 1 - \beta = 0.05$. This conservative behaviour of the test is due to the following: If the sample size is very low, the number of possible $\hat{\gamma}$ -estimations is limited. Since any randomized $\hat{\gamma}$ -value which is equal to the observed approximation quality counts for randomness, the hypothesis “casualness” will get a bit more probability than it should. If the sample size gets larger and $\gamma = 0.0$ holds, it will be very unlikely that an observation shows an approximation quality

Figure 1: EXPECTATION OF APPROXIMATION QUALITY,
GIVEN SAMPLE SIZE AND γ

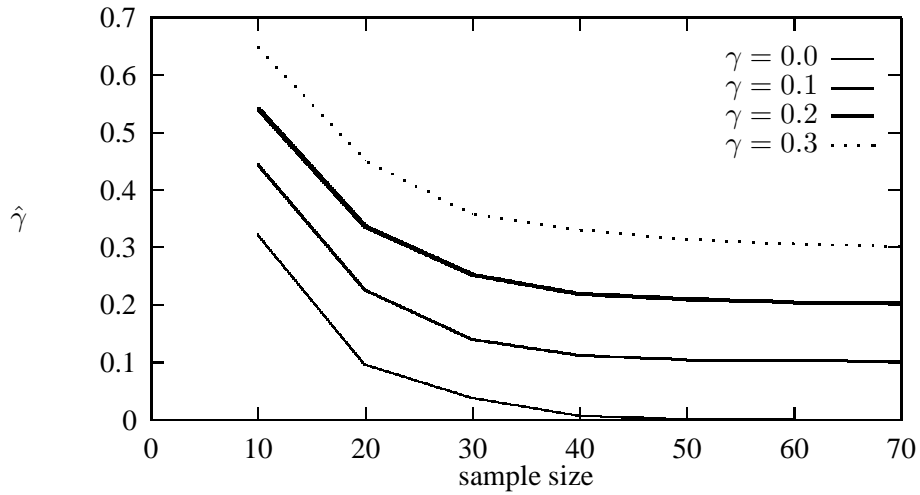
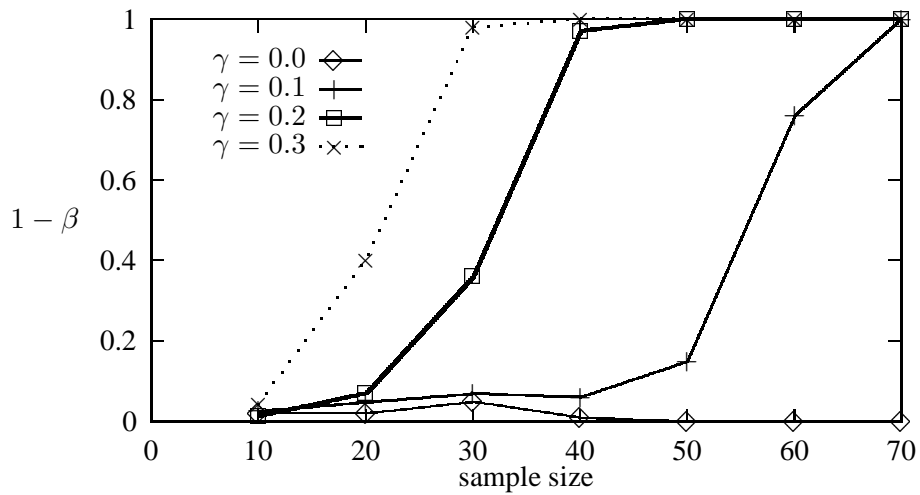


Figure 2: TEST CHARACTERISTIC OF THE RANDOMIZATION PROCEDURE ($\alpha = 5\%$)



$\hat{\gamma} > 0.0$ and because $\hat{\gamma} = 0.0$ is the minimal empirical outcome, the power of the test will approach 0 for larger sample sizes.

Although the behaviour of the test characteristic given $\gamma = 0.0$ is somewhat peculiar, the power curves

of an effect $\gamma > 0.0$ show that the randomization test has a reasonable power – at least in the chosen situation.

Inspecting the power curves we observe that an effect of 30% stable rules – i.e. $\gamma = 0.3$ – will result in a steep power curve.

Obviously, 30% is not 100% and therefore, it should be noted that the proposed procedure is an omnibus type test, and a significant result may **not** be interpreted in the sense that the given rule system is consistent, but that there (mutually) exists a subset of consistent rules within the given rule system. In other words: A significant test result is a minimal requirement for a rule system!

3.3 Computational considerations

It is well known that randomization is a rather expensive procedure, and one might have objections against this technique because of its cost in real life applications. However, we think that a simulation of 1000 randomized $(\sigma(Q), P)$ -assignments will be good enough to get an impression of the probability of casualness in the given sample. If $f(N)$ is the time complexity for performing the computation of γ , the time complexity of the simulation based randomization procedure is $1000f(N)$. This is not too bad, if we compare this time complexity with the one of finding a reduct in a set of $|Q|$ attributes, which is known to be NP-hard in the attribute number.

Let $g(|Q|)$ be the number of evaluated sets of attributes for searching rough reducts of the type $R \rightarrow P$ ($R \subseteq Q$), we need $g(|Q|) \cdot f(N)$ computations of approximation qualities $\gamma_R(P)$; $R \subseteq Q$. Therefore, the cost of the simulation based randomization procedure is well within the complexity of the whole rough set approach. If randomization is too costly for a data set, RSDA itself will not be applicable in this case.

It is an additional task to investigate the possibilities to speed up the computation of the significance of a given rule system. Some simple short cuts such as a check whether the entropy of the Q partition is near $\log_2(N)$ may avoid superfluous computation. Furthermore, simple and less costly procedures can be used to transform the raw data into a two-dimensional contingency table of cross-classifying θ_Q by θ_P . Simulations – or even exact methods analogous to those of Metha & Hilton (1993) – can be done more efficiently using the computed contingency table. For our re-analysis of the published data sets below it was not necessary to speed up the computations.

3.4 Conditional casual attributes

We call an attribute q within a minimal determining set Q for P *conditional casual*, if there are only a few observation in which the attribute q is needed to predict P . This will be made concise below.

Example 2. Consider the following information system:

U	q_1	q_2	p	U	q_1	q_2	p
1	0	0	0	5	1	0	1
2	0	2	0	6	1	2	1
3	0	2	0	7	1	2	1
4	1	1	0	8	0	1	1

The rule $\{q_1, q_2\} \rightarrow p$ is perfect, and we also have $\gamma_{q_1}(p) = \gamma_{q_2}(p) = 0$. However, the influences of the attributes q_1 and q_2 differ: Whereas attribute q_1 is essential to predict p , the attribute q_2 is needed only to explain the two additional elements 4 and 8. \square

As in the preceding section, our statistical approach is to compare the actual $\gamma_Q(P)$ with the results of a random system: For each permutation σ of U and each $q \in Q$ we obtain a new information function $f^{\sigma,q}$ by setting

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

The resulting approximation quality of P by Q is denoted by $\gamma_{Q,\sigma(q)}(P)$, and the distribution of the prediction success is given by the set

$$\mathcal{R}_{P,Q,q} \stackrel{\text{def}}{=} \{\gamma_{Q,\sigma(q)}(P) : \sigma \text{ a permutation of } U\}.$$

As above, if the position of $\gamma_{Q,\sigma(q)}(P)$ is in the upper 5% region, 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.

In rough set analysis, the decline of the approximation quality when omitting one attribute is usually used to determine whether an attribute within a minimal determining set is of high value for the prediction. This approach does not take into account that the decline of approximation quality may be due to chance.

Example 3. The following example 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 rough analysis methods.

U	q	r_1	r_2	r_3	p	U	q	r_1	r_2	r_3	p
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 $q \rightarrow p$ has the approximation quality $\gamma_q(p) = 0.5$. Assume that an additional attribute r is conceptualized in three different ways:

- A fine grained measure r_1 using 8 categories,

- A medium grained description r_2 using 4 categories.
- A coarse description r_3 using 2 categories, and

For $1 \leq i \leq 3$ we have $\gamma_{\{q,r_i\}}(p) = 1$, so that each of these approximations is perfect. If we regard $\gamma_q(p) = 0.5$ as the value of the decline of the approximation quality when leaving out attribute r_i in the prediction of p , we have a situation in which standard rough set dependency analysis does not distinguish between the alternate descriptions with respect to the additional attribute r_i , $1 \leq i \leq 3$.

If we consider the expectation $E[\gamma_{q,\sigma(r_i)}(p)]$, we observe that

$$\begin{aligned} E[\gamma_{q,\sigma(r_1)}(p)] &= 1, \\ E[\gamma_{q,\sigma(r_2)}(p)] &= 0.88, \\ E[\gamma_{q,\sigma(r_3)}(p)] &= 0.624. \end{aligned}$$

The statistical approach offers additional information to evaluate the increase of the approximation quality, if we add one of the r_i attributes to the left side of the prediction rules.

- Any attribute s with the same frequency distribution as the values $f(x, r_1)$, $x \in U$, is expected to have approximation quality $\gamma_{\{q,s\}}(p) = 1$. Therefore we cannot trust the rules derived from the description $\{q, r_1\} \rightarrow p$, because the attribute r_1 is exchangeable with any random generated attribute $s = \sigma(r_1)$.
- The expectation of a random generated rule system with an attribute $s = \sigma(r_3)$ is only $\gamma_{\{q,s\}}(p) = 0.624$, and thus by far smaller than the observed value $\gamma_{\{q,r_3\}}(p) = 1$.
- The result of the 4 category example is in between.

Whereas the statistical evaluation of the additional predictive power of the three chosen attribute differs, the analysis of the decline of the approximation quality tells us nothing about these differences. \square

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.

3.5 Cross validation of learned dependencies

If rough set analysis is used to learn the best subset of Ω to determine P , a simple randomization procedure is not sufficient, because it does not reflect the optimization of the learning procedure.

A simple approach is to split U into a learning subset and a test subset of objects. Within the learning subset, the testing procedures may be used as a guide for including or eliminating attributes. Within the test subset the same procedure can be used to validate the chosen attributes.

If the test procedure shows a significant result, the prediction using the attributes from the learning set is validated, because the attributes show predictive power in another independent set of objects. If the

test procedure does not show a significant result, there are too few rules which can be used to predict the decision attributes from the learned attributes.

A significant result is a minimal requirement for checking the predictive power of the reduct $R \rightarrow P$ under study. A significant result should be interpreted as “some of the rules within the rule system $R \rightarrow P$ are consistent in the test subset”. Note, that these rules need not be the same as those in the learning subset! Therefore, a significant result using the test set of objects is not enough to validate the rules derived from the learned attributes.

To test the stability of rule, we split U into learning (e.g. the first half of the data set) and test objects (e.g. the second half of the data set) and use the split as an additional prediction attribute (e.g. “time”). If the additional attribute is not conditional casual, the learning rules distinguish between learning and test subset and we need the additional attribute to describe the rules. Therefore, the rules differ between learning set and test set. If the additional attribute is conditional casual, the hypothesis that the rules in both sets of objects are identical should be kept. An example of this approach is given in Chapter 4.1 (Table 5).

4 Reanalysis of sample information systems

4.1 Duodenal ulcer data

All data used in this paper are obtainable from `ftp://luce.psycho.uni-osnabrueck.de/home/roughdat/data.zip`, and all calculations are done using the GROBIAN system of Düntsch et al. (1996) which, in turn, uses some routines from Gwaryś & Sienkiewicz (1993).

One of the first rough set analyses published was the study of Pawlak et al. (1986) which describes patients after highly selective vagotomy (HSV) for duodenal ulcer. An enhanced data set was used in Słowiński & Słowiński (1990) and Słowiński (1992a), and this data set will be used in the sequel.

The attribute “Visick grading” (Attr. 12) determines a partition of the set of patients. Pawlak et al. (1986) obtained – using rough set analysis – that the attribute set R , consisting of

- 3: Duration of disease
- 4: Complication
- 5: Basic HCI concentration
- 6: Basic Vol. of gastric juice
- 9: Stimulated HCI concentration
- 10: Stimulated Vol. of gastric juice

suffices to predict attribute 12 (“Visick grading”). Based on the decline of the approximation quality it was speculated that the attribute sets

$$A \stackrel{\text{def}}{=} \{4, 5, 6, 9, 10\}, B \stackrel{\text{def}}{=} \{3, 4, 6, 9, 10\}, \text{ or } C \stackrel{\text{def}}{=} \{3, 4, 5, 6, 10\}$$

are candidates for future research.

The results of the randomization based on 1000 simulations for each test are given in Table 3. Col. 1 shows the attributes under consideration, col. 2 the observed approximation quality γ of this set, col. 3 the estimated position of γ in the distribution of the random matching assumption, and col. 4 the estimated 5% cutpoint in the distribution of gamma assuming random matching.

Table 3: REANALYSIS OF THE DUODENAL ULCER DATA, I

Attributes	γ_{obs}	$\hat{p}(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}} H)$	$\gamma_{\mathcal{R}}(\alpha = 5\%)$	interpretation
3,4,5,6,9,10	0.795	0.013	0.770	not casual
.,4,5,6,9,10 (A)	0.590	0.153	0.623	casual
3,.,5,6,9,10	0.516	0.199	0.557	casual
3,4,.,6,9,10 (B)	0.680	0.018	0.656	not casual
3,4,5,.,9,10	0.549	0.084	0.556	casual
3,4,5,6,.,10 (*)	0.631	0.008	0.590	not casual
3,4,5,6,9,.(C)	0.648	0.011	0.607	not casual

We observe that with this data set, the prediction success of the attribute set $\{3, 4, 5, 6, 9, 10\}$ is satisfactory. The proposed attribute sets B and C are not casual, whereas the proposed attribute set A is casual. Furthermore, one interesting attribute set (indexed by *) has been overlooked.

The analysis of attributes within R (Table 4) are checked using the technique of determining the conditional casualness. The underlined attribute in col. 1 is the attribute under study.

Table 4: REANALYSIS OF THE DUODENAL ULCER DATA, II

Attribute.	decline of γ_{obs}	overall γ_{obs}	$\hat{p}(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}} H)$	$\gamma_{\mathcal{R}}(\alpha = 5\%)$	interpretation
<u>3,4,5,6,9,10</u>	0.590	0.795	0.182	0.828	cond. casual
<u>3,4,5,6,9,10</u>	0.516	0.795	0.099	0.811	cond. casual
3,4, <u>5</u> ,6,9,10	0.680	0.795	0.394	0.844	cond. casual
3,4,5, <u>6</u> ,9,10	0.549	0.795	0.107	0.811	cond. casual
3,4,5,6, <u>9</u> ,10	0.631	0.795	0.127	0.811	cond. casual
3,4,5,6,9, <u>10</u>	0.648	0.795	0.310	0.844	cond. casual

The astonishing result: All attributes are conditional casual within R . This means that there are always only a few of the 122 observations which can be predicted additionally by introducing the attribute under study into the set. If we doubled all observations and analysed the set of 244 objects, no attribute would be conditional casual.

Thus, one could argue that the number of observations in the duodenal ulcer information system is too small to determine influences of the attributes within R .

The attribute set discussed in Pawlak et al. (1986) was based on a reduct searching procedure. In order to discuss the cross validation procedure, we split the data set into 2 subsets containing 61 cases each. The proportion of the categories of the dependent attributes (Attr. 12) are matched in both subsets; the lower subject codings are gathered in the learning set, the higher ones in the test set. Tab. 5 shows the result of the cross validation procedure.

In the learning set, the attributes 3, 4, 5, and 6 show a quite reasonable result (s. Tab. 5). This result cannot be replicated in the test set. Putting learning data set and test data set together shows a significant influence of the data set coding (“time”; 1 = first half of the data set; 2 = second half of the data set).

The random matching analysis of the attribute sets shows that the overall success of R is satisfactory and that some – but not all – speculations about reducing R are valid. Furthermore, the result suggests a reduction of the number of attributes within R , because all attributes are conditional casual. Additionally, the cross-validation procedure shows a huge internal heterogeneity of the data set.

4.2 Earthquake data

In Teghem & Benjelloun (1992), the authors search for premonitory factors for earthquakes by emphasizing gas geochemistry. The partition attribute (attribute 16) was the seismic activity on 155 days measured on the Richter scale. The other attributes were radon concentration measured at 8 different locations (attributes 1-8) and 7 measures of climatic factors (attributes 9-15). A problem with the information system was that it has an empty core with respect to attribute 16, and that an evaluation of some reducts turned out to be difficult.

The statistical evaluation of some of the information systems proposed by Teghem & Benjelloun (1992) gives us additional insights (Tab. 6).

We see that the proposed set $\{1, 2, 3, 6, 12\}$ is casual, although it is a reduct (and thus has perfect approximation quality), and that $\{1, 4, 6\}$ is casual, too. Based on the results of our statistical evaluation procedure, the most promising model discussed by the authors seems to be the reduct $\{1, 2, 6, 8\}$, or, if cheaper measurement equipment is preferred, a choice of the measurement locations $\{1, 2\}$.

4.3 Rough set analysis of Fisher’s Iris Data

Teghem & Charlet (1992) use the famous Iris data first published by Fisher (1936) to show the applicability of rough set dependency analysis for problems normally treated by discriminant analysis. The set U consists of 150 flowers characterized by five attributes namely,

1. Petal length,
2. Petal width,
3. Sepal length,
4. Sepal width, and

Table 5: REANALYSIS OF THE DUODENAL ULCER DATA, III

Learning Set		
Variables in system	γ_{obs}	$\hat{p}(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}} H)$
3,4,5,6	0.82	0.01
<u>3</u> ,4,5,6	0.32	0.02
3, <u>4</u> ,5,6	0.30	0.01
3,4, <u>5</u> ,6	0.59	0.05
3,4,5, <u>6</u>	0.43	0.17

Test Set		
Variables in system	γ_{obs}	$\hat{p}(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}} H)$
3,4,5,6	0.39	0.46
<u>3</u> ,4,5,6	0.31	0.55
3, <u>4</u> ,5,6	0.21	0.63
3,4, <u>5</u> ,6	0.21	0.35
3,4,5, <u>6</u>	0.34	0.59

Learning & Test Set		
Variables in system	γ_{obs}	$\hat{p}(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}} H)$
3,4,5,6,time	0.61	0.01
<u>3</u> ,4,5,6,time	0.32	0.03
3, <u>4</u> ,5,6,time	0.25	0.03
3,4, <u>5</u> ,6,time	0.40	0.05
3,4,5, <u>6</u> ,time	0.39	0.16
3,4,5,6, <u>time</u>	0.43	0.05

5. A partition attribute.

Table 7 validates the argument that only the attribute set $\{3, 4\}$ should be used to predict the partition attribute.

5 Conclusion

Gathering evidence in procedures of Artificial Intelligence should not be based upon casual observations. Our approach shows how – in principle – a system using the rough set dependency analysis will defend itself against randomness.

The reanalysis of three published data sets shows that there is an urgent need for such a technique: Parts of the claimed results using the first two data sets are invalidated, some promising dependencies

Table 6: REANALYSIS OF THE EARTHQUAKE DATA

Variables in system	γ_{obs}	$\hat{p}(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}} H)$
model: 1,2	0.83	0.01
<u>1</u> ,2	0.46	0.01
1, <u>2</u>	0.56	0.01
model: 1,4,6	0.85	0.63
model: 1,2,4,5	1.00	0.02
<u>1</u> ,2,4,5	0.95	0.33
1, <u>2</u> ,4,5	0.86	0.02
1,2, <u>4</u> ,5	0.95	0.26
1,2,4, <u>5</u>	0.92	0.08
model: 1,2,4,6	1.00	0.01
<u>1</u> ,2,4,6	0.90	0.10
1, <u>2</u> ,4,6	0.85	0.01
1,2,4,6	0.93	0.11
1,2,4, <u>6</u>	0.92	0.08
model: 1,2,6,8	1.00	0.01
<u>1</u> ,2,6,8	0.92	0.09
1, <u>2</u> ,6,8	0.88	0.05
1,2, <u>6</u> ,8	0.93	0.06
1,2,6, <u>8</u>	0.92	0.09
model: 1,2,3,6,12	1.00	0.18

are overlooked and, as we show using data of Study 1, our proposed cross-validation technique offers a new horizon for the interpretation. Concerning Study 3, the conclusions of the authors are validated.

As we demonstrate above in Study 2, the proposed statistical evaluation of rough set dependencies helps even in an empty core situation, but it is also applicable if many random errors contaminate the data as in Study 3.

References

- Dütsch, I., Gediga, G. & Jütting, A. (1996). GROBIAN – An engine for rough set data analysis. In *Proceedings of the First International Conference on Practical Aspects of Knowledge Management*, Basel.
- Edgington, E. S. (1987). Randomization Tests, vol. 31 of *Statistics: Textbooks and Monographs*. New York and Basel: Marcel Dekker.
- Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Table 7: REANALYSIS OF FISHER'S IRIS DATA

Variables in system	γ_{obs}	$\hat{p}(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}} H)$
model: 1,2,3,4	0.78	0.01
<u>1</u> ,2,3,4	0.75	1.00
1, <u>2</u> ,3,4	0.77	0.94
1,2, <u>3</u> ,4	0.72	0.71
1,2,3, <u>4</u>	0.62	0.01
model: 1,3,4	0.77	0.01
<u>1</u> ,3,4	0.75	0.30
1, <u>3</u> ,4	0.68	0.03
1,3, <u>4</u>	0.61	0.01
model: 3,4	0.75	0.01
<u>3</u> ,4	0.67	0.01
3, <u>4</u>	0.59	0.01

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

Gwaryś, M. & Sienkiewicz, J. (1993). Rough set library, Version 2.0. User manual, Warsaw University of Technology.

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 (1992b), 251–265.

Krusińska, E., Słowiński, R. & Stefanowski, J. (1992b). Discriminant versus rough set approach to vague data. *Appl. Stochastic Models and Data Analysis*, **8**, 43–56.

Manly, B. F. J. (1991). *Randomization and Monte Carlo Methods in Biology*. London: Chapman and Hall.

Metha, C. R. & Hilton, J. F. (1993). Exact power of conditional and unconditional tests: Going beyond the 2x2 contingency table. *The American Statistician*, **47**, 91–98.

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

Pawlak, Z., Słowiński, K. & Słowiński, R. (1986). Rough classification of patients after highly selective vagotomy for duodenal ulcer. *Internat. J. Man-Mach. Stud.*, **24**, 413–433.

Słowiński, K. (1992a). Rough classification of HSV patients. In Słowiński (1992b), 77–94.

- Słowiński, K. & Słowiński, R. (1990). Sensitivity analysis of rough classification. *Internat. J. Man-Mach. Stud.*, **32**, 693–705.
- Słowiński, R. (1992b). Intelligent decision support: Handbook of applications and advances of rough set theory, vol. 11 of *System Theory, Knowledge Engineering and Problem Solving*. Dordrecht: Kluwer.
- Teghem, J. & Benjelloun, M. (1992). Some experiments to compare rough sets theory and ordinal statistical methods. In Słowiński (1992b), 267–284.
- Teghem, J. & Charlet, J.-M. (1992). Use of “rough sets” method to draw premonitory factors for earthquakes by emphasizing gas geochemistry: the case of a low seismic activity context in Belgium. In Słowiński (1992b), 165–179.
- Tsumoto, S. & Tanaka, H. (1996). A common algebraic framework for empirical learning methods based on rough sets and matroid theory. *Fundamenta Informaticae*, **27**, 273–288.
- 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.