Methoδos Primers Series: The aim of this series is to make available concise introductions to topics in Methodology, Evaluation, Psychometrics, Statistics, Data Analysis at an affordable price. Each volume is written by experts in the field, guaranteeing a high scientific standard.

δ

2

Primers Series, Vol. 2

About the book

This is not the first book on rough set analysis and certainly not the first book on knowledge discovery algorithms, but it is the first attempt to do this in a non-invasive way. In this book the authors present an overview of the work they have done in the past seven years on the foundations and details of data analysis. It is a look at data analysis from many different angles, and the authors try not to be biased for – or against – any particular method. This book reports the ideas of the authors, but many citations of papers on Rough Set Data Analysis in knowledge discovery by other research groups are included as well.

About the authors:

Ivo Düntsch is Professor for Computing Science at the Faculty of Informatics of the University of Ulster. His main interests are algebraic logic and the modelling of knowledge under incomplete information.

Günther Gediga teaches in the Faculty of Psychology at the Universität Osnabrück. His main interests are the mathematical foundations of Psychometrics and Data Analysis, a task which needs a precise description of basic mathematical ideas such as those presented in this book.

Düntsch & Gediga Rough set data analysis: A road to non-invasive data analysis

Ivo Düntsch & Günther Gediga

Rough set data analysis

A road to non-invasive knowledge discovery

ISBN
1-903280-01-X

10£ 15€

Methoδos

Methoδos Primers, Vol. 2

The aim of the Methoδos Primers series is to make available concise introductions to topics in

Methodology, Evaluation, Psychometrics, Statistics, Data Analysis

at an affordable price. Each volume is written by experts in the field, guaranteeing a high scientific standard.

Methoδos Publishers (UK)
Methoδos Verlag (D)

Rough set data analysis:
A road to non-invasive knowledge discovery

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

Günther Gediga
FB Psychologie / Methodenlehre
Universität Osnabrück
49069 Osnabrück, Germany
Guenther@Gediga.de

First published in 2000
by Methoδos Publishers (UK),
24 Southwell Road
Bangor, BT20 3AQ

ISBN 190328001X
A CIP record for this book is available from the British Library.

Typeset in Times New Roman 10pt.
Manufactured from camera-ready copy supplied by the authors by
Sächsisches Digitaldruck Zentrum
Tharandter Straße 31-33
D-01159 Dresden
Germany

Methoδos Publishers (UK), Bangor
Methoδos Verlag (D), Bissendorf

Contents

Preface

This is not the first book on rough set analysis and certainly not the first book on knowledge discovery algorithms, but it is the first attempt to do this in a non-invasive way. The term *non-invasive* in connection with knowledge discovery or data analysis is new and needs some introductory remarks. We – Ivo Düntsch & Günther Gediga – have worked from about 1993 on topics of knowledge discovery and/or data analysis (both topics are sometimes hard to distinguish), and we felt that most of the common work on this topics was based on at least discussable assumptions. We regarded the invention of Rough Set Data Analysis (RSDA) as one of the big events in those days, because, at the start, RSDA was clearly structured, simple, and straightforward from basic principles to effective data analysis. It is our conviction that a model builder who uses a structural and/or statistical system should be clear about the basic assumptions of the model. Furthermore, it seems to be a wise strategy to use models with only a few (pre-)assumptions about the data. If both characteristics are fulfilled, we call a modelling process *non-invasive*. This idea is not really new, because the "good old non-parametric statistics" approach based on the motto of Sir R. A. Fisher

Let the data speak for themselves,

can be transferred to the context of knowledge discovery. It is no wonder that e.g. the randomisation procedure (one of the flagships of non-parametric statistics) is part of the non-invasive knowledge discovery approach.

In this book we present an overview of the work we have done in the past seven years on the foundations and details of data analysis. During this time, we have learned to look at data analysis from many different angles, and we have tried not to be biased for - or against - any particular method, although our ideas take a prominent part of this book. In addition, we have included many citations of papers on RSDA in knowledge discovery by other research groups as well to somewhat alleviate the emphasis on our own work.

We hope that the presentation is neither too rough nor too fuzzy, so that the reader can discover some knowledge in this book.

Jordanstown, Osnabrück, April 2000

Ivo Düntsch & Günther Gediga

Chapter 1

Introduction

According to the widely accepted description of Fayyad et al. [41], the (iterative) process of knowledge discovery in databases (KDD) consists of the following steps:

KDD 1. Developing an understanding of the application domain, the relevant prior knowledge, and the goal(s) of the end-user.

KDD 2. Creating or selecting a target data set.

KDD 3. Data cleaning and preprocessing; this step includes, among other tasks, removing noise or accounting for noise, and imputation of missing values.

KDD 4. Data reduction: Finding useful features to represent the data depending on the goal of the task. This may include dimensionality reduction or transformation.

KDD 5. Matching the goals to a particular data mining method such as classification, regression, clustering etc.

KDD 6. Model and hypothesis selection, choosing the data mining algorithm(s) and methods to be used for for searching for data patterns.

KDD 7. Data mining.

KDD 8. Interpreting mined patterns.

KDD 9. Acting on discovered knowledge.

The usual starting situation in data mining is a vast amount of data, often in the form of relational tables, from which useful knowledge is to be gathered – information which is not visible to the naked eye. Much of data mining consists of classification or clustering tasks, in other words, of supervised or unsupervised learning, and learning the rules which are associated with the data classification. This is certainly not new; taxonomy has played in major part in the natural sciences for a long time, and many statistical methods are available

to accomplish these tasks. However, one needs to keep in mind that most statistical methods were invented to work well with *experimental* data (which, of course, does not mean that it works exclusively with such data, an example being national statistics data sets), while KDD handles *observed* data. Therefore, it is at first glance not clear whether or how statistical methods are suitable for data mining tasks. Taking the description above as a model for the KDD process, one sees that KDD and statistical modelling are, in a way, complementary as indicated in Table 1.1. Methods which are more specific to KDD, respectively machine

Table 1.1: KDD & Statistical models

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

learning, include decision trees [97, 98] and inductive logic programming [73].

All statistical and KDD methods make external model assumptions. A typical example is

> "We will consider rectangular datasets whose rows can be modelled as independent, identically distributed (iid) draws from some multivariate probability distribution. . . . We will consider three classes of distributions f:
>
> 1. the multivariate normal distribution;
> 2. the multinomial model for cross-classified categorical data, including log-linear models; and
> 3. a class of models for mixed model and categorical data . . . " [106].

Unlike in the previous quote, model assumptions are not always spelled out "in toto", and thus, it is not clear to an observer on what basis and with what justification a particular model is applied. The assumption of representativeness, for example, is a problem of any analysis in most real life data bases. The reason for this is the huge state complexity of the space of possible rules, even when there is only a small number of features (Table 1.2).

Similar problems occur in other areas; for example, there are about 10^{23} grammatically correct English sentences of twenty words or less [16]. These problems are not particular to "hard" statistical methods, but also apply to a "soft" approach to KDD:

> "Soft computing differs from conventional (hard) computing in that, unlike hard computing, it is tolerant of imprecision, uncertainty and partial truth. In effect, the role model for soft computing is the human mind. The guiding principle of soft computing is: Exploit the tolerance for imprecision, uncertainty and partial

Table 1.2: State complexity of a system with a moderate number of features

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

> truth to achieve tractability, robustness and low solution cost ... The principal components of soft computing are fuzzy logic, neural network theory, and probabilistic reasoning. [129]

All of these "soft" methods require "hard" parameters outside the observed phenomena – membership degrees, prior probabilities, parameters for differential equations – the origin of which is not always clear. Indeed, the following words of caution by Cohen [12] seem appropriate to mention:

> "Mesmerized by a single-purpose, mechanised 'objective' ritual in which we convert numbers into other numbers and get a yes-no answer, we have come to neglect close scrutiny of where the numbers come from".

An example from the contemporary literature has received some prominence, where a result in form of a number does not give universal satisfaction:

> " 'Forty two!' yelled Loonquawl, 'Is that all you've got to show for seven and a half million years' work?' 'I checked it very thoroughly,' said the computer, 'and that quite definitely is the answer. I think the problem is, to be quite honest with you, that you've never actually known what the question is.' "[2]

Given that the methodology of a KDD process is different from statistical data analysis, and that the use of statistical models may raise more questions than it answers, one can pursue the idea of minimising model assumptions and start with what is there – the observed data.

In this book, we advocate and develop a non–invasive approach to data analysis, whose motto[1] is

Let the data speak for themselves,

[1] attributed to R.A Fisher by Jaynes [57], p. 641

and which

(1.1.1)
Uses minimal model assumptions by drawing all parameters from the observed data,

(1.1.2)
Admits ignorance when no conclusion can be drawn from the data at hand.

This aspect is in contrast to most statistical techniques, not excluding the non-parametric ones. Even the bootstrap [discussed in the rough set context in 122] needs some parametric assumptions, because one has to assume that the percentages of the observed equivalence classes are suitable estimators of the latent probabilities of the equivalence classes in the population.

We hope to show in this book that, given minimal assumptions, the data tell us much more than may seem at first glance, and that additional stringent assumptions are often neither necessary nor, indeed, desirable. These can be applied in a second step if the non-invasive analysis offers conclusions which are not sufficiently sharp for the intention of the researcher.

We will build our presentation on the paradigm of rough set data analysis (RSDA) [89] which draws all its information from the given data. RSDA is mainly applied to information gathered in data tables. However, this is by far not the only knowledge operationalisation to which the non-invasive paradigm can be applied; in Chapter 9 we shall explore applications from other fields such as spatial reasoning and psychometric modelling.

Even though RSDA, as a symbolic method, uses a only few parameters which need simple statistical *estimation* procedures, its results should be controlled using statistical *testing* procedures, in particular, when they are used for modelling and prediction of events. We agree with Glymour et al. [49] that

> " Data mining without proper consideration of the fundamental statistical nature of the inference problem is indeed to be avoided."

The problem here, of course, is not to allow subjective model assumptions to creep in through the back door, when testing procedures are applied. We believe that the procedures described below for significance testing, data discretisation, and model selection satisfy our goal of minimising the use of external parameters, while still delivering good results.

In this book, we collect in a unified way our own work on non-invasive data analysis based on the rough set paradigm. As far as a description of RSDA is concerned, this book is incomplete. There are many important different strands of RSDA which we will not present in detail. We do, however, provide pointers to further literature on topics which we have not covered, but which are in the mainstream of RSDA research.

Chapter 2

Data models and model assumptions

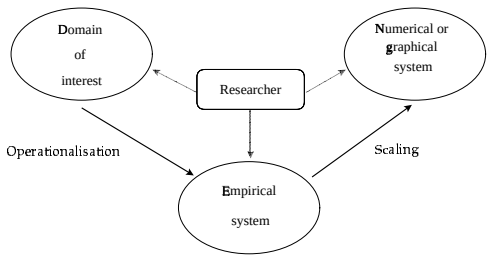
Following Gigerenzer [48], we assume that a *data model* has the following parts (see Figure 2.1):

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

In the centre of the modelling process is

4. The researcher.

Figure 2.1: The data model



The role of the researcher cannot be overemphasised. It is (s)he who chooses the domain of investigation, the data sample to be studied and how to study it, in other words, the operationalisation and the numerical model.

A simple example of a data model in this sense is the following: The domain of interest is the change of temperature. An empirical system is the observation of expansion or contraction of mercury as the temperature changes, and a representation system is the measurement of the behaviour of mercury on a scale such as Celsius, Reaumier, or Fahrenheit. [43]

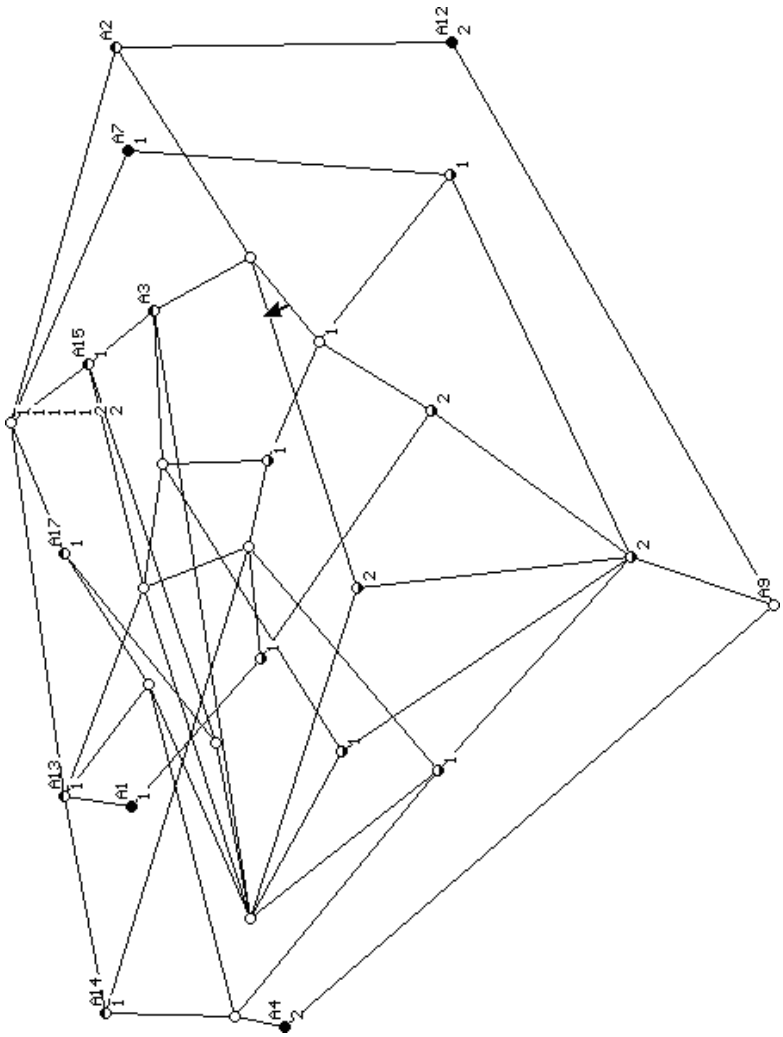
As another example, consider the situation that the knowledge state of individuals in a certain area is to be assessed, which is our domain of interest $\mathcal{D}$. The empirical system consists of the individuals and problems which they are asked to solve. These problems are given by an expert who assumes that they constitute a true operationalisation of the real knowledge states of the individuals. A numerical system for this domain consists of, for example, the test scores achieved by the students.

Not every representation system needs to be numerical. Into the empirical system above, we can introduce two relations: For two problems p, q we say that q *is harder than* p if every individual who solves q also solves p, and there is at least one individual who solves p but not q. For two individuals A, B, we say that A *is better than* B, if A solves all problems which B can solve, and at least one additional problem. A consolidated (graphical) representation is, for example, a weighted line diagram showing these relations such as Figure 2.2 [31].

The diagram should be read as follows:

- Students of two groups – coded by 1 and 2 – and items – coded by A2, A4... – are displayed in one line diagram. Sometimes there are multiple entries; e.g. 1111122 at the top of the line diagram means that there are five students from group 1 and two students from group 2 at this position.
- If we compare students an ascending line from student x to student y means *y is better than x,* since y was able to solve each problem which x could solve, as well as (at least) one additional problem; for example the students at the top of line diagram are better than any other student in the test.
- Ascending lines between problem nodes should be read as "is harder than", e.g. A_2 (on the right) is harder than A_{12}, and both are harder than A_9. The reason: Every student solved A_9 and there is one student who solves $A_1 2$ but who did not solve A_2.
- An ascending line from a subject x to an item q indicates that x was not able to solve q; for example, the student from group 1 coded at item $A_1 2$ was not able to solve problem A_2.

Figure 2.2: Representation of assessment



In traditional statistical modelling, operationalisation and scaling are frequently combined into one procedure. The choice of an empirical model is often determined by the numerical model which is to be applied. But, as we shall see, operationalisation and (possible) scaling are not independent in the KDD process either.

The first step of the KDD process aims at minimising uncertainty arising from operationalisation, and KDD 2 coincides with the choice by the researcher of a domain of interest.

The next step, KDD 3, is concerned with making the data suitable for further analysis e.g by removing noise and imputing missing data. In order to perform this task, we have to know what we mean by noise, and therefore, we must have decided already at this stage on

some kind of numerical system which implies the choice of model assumptions and suitable hypotheses. Similarly, the dimensionality reduction of KDD 4 presupposes the choice of model for a numerical system. It follows that at least the hypothesis and model selection of KDD 6 must take place right after KDD 2 in order to avoid implicit and unstated model assumptions, and not fall into the trap of circular reasoning.

Not clearly separating the operationalisation and scaling processes may result in unstated (or overlooked) model assumptions which may compromise the validity of the result of the analysis. We invite the reader to consult [53] for an indication of what can go wrong when statistical models are applied which are not in concordance with the objectives of the research (if these are known). In particular, all we can hope for is an approximation of the reality that models are supposed to represent, and that there is no panacea for all situations.

The operationalisation is the first source of uncertainty: One question is whether the elements and relations of the empirical model $\mathcal{E}$ are representative for the domain $\mathcal{D}$, another whether the choice of attributes covers the relevant aspects of $\mathcal{D}$. These choices are usually made by a human expert, and thus, to some degree subjective ("model selection bias"). Too many attributes may overfit a model, and too few may result in unexplainable behaviour caused by latent attributes. The choice of each of the parts of the model is a pragmatic decision by researchers: How they want to represent the properties and dependencies of real life criteria in the best possible way, according to their present objectives and their state of knowledge about the world.

In the present context we assume that the operationalisation of data is sufficiently valid to give us a sound basis for analysis, and we do not query how this came about. We take subjectivity at this stage for granted as a fact of life, and start from the empirical model.

Chapter 3

Basic rough set data analysis

3.1 Fundamentals

Rough set theory has been introduced in the early 1980s by Z. Pawlak [89], and has become a well researched tool for knowledge discovery. The basic assumption of RSDA is that information is presented and perceived up to a certain granularity:

> "The information about a decision is usually vague because of uncertainty and imprecision coming from many sources ... Vagueness may be caused by *granularity* of representation of the information. Granularity may introduce an ambiguity to explanation or prescription based on vague information" [91].

In contrast to other methodologies, the original rough set approach uses only the knowledge presented by the data itself, and does not rely on outside statistical or other parameters or assumptions.

The most important areas which RSDA addresses are

- Describing object sets by attribute values,
- Finding dependencies between attributes,
- Reducing attribute descriptions,
- Analysing attribute significance,
- Generating decision rules.

Among others, RSDA has connections to fuzzy sets [23], genetic algorithms [9, 127], evidence theory [63, 111, 113], statistical methods [10, 58], and information theory [33, 126].

Numerous applications as well as the theoretical background of recent enhancements of RSDA can be found in [65, 86, 93, 94].

Recall that an equivalence relation[1] θ on a set U is a binary relation on U which satisfies the following for all $x, y, z \in U$:

$x\theta y$	reflexivity,
$x\theta y \iff y\theta x$	symmetry,
$x\theta y$ and $y\theta z \Rightarrow x\theta z$	transitivity.

If x in U, then $\theta x = \{y \in U : y\theta x\}$ is the *equivalence class of* x (with respect to θ).

The granularity of information in a chosen situation can be described by an equivalence relation, up to the classes of which objects are discernable. In our context, these are also called *indiscernability relations*. What happens within the classes is not part of our knowledge. However, we are able to count the occurrences of indiscernable objects: Even though we may not be able to distinguish x, y with respect to the equivalence relation at hand, we can nevertheless recognise that they are different elements. In a sense, this is like looking at identical twins: We know there are two of them, but we cannot tell one from another.

More formally, we call a pair $\langle U, \theta\rangle$ an *approximation space*, where U is a finite set, and θ is an equivalence relation on U. If $X \subseteq U$, then we know the elements of X only up to the (union of) classes of θ. This leads to the following definitions:

$$\underline{X}_\theta = \{x \in U : \theta x \subseteq X\} \tag{3.3.1}$$

is the *lower approximation of* X, and

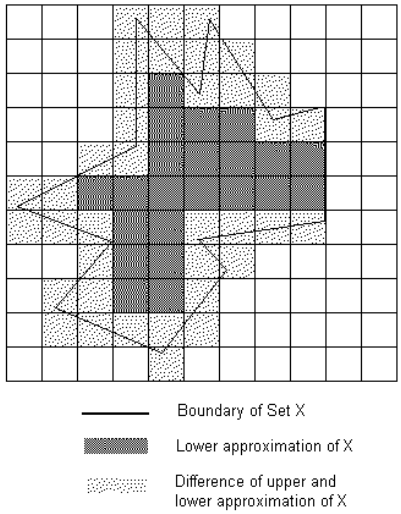
$$\overline{X}^\theta = \{x \in U : \theta x \cap X \neq \emptyset\} \tag{3.3.2}$$

is the *upper approximation of* X with respect to θ. Here, $\theta(x) = \{y \in U : x\theta y\}$ is the equivalence class of x. If θ is understood, we shall usually omit the subscript (superscript). These operators can be interpreted as follows: If $\theta(x) \subseteq X$, then we know for certain that $x \in X$, if $\theta(x) \subseteq U \setminus \overline{X}$ we are certain that $x \notin X$. In the *area of uncertainty* – also called *boundary* – $\overline{X} \setminus \underline{X}$ we can make no sure prediction since $\theta(x)$ intersects both X and $U \setminus X$. Observe that the bounds of the equivalence classes are crisp: We can recognise whether an element x of U is in a class of θ or not (see Fig. 3.1).

We note in passing that the lower (upper) approximation in $\langle U, \theta\rangle$ is the interior (closure) operator of the topology on U whose non empty open sets are the union of equivalence classes. This topology is sometimes called *Pawlak topology* [e.g. by 64]; the terminology seems somewhat unfortunate, since it has been known for some time – since before the advent of rough sets – that on a finite set U there are natural bijective correspondences among

[1] For basic concepts of sets and relations we invite the reader to consult [36]

Figure 3.1: Rough approximation



- The set of all equivalence relations on U,
- The set of topologies on U in which each closed set is open,
- The set of all regular (not necessarily T_1) topologies on U,

see for example [55] and the references therein.

A *rough set* (of the approximation space $\langle U, \theta \rangle$ is a pair of the form $\langle \underline{X}, \overline{X} \rangle$, where $X \subseteq U$.
A subset X of U is called *definable* if $\underline{X} = \overline{X}$. In this case, X is empty or a union of equivalence classes of θ, and the area of uncertainty is $\emptyset$.

The collection of subsets of a set U forms a Boolean algebra under the operations $\cup, \cap, -$. With rough sets we would expect a different behaviour, since, for example, it is not immediately clear what should be understood by a statement such as

$$x \text{ is not an element of a rough set } A = \langle \underline{X}, \overline{X} \rangle.$$

We consider two possibilities of interpretation:

- x is not in the lower approximation of X,
- x is not in the upper approximation of X.

This leads to a structure which has operations that behave like intersection and union, and also has the two forms of negation listed above.

More concretely, the collection of all rough sets on a set U can be made naturally into an algebraic structure $\mathfrak{R}$ as follows [14, 95]:

$$\begin{aligned}
\langle \underline{X}, \overline{X} \rangle + \langle \underline{Y}, \overline{Y} \rangle &\stackrel{\text{def}}{=} \langle \underline{X \cup Y}, \overline{X \cup Y} \rangle, \\
\langle \underline{X}, \overline{X} \rangle \cdot \langle \underline{Y}, \overline{Y} \rangle &\stackrel{\text{def}}{=} \langle \underline{X \cap Y}, \overline{X \cap Y} \rangle, \\
\langle \underline{X}, \overline{X} \rangle^{*} &\stackrel{\text{def}}{=} \langle -\overline{X}, -\overline{X} \rangle, \\
\langle \underline{X}, \overline{X} \rangle^{+} &\stackrel{\text{def}}{=} \langle -\underline{X}, -\underline{X} \rangle \\
\mathbf{0} &\stackrel{\text{def}}{=} \langle \emptyset, \emptyset \rangle, \\
\mathbf{1} &\stackrel{\text{def}}{=} \langle U, U \rangle
\end{aligned}$$

With these operations, $\mathfrak{R}$ becomes a regular double Stone algebra - a distributive lattice with two additional operations

$$\begin{aligned}
^{*} &: \text{ Pseudocomplement}, \\
^{+} &: \text{ Dual pseudocomplement},
\end{aligned}$$

in which an element x is uniquely identified by x^{+} and x^{*}.

These algebras can serve as semantic models for three valued Łukasiewicz logic [25]. The connections between algebra and logic of rough set systems are explored in some detail in [88]. A different logical approach to the rough set model is given by [84] and [66]: There, the lower approximation is considered a necessity operator $\Box$, and the upper approximation a possibility operator $\Diamond$.

It may be also worthy of mention that RSDA can be interpreted in Shafer's evidence theory [109] in such a way that beliefs are obtained internally from the lower approximation of a set, and plausibility from its upper approximation [111, 113].

Another possible of expressing roughness is by the *disjoint representation*: Instead of $\langle \underline{X}, \overline{X} \rangle$ we consider pairs of disjoint sets such as $\langle \underline{X}, -\overline{X} \rangle$. If, say, $\langle A, B \rangle$ is such a pair, then, in terms of machine learning, we may interpret A as the set of positive examples, B as the set of negative examples, and the rest as the region where anything is possible. Each crisp set $\langle C, -C \rangle$ with $A \subseteq C,\ B \subseteq -C$ is a possible concept to be learned. The corresponding algebraic structures are semi-simple Nelson algebras, and we invite the reader to consult [88] for an in-depth study of this connection.

3.2 Approximation quality

Let us fix an approximation space $\langle U, \theta \rangle$. The main statistical tool of RSDA is the *approximation quality function* $\gamma : 2^U \to [0,1]$: If $X \subseteq U$, then

$$\gamma(X) = \frac{|\underline{X}| + |U \setminus X|}{|U|}, \tag{3.3.3}$$

which is just the ratio of the number of certainly classified elements of U to the number of all elements of U. If $\gamma(X) = 1$, then $X = \underline{X} = \overline{X}$, and X is definable.

It was shown in [35] that the γ approximation is a manifestation of the underlying statistical principle of RSDA, namely, the *principle of indifference*: Within each equivalence class, the elements are assumed to be randomly distributed.

While the approximation quality γ measures the global classification success in terms of the equivalence classes, one can use the same principle for elements and define *rough membership functions* [90]:

For each $X \subseteq U$, let $\mu_X : U \to [0,1]$ be a function defined by

$$\mu_X(x) = \frac{|\theta x \cap X|}{|\theta x|}. \tag{3.3.4}$$

It is easy to see that for all $X, Y \subseteq U$,

$$\mu_X(x) = \begin{cases} 1, & \text{iff } x \in \underline{X}, \\ 0, & \text{iff } x \notin \overline{X}, \end{cases} \tag{3.3.5}$$

$$\mu_{U \setminus X}(x) = 1 - \mu_X(x), \tag{3.3.6}$$

$$\mu_{X \cup Y}(x) \geq \max(\mu_X(x), \mu_Y(x)), \tag{3.3.7}$$

$$\mu_{X \cap Y}(x) \leq \min(\mu_X(x), \mu_Y(x)). \tag{3.3.8}$$

The cases, where equality holds in (3.3.7) and (3.3.8) - and when, consequently, μ_X is a fuzzy membership function, as well as an efficient algorithm to compute rough membership functions are given in [90]. Unlike fuzzy membership, the μ_X values are obtained from the internal structure of the data, and no outside information is needed.

Rough sets and fuzzy sets are geared towards different situations: While at the heart of RSDA is the concept of granularity, mathematically expressed by classes with crisp boundaries, within which no information is available, fuzzy sets describe vagueness of a concept where boundaries among classes are ill–defined. Frameworks to combine the two points of view have been proposed, among others, in [23, 117, 128]. Hybrid systems, where RSDA is used as a pre-processing device for fuzzy methods have been given, for example, in [92, 121]

3.3 Information systems

In most cases, we will not consider only one feature of the objects. An important property of the domain of interest is that it has (possibly) infinitely many dimensions, and that it usually is not well circumscribed in advance. Statistical modelling usually takes as few variables (features) as possible to describe a situation and predict new cases. In contrast – keeping in mind that the choice of dimensions is subjective and may differ from one researcher to the next – we advocate a top down approach in considering in the initial stage as many features as cost and time restraints allow. Later, we will show how to reduce the number of attributes.

The operationalisation of choice is the very basic (and simple) form of an

$$\text{OBJECT} \mapsto \text{ATTRIBUTE}$$

relationship which is at the heart of almost all statistical modelling (and the empirical model of relational databases). More formally, an *information system* $\mathcal{I}$ is a structure $\langle U, \Omega, (V_a)_{a\in\Omega}\rangle$ such that

- U is a finite set of objects.
- Ω is a finite set of mappings $a : U \to V_a$; each $a \in \Omega$ is called an *attribute*.
- V_a is the set of *attribute values* of attribute a.

Since U and Ω are assumed to be finite, we can picture an information system as a matrix such as Table 3.1, which shows part of the famous Iris data [42]. The operationalisation

Table 3.1: Fisher's Iris data [42]

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

assumes the "nominal scale restriction" that

(3.3.9) Each object has exactly one value for each attribute at a given point in time.

(3.3.10) The observation of this value is without error.

Another operationalisation which allows semantically richer situations is given in [37], and it will be discussed in Chapter 9.

3.4 Indiscernability relations

Each attribute set Q determines in a natural way an equivalence relation θ_Q on U, called an *indiscernability relation*: Two elements are in the same equivalence class, if they have the same values under the attributes contained in Q:

(3.3.11) $x \equiv_{\theta_Q} y$ if and only if $a(x) = a(y)$ for all $a \in Q$.

In other words, x and y cannot be distinguished with the attributes in Q.

Each class of θ_Q is determined by exactly one feature vector $\langle t_a \rangle_{a \in Q} \in \prod_{a \in Q} V_a$, and it contains exactly those objects which are described by this vector. These feature vectors will also be called *granules*. With some abuse of language, we denote the granule belonging to $x \in U$ by $Q(x)$, i.e.

(3.3.12) $$Q(x) = \langle a(x) \rangle_{a \in Q}.$$

We will sometimes write $\vec{x}_Q$ instead of $Q(x)$; if $Q = \Omega$, we will just write $\vec{x}$.

The finest equivalence relation which the system gives us is θ_Ω. If the system comes from a database table, then, by the constraints of the relational model, θ_Ω is the identity relation.

The following result shows that increasing the number of attributes in an attribute set leads to a finer partition:

Proposition 3.1. *Let $Q \subseteq P \subseteq \Omega$. Then, $\theta_P \subseteq \theta_Q$.*

Proof. Let $x \theta_P y$; then, by definition of θ_P, we have $a(x) = a(y)$ for all $a \in P$. Since $Q \subseteq P$ by the hypothesis, it follows that in particular $a(x) = a(y)$ for all $a \in Q$, and therefore, $x \theta_Q y$. □

Let d be a new attribute with a set of values V_d and information function $d : U \to V_d$, and suppose that $\emptyset \neq Q \subseteq \Omega$. The aim is to relate the values an object has with respect to the attributes of Q to its value with respect to d. The new attribute is called the *dependent attribute* or *decision attribute*; the elements of Q are called *independent* or *condition* attributes. The structure $\mathbf{D} = \langle \mathcal{I}, d \rangle$ is called a *decision system*; it is called *consistent* if

(3.3.13) $(\forall x, y \in U)[(\forall a \in \Omega) a(x) = a(y)$ implies $d(x) = d(y)]$.

Table 3.2: Television sets

Type	Price	Guarantee	Sound	Screen	d
1	high	24 months	Stereo	76	high
2	low	6 months	Mono	66	low
3	low	12 months	Stereo	36	low
4	medium	12 months	Stereo	51	high
5	medium	18 months	Stereo	51	high
6	high	12 months	Stereo	51	low

In other words, $\mathbf{D}$ is consistent if objects which have the same description under Ω have the same value under f_d. The condition attributes in Table 3.1 are petal length, petal width, sepal length, sepal width, and the decision attribute is the class of the specimen.

The following description of television sets given in Table 3.2 shall serve as another example [34]. Here, $\Omega = \{\text{Price, Guarantee, Sound, Screen}\}$, and d is the decision attribute, which indicates a customer's decision to buy.

Since θ_Ω is the identity relation on U, the decision system trivially fulfils (3.3.13), and therefore it is consistent. The partition induced by θ_d consists of the sets

$$\{1, 4, 5\},\ \{2, 3, 6\}. \tag{3.3.14}$$

3.5 Feature selection

The next question which we investigate is whether it is possible to express the d-value of x by the values of x for some attribute set $Q \subsetneq \Omega$. In other words, we are interested in rules of the form

$$(\forall x \in U)(\forall a \in Q)[a(x) = t_a \text{ implies } d(x) = s]. \tag{3.3.15}$$

This constitutes a mechanism of feature selection which is lossless in the sense that the value of each x on d is uniquely determined by its values on Q; in other words, we have a functional dependency. The algebraic equivalent of (3.3.15) is

$$\theta_Q \subseteq \theta_d. \tag{3.3.16}$$

If this is the case, we say that d is *dependent on* Q, and write this dependency as $Q \Rightarrow d$. For the algebraic properties of dependencies we refer the reader to [28, 83].

One major task of the original RSDA was to find minimal sets Q with property (3.3.15). This leads to the following concept: A *reduct for* d is a set Q of attributes such that $Q \Rightarrow d$, and

for which

(3.3.17) If $S \subsetneq Q$, then $S \not\Rightarrow d$.

In other words, Q determines d, and each proper subset of Q does not. A reduct for d is also called a *relative reduct* to emphasise its dependence on a decision attribute. A *minimal reduct* is a reduct with a minimal number of elements.

More generally we will call $Q \subseteq \Omega$ a *reduct* (of the information system $\mathcal{I}$), if Q is minimal with respect to

(3.3.18) $$\theta_Q = \theta_\Omega.$$

Reducts correspond to keys of a relational database; consequently, as was pointed out in [100] the problem of finding a reduct of minimal cardinality is, in general, NP-hard, and finding all reducts has exponential complexity [115].

Reducts are usually not unique, so that the problem arises which one is most suitable to express the situation. If we use the given data as a set of examples which we want to generalise, this implies the question which reduct generates the rules which can be most effectively used for prediction. We will discuss this question further in Chapter 6.

The intersection of all reducts is called the *core*; each element is of the core is called *indispensable*. Note that if a is in the core, then $a \in P$ for every $P \subseteq \Omega$ for which $\theta_P = \theta_\Omega$.

Let us determine the reducts of the television set example with respect to d. We want to find sets $Q \subseteq \Omega$ of minimal cardinality which satisfy (3.3.15), or, equivalently, (3.3.16). The latter, in turn, is equivalent to

(3.3.19) Each class of θ_Q is a subset of a class of θ_d.

This is the condition we shall use. The classes of the relations θ_Q, where Q has exactly one element are as follows:

$$\begin{aligned}
\text{Price: } & \{1,6\},\{2,3\},\{4,5\} \\
\text{Guarantee: } & \{1\},\{2\},\{3,4,6\},\{5\} \\
\text{Sound: } & \{1,3,4,5,6\},\{2\} \\
\text{Screen: } & \{1\},\{2\},\{3\},\{4,5,6\}.
\end{aligned}$$

For each such θ_Q there is a class which intersects two classes of θ_d, so none of these can be

a reduct. Thus, we consider the sets with two attributes:

$$\begin{aligned}
\text{Price, Guarantee: } & \text{Identity} \\
\text{Price, Sound: } & \{1,6\},\{2\},\{3\},\{4,5\} \\
\text{Price, Screen: } & \{1\},\{2\},\{3\},\{4,5\},\{6\} \\
\text{Guarantee, Sound: } & \{1\},\{2\},\{3,4,6\},\{5\} \\
\text{Guarantee, Screen: } & \{1\},\{2\},\{3\},\{4,6\},\{5\} \\
\text{Sound, Screen: } & \{1,6\},\{2\},\{3\},\{4,5\}.
\end{aligned}$$

This gives us two reducts for d, namely

$$\begin{aligned}
R_1 &= \{\text{Price, Guarantee}\}, \\
R_2 &= \{\text{Price, Screen}\}.
\end{aligned}$$

If $Q \subseteq \Omega$ has three elements and contains neither R_1 nor R_2, then $Q = \{\text{Guarantee, Sound, Screen}\}$. θ_Q has the classes

$$\{1\},\{2\},\{3\},\{4,6\},\{5\}.$$

Since $\{4,6\}$ is not a subset of a class of θ_d, it cannot be reduct for d.

If we want to find the reducts of the system $\mathcal{I}$ (that is, without considering the decision attribute d), we need to find those $Q \subseteq \Omega$ which are minimal with respect to $\theta_Q = \theta_\Omega$; recall that in our example θ_Ω is the identity relation on U. Since no single attribute leads to the identity, we look at the sets containing two attributes, and find the partitions listed below. There, the abbreviations Pr,Gu,So,Sc have the obvious meaning, and we list only non-singleton classes.

$$\begin{aligned}
&\text{Pr, Gu :} \\
&\text{Pr,So :}\{1,6\},\ \{4,5\} \\
&\text{Pr,Sc :}\{4,5\} \\
&\text{Gu,So :}\{3,4,6\} \\
&\text{Gu,Sc :}\{4,6\} \\
&\text{So,Sc :}\{4,5,6\}.
\end{aligned}$$

This gives us the reduct $\{$Pr,Gu$\}$. Since no other reduct can contain $\{$Pr,Gu$\}$ as a subset, we need to check only the following cases:

$$\begin{aligned}
\text{Pr,So,Sc :}&\{4,5\} \\
\text{Gu,So,Sc :}&\{4,6\}
\end{aligned}$$

Thus, $\{$Pr,Gu$\}$ is the only reduct, and thus, it is the core as well.

The concept of a (perfect) reduct can be generalised as follows: A set $Q \subseteq \Omega$ is called an ϵ – *reduct with respect to* d (or just an ϵ – *reduct* if d is understood) if it is minimal with respect to the property

$$|\gamma(Q \rightsquigarrow d) - \gamma(\Omega \rightsquigarrow d)| \leq \epsilon. \tag{3.3.20}$$

If $\epsilon = 0$, we have a reduct as defined above. Observe, that the choice of ϵ is made by the researcher, and depends on the willingness to allow imprecision. Finding ϵ – reducts is computationally expensive [115], and great efforts are being made to find appropriate heuristic methods for reduct computation [e.g. 9, 127].

3.6 Discernability matrices and Boolean reasoning

A transparent method of finding reducts is a cross-classification of objects by assigning to each pair $\langle x, y\rangle \in U^2$ the set $\delta(x, y)$ of all those attributes a for which $a(x) \neq a(y)$ [115]. The result is called a *discernability matrix*. Since the assignment $\langle x, y\rangle \to \delta(x, y)$ is obviously symmetric and $\delta(x, x) = \emptyset$, we need only record the upper triangle. For our television set example, the indiscernability matrix (without the decision attribute d) is given in Table 3.3.

Table 3.3: Discernability matrix

	2	3	4	5	6
1	Pr,Gu,So,Sc	Pr,Gu,Sc	Pr,Gu,Sc	Pr,Gu,Sc	Gu,Sc
2		Gu,So,Sc	Pr,Gu,So,Sc	Pr,Gu,So,Sc	Pr,Gu,So,Sc
3			Pr,Sc	Pr,Gu,Sc	Pr,Sc
4				Gu	Pr
5					Pr,Gu

Proposition 3.2. *[115]*

1. *The core of* $\mathcal{I}$ *is the set*

$$\{q \in \Omega : \delta(x, y) = \{q\} \text{ for some } x, y \in U\}.$$

2. $P \subseteq \Omega$ *is a reduct of* $\mathcal{I}$ *if and only if* P *is minimal with respect to the property*

$$P \cap \delta(x, y) \neq \emptyset \tag{3.3.21}$$

for all $x, y \in U,\ \delta(x, y) \neq \emptyset$.

Proof. 1. "$\subseteq$": Let $a \in \Omega$ be in the core $\mathcal{C}$ of $\mathcal{I}$. By Proposition 3.1 we have $\theta_\Omega \subseteq \theta_{\Omega\setminus\{a\}}$, and, since $a \in \mathcal{C}$, we must have strict inclusion. Thus, there are $x, y \in U$ which are in different classes of θ_Ω, but in the same class of $\theta_{\Omega\setminus\{a\}}$. It follows that only a can distinguish x and y, and hence, $\delta(x,y) = \{a\}$.

"$\supseteq$": Suppose that $\delta(x,y) = \{a\}$ for some $x, y \in U$. Then, a is the only attribute which distinguishes x from y, and therefore, $a \in \mathcal{C}$.

2. Let $M(\Omega)$ be the set of all subsets of Ω which minimally satisfy (3.3.21).

"$\Rightarrow$: Suppose that P is a reduct of $\mathcal{I}$, i.e. $\theta_P = \theta_\Omega$ and P is minimal with this property. Assume first that $P \cap \delta(x,y) = \emptyset$ for some , $y \in U$ with $\delta(x,y) \neq \emptyset$. Then, no element of P distinguishes x and y; since $\theta_P = \theta_\Omega$ it follows that θ_Ω does not distinguish these elements. On the other hand, there is some $a \in \delta(x,y)$, so that x and y are in different classes of θ_a. Since $\theta_\Omega \subseteq \theta_a$ by Proposition 3.1, θ_Ω must distinguish x and y as well, contradicting our assumption.

Next, we show the minimality of P with respect to (3.3.21). Assume that $Q \subsetneq P$ and Q satisfies (3.3.21). Since P is a reduct, i.e. minimally satisfying (3.3.18), there are $x, y \in U$ which are separated by θ_P but not by θ_Q. But then $Q \cap \delta(x,y) = \emptyset$, contradicting the assumption that Q satisfies (3.3.21).

"$\Leftarrow$": Suppose that $P \subseteq \Omega$ minimally satisfies (3.3.21). Assume that $\theta_\Omega \subsetneq \theta_P$. Then, there are $x, y \in U$ such that θ_Ω separates x and y, i.e. in particular $\delta(x,y) \neq 0$, but $P \cap \delta(x,y) = \emptyset$, a contradiction. Next, assume that $\theta_Q = \theta_\Omega$ for some $Q \subseteq P$. If $Q \cap \delta(x,y) = \emptyset$ for some $x, y \in U$ with $\delta(x,y) \neq \emptyset$, then θ_Q does not separate x and y. However, $since \theta_Q = \theta_\Omega$, this contradicts $\delta(x,y) \neq \emptyset$. Therefore, Q satisfies (3.3.21), and the minimality of P implies that $Q = P$. It follows that P is a reduct. □

For our example, we can read off Table 3.3 that {Pr,Gu} is the only reduct.

Associated with a discernability matrix is a *discernability function*, which is a frequently (and successfully) used tool to handle reducts [115]. First, we need some preparation from *Boolean reasoning*: Suppose that $\mathbf{2} = \langle \{0,1\}, \wedge, \vee, - \rangle$ is the two element Boolean algebra. A *Boolean function* is a mapping $f : \mathbf{2}^n \to \mathbf{2}$, where $1 \leq n$, and $\mathbf{2}^n = \underbrace{\mathbf{2} \times \mathbf{2} \times \cdots \times \mathbf{2}}_{n-\text{times}}$. If $\vec{x}, \vec{y} \in \mathbf{2}^n$ we say that $\vec{x} \leq \vec{y}$ if $x_i \leq y_i$ for all $1 \leq i \leq n$. A Boolean function $f : \mathbf{2}^n \to \mathbf{2}$ is called *monotone*, if $\vec{x} \leq \vec{y}$ implies $f(\vec{x}) \leq f(\vec{y})$. If $V = \{y_i : 1 \leq i \leq n\}$ is a set of variables, and $T \subseteq V$, we call T an *implicant of* f, if for any valuation of $\vec{y} \in \mathbf{2}^n$

(3.3.22) $$y_i = 1 \text{ for all } x_i \in T \text{ implies } f(\vec{y}) = 1.$$

Observe that we can regard the left hand side of (3.3.22) as a conjunction, and we can equivalently write

(3.3.23) $$\bigwedge T = 1 \Rightarrow f(\vec{y}) = 1.$$

Thus, an implicant gives us a sufficient condition for $f(\vec{y}) = 1$. A *prime implicant of* f is a subset T of V such that T is an implicant, but no proper subset of T has this property.

Suppose that $\Omega = \{q_1, \ldots, q_n\}$, and $U = \{x_1, \ldots, x_m\}$. For each $q_i \in \Omega$, we let q_i^* be a variable, and, for $\delta(x_i, x_j) \neq \emptyset$, we define $\delta^*(x_i, x_j) = \bigvee\{q_r^* : q_r \in \delta(x_i, x_j)\}$. The *discernability function of* $\mathcal{I}$ is the formal expression

$$\Delta(q_1^*, \ldots, q_n^*) = \bigwedge\{\delta^*(x_i, x_j) : 1 \leq i \lneq j \leq m,\ \delta(x_i, x_j) \neq \emptyset\}. \tag{3.3.24}$$

The discernability function of Table 3.3 is

$$\begin{aligned}\Delta(\mathrm{Pr}^*, \mathrm{Gu}^*, \mathrm{So}^*, \mathrm{Scr}^*) = {} & (\mathrm{Pr}^* \vee \mathrm{Gu}^* \vee \mathrm{So}^* \vee \mathrm{Sc}^*) \wedge (\mathrm{Pr}^* \vee \mathrm{Gu}^* \vee \mathrm{Sc}^*) \wedge (\mathrm{Pr}^* \vee \mathrm{Gu}^* \vee \mathrm{Sc}^*) \\ & \wedge (\mathrm{Pr}^* \vee \mathrm{Gu}^* \vee \mathrm{Sc}^*) \wedge (\mathrm{Gu}^* \vee \mathrm{Sc}^*) \wedge (\mathrm{Gu}^* \vee \mathrm{So}^* \vee \mathrm{Sc}^*) \\ & \wedge (\mathrm{Pr}^* \vee \mathrm{Gu}^* \vee \mathrm{So}^* \vee \mathrm{Sc}^*) \wedge (\mathrm{Pr}^* \vee \mathrm{Gu}^* \vee \mathrm{So}^* \vee \mathrm{Sc}^*) \\ & \wedge (\mathrm{Pr}^* \vee \mathrm{Gu}^* \vee \mathrm{So}^* \vee \mathrm{Sc}^*) \wedge (\mathrm{Pr}^* \vee \mathrm{Sc}^*) \wedge (\mathrm{Pr}^* \vee \mathrm{Gu}^* \vee \mathrm{Sc}^*) \\ & \wedge (\mathrm{Pr}^* \vee \mathrm{Sc}^*) \wedge \mathrm{Gu}^* \wedge \mathrm{Pr}^* \wedge (\mathrm{Pr}^* \vee \mathrm{Gu}^*).\end{aligned}$$

The connection between reducts and the discernability function is as follows [115]:

Proposition 3.3. *Q is a reduct of $\mathcal{I}$ if and only if Q is a prime implicant of Δ.*

Proof. "⇒": Suppose that Q is a reduct of $\mathcal{I}$, and let $\vec{y} \in \mathbf{2}^n$ be a valuation of $\{q_1^*, \ldots, q_n^*\}$ such that $y_i = 1$ for all $q_i \in Q$. We first show that Q is an implicant of Δ. Assume that $\Delta(\vec{y}) = 0$. Then, by definition (3.3.24), there are $1 \leq i \lneq j \leq m$ such that $x_i, x_j \in U$, $\delta(x_i, x_j) \neq \emptyset$ and $x_k = 0$ for all $q_k \in \delta x_i, x_j$. It follows that $\delta(x_i, x_j) \cap Q = \emptyset$, and therefore, Q does not distinguish between x_i and x_j. Since Q is a reduct, we have, in particular, $\theta_Q = \theta_\Omega$, so that, in fact, x_i and x_j cannot be distinguished by any attribute in Ω. Hence, $\delta(x_i, x_j) = \emptyset$, contrary to our assumption.

Suppose that $P \subseteq Q$ is an implicant of Δ, and assume there are $x_i, x_j \in U$ such that $x_i \theta_P x_j$, and Ω distinguishes x_i and x_j, i.e. $\delta(x_i, x_j) \neq \emptyset$. It follows from $x_i \theta_P x_j$ that $P \cap \delta(x_i, x_j) = \emptyset$. Let $\vec{y}$ be a valuation such that

$$y_k = \begin{cases} 1, & \text{if } q_k \in P, \\ 0, & \text{otherwise.} \end{cases} \tag{3.3.25}$$

Then, $\bigwedge\{y_k : q_k \in P\} = 1$, while $\Delta(\vec{y}) = 0$, contradicting the assumption that P is an implicant of Δ. It follows that $\theta_P \subseteq \theta_\Omega$, and the fact that Q is a reduct implies $P = Q$.

"⇐": Suppose that P is a prime implicant of Δ, and let $x_i \theta_P x_j$. Then, $P \cap \delta(x_i, x_j) = \emptyset$. Assume that Ω distinguishes x_i and x_j. Choose a valuation $\vec{y}$ as in (3.3.25). By the same argument as above, we arrive at a contradiction. Since $\theta_P = \theta_\Omega$, P contains a reduct Q. It is straightforward to see that Q is an implicant, and therefore, since P is prime, we have $Q = P$. □

This shows that finding reducts is equivalent to finding the prime implicants of certain Boolean functions. Much work has been done on employing methods of Boolean reasoning for reduct determination and rule finding, and we refer the reader to [77–81, 112, 115].

3.7 Rules

A local condition for dependency of d on Q can be defined by calling a class X of θ_Q *d - deterministic*, if it is contained in a class of θ_d. Then, $Q \Rightarrow d$ if and only if each class of θ_Q is d – deterministic.

Suppose that the class X of θ_Q corresponds to the feature vector $\langle t_a \rangle_{a \in Q}$, and that X intersects, say, the classes $Y_0, Y_1, \ldots, Y_k$, which are associated with the values $s_0, s_1, \ldots, s_k \in V_d$. In this case, we obtain the rule

$$(\forall x \in U)[(\forall q \in Q)a(x) = t_a \text{ implies } d(x) = s_0 \text{ or } \ldots \text{ or } d(x) = s_k]. \tag{3.3.26}$$

We denote the collection of rules of the form (3.3.26) by $Q \rightsquigarrow d$, and, with some abuse of language call $Q \rightsquigarrow d$ a rule as well.

In the television set example *Price* and *Guarantee* completely determine consumer behaviour. This leads to the rules

$$\text{If Guarantee} \begin{cases} = 6 \text{ months:} & \text{low} \\ = 12 \text{ months:} & \text{If Price} = \begin{cases} \text{medium:} & \text{high} \\ \text{otherwise:} & \text{low} \end{cases} \\ \geq 18 \text{ months:} & \text{high.} \end{cases}$$

Observe that the rule system is deterministic. If we only consider the attribute *Guarantee*, then we obtain the rule system

$$\text{If Guarantee} \begin{cases} = 6 \text{ months:} & \text{low} \\ = 12 \text{ months:} & \text{low or high} \\ \geq 18 \text{ months:} & \text{high.} \end{cases}$$

In general, rules obtained from a decision system have the form $\alpha \rightsquigarrow \beta$, where α is a positive Boolean combination of descriptors of the form $a(x) = t$, and β is a disjunction of descriptors of the form $d(x) = s$. One can optimise these rules in two ways: Minimising the number of independent attributes appearing in α, or minimising the number of descriptors. There is a substantial body of literature on this topic, and we point the reader to [6] for details and further references.

3.8 Approximation quality of attribute sets

Even though RSDA is a symbolic method of analysis, it uses counting information provided by the classes of the equivalence relations under consideration. The basic statistic used in RSDA is an extension of Definition 3.3.3 of γ to the partitions obtained from Q and d:

$$\gamma(Q \rightsquigarrow d) = \frac{|\bigcup\{X : X \text{ is a } d\text{ – deterministic class of } \theta_Q\}|}{|U|}. \tag{3.3.27}$$

$\gamma(Q \rightsquigarrow P)$ is called the *approximation quality of Q with respect to d*. It measures the relative frequency of those objects with can be correctly classified as being in a class of θ_d with the knowledge given by the attributes in Q. If $\gamma(Q \rightsquigarrow d) = 1$, then the approximation quality is perfect, and d is dependent on Q.

Note that the approach is quite different to other non-parametric techniques such as the k-nearest neighbour or the kernel density estimation. The k-nearest neighbour technique needs an external definition of a distance or a similarity to start the grouping procedure (which is an extra step with many degrees of freedom) and to describe dependencies among attributes. Such an indirect way to describe dependencies is not needed in the rough set approach. The kernel density estimation needs the additional assumption of the family of the kernel density and the fixing of some parameters (e.g. the unknown variance of a normal kernel). Both assumptions cannot be justified by the data alone, but are part of the numerical model which is a subjective construction of the researcher.

The approximation quality is traditionally used in RSDA as a measure of "goodness of fit" of attribute reduction, and (relative) reducts in particular are regarded as an optimal solution. This is certainly true for the static case of a non-changing data system. However, we shall show below that, in case of prediction, this interpretation has to be taken with some care, as it is conditional on the chosen reduct, and, furthermore, does not take into account possible random influences. Reducts need not be stable under small variations of data [7], and "perfect" rules obtained from reducts need not be statistically significant [30]. We shall address the latter problem in Chapter 4, and propose a different criterion for attribute reduction and model selection in Chapter 6.

Chapter 4

Rule significance

4.1 Significant and casual rules

Any rule based inference system must consider the fact that a rule may be due to chance, and that a rule which is logically valid need not be significant. This is not important, if one considers a static database. However, if the rules are to be used for prediction or classification of unseen data, then the use of random rules is questionable:

> "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". [54]

We observe that γ is not a good indicator, since $\gamma(Q \rightsquigarrow d) = 1$ if, for example, each rule is based on exactly one object. There are numerous statistical methods to test a given hypothesis. Most of these, however, depend on an assumed sampling distribution, which is something we want to avoid in our non-invasive context. If we want to test the significance of rules we must take care that no specific model assumptions are used which are not justified by the data at hand.

One method which is applicable for any kind of data sample and any distribution is *randomisation*:

- "A randomisation test is a procedure that involves comparing a test statistic with a distribution that is generated by randomly reordering the data values in some sense.
- Randomisation tests have the advantage of being valid with non-random samples and allowing the user to choose a test statistic that is appropriate for the particular situation being considered." [70]

In our particular case, the test statistic is the approximation quality γ, and the null hypothesis H_0

The rule $Q \rightsquigarrow d$ is due to chance

is tested by comparing its approximation quality to the approximation quality of the rules obtained by reordering randomly the feature vectors for objects, while keeping the decision values constant.

More formally, let Σ be the set of all permutations of U, and $\sigma \in \Sigma$. We define new attribute functions a^σ by

$$a^\sigma(x) \stackrel{\text{def}}{=} \begin{cases} a(\sigma(x)), & \text{if } a \in Q, \\ a(x), & \text{otherwise.} \end{cases} \tag{4.4.1}$$

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

The value

$$p(\gamma(Q \rightsquigarrow d)|H_0) = \frac{|\{\gamma(Q^\sigma \rightsquigarrow d) \geq \gamma(Q \rightsquigarrow d) : \sigma \in \Sigma\}|}{|U|!} \tag{4.4.2}$$

measures the significance of the observed approximation quality. If $\alpha = p(\gamma(Q \rightsquigarrow d)|H_0)$ is low, traditionally below 5%, then the rule $Q \rightsquigarrow d$ is deemed *significant*, and the H_0 hypothesis can be rejected. Otherwise,if $\alpha \geq 0.05$, we call $Q \rightsquigarrow d$ a *casual rule*. We would like to emphasise that failure to reject the null hypothesis does not mean that it is true, and that α does not signify the probability of the null hypothesis, but the probability of the rule, given that H_0 is true.

The cut-off value of 5% is somewhat arbitrary and can be viewed as a minimal requirement for a valid rule. In case of a "significant" result – which means that the probability of the same or a better result can be achieved by random is lower than 5% – we can safely assume that the rule is – at least in parts – not an arbitrary one. It should be noted that a fixed level of significance (such as the chosen 5%) can be obtained more easily, when the sample size grows [74]; in other words, the relative size of the systematic part in a large sample may be lower than in a small sample for a result to be significant.

As an example, consider the following (contrived) information system [30]:

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

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

Now suppose that we have collected three additional observations:

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

To decide whether the given rule is casual under the statistical assumption, we have to consider all $6! = 720$ possible permuted information systems and the approximation qualities of the randomised rules. This distribution is shown in Table 4.1, with $\alpha = p(\gamma(\{x_1, x_2\} \rightsquigarrow d)|H_0)$. Given the 6-observation example, the probability of obtaining a perfect approxima-

Table 4.1: Results of randomisation analysis; 6 observations

γ	No of cases	α	Example of σ
1.00	48	0.067	$1, 1', 2, 2', 3, 3'$
0.33	288	0.467	$1, 1', 2, 3, 2', 3'$
0.00	384	1.000	$1, 2, 2', 3, 1', 3'$

tion of d by $\{x_1, x_2\}$ under the assumption of random matching, is 0.067 which is smaller than in the 3–observation example, but, using conventional $\alpha = 0.05$, not convincing enough to decide that the rule is sufficiently significant to be not casual.

The example shows that even the optimal value $\gamma = 1$ does not protect against the simple hypothesis that the rule may be drawn by random. Additionally it shows how the insignificance can be changed to significance, if the number of cases per granule is increased. On the other hand, a rule with a lower value of γ may be significant:

U	x_1	x_2	d	U	x_1	x_2	d
1	0	1	0	5	1	4	1
2	0	2	0	6	1	5	2
3	0	2	0	7	1	6	3
4	0	3	0	8	1	7	4

The value of $\gamma(\{x_1\} \rightsquigarrow d)$ is 0.5, which is weak by traditional rough set standards. On the other hand, within the 8! possible permutations there are only $2 \times 4!$ results which have achieved the same γ value, resulting in $p(\gamma(\{x_1\} \rightsquigarrow d)|H_0) = 0.0286 < 5\%$. The value of $\gamma(\{x_2\} \rightsquigarrow d)$ is 1, but the rule is casual.

4.2 Conditional significance

We can use a similar randomisation test to determine the influence of one attribute on the classification success. In traditional RSDA, the decline of the approximation quality when omitting one attribute is usually used to determine whether an attribute within a reduct is of high value for the prediction. This approach does not take into account the possibility that the decline of approximation quality may be due to chance.

Suppose that we want evaluate the contribution of $q \in Q$ to the rule $Q \rightsquigarrow d$. As in (4.4.1), our statistical approach is to compare the actual $\gamma(Q \rightsquigarrow d)$ with the results of a random system. Let H_0, q be the hypothesis

The influence of q on $\gamma(Q \rightsquigarrow d)$ is due to chance.

If σ is a permutation of U and $a \in Q$ we define

$$a^{\sigma,q}(x) \stackrel{\text{def}}{=} \begin{cases} a(\sigma(x)), & \text{if } a = q, \\ a(x), & \text{otherwise.} \end{cases} \tag{4.4.3}$$

Thus, we randomise the values for q within the attribute set Q. The significance of the influence of q is measured by

$$p(\gamma(Q \rightsquigarrow d)|H_0, q) = \frac{|\{\gamma(Q^{\sigma,q} \rightsquigarrow d) \geq \gamma(Q \rightsquigarrow d) : \sigma \in \Sigma\}|}{|U|!} \tag{4.4.4}$$

Here, $Q^{\sigma,q} \rightsquigarrow d$ is the rule in the permuted information system. With the same caveat regarding the 5% cutoff as above, we call q *conditional casual* (in $Q \rightsquigarrow d$), if $0.05 \leq p(\gamma(Q \rightsquigarrow d)|H_0, q)$.

The following example from [30] shows that, depending on the nature of an attribute, statistical evaluation leads to different expectations of the change of approximation quality which are not visible under ordinary RSDA.

Table 4.2: Conditional casualness

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

Consider the information system of Table 4.2. The prediction rule $q \rightsquigarrow d$ has the approximation quality $\gamma(q \rightsquigarrow p) = 0.5$. Now, suppose that an additional attribute r is conceptualised in three different ways:

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

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

In order to show that there is a need for testing the significance of the drop from 1 to 0.5, we can look at the statistical expectations of the prediction success, if we fix the attribute q and have a "random" attribute r_i.

If we do so, we consider the expectation $E[\gamma(\{q, \sigma(r_i)\} \rightsquigarrow d)]$, and we observe that

$$E[\gamma(\{q, \sigma(r_1)\} \rightsquigarrow d)] = 1,$$
$$E[\gamma(\{q, \sigma(r_2)\} \rightsquigarrow d)] = 0.88,$$
$$E[\gamma(\{q, \sigma(r_3)\} \rightsquigarrow d)] = 0.624.$$

The increase from 0.5 to 1 is due to random influences, if we use attribute r_1. Therefore, the drop from 1 to 0.5 cannot be significant, if we eliminate r_1 from the attribute set. Using attribute r_3, the expectation of the prediction quality given a random representation of r_3 and q is only 0.624 and it is very likely that the drop from 1 to 0.5 when eliminating r_3 from the attribute set is significant. The quality of r_2 is in between that of the other two, and must be a subject of statistical testing as well.

The statistical approach offers additional information in the evaluation of the increase (or drop) of the approximation quality, if we add (or remove) one of the r_i attributes to (or from) the left side of the prediction rules.

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

4.3 Sequential randomisation

We see from the denominator $|U|!$ of (4.4.2) that the computational cost of obtaining the significance is feasible only for small values of $|U|$, and more sophisticated tools are needed to apply the randomisation test. A fairly simple method of shortening the processing time of the randomisation test is the adaptation of a sequential testing scheme to the given situation. Because this sequential testing scheme can be used as a general tool in randomisation analysis, we present the procedure in a more general way.

Suppose that θ is a a statistic with realizations θ_i, and that we have fixed a realization θ_c. We can think of θ_c as $\gamma(Q \rightsquigarrow d)$ and θ_i as $\gamma(Q^\sigma \rightsquigarrow d)$. An evaluation of the hypothesis $\theta \geq \theta_c$ given the null hypothesis H_0 can be done by using a sample of size n from the θ distribution, and counting the number k of θ_i for which $\theta_i \geq \theta_c$. The evaluation of $p(\theta \geq \theta_c|H_0)$ can now be done by the estimator $\hat{p}_n(\theta \geq \theta_c|H_0) = \frac{k}{n}$, and the comparison $\hat{p}_n(\theta \geq \theta_c|H_0) \lneq \alpha$ will be performed to test the significance of the statistic. For this to work we have to assume that the simulation is asymptotically correct, i.e. that

$$\lim_{n\to\infty} \hat{p}_n(\theta \geq \theta_c|H_0) = p(\theta \geq \theta_c|H_0). \tag{4.4.5}$$

Assuming independence of the draws – which in our situation is no restriction –, the results of the simulation k out of n can be described by a binomial distribution $p^k(1-p)^{n-k}$ with parameter $p = p(\theta \geq \theta_c|H_0)$. The fit of the approximation of $\hat{p}_n(\theta \geq \theta_c|H_0)$ can be determined by the confidence interval of the binomial distribution.

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

$$H_b : p(\theta \geq \theta_c|H_0)) \in [0, \alpha) \tag{4.4.6}$$

$$H_a : p(\theta \geq \theta_c|H_0)) \in [\alpha, 1] \tag{4.4.7}$$

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

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

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

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

Given a sample of n observations with k observations which count against H_0 and $n - k$ observations which count towards H_0, we can use the the Wald-procedure [123], which defines the likelihood ratio

$$LQ(n,k) = \frac{\sup_{p\in[0,\alpha)} p^k(1-p)^{n-k}}{\sup_{p\in[\alpha,1]} p^k(1-p)^{n-k}}, \tag{4.4.8}$$

and we obtain the following approximative sequential testing scheme:

1. If
$$LQ(n,k) \lneq \frac{s}{1-r},$$
then H_a is true with probability at most s.

2. If
$$LQ(n,k) \gneq \frac{1-s}{r},$$
then H_b is true with probability at most r.

3. Otherwise
$$\frac{s}{1-r} \leq LQ(n,k) \leq \frac{1-s}{r},$$
and no decision with precision (r, s) is possible. Hence, the simulation must continue.

Within this testing procedure the number n of observations is treated as a random variable. If we fix the critical value α, and the precision r, s, any (n, k)-combination has its own $LQ(n, k)$ value and therefore, one of the alternatives must hold. Because the procedure can be used in an iterative manner, we will stop the simulation, if alternative (1) or (2) occurs. It is well known [123] that this procedure is time saving in comparison to testing schemes in which n is fixed.

With this procedure, which is implemented in our rough set engine GROBIAN[1] [29], the computational effort for the significance test in most cases decreases dramatically, and a majority of the tests need less than 100 simulations.

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

Chapter 5

Data discretisation

5.1 Classificatory discretisation

A numerical attribute usually gives rise to many small classes of objects, which, in turn, lead to rules whose significance is below an acceptable level. Collecting numerical values into classes such as intervals or ranges of values is usually called *discretisation* or *horizontal compression*. There are well established methods of achieving a reduction of the number of classes by discretisation which need extra parameters such as minimum size, Euclidean distances, independence degrees or other measures. For an overview of discretisation methods in data mining we invite the reader to consult [78].

Within our non-invasive paradigm we need to find other procedures which may be used, when the numerical information needed for the common discretisation methods is not available or cannot be justified. Such a method was presented in [32]; before we formally describe it we shall look at an example to explain the idea behind the procedure.

Table 5.1: Heart attack information system I

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

Consider the information system of Table 5.1 which we interpret as follows:

- $x_1, \ldots, x_8$ are patients.
- The attribute m is a combined measure of medical indicators for the risk of a heart attack, while p is a combined measure of psychological indicators, see e.g. [26, 103].

- The values of the risk measures are

  1. No risk, 2 – Small risk, 3 – Medium risk, 4 – High risk, 5 – Very high risk.

- The decision variable H is the observation of a heart attack within a predefined time span coded as

  1 – Heart attack, 0 – No heart attack

It is easily seen that $\gamma(\{m,p\} \rightsquigarrow H) = 1$, and thus, the rule is logically valid. One the other hand, the significance analysis described in Chapter 4 shows that the probability to obtain this rule by chance is close to 100%. This is surprising, because the dependency

(5.5.1) High medical or high psychological risk leads to heart attack

is obviously present. But note that (5.5.1) uses far less information than is present in the Table, since there are only the two risk values {high, not high} instead of the five values in the system. If we look more closely at Table 5.1, we observe that for all $x \in U$,

$$m(x) \in \{4,5\} \text{ implies } H = 1,$$
$$p(x) \in \{4,5\} \text{ implies } H = 1,$$

and that conversely

$$H = 1 \text{ implies } m(x) \in \{4,5\} \text{ or } p(x) \in \{4,5\}.$$

If consequently we recode in both V_m and V_u

$$1,2,3 \to 0,\ 4,5 \to 1,$$

we obtain the system shown in Table 5.2.

Table 5.2: Heart attack information system II

U	m	p	H	U	m	p	H
x_1	0	0	0	x_5	0	1	1
x_2	0	0	0	x_6	1	0	1
x_3	0	0	0	x_7	0	1	1
x_4	0	0	0	x_8	1	1	1

We still have the rule $\{m,p\} \rightsquigarrow H$; the significance analysis, however, shows that the chance to get the same result by random is about 2.8%. Hence, this dependency can be considered

significant. The higher statistical strength of the prediction given in the recoded system is due to fact that the risk groups 1, 2, and 3 are identified, as well as the 4 and 5 risk groups. The differences within these risk groups are neglected, and only the difference between the recoded risk groups remains as a characteristic of the set of prediction attributes $Q = \{m, p\}$. This leads to a duplication of rule instances which influences the statistical significance in a positive way.

The general idea uses a binarisation of the information system: For each attribute $q \in \Omega$ let V_q^+ be the set of attribute values which are actually taken by some $x \in U$. For each $t \in V_q^+$ we define a new attribute function $q^t : U \to \{0, 1\}$ by setting

$$q^t(x) = \begin{cases} 1, & \text{if } q(x) = t, \\ 0, & \text{otherwise.} \end{cases}$$

We let $\Omega_q = \{q^t : t \in V_q^+\}$, and $\mathcal{I}_B$ be the information system with attribute set $\Omega^+ = \bigcup_{q \in \Omega} \Omega_q$. The binarisation of the heart attack system is shown in Table 5.3.

Table 5.3: The binarised system $\mathcal{I}_B$

U	m					p					H
	m^1	m^2	m^3	m^4	m^5	p^1	p^2	p^3	p^4	p^5	
x_1	1	0	0	0	0	0	0	1	0	0	0
x_2	0	0	1	0	0	0	1	0	0	0	0
x_3	0	1	0	0	0	1	0	0	0	0	0
x_4	0	0	1	0	0	0	0	1	0	0	0
x_5	0	1	0	0	0	0	0	0	1	0	1
x_6	0	0	0	1	0	1	0	0	0	0	1
x_7	1	0	0	0	0	0	0	0	0	1	1
x_8	0	0	0	0	1	0	0	0	1	0	1

Once the system has been binarised, we do the following:

1. For each $q \in \Omega$ and each $a \in V_d^+$ find all $t \in V_q^+$ for which

   $$(\forall x \in U)[q^t(x) = 1 \text{ implies } d(x) = a]. \tag{5.5.2}$$

2. Let $M^{q,a}$ be the set of all these binary attributes, and define a new (binary) attribute $m^{q,a}$ by

   $$m^{q,a}(x) = 1 \iff q^t(x) = 1 \text{ for some } t \in M^{q,a}, \tag{5.5.3}$$

   $$\iff \max_{t \in M^{q,a}} q^t(x) = 1, \tag{5.5.4}$$

We now replace all attributes $q^t,\ t \in M^{q,a}$ simultaneously by the attribute $m^{q,a}$. This step effectively collects all attribute values of q which do not split the decision class belonging to a into a single attribute value.

For our example, we obtain the reduced information system as shown in Table 5.4 after this step.

Table 5.4: Reduced binary system $\mathcal{I}_B^r$

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

3. We now define a new information system $\mathcal{I}^* = \langle U, \Omega^*, (V_q^*)_{q\in\Omega}\rangle$ as follows:

$$V_q^* = \{m^{q,a} : a \in V_d^+\} \cup V_q \setminus \bigcup_{a\in V_d^+} M^{q,a}.$$

For each $x \in U$ there is exactly one $t_x \in V_q^*$ such that $q^t(x) = 1$ by definition of V_q^*, and we set $q^*(x) = t_x$.

In our example system, the values for m have been reduced to 4, and those for p to 3. As the algorithm shows, this type of data compression is not expensive and easily implemented.

In a further (expensive) step, one can consider minimal reducts of I_B^r before collecting attribute values; this will usually reduce the number of attribute values even further. In our example system, the unique reduct is the set $\{m_{45}, p_{45}\}$, so that in the optimal reduction each attribute has only two values as shown in Table 5.2.

The question arises, whether these transformations keep the dependency structure of the original system, and whether it has an effect on the significance of the attributes. This is answered by the following result from [32]:

Proposition 5.1. *Let $\mathcal{I}^*$ be obtained from the decision system $\mathcal{I}$ as described above. Let $\emptyset \neq Q \subseteq \Omega$, and Q^* be the corresponding attribute set from $\mathcal{I}^*$. Then,*

1. $\gamma(Q \rightsquigarrow d) = \gamma(Q^* \rightsquigarrow d)$.

2. $p(\gamma(Q \rightsquigarrow d)|H_0) \geq p(\gamma(Q^* \rightsquigarrow d)|H_0)$

Proof. 1. Suppose that $\mathcal{P}(Q)$ ($\mathcal{P}(d)$) is the set of all classes of θ_Q (θ_d). Recall that

$$\gamma(Q \rightsquigarrow d) = \frac{\sum |\{X : X \text{ is a } d\text{-deterministic class of } \theta_Q\}|}{|U|}.$$

If Y is a class of $\mathcal{P}(d)$, then

$$Z \stackrel{\text{def}}{=} \bigcup \{X \in \mathcal{P}(Q) : X \subseteq Y\}$$

contains exactly those elements of U which contribute to the Q-deterministic part of Y. Since Z is a class of Q^*, and every d–deterministic class of θ_{Q^*} has this form, the conclusion follows.

2. First, we observe that because attribute values are identified in the the filtration process, for each $R \subseteq \Omega$, each class of θ_{R^*} is a union of classes of θ_R. Thus, given any $\sigma \in \Sigma$, the rule $Q^\sigma \rightsquigarrow d$ will have at least as many deterministic cases as $Q^{*\sigma} \rightsquigarrow d$. It follows that $\gamma(Q^\sigma \rightsquigarrow d) \geq \gamma(Q^{*\sigma} \rightsquigarrow d)$. Thus, for every $\sigma \in \Sigma$ with $\gamma(Q^{\sigma *} \rightsquigarrow d) \geq \gamma(Q^* \rightsquigarrow d)$ we have

$$\gamma(Q^\sigma \rightsquigarrow d) \geq \gamma(Q^{*\sigma} \rightsquigarrow d) \geq \gamma(Q^* \rightsquigarrow d) = \gamma(Q \rightsquigarrow d),$$

the latter by 1. Hence, the numerator of the right hand side of (4.4.2) for $Q \rightsquigarrow d$ is at least as large as that for $Q^* \rightsquigarrow d$, whence the conclusion follows. □

Thus, the approximation quality is the same, and the rule significance is not worse than before. Indeed, in all cases we have tested, there was a remarkable increase of the significance, even in cases where the attributes were highly continuous. In [10] we have discretised Fisher's Iris data with our non-numerical method; the results are shown in Table 5.5. We list the resulting number of classes, and in brackets the number of classes of the un-discretised data. Observe the dramatic fall in the number of classes of the petal attributes.

This classificatory method of discretisation can be extended by considering a generalised type of information system [124]: A *multi-valued information system* is a structure $\langle U, \Omega, (V_a)_{a \in \Omega} \rangle$ such that

- U is a finite set of objects.
- Ω is a finite set of mappings $a : U \to 2^{V_a}$; each $a \in \Omega$ is called a *multi-valued attribute*.
- V_a is the set of *attribute values* of attribute a.

A *hypertuple* is an element of $\prod_{a \in \Omega} 2^{V_a}$, and a *hypergranule* is a hypertuple t which is equal to some $\Omega(x)$, i.e. $t = \langle a(x) \rangle_{a \in \Omega}$. A *hyperrelation* is a collection of hypertuples. A

Table 5.5: Non-numerical discretisation

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	

simple tuple is a hypertuple in which each component is a singleton set. We can regard each (single-valued) informations system $\mathcal{I}$ as a multi-valued one $\mathcal{I}^m$ by considering the singleton $\{a(x)\} \subseteq V_a$ instead of $a(x) \in V_a$. Thus, a granule of $\mathcal{I}$ becomes a simple hypergranule in $\mathcal{I}^m$, and we shall usually identify the two.

Hypertuples can be ordered by setting

$$t_1 \leq t_2 \iff t_1(a) \subseteq t_2(a) \text{ for all } a \in \Omega. \tag{5.5.5}$$

If R is a hyperrelation, we let $\downarrow R = \{t \in \prod_{a\in\Omega} 2^{V_a} : t \leq r \text{ for some } r \in R\}$. We also define $t_1 \vee t_2$ by setting

$$(t_1 \vee t_2)(a) = t_1(a) \cup t_2(a). \tag{5.5.6}$$

Clearly, $t_1 \vee t_2$ is the supremum of $\{t_1, t_2\}$ with respect to $\leq$. If $T, R \subseteq \prod_{a\in\Omega} 2^{V_a}$ we let $T \vee R = \{t \vee r : t \in T,\ r \in R\}$.

Now, suppose that d is a single-valued decision attribute and M is a class of θ_d. A hypertuple $t \in \prod_{a\in\Omega} 2^{V_a}$ is called *equilabelled with respect to* M if for all granules $\Omega(x)$

$$\text{If } \Omega(x) \leq t, \text{ then } x \in M. \tag{5.5.7}$$

Observe that each granule is equilabelled with respect to some class of θ_d. Each equilabelled hypergranule can replace a number of granules without destroying the class information of θ_d. The aim is now to find the maximal equilabelled hypergranules; this will give us an optimal data compression. Clearly, we can do this by considering the classes of θ_d one at a time; thus, let M be such a class, and $\mathcal{E}$ be the set of all hypertuples which are equilabelled with respect to M. Furthermore, let H be the set of its maximal elements. A simple algorithm to find H was given in [124]:

1. $C_1 \stackrel{\text{def}}{=} M$.
2. $C_{k+1} \stackrel{\text{def}}{=}$ The set of maximal elements of $[\downarrow (C_k + M)] \cap \mathcal{E}$.

It can be shown by elementary lattice theoretic arguments that the algorithm becomes constant from some C_n on, and that $C_n = H$. The result of this algorithm with respect to the Iris data is given in Table 5.6.

Table 5.6: Hypergranules for the Iris data

Attribute							Class
Sepal length		Sepal width		Petal length		Petal width	
$\{43, \ldots, 58\}$	×	$\{23, 29, \ldots, 44\}$	×	$\{10, \ldots, 17, 19\}$	×	$\{1, \ldots, 6\}$	Setosa
$\{49, \ldots 52, 54, \ldots, 70\}$	×	$\{20, 22, \ldots, 34\}$	×	$\{30, 33, 35, \ldots, 49\}$	×	$\{10, \ldots, 16\}$	Versicolor
$\{56, \ldots, 59, 61, \ldots, 69, 71, \ldots, 74, 76, 77, 79\}$	×	$\{25, \ldots, 34, 36, 38\}$	×	$\{49, \ldots, 61, 63, 64, 66, 67, 69\}$	×	$\{18, \ldots, 25\}$	Virginica
$\{61, 63\}$	×	$\{26, 28\}$	×	$\{51, 56\}$	×	$\{14, 15\}$	Virginica
$\{60\}$	×	$\{30\}$	×	$\{48\}$	×	$\{18\}$	Virginica
$\{72\}$	×	$\{30\}$	×	$\{58\}$	×	$\{16\}$	Virginica
$\{62\}$	×	$\{28\}$	×	$\{48\}$	×	$\{18\}$	Virginica
$\{60\}$	×	$\{22\}$	×	$\{50\}$	×	$\{15\}$	Virginica
$\{49\}$	×	$\{25\}$	×	$\{45\}$	×	$\{17\}$	Virginica
$\{60\}$	×	$\{27\}$	×	$\{51\}$	×	$\{16\}$	Versicolor
$\{67\}$	×	$\{30\}$	×	$\{50\}$	×	$\{17\}$	Versicolor
$\{59\}$	×	$\{32\}$	×	$\{48\}$	×	$\{18\}$	Versicolor

Observe that the table contains only 12 entries instead of the original 150. The hypergranules can now be used to classify unseen elements, and we refer the reader to [124] for further details.

5.2 Discretisation of real valued attributes

A different approach to discretisation when the attributes are real-valued can be taken by searching for hyperplanes which determine optimal intervals for discretisation [76, 80, 81]. To explain the principle, we shall only present the case of cuts, i.e. hyperplanes parallel to an axis, as presented in [81], and leave the reader to consult the cited references for the more general case.

We suppose that each attribute domain a is a left-closed, right-open interval of the real line, i.e. $V_a = [l_a, r_a)$ for some $l_a, r_a \in \mathbb{R},\ l_a \lneq r_a$. The idea is to partition each V_a into a set of subintervals, which can replace the original attribute values, and thus lead to a reduction of taken values.

A *sequence of cuts for* $a \in \Omega$ is a sequence $C_a = \{c_i^a : i \leq t(a)\}$ of numbers such that

$$l_a = c_0^a \lneq c_1^a \cdots \lneq c_{t(a)}^a = r_a. \tag{5.5.8}$$

More generally, we will call pair $\langle a, c\rangle$ a *cut*, if $c \in C_a$, and with some abuse of language, we identify C_a with $\{\langle a, c\rangle : c \in C_a\}$. Similarly, we identify a family $\mathcal{C} = \{C_a : a \in \Omega\}$ of sequences of cuts with the set $\{\langle a, c\rangle : a \in \Omega,\ c \in C_a\}$, and call it briefly a *family of cuts*. Observe that by our definition, $\langle a, l_a\rangle, \langle a, r_a\rangle \in \mathcal{C}$ for all $a \in \Omega$.

Each sequence C_a of cuts for a determines a partition $\mathcal{P}_{C_a}$ of V_a by

$$V_a = [c_0^a, c_1^a) \cup [c_1^a, c_2^a) \cup \cdots \cup [c_{n-1}^a, c_{t(a)}^a). \tag{5.5.9}$$

Conversely, each partition $\mathcal{P}_a$ of V_a into left-closed, right-open intervals determines a sequence of cuts for a in a natural way.

If $\mathcal{C} = \{\langle a, c\rangle : a \in \Omega, c \in C_a\}$ and $\mathcal{C}' = \{\langle a, c'\rangle : a \in \Omega, c \in C'_a\}$ are families of cuts, we write $\mathcal{C} \leq \mathcal{C}'$, if $\mathcal{C} \subseteq \mathcal{C}'$. If $\mathcal{C} \lneq \mathcal{C}'$, then, at least one C_a contains less cuts that C'_a, and therefore, the partition determined by C_a will be coarser than that of C'_a.

Given a family $\mathcal{C}$ of cuts, we obtain a new information system $\mathcal{I}^{\mathcal{C}} = \langle U, \Omega^{\mathcal{C}}, \langle V_{a^c}\rangle_{a^c \in \Omega^c}\rangle$, where

1. $\Omega^{\mathcal{C}} = \{a^{\mathcal{C}} : a \in \Omega\}$.
2. For each $a^{\mathcal{C}} \in \Omega^{\mathcal{C}}$,
   (a) $V_a^{\mathcal{C}} = \{i \in \mathbb{N} : 1 \leq i \leq t(a)\}$.
   (b) $a^{\mathcal{C}}(x) = i$ if and only if $a(x) \in [c_{i-1}^a, c_i^a)$.

Suppose that d is a decision attribute. We say that $\mathcal{C}$ *is consistent with* d, if for all $x_i, x_j \in U$,

$$\Omega_{\mathcal{C}}(x_i) = \Omega_{\mathcal{C}}(x_j) \Rightarrow d(x_i) = d(x_j). \tag{5.5.10}$$

$\mathcal{C}$ is called *irreducible* (with respect to d), if $\mathcal{C}$ is consistent with d and no $\mathcal{C}' \lneq \mathcal{C}$ has this property. The aim is now, to find irreducible sets of cuts. In other words, we are looking for a consistent family of cuts, leading to partitions of the sets V_a into left-closed, right open intervals, such that joining any two intervals would make the system inconsistent. Observe that the sets C_a are usually not independent, and that it is possible for a consistent $\mathcal{C}$ to have, $a^{\mathcal{C}}(x_i) = a^{\mathcal{C}}(x_j)$ and $d(x_i) \neq d(x_j)$ for some $a \in \Omega$. This is analogous to the fact that, in determining dependencies, we have for all $P, Q \subseteq \Omega$

$$P \subseteq Q \Rightarrow \theta_Q \subseteq \theta_P.$$

Recalling the connection between reducts and Boolean reasoning (Proposition 3.3), the following comes as no surprise [81]:

Proposition 5.2. *Finding an irreducible family of cuts for a decision system is polynomially equivalent to finding a prime implicant of a monotone Boolean function in conjunctive normal form.*

Thus, finding an irreducible family of cuts is NP-complete [115]; effective heuristics are described in [75, 79]. An overview and detailed examples of these procedures can be found in [78].

The hyperplane approach uses numeric information of the attribute domain, which the discretisation of [32, 124] does not take into account. It is possible to include further restrictions on hypertuples in [124] to mimick a hyperplane algorithm as well. Therefore, the method of [124] is the more general approach, and it can be tailored to other restrictions on the data set as well.

Chapter 6

Model selection

In the static case, reducts describe the data error free. However, there are usually many reducts and the question arises which of these is, in a sense, the best one. The criterion for this can be, for example, the cost of determining attribute values, so that the best reduct is one which minimises these costs. In a dynamic situation, a best attribute set might be one which has a high classification quality with respect to unseen cases. It has been known for some time that reducts, which stem from static databases are not always the best choice when it comes to prediction. In this chapter, we describe two approaches to the choice of attribute sets for prediction: The dynamic reducts of [7], and the entropy based measure of [33] which does not use reducts at all.

6.1 Dynamic reducts

Dynamic reducts aim to improve the prediction quality of rules generated by a reduct by measuring this quality over a number of randomly generated sub-universes of the domain of interest:

> "The underlying idea of dynamic reducts stems from the observation that reducts generated from information systems are unstable in the sense that they are sensitive to changes in the information system introduced by removing a randomly chosen set of objects. The notion of dynamic reduct encompasses the stable reducts, i.e. reducts that are the most frequent reducts in random samples created by subtables of the given decision table" [114].

For $\mathcal{I} = \langle U, \Omega, V_q, f_q, d\rangle_{q\in\Omega}$ and $U' \subseteq U'$, we call $\mathcal{I}' = \langle U', \Omega, V_q, f_q, d\rangle_{q\in\Omega}$ a *subsystem of* $\mathcal{I}$. If $\mathbf{F}$ is a family of subsystems of $\mathcal{I}$, then

$$DRed(I, \mathbf{F}) = Red(\mathcal{I}) \cap \bigcap\{\mathcal{J} : \mathcal{J} \in \mathbf{F}\} \tag{6.6.1}$$

is called the *family of* $\mathbf{F}$*-dynamic reducts of* $\mathcal{I}$. It is easily seen that in most cases, this is too restrictive for practical use, and thus, a threshold is introduced: Let $0 \leq \epsilon \leq 1$. The *family of* $(\epsilon, \mathbf{F})$*-dynamic reducts of* $\mathcal{I}$ is defined by

$$D_\epsilon Red(I, \mathbf{F}) = \{Q \in Red(\mathcal{I}) : 1 - \epsilon \leq s_{\mathbf{F}}(Q)\} \tag{6.6.2}$$

where

$$s_{\mathbf{F}}(Q) = \frac{|\{\mathcal{J} \in \mathbf{F} : Q \in Red(\mathcal{J})\}|}{|\mathbf{F}|} \tag{6.6.3}$$

is the $\mathbf{F}$- *stability coefficient of* Q. Model selection proceeds in four steps:

1. Choose numbers n, k_j, $j \leq n$, $1 \leq k_j \leq \frac{|U|}{2}$, and a threshold ϵ.
2. For each $1 \leq j \leq n$ generate a subsystem $\mathcal{I}_j$ of $\mathcal{I}$ by randomly deleting a k_j objects of U, and set $\mathbf{F} = \{\mathcal{I}_j : 1 \leq j \leq n\}$.
3. Find the reducts for $\mathcal{I}$and each $\mathcal{I}_j$.
4. Choose all reducts Q with $1 - \epsilon \leq s_{\mathbf{F}}(Q)$ for further processing.

From these "true dynamic reducts", decision rules for classification are computed, and the final decision is taken by "majority voting".

The method of dynamic reducts employs a kind of internal cross-validation in order to improve the external prediction quality. We observe that the researcher has to make some subjective choices in step 1. of the procedure, which are not contained in the data. The huge complexity of step 3. forces applications of heuristic techniques, such as combinatorial or genetic algorithms. Extensive experiments reported in [5] show that the dynamic reduct approach fares considerably better than the traditional RSDA method and compares well with customary procedures. For extensions of the method and similar approaches we refer the reader to [5, 82, 114].

6.2 Rough entropy measures

A totally different route was proposed in [33], which is not based on reducts, but on information theoretic entropy measures. The rationale behind this is the observation that reduct based prediction rules are relative to the chosen reduct, and only measure the uncertainty of the prediction, while not taking into account the predictor variables. Thus, a comparison between the prediction qualities based on different reducts does not view the full picture. In order to obtain an unconditional measure of prediction success of a set Q of predictor attributes, one has to combine

1. The complexity $H(Q)$ of coding the hypothesis Q.
2. The conditional coding complexity $H(d|Q)$ of d, given by the values of attributes in Q

into one measure $H(Q \rightsquigarrow d)$.

This approach is in the spirit of the minimum description length principle (MDLP) of [101, 102], and will be done by suitable entropy functions.

Information theoretic entropy is a purely syntactic measurement of the coding effort of a sequence of bits, knowing their probability. It answers the question

- How many binary questions must we pose (optimally and in the long run), if we do clever guessing by using the knowledge about the probability distribution p?

The famous coding theorem [110] shows that the entropy function is the lower bound of the coding complexity of events given their probabilities. Entropy-related functions have been widely used as objective measures in many fields, and we point the interested reader to [62].

Let $\mathcal{P}$ be a partition of U with classes $X_i, i \leq k$, each having cardinality r_i. In compliance with the principle of indifference we assume that the elements of U are randomly distributed within the classes of $\mathcal{P}$, so that the probability of an element x being in class X_i is just $\frac{r_i}{n}$. We define the *entropy* of $\mathcal{P}$ by

$$H(\mathcal{P}) \stackrel{\text{def}}{=} \sum_{i=0}^{k} \frac{r_i}{n} \cdot \log_2(\frac{n}{r_i}). \tag{6.6.4}$$

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

The entropy estimates the mean number of comparisons minimally necessary to retrieve the equivalence class information of a randomly chosen element $x \in U$. We can also think of the entropy of $\mathcal{P}$ as a measure of granularity of the partition: If there is only one class, then $H(\mathcal{P}) = 0$, and if $\mathcal{P}$ corresponds to the identity ϖ, then $H(\mathcal{P})$ reaches the maximum $\log_2(|U|)$. In other words, with the universal relation there is no information gain, since there is only one class and we always guess the correct class of an element; if the partition contains only singletons, the inclusion of an element in a specific class is hardest to predict, and thus the information gain is maximized.

The assumption which we make in choosing a predictor set Q is, in a way, a worst case scenario: We suppose that the deterministic classes of θ_Q give us reliable information, while everything outside these classes is the result of a random process which is totally unknown

to the researcher. Thus, in setting up the partition which determines the desired measure $H(Q \rightsquigarrow d)$, we keep the deterministic classes of θ_Q and put every element of U which is not in one of these classes into its own class; in this way, the entropy is maximised. This *maximum entropy principle* is well suited to our non-invasive philosophy:

> "Although there may be many measures μ that are consistent with what we know, the *principle of maximum entropy* suggests that we adopt that μ^* which has the largest entropy among all the possibilities. Using the appropriate definitions, it can be shown that there is a sense in which this μ^* incorporates the 'least' additional information" [56].

Indeed, in choosing $H(Q \rightsquigarrow d)$ the way we do, we assume representativeness only of the deterministic classes of θ_Q, and admit total ignorance otherwise. This is in contrast to the information gain used in machine learning which assumes that all classes are representative, and which furthermore suffers from the symmetry of the measure.

More formally, we suppose that $X_i, 1 \leq i \leq t$, are the classes of θ_Q each having cardinality r_i. A class is X_i is deterministic with respect to d if and only if $1 \leq i \leq c$, and we denote by V the union of the deterministic classes. Furthermore, let $|U| = n$ and $\hat{\pi}$ be the probability measure associated with θ_Q, i.e.

$$\hat{\pi}_i = \frac{|X_i|}{|U|} = \frac{r_i}{n}.$$

Then, we obtain the entropy of θ_Q as

$$H(Q) = \sum_{i=1}^{t} \hat{\pi}_i \cdot \log_2 \frac{1}{\hat{\pi}_i} = \sum_{i=1}^{t} \frac{r_i}{n} \cdot \log_2 \frac{n}{r_i}. \tag{6.6.5}$$

In order to keep only the deterministic classes of θ_Q, we define a new equivalence relation θ_Q^+ on U by

$x \equiv_{\theta_Q^+} y$ if and only if $x = y$ or there exists some $1 \leq i \leq c$ such that $x, y \in X_i$.

Its associated probability distribution is given by $\{\hat{\psi}_i : 1 \leq i \leq (c + |U \setminus V|)\}$ with

$$\hat{\psi}_i \stackrel{\text{def}}{=} \begin{cases} \hat{\pi}_i, & \text{if } 1 \leq i \leq c, \\ \frac{1}{n}, & \text{otherwise.} \end{cases} \tag{6.6.6}$$

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

$$H(Q \rightsquigarrow d) = \sum_i \hat{\psi}_i \cdot \log_2 (\frac{1}{\hat{\psi}_i}).$$

The maximum for $H(Q \rightsquigarrow d)$ occurs when

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

In both cases we have $H(Q \rightsquigarrow d) = \log_2(n)$.

Next, we define the *normalized relative deterministic prediction success* $S(Q \rightsquigarrow d)$, which we will call *normalised rough entropy* (NRE): First, let θ_d be the identity ϖ, so that $H(d) = \log_2(n)$. Then,

$$S(Q \rightsquigarrow d) \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } \theta_Q = \varpi, \\ 0, & \text{otherwise.} \end{cases} \tag{6.6.7}$$

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

$$S(Q \rightsquigarrow d) \stackrel{\text{def}}{=} 1 - \frac{H(Q \rightsquigarrow d) - H(d)}{\log_2(n) - H(d)}, \tag{6.6.8}$$

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

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

The aim of model selection is now to minimise rough entropy, or, equivalently, to maximise NRE; note that this is independent of the reduct criterion. In order to gauge the prediction quality of the entropy based model selection, we have compared its performance on 14 published data sets[1] with the well known machine learning algorithm C4.5 as reported in [99]. The validation by the training set – testing set method was performed by splitting the full data set randomly into two equal sizes 100 times, assuming a balanced distribution of training and testing data (TT2 method). The mean error value is our measure of prediction success. We choose only half of the set for training purposes in order to have a basis for testing the predictive power of the resulting attribute sets. Because all data sets contained continuous attributes and most of them missing values as well, a preprocessing step was necessary to apply the algorithm to these data sets. Missing values were replaced by the mean value in case of ordinal attributes, and by the most frequent value otherwise. The preprocessing of the continuous

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

data was done by three different global discretisation methods: The first method performs the global filtering method described in Section 5.1 which influences the NRE, but does not affect γ, and thus has no influence on the dependency structure. This results in minimal granularity of attributes with respect to the decision attribute. The other two discretization methods cluster the values of an attribute into ten, resp. five, classes with approximately the same number of objects. The discretization method can be refined by transforming the methods of local discretisation of continuous attributes given in [11] and [22] to our entropy measure.

Table 6.1: $H(Q \rightsquigarrow d)$ and C4.5

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

In Table 6.1 we list the basic parameters of the data sets, and compare the results of our method with the C4.5 performance given in [99]. This has to be taken with some care, since Quinlan uses 10-fold cross validation (CV10) on data sets optimized by

> " ... dividing the data into ten blocks of cases that have similar size and class distribution" [99, p.81, footnote 3.].

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

The column "No. of pred. attr." records the number of attributes which are actually used for prediction; this is in most cases considerably less than the number of all attributes. The results

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

6.3 Entropy measures and approximation quality

In this Section we will exhibit some properties of the conditional entropy $H(d|Q)$, and show where the approximation quality γ, the traditional performance measure of RSDA, is positioned in our entropy based method. We will use the parameters of the preceding Section; also, let $\gamma = \gamma(Q \rightsquigarrow d)$, and note that $\gamma = \frac{|V|}{n}$, and therefore, $1 - \gamma = \frac{|U \setminus V|}{n}$.

We first find the conditional entropy $H(d|Q)$ by

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

The following proposition gives the bounds within which $H(d|Q)$ varies:

Proposition 6.1. $(1-\gamma) \leq H(d|Q) \leq (1-\gamma)\log_2(n - |V|)$.

Proof. First, observe that $\log_2(n - |V|) \geq \log_2(2) = 1$. The minimum value of $\sum_{i>c} \hat{\pi}_i \cdot \log_2(\hat{\pi}_i)$ is obtained when $c = t - 1$, and in this case,

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

Therefore,

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

For the other direction, we first note that each nondeterministic class X has at least two elements, and that $\sum_{i>c}\hat{\pi}_i\cdot\log_2(\frac{1}{\hat{\pi}_i})$ has a maximum if either each such class has exactly two elements, or all but one class have two elements and one class has three elements. Since the value of $\sum_{i>c}\hat{\pi}_i\cdot\log_2(\frac{1}{\hat{\pi}_i})$ is greater in the first case, we assume w.l.o.g. that $n-|V|$ is even, so that

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

Therefore,

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

which proves our claim. □

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

Proposition 6.2. *If $Q \subseteq R$, then $H(d|R) \leq H(d|Q)$.*

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

$$H(R \rightsquigarrow d) \leq H(Q \rightsquigarrow d).$$

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

A similar result does not hold for $H(Q \rightsquigarrow d)$ as the example given in Table 6.2 shows: There,

$$H(\{q_1\} \rightsquigarrow \{p\}) = 1.5 < 2 = H(\{q_1, q_2\} \rightsquigarrow \{p\}) = H(\{q_2\} \rightsquigarrow \{p\}).$$

Table 6.2: $H(Q \rightsquigarrow d)$

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

The question arises, where the approximation function γ is positioned in this model. Proposition 6.1 shows that, for fixed Q,

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

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

Proposition 6.3. $\gamma(Q \rightsquigarrow d) < \gamma(R \rightsquigarrow d) \iff H_{\mathsf{max}}(d|R) < H_{\mathsf{max}}(d|Q)$.

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

$$\begin{aligned} H_{\mathsf{max}}(d|R) &= (1 - \gamma(R \rightsquigarrow d)) \cdot \log_2(n - |V_{R \rightsquigarrow d}|), \\ &< (1 - \gamma(Q \rightsquigarrow d)) \cdot \log_2(n - |V_{Q \rightsquigarrow d}|), \\ &= H_{\mathsf{max}}(d|Q), \end{aligned}$$

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

$$k \cdot \log_2 k < (k + 1) \cdot \log_2(k + 1). \tag{6.6.9}$$

We can also assume that $0 < H_{\mathsf{max}}(d|R)$, so that $U \setminus V_{R \rightsquigarrow d} \neq \emptyset$. Now,

$$\begin{aligned} &H_{\mathsf{max}}(d|R) < H_{\mathsf{max}}(d|Q) \\ &\Rightarrow (1 - \gamma(R \rightsquigarrow d)) \cdot \log_2(n - |V_{R \rightsquigarrow d}|) < (1 - \gamma(Q \rightsquigarrow d)) \cdot \log_2(n - |V_{Q \rightsquigarrow d}|) \\ &\Rightarrow (n - |V_{R \rightsquigarrow d}|) \cdot \log_2(n - |V_{R \rightsquigarrow d}|) < (n - |V_{Q \rightsquigarrow d}|) \cdot \log_2(n - |V_{Q \rightsquigarrow d}|) \\ &\Rightarrow (n - |V_{R \rightsquigarrow d}|) < (n - |V_{Q \rightsquigarrow d}|) \text{ by (6.6.9)} \\ &\Rightarrow |V_{Q \rightsquigarrow d}| < |V_{R \rightsquigarrow d}| \\ &\Rightarrow \gamma(Q \rightsquigarrow d) < \gamma(R \rightsquigarrow d). \end{aligned}$$

This completes the proof. □

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

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

Proposition 6.2 does not extend to the hypothesis $\gamma(Q \rightsquigarrow d) \lneq \gamma(R \rightsquigarrow d)$, and thus, a result similar to 6.3 does not hold, as the following example shows: Consider the equivalence relations $\theta_d, \theta_Q, \theta_R$ with the following partitions:

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

Then,

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

On the other hand,

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

The preceding results show that RSDA which tries to maximize γ is a procedure to minimize the maximum of the conditional entropy of rough prediction. This shows clearly that, unlike $H(Q \rightsquigarrow d)$, the traditional γ is a conditional measure, and thus,its application in comparing the quality of approximation for different attribute sets has to be taken with care.

Chapter 7

Probabilistic granule analysis

This chapter is concerned with a probabilistic counterpart to the rough set model, which shares the idea of predicting a dependent variable by granules of knowledge, and which can be called a non-invase technique as well, since it is a distribution free type of analysis [45].

As we have shown, RSDA concentrates on finding deterministic rules for the description of dependencies among attributes. Once a rule is found, it is assumed to hold without any error by the nominal scale assumption (3.3.10). If a measurement error is assumed to be an immeasurable part of the data – as e.g. statistical procedures do – the pure RSDA approach will not produce acceptable results, because "real" measurement errors cannot be explained by any rule.

7.1 The variable precision model

There are several possibilities to reduce the precision of prediction to cope with measurement error. One possibility is the *variable precision rough set model* [130], which assumes that rules are valid only within a certain part of the population. The main tool of the model is a precision parameter β which expresses a bound for misclassification. Let $X, Y \subseteq U$; the *relative degree of misclassification of X with respect to Y* is defined by

$$c(X,Y) = \begin{cases} 1 - \frac{|X \cap Y|}{|X|}, & \text{if } X \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases} \tag{7.7.1}$$

We observe that $c(X,Y) = 0$ if and only if $X \subseteq Y$. If $0 \leq \beta \lneq 0.5$ we set

$$X \overset{\beta}{\subseteq} Y \iff c(X,Y) \leq \beta. \tag{7.7.2}$$

The number β measures the acceptable relative error and it is called the *precision parameter*. . If $Q \subseteq \Omega$ and d is a decision attribute, we say that a class X of θ_Q is β-*deterministic*, if there is some class Y of θ_d such that $X \overset{\beta}{\subseteq} Y$. We will denote the union of all β-deterministic classes by $D(Q, d, \beta)$. The attribute d is called β-*dependent on* Q, if $D(Q, d, \beta) = U$.

The β-*approximation quality* of a rule $Q \rightsquigarrow d$ is now defined as

$$\gamma(Q \overset{\beta}{\rightsquigarrow} d) = \frac{|D(Q, d, \beta)|}{|U|}. \tag{7.7.3}$$

The advantage of this approach is that it uses only two parameters (the precision parameter β and the approximation quality γ) to describe the quality of a rule system; the disadvantages are that precision and γ are partially exchangeable, and that there is no theoretical background to judge which combination is best suited to the data. Furthermore, the β-dependency does not have the nice structural properties as the normal rough set dependencies; they are, for example, not transitive.

7.2 Replicated decision systems

In order to formulate a probabilistic version of RSDA, which is able to handle measurement errors as well, we enhance some of the concepts defined before.

There are two different aspects of errors, which have to be modelled: First, the concept of "error" is quite different in the lower and upper approximation. Whereas the lower bound is based on a logical conjunction, which can be expected to be relative robust against random influences, the upper bound is a disjunction and therefore much more sensitive if random processes cause errors. The second aspect is that rules may not be stable, and therefore a replication of a decision system might look quite different from the original one, even if the underlying structure is unchanged. As we will see below, both problems can be tackled, if we use a well known trick to estimate reliability and / or stability: The assumption that an identical decision system can be observed (at least) twice.

A *replicated decision system* $\mathcal{D}$ is a structure $\langle U, \Omega, Y, (V_a)_{a\in\Omega}\rangle$, where

- $\langle U, \Omega, (V_a)_{a\in\Omega}\rangle$ is an information system,
- $Y = \{y_1, \ldots, y_S\}$ is a set of replicated decision attributes, explained more fully below.

The introduction of replicated decision variables offers the opportunity to control the effect of a measurement error: The smaller the agreement among multiple replications of the decision attribute, the more measurement error has to be assumed. This concept of replicated measurements is a way to estimate the reliability of, for example, psychometric tests, using

the retest-reliability estimation, which in turn uses a linear model to estimate reliability and error of measurement as well.

We denote the set of granules by $G = \{\vec{a_1}, \dots, \vec{a_M}\}$; in other words,

$$G = \{\Omega(x) : x \in U\}, \tag{7.7.4}$$

see (3.3.12). Each replica of the decision attribute takes the values $V_Y = \{r_1, r_2, \dots, r_Y\}$. The classes of θ_{y_t} are denoted by $M_{t,1}, \dots, M_{t,r_Y}$; for each granule $\vec{a}_i$, we let $\xi(i,t,j)$ be the number of objects described by $\vec{a}_i$ which are in class $M_{t,j}$. In other words,

$$\xi(i,t,j) = |\{x \in U : \Omega(x) = \vec{a}_i \text{ and } x \in M_{t,j}\}| \tag{7.7.5}$$

We also let

$$\nu(\vec{a}_i) = |\{x \in U : \Omega(x) = \vec{a}_i\}|. \tag{7.7.6}$$

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

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

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

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

An example of the parameters of a decision system with one replica of the decision attribute is shown in Table 7.1. The example given in Table 7.1 shows that indeterministic rules alone

Table 7.1: A replicated decision system

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

do not use the full information given in the database. There is no deterministic rule to predict a value of the decision attribute y_1, given a value of the independent attributes $\langle a_1, a_2 \rangle$: Both indeterministic rules will predict both possible outcomes in the decision variable. The pure

rough set approach now concludes that no discernable assignment is possible. But if we inspect the table, we see that the error of assigning $\langle 0,1\rangle$ to 1 is small (1 observation) and that $\langle 1,0\rangle \mapsto 2$ is true up to 2 observations.

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

$$Y = \sum_{1\leq r\leq R} \omega_r Y_r, \tag{7.7.7}$$

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

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

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

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

Probabilistic prediction rules are of the form

$$\vec{a}_j \rightsquigarrow Y_{g(j)},\ 1 \leq j \leq M,$$

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

In the example of Table 7.1 there are two possibilities for R, and we use maximum likelihood to optimise the binomial distribution, the application of which is straightforward, if we additionally assume that the observations stem from a simple sampling scheme. In case $R = 1$, both granules use the same distribution Y_1. In this case, the likelihood function $L_1 = L(Y_1|\langle 0,1\rangle, \langle 1,0\rangle)$ is given by

$$L_1 = \binom{16}{9} \pi^9 (1-\pi)^7 \tag{7.7.8}$$

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

$$\langle 0,1\rangle \text{ or } \langle 1,0\rangle \rightsquigarrow \{\langle 1, \tfrac{9}{16}\rangle, \langle 2, \tfrac{7}{16}\rangle\}. \tag{7.7.9}$$

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

$$L_2 = \binom{6}{1}\pi_1^1(1-\pi_1)^5\binom{10}{8}\pi_2^8(1-\pi_2)^2. \tag{7.7.10}$$

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

$$\begin{cases} \langle 0, 1\rangle \rightsquigarrow \{\langle 1, \frac{5}{6}\rangle, \quad \langle 2, \frac{1}{6}\rangle\}, \\ \langle 1, 0\rangle \rightsquigarrow \{\langle 1, \frac{2}{10}\rangle, \quad \langle 2, \frac{8}{10}\rangle\}. \end{cases} \tag{7.7.11}$$

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

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

$$L_1(\max) = \hat{\pi}^9(1-\hat{\pi})^7 \qquad = 0.0000173, \tag{7.7.12}$$

$$L_2(\max) = \hat{\pi}_1^1(1-\hat{\pi}_1)^6\hat{\pi}_2^8(1-\hat{\pi}_2)^2 \qquad = 0.0003746, \tag{7.7.13}$$

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

There are – at least – two standard procedures for model selection, which are based on the likelihood and the number of parameters: The Akaike Information Criterion (AIC) [3] and the Schwarz Information Criterion (SIC) [107].

If $L(\max)$ is the maximum likelihood of the data, P the number of parameters, and K the number of observations, these are defined by

$$AIC = 2(P - \ln(L(\max))) \tag{7.7.14}$$

$$SIC = 2\left(\frac{\ln(K)}{2}\cdot P - \ln(L(\max))\right). \tag{7.7.15}$$

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

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

$$\begin{aligned} AIC(L_1(\max)) &= 2(1 - \ln(0.0000173)) &&= 23.930, \\ SIC(L_1(\max)) &= 2(\frac{\ln(16)}{2} - \ln(0.0000173)) &&= 24.702. \end{aligned}$$

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

$$\begin{aligned} AIC(L_2(\max)) &= 2(3 - \ln(0.0003746)) &&= 21.779 \\ SIC(L_2(\max)) &= 2(3\ln(16) - \ln(0.0003746)) &&= 24.090. \end{aligned}$$

and we can conclude that the rule – system (7.7.11) is better suited to the data then the simple 1-rule – system (7.7.9).

7.3 An algorithm to find probabilistic rules

The algorithm of finding probabilistic rules starts by searching for the optimal granule mapping based on a set Ω of (mutually) predicting attributes and a set Y of replicated decision attributes. Finding the best mapping is a combinatorial optimisation problem, which can be approximated by hill-climbing methods, whereas the computation of the maximum likelihood estimators, given a fixed mapping g, is straightforward: One computes the multinomial parameters $\hat{\pi}_t(i_k)$ of the samples i defined by g for every replication y_t of Y and every value $r_k \in \{r_1, \ldots, r_Y\}$, and computes the mean value

$$\hat{\pi}(i_k) = \frac{\sum_{t=1}^{s} \hat{\pi}_t(i_k)}{s}, \tag{7.7.16}$$

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

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

the computation of the AIC is now possible.

An algorithm to find the probabilistic rules based on AIC optimization is given in Table 7.2 on the facing page. The adapatation of similar optimization criteria like SIC or a function f(AIC,SIC) is straightforward. The result of algorithm offers the most parsimonious description of a probabilistic rule system (in terms of AIC). In order to reduce the number of independent attributes within the rules, a classical RSDA reduct analysis of these attributes can be applied, using the results of the mapping g as a decision attribute.

Table 7.2: Rule finding algorithm

$R := 0,\ \Delta(AIC) = 1.$
while $R \lneq M$ and $\Delta(AIC) \gneq 0$ **do**
  $R := R + 1$
  for all y_t **do**
    Find $h^t_{\max}(R)$ and its associated multinomial parameters
    $\hat{\pi}^t_{i,k}, 1 \leq i \leq M,\ 1 \leq k \leq Y.$
  end for
  for all $1 \leq i \leq M,\ 1 \leq k \leq Y$ **do**
    $\hat{\pi}_{i,k} := \frac{\sum_{t=1}^{S} \hat{\pi}^t_{i,k}}{S}$
  end for
  Compute $L_R(\max)$ from the $\hat{\pi}_{i,k}$.
  Compute AIC_R for $L_R(\max)$.
  if $R = 1$ **then**
    $\Delta(AIC) := AIC_1$
  else
    $\Delta(AIC) := AIC_{R-1} - AIC_R$
  end if
end while

7.4 Unsupervised learning and nonparametric distribution estimates

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

The Figures 7.1 and 7.2 show the result of the mixture analysis based on granules using the mixture

$$Y = \frac{1}{2}N(-2.0, 1.0) + \frac{1}{2}N(0.0, 1.0). \tag{7.7.17}$$

$N(\mu, \sigma)$ is the normal distribution with parameters μ and σ. 1024 observations per replication were simulated; one simulation was done with 2 replications (Figure 7.1), and another

with 5 replications (Figure 7.2). The simulated data were grouped into 32 intervals with approximately the same frequencies in the replications, and the searching algorithm outlined above was applied.

Figure 7.1: Nonparametric estimates of a $\frac{(N(-2,1)+N(0,1))}{2}$ mixture distribution (2 replications; lines denotes theoretical distributions)

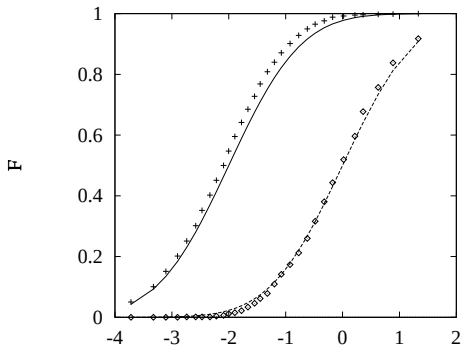
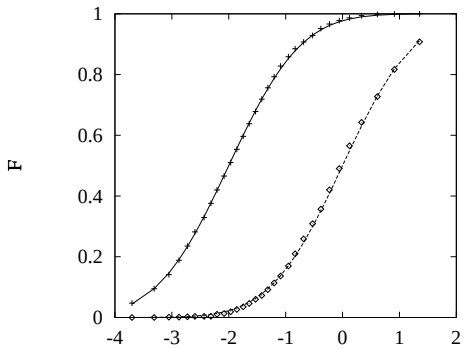


Figure 7.2: Nonparametric estimates of a (N(-2,1)+N(0,1))/2 mixture distribution (5 replications; lines denotes theoretical distributions)



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

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

The next numerical experiment was performed with the Iris data [42]. It is well known (e.g.

[10]) that Sepal width attribute is not very informative; therefore we shall skip it for the subsequent analysis.

If we assume that the three remaining attributes measure the same variable up to some scaling constants, we can use the z–transformed attributes as a basis for the analysis. The unsupervised AIC search algorithm clearly votes for three classes in the unknown joint dependent attribute. If we use the estimated distribution functions (Figures 7.3, 7.4, 7.5) for the classification of the elements, we find a classification quality of about 85% (Table 7.3), which is not too bad for an unsupervised learning procedure.

Table 7.3: Iris: Classification results

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

The procedure does not offer only classification results, but also estimators of the distributions of dependent attributes within the groups without having a prior knowledge about the group structure. The Figures 7.3, 7.4, 7.5 compare three estimated distributions with the respective the distributions of three (normalised) variables within the groups. The results show that the "Sepal length" attribute does not fit very well and that the estimated distributions summarise this aspect of both "Petal" measures.

Figure 7.3: Setosa distributions of 3 attributes and its estimation

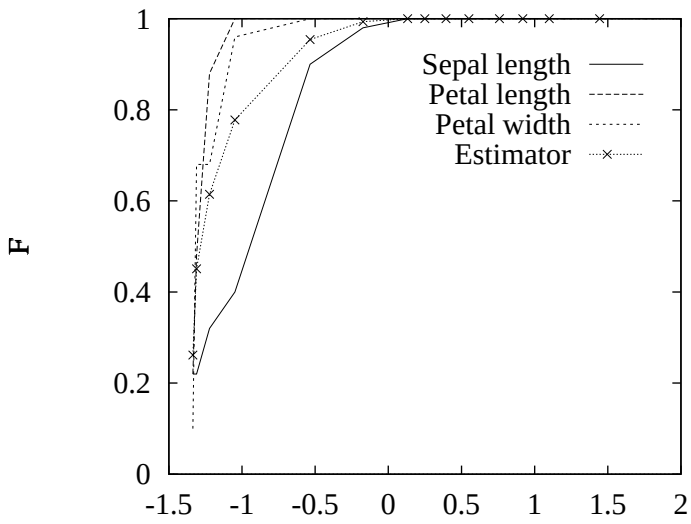


Figure 7.4: Versicolor distributions of 3 attributes and its estimation

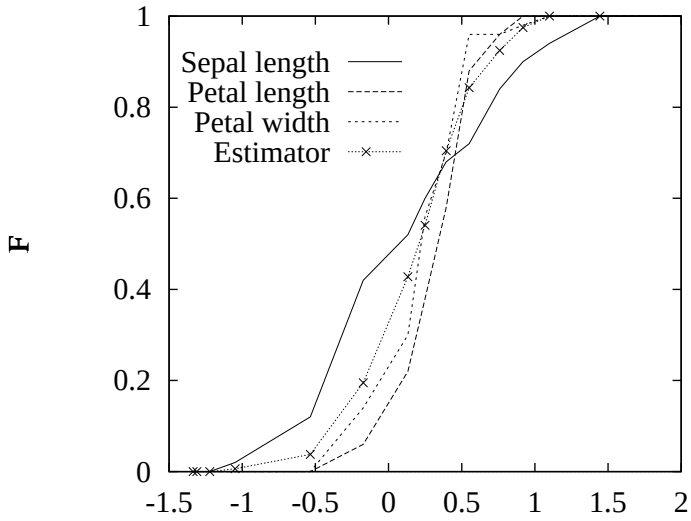
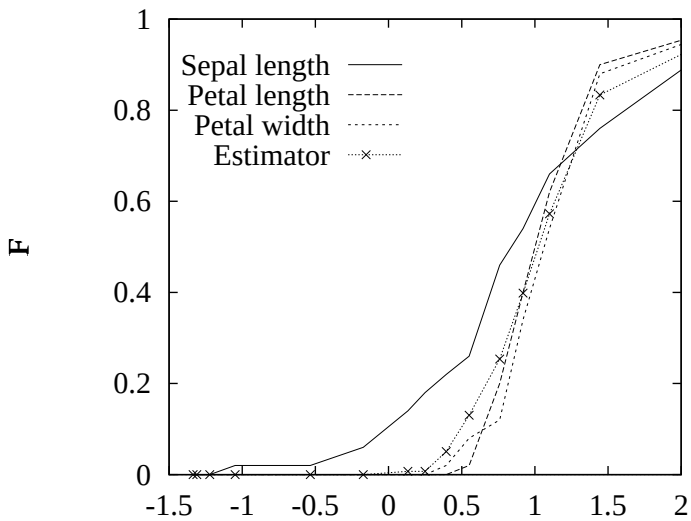


Figure 7.5: Virginica distributions of 3 attributes and its estimation



Chapter 8

Imputation

8.1 Statistical procedures

In real life, observed data is seldom complete. There are several ways to "impute" missing values, most of which are based on statistical procedures. The basic idea of these procedures is the estimation of the missing values by minimising a loss function such as the least square measure or the negative likelihood. For an overview of the current practice of working with missing data we invite the reader to consult [1]; for a more complete treatment we recommend the excellent book [106].

Consider the information system in Table 8.1. We write ? in cell $\langle x, a \rangle$, if a is not defined at x. In our interpretation, the ? is a placeholder for any value in V_a.

Table 8.1: Missing data table

	a_1	a_2	a_3	a_4
x_1	5.00	4.00	3.00	2.00
x_2	?	3.00	2.00	1.00
x_3	2.00	?	4.00	5.00
x_4	?	2.00	?	3.00
x_5	2.00	2.00	?	?
x_6	5.00	4.00	3.00	2.00
x_7	3.00	2.00	1.00	1.00
x_8	3.00	2.00	5.00	?

Table 8.2: Single imputation via iterated linear regression

	a_1	a_2	a_3	a_4
x_1	5.00	4.00	3.00	2.00
x_2	4.30	3.00	2.00	1.00
x_3	2.00	2.00	4.00	5.00
x_4	2.39	2.00	2.28	3.00
x_5	2.00	2.00	1.62	3.11
x_6	5.00	4.00	3.00	2.00
x_7	3.00	2.00	1.00	1.00
x_8	3.00	2.00	5.00	4.17

In order to cope with the missing data entries several strategies can be used. The simplest one is ignorance, which means that any observation with missing entries is skipped from further

analysis. There is well-documented evidence to show that ignorance is usually a bad strategy [69].

A second stream of methods uses a *single imputation* strategy. Here, a missing value is replaced by the best suited replacement, where "best suited" is defined in terms of a statistical model. Such models usually make some distributional assumption such as a multinomial or a multivariate normal distribution, e.g. the treatment of missing data in the AMOS system [4], or the EQS system [8], or they use mixed multinomial/normal models [106].

Even though the approach has come under criticism [e.g. in 50], we use the *iterated linear regression algorithm for single imputation* as an illustrative example to show how missing data can be replaced by a simple statistical technique.

With the data of Table 8.1, we result in the underlined imputed values in Table 8.2. This method estimates the missing values by linear regression of the other variables in a form such as

$$a_1(x_3) = b_0 + b_2 a_2(x_3) + b_3 a_3(x_3) + b_4 a_4(x_3),$$

where b_0, b_2, b_3, b_4 are constants which are optimal to fit the regression line for the prediction of a_1 by the attributes a_2, a_3, a_4. Because the replacement of the missing values has changed after one cycle, many iteration steps are necessary to result in a stable optimal configuration. Furthermore, the critical assumption of this imputation model, namely, that there is a linear relationship among the variables, cannot be assumed in general. Even if the relationship is linear, the procedure faces another problem: The existence of only one outlier will bias the estimation of the b-values dramatically.

Of course, many non-linear relationships can be modelled simply by a non-linear transformation of the variables, but a sound model of the data is necessary to achieve a good description of the missing value. A badly chosen model can make the things even worse than the ignorance strategy, and hence, statistical imputation should be used with care. Detailed discussions of the interplay between the model used for imputation and the model used for analysis can be found in [71] and [105].

A third stream of methods employs the *multiple imputation* strategy [104]. Here, the data set is replaced by a number of "mutual" data sets in order to simulate the uncertainty about the missing values. Every data set can then be used within the data analysis, and a model based aggregation scheme enables the combination of the results.

The famous EM-algorithm [15] was one of the first effective approaches to handle missing data problems on the basis of the likelihood measure. One drawback of the EM algorithm is that it is slow and costly. Furthermore, it presupposes a restricted model class for the data, namely, when the distribution of the data is assumed to be a sample from a fixed distribution family such as the multivariate normal distribution. This is problematic, because the model assumption itself influences the estimation of the missing values.

8.2 Imputation from known values

In our non-invasive setup we do not have the tools of statistical analysis such as loss functions or a likelihood function; therefore, other optimisation criteria must be used. A simple criterion is the demand that the rules of the system should have a maximum in terms of consistency with respect to known values. If we fill a missing entry with a value, we should result in a rule which is consistent with the other rules of the system. Our algorithm imputes missing values in a granule $\vec{x}$ by presenting a list of possible values drawn from the set of all granules $\vec{y}$ which do not contradict $\vec{y}$, i.e. they have the same entries wherever both are defined. One feature of the procedure is that it does not inter-/extrapolate into unknown regions like the regression method, because there is no mechanism for inter-/extrapolation. The procedure uses only rules and dependencies within the observed body of the data, and therefore missing values can only replaced by known facts. Further consistent completions are usually possible, but cannot be drawn from existing values in other granules. This is typical for the cautious approach of non-invasive data analysis: If the procedure cannot replace a missing value it will signal a "do not know"- sign, which may be more reliable (and honest) than any extrapolation based on strong model assumptions.

We do not differentiate between independent attributes and a decision attribute, since we want to achieve consistency everywhere. Decision rules obtained from partial systems have been investigated in [59].

In order to formalise this situation we say that an information system $\mathcal{I}$ is *partial*, if the attributes are allowed to be partial functions. In other words, an attribute $a \in \Omega$ is a function $a : \mathrm{dom}(a) \to V_a$, where $\mathrm{dom}(a)$ is a subset of U, called the *domain of* a.

An *extension of* $\mathcal{I}$ is an information system, in which each attribute is an extension of an attribute of $\mathcal{I}$. More formally, $\mathcal{I}' = \langle U, \Omega', (V_a)_{a\in\Omega}\rangle$ such that the assignment $a \mapsto a'$ is a bijection from $\Omega \to \Omega'$, and a' is an extension of a. . A *completion of* $\mathcal{I}$ is an extension of $\mathcal{I}$ in which each attribute is a total function.

For each $x \in U$ and each $\emptyset \neq Q \subseteq \Omega$ let

$$\mathrm{rel}_Q(x) = \{a \in Q : x \in \mathrm{dom}\, a\}$$

be the set of *Q-relevant attributes for x*. For each $\emptyset \neq Q \subseteq \Omega$ we define a relation $Q_{\mathcal{I}}$ on U by

$$xQ_{\mathcal{I}}y \iff a(x) = a(y) \text{ for all } a \in \mathrm{rel}_Q(x) \cap \mathrm{rel}_Q\, y. \tag{8.8.1}$$

This relation has also been used in [59].

If $xQ_{\mathcal{I}}y$, we say that x and y are *consistent*. This terminology is justified by the fact that $xQ_{\mathcal{I}}y$ just in case that whenever a is defined on both x and y, it does not distinguish between

them. For example, x and y with

$$\vec{x}_Q = \langle 1, ?, 3\rangle,\ \vec{y}_Q = \langle 1, 4, ?\rangle$$

are consistent, while in case

$$\vec{x}_Q = \langle 1, ?, 3\rangle,\ \vec{y}_Q = \langle 1, ?, 2\rangle$$

they are not. Consistency is a generalisation of indiscernability used in RSDA: Two objects x, y are Q-indiscernable, if $\mathrm{rel}_Q(x) = Q = \mathrm{rel}_Q(y)$, and their induced granules are equal. The granules of two consistent objects can be made equal on the union of their relevant attributes by filling in missing values in one granule by values which are defined in the other granule.

It is clear from the definition that $Q_{\mathcal{I}}$ is reflexive and symmetric, but not necessarily transitive. Such relations are usually called *tolerance relations* or *similarity relations*. For $x \in U$ we set $Q_{\mathcal{I}}(x) = \{y \in U : xQ_{\mathcal{I}}y\}$. The sets $Q_{\mathcal{I}}(x)$ are called *similarity classes*. Clearly,

$$Q_{\mathcal{I}}(x) = \{y \in U : (\forall a \in Q)[a(x) = a(y) \text{ or } a(x) \text{ is not defined or } a(y) \text{ is not defined}]\},$$

We call $x \in U$ *a-casual* (with respect to Q and $\mathcal{I}$), if $a \in Q$, and

$$(\forall y)[y \in Q_{\mathcal{I}}(x) \Rightarrow y \notin \mathrm{dom}(a)]. \tag{8.8.2}$$

In this case, there is no information for x with respect to attribute a from any granule compatible with $\vec{x}_Q$.

For Table 8.1 we have the following similarity classes:

$$\begin{array}{ll}
\Omega_{\mathcal{I}}(x_1)=\{x_1, x_6\}, & \Omega_{\mathcal{I}}(x_5)=\{x_4, x_5\},\\
\Omega_{\mathcal{I}}(x_2)=\{x_2\}, & \Omega_{\mathcal{I}}(x_6)=\{x_1, x_6\},\\
\Omega_{\mathcal{I}}(x_3)=\{x_3, x_5\}, & \Omega_{\mathcal{I}}(x_7)=\{x_7\},\\
\Omega_{\mathcal{I}}(x_4)=\{x_4, x_5, x_8\}, & \Omega_{\mathcal{I}}(x_8)=\{x_4, x_8\}.
\end{array}$$

We observe that x_2 is a_1-casual.

It is our aim to transform a partial system $\mathcal{I}$ into a system without missing values. If the granule $\vec{x}_Q$ has a missing value at, say, $a \in Q$, we will try to impute it from the a-values of the objects in the similarity class of x.This will not always be possible, and, if it is, there may not be a unique value. Thus, the result of the imputation process will in some (or many) cases be a list of values from which a value may be picked, possibly by other methods, without violating the consistency.

The next result is crucial for our imputation algorithm:

Lemma 8.1. *If* $\mathrm{rel}_Q(x) \subseteq \mathrm{rel}_Q(y)$ *and* $a(x) = a(y)$ *for all* $a \in \mathrm{rel}_Q(x)$, *then* $Q_{\mathcal{I}}(y) \subseteq Q_{\mathcal{I}}(x)$.

Proof. Let $\mathrm{rel}_Q(x) \subseteq \mathrm{rel}_Q(y)$, and assume $Q_{\mathcal{I}}(y) \not\subseteq Q_{\mathcal{I}}(x)$. Then, there is some $z \in U$ such that

1. Whenever $a \in Q$ is defined for y and z, then $a(y) = a(z)$.
2. There is some $a_0 \in Q$ such that $a_0(x)$ and $a_0(z)$ exist, and $a(x) \neq a(z)$.

By the assumption $\mathrm{rel}_Q(x) \subseteq \mathrm{rel}_Q(y)$, a_0 is defined for y as well. Since $a_0(y) = a_0(z)$ by 1., we have $a_0(x) \neq a_0(y)$, contradicting $a(x) = a(y)$ for all $a \in \mathrm{rel}_Q(x)$. □

Next, we define a mapping $m : U \times \Omega \to \bigcup_{a \in \Omega} 2^{V_a}$ which will give us the possible imputable values by collecting for each $x \in U$ and each $a \in \Omega$ those entries which appear as entries $a(y)$ in the granules induced by a $y \in U$ which is consistent to x.

(8.8.3)

$$m(x,a) = \begin{cases} a(x), & \text{if } a(x) \text{ is defined,} \\ \{a(y) : y \in Q_{\mathcal{I}}(x)\}, & \text{if } a \text{ is not defined at } x, \text{ but } a \text{ is defined for some } y \in Q_{\mathcal{I}}(x), \\ ?, & \text{otherwise.} \end{cases}$$

We see that m leaves unique values alone; furthermore, if a is not defined at any $y \in Q_{\mathcal{I}}(x)$, i.e. if x is a – casual, then we will not be able to fill the entry $\langle x, a\rangle$; in this case, there is no "collateral knowledge" for $\langle x, a\rangle$. This is the case, when a rule is based on only one granule; we have discussed this briefly in [33].

Based on Lemma 8.1, we can now give a non-invasive imputation algorithm.

Algorithm 1. Define a sequence of information systems as follows:

1. $\mathcal{I}_0 = \mathcal{I}$.
2. Suppose that $\mathcal{I}_k = \langle U, \Omega_k, \{V_{a^k} : a^k \in \Omega_k\}\rangle$ is defined for some $k \geq 0$.
   (a) Find the similarity classes $Q_{\mathcal{I}_k}(x)$.
   (b) For each $a^k \in \Omega_k$, $x \in U$, let

$$a^{k+1}(x) = \begin{cases} m(x, a^k), & \text{if } |m(x, a^k)| = 1, \\ ?, & \text{otherwise.} \end{cases}$$

   (c) Set $\Omega_{k+1} \stackrel{\text{def}}{=} \{a^{k+1} : a^k \in \Omega_k\}$ and $V_{a^{k+1}} \stackrel{\text{def}}{=} V_{a^k}$.

With this procedure, we successively extend the attribute mappings; in other words, we increase (or leave constant) $\mathrm{rel}_\Omega(x)$. Lemma 8.1 now tells us that there is a k such that $Q_{\mathcal{I}_k}(x) = Q_{\mathcal{I}_{k+1}}(x)$ for all $x \in U$. At this step we report the matrix $\langle m(x,a)\rangle_{a\in\Omega_k}$. If $m(x,a)$ has more than one element, this set will give us the possibilities for value $a(x)$, based on previous experience.

Except for those entries for which there is no collateral knowledge, the first step of the algorithm will put a list of possible values into every other cell, so that, in the end, all ? will be removed in these cells. Thus, in removing ? entries, we cannot get any better.

One might argue that there is a bias towards one element sets $m(x,a)$, since we always fill those in first, and then compute the similarity classes. If we have committed ourselves to minimise the number of remaining ? and fill in whatever we can, we have no other choice: We must impute singletons first, since they are all we have in such an instance. As remarked above, this procedure leads to the least number of remaining ? cells.

The steps taken by this method for the example data are given in Table 8.3. We have kept $m(x,a)$ to indicate how the compatibilities change after we have extended attribute functions in one step.

Table 8.3: Non-invasive imputation I

	a_1	a_2	a_3	a_4
x_1	5.00	4.00	3.00	2.00
x_2	?	3.00	2.00	1.00
x_3	2.00	$\{a_2(x_5)\}$	4.00	5.00
x_4	$\{a_1(x_5), a_1(x_8)\}$	2.00	$\{a_3(x_8)\}$	3.00
x_5	2.00	2.00	$\{a_3(x_3)\}$	$\{a_4(x_3), a_4(x_4)\}$
x_6	5.00	4.00	3.00	2.00
x_7	3.00	2.00	1.00	1.00
x_8	3.00	2.00	5.00	$\{a_4(x_4)\}$

After the first step we note that we cannot make the table complete, since x_2 is a_1-casual. There is some ambiguity for the observations x_4 and x_5 which cannot be resolved in the first step. The reason is that there are – at this step – consistent replacements of the missing values in x_3, x_5, and x_8 respectively, which allow us to build a suitable granule for the prediction of the missing values of x_4 and x_5. Since the similarity classes are reduced in step 2, there are less possibilities for replacement, and, indeed, all ambiguities can be resolved.

As another example, consider the data in Table 8.6 taken from [59].

Table 8.4: Non-invasive imputation II

	a_1	a_2	a_3	a_4
x_1	5.00	4.00	3.00	2.00
x_2	?	3.00	2.00	1.00
x_3	2.00	{2.00}	4.00	5.00
x_4	$\{a_1(x_8)\}$	2.00	{5.00}	3.00
x_5	2.00	2.00	{4.00}	$\{a_4(x_3)\}$
x_6	5.00	4.00	3.00	2.00
x_7	3.00	2.00	1.00	1.00
x_8	3.00	2.00	5.00	{3.00}

Table 8.5: Non-invasive imputation III: Final state

	a_1	a_2	a_3	a_4
x_1	5.00	4.00	3.00	2.00
x_2	?	3.00	2.00	1.00
x_3	2.00	{2.00}	4.00	5.00
x_4	{3.00}	2.00	{5.00}	3.00
x_5	2.00	2.00	{4.00}	{5.00}
x_6	5.00	4.00	3.00	2.00
x_7	3.00	2.00	1.00	1.00
x_8	3.00	2.00	5.00	{3.00}

Table 8.6: Car data I

Car	Price	Mileage	Size	Max-Speed
1	high	low	full	low
2	low	?	full	low
3	?	?	compact	low
4	high	?	full	high
5	?	?	full	high
6	low	high	full	?

Table 8.7: Car data II

Car	Price	Mileage	Size	Max-Speed
1	high	low	full	low
2	low	high	full	low
3	?	?	compact	low
4	high	high	full	high
5	high	high	full	high
6	low	high	full	low

We have

$$\Omega_{\mathcal{I}}(1) = \{1\}, \Omega_{\mathcal{I}}(2) = \{2, 6\}, \Omega_{\mathcal{I}}(3) = \{3\},$$
$$\Omega_{\mathcal{I}}(4) = \{4, 5\}, \Omega_{\mathcal{I}}(5) = \{4, 5, 6\}, \Omega_{\mathcal{I}}(6) = \{2, 5, 6\}.$$

Observe that attribute 3 is price- and mileage-casual. The result of our imputation procedure is given in Table 8.7. The original in [59] has an additional decision attribute as shown in Table 8.8. The similarity classes for the new table are as above, except for

Table 8.8: Extended car data I

Car	Price	Mileage	Size	Max-Speed	d
1	high	low	full	low	good
2	low	?	full	low	good
3	?	?	comp.	low	poor
4	high	?	full	high	good
5	?	?	full	high	excellent
6	low	high	full	?	good

Table 8.9: Extended car data II

Car	Price	Mileage	Size	Max-Speed	d
1	high	low	full	low	good
2	low	high	full	low	good
3	?	?	comp.	low	poor
4	high	?	full	high	good
5	?	?	full	high	excellent
6	low	high	full	low	good

$$\Omega_{\mathcal{I}}(4) = \{4\}, \Omega_{\mathcal{I}}(5) = \{5\},$$

due to the separating of 4 and 5 by the new attribute.

In the study [59], imputation criteria are derived from the complexity of rules which are obtained from the incomplete system. These constructions are somewhat involved, and we refer the reader to [59] for details. The connection between the rule complexity and the quality of imputation, however, is not explained there.

A simulation study carried out in [44] has shown that the number of possible replacements is negatively correlated with state complexity. This is a price one has to pay if one wants to know the possible (and not the probable best) replacements: If the state complexity grows, the chance of mutual replacements which are in the underlying distribution of the data rises; if the number of missing values grows as well, the risk of resulting in many replacements grows multiplicatively.

A direct use of the results of this procedure in further statistical analysis will usually not be feasible, because the number of possible data sets will be by far too high to be of any practical value in, say, a multiple imputation procedure, which needs only a few number of imputed data sets ([104, p. 117], [106]). The intension of the proposed procedure is to inform the user of what might happen if missing values are imputed, which is a different goal to than to find a (statistical) procedure to estimate a model among variables. This interplay of non-invasive computing and more demanding statistical modelling is intended: Non-invasive computing shows which results are possible from the obtained data – statistical modelling offers the most probable solution of the problem.

Chapter 9

Beyond rough sets

We have said at the beginning of this book that we regard the rough set method as a paradigm which is based on the principles (1.1.1) and (1.1.2). These principles are, of course, not only applicable to the traditional realm of rough sets - attribute reduction and rule generation from information systems - but also in other situations, for example, approximation of regions in spatial reasoning and psychometric skill theory. In this Chapter, we shall explore a few other scenarios for the RSDA principles.

9.1 Relational attribute systems

Recall that the model assumptions of the OBJECT $\rightarrow$ ATTRIBUTE operationalisation were

1. Each object has exactly one value for each attribute at a given point in time.
2. The observation of this value is without error.

While we have addressed the second condition in Chapter 7, we shall discuss in this Section how the first condition can be extended.

A natural way to relax the uniqueness of assignment is to allow a set of attribute values to be associated with an object via an attribute function, so that each attribute is a mapping

$$a : U \rightarrow 2^{V_a}.$$

We have already encountered these structurs on p. 45 in Chapter 5.1. These *multi-valued information systems* were introduced by Lipski [67, 68] under the name of *systems with incomplete information*. They are also used in symbolic data analysis, e.g. [17, 18, 96] and in rough set-based data analysis, e.g. [86, 87, 124]. .

With such systems, things are not as straightforward as they seem. Consider an attribute a which is interpreted as "Languages spoken", and suppose that

$$a(\text{Tom}) = \{\text{French, German, English}\}.$$

There are many possible interpretations, of which we list just four:

(9.9.1)
Tom speaks French or German or English (disjunctive, non-exclusive).

(9.9.2)
Tom speaks exactly one of French, German, English (disjunctive, exclusive).

(9.9.3)
Tom speaks French and German and English, and possibly others (conjunctive, non-exclusive).

(9.9.4)
Tom speaks French and German and English, and no other languages (conjunctive, exclusive).

It is therefore necessary to add semantic information to the design of the system, which needs to be fulfilled regardless of the concrete model, and in the sequel we shall present the approach taken in [37].

The basic idea is to replace the attribute functions by attribute relations. In this way, totality ("Each object has at least an attribute value") and uniqueness ("Each object has at most one attribute value") are special cases of the more general situation. We can also use several relations for the same attribute which can differentiate between different situations. For example, we can interpret

$$xI_at \iff x \text{ certainly speaks language } t,$$
$$xB_at \iff x \text{ possibly speaks language } t.$$

Totality, uniqueness, (non-)exclusiveness can all be expressed as relational constraints, for which we need some preparation. Suppose that $R \subseteq A \times B$. We let $\text{ran}(R) = \{y \in B : xRy \text{ for some } x \in A\}$ be the *range of* R, and $\text{dom}(R) = \{x \in A : xRy \text{ for some } y \in B\}$ be the *domain of* R. For each $x \in A$, we let $R(x) = \{y \in B : xRy\}$. $R^\smile$ is the *converse of* R, i.e.

$$xR^\smile y \iff yRx. \tag{9.9.5}$$

If $R \subseteq A \times B,\ S \subseteq B \times C$, then $R;S$ is the *composition of* R *and* S, i.e.

$$x(R;S)y \iff (\exists z \in B)[xRz \text{ and } zSy]. \tag{9.9.6}$$

We denote by 1_A the *universal relation on* A

$$1_A = \{\langle x, y\rangle : x, y \in A\}, \tag{9.9.7}$$

and by $1'_A$ the *identity relation on* A

$$1'_A = \{\langle x, x\rangle : x \in A\}. \tag{9.9.8}$$

Now, "I_a is a function" (i.e. an attribute in the conventional sense) is equivalent to the equations

$$I_a; 1_{V_a} = U \times V_a \qquad I_a \text{ is total,} \tag{9.9.9}$$

$$I_a{}^{\smile}; I_a \subseteq 1'_{V_a} \qquad I_a \text{ is unique.} \tag{9.9.10}$$

This leads to our main definition: A *relational attribute system* (RAS) is a structure

$$\langle U, \Omega, \langle \mathcal{R}_a\rangle_{a\in\Omega}, \langle V_a\rangle_{a\in\Omega}, \Delta\rangle$$

such that

1. U is a finite set of objects.
2. Ω is a finite set of attribute names.
3. $R \subseteq U \times V_a$ for each $a \in \Omega$ and each $R \in \mathcal{R}_a$.
4. Δ is a set of constraints.

We do not want to put restrictions on Δ; the constraints will often be expressible in first order logic or, more specific, as relational equations; examples for constraints are (9.9.9) and (9.9.10).

For simplicity, we will restrict ourselves in the sequel to the case of just one attribute name a, and two relations $I, B \in \mathcal{R}_a$, which we will interpret as

$$xIt \qquad \Longleftrightarrow x \text{ certainly has } a\text{-property } t,$$

$$xBt \qquad \Longleftrightarrow x \text{ possibly has } a\text{-property } t.$$

The relation I enables us to express conjunctive conditions, while B allows disjunctive information to be expressed.

In order to facilitate notation, we suppose that we have a matrix M in which the columns are labelled with the elements of V_a, and the rows with the elements of U. If xIt, we indicate this by writing ♣ into cell $\langle x, t\rangle$, and if xBt, we write ♢.

We want to point out that in this setup we only express positive knowledge. Therefore, the absence of ♣ or ♢ in cell $\langle x, t\rangle$ does not mean that x does not have property t; this could be expressed by introducing a third relation.

A general constraint that we make is

(9.9.11) $$I \cap B = \emptyset.$$

Therefore, we regard uncertainty as strict in the sense that xBt means that we regard x possibly having property t, but we ceratinly do not know for certain. In matrix form, (9.9.11) can be expressed as

(9.9.12) Each cell of M contains at most one entry.

In rough set theory, two objects in a single-valued information system are called indiscernable, if they have the same feature vector. In a multivalued system there are other possibilities which use set theoretic relations on the sets $a(x)$. This leads to the *information relations* first studied in [85]. Our relational setting extends these relations in the following way: We will consider the relations

(9.9.13) $$=,\ \subsetneq,\ \supsetneq,\ O,\ D,$$

where for a set M and subsets t, u of M,

$tOu \iff t \cap u \neq \emptyset,$ and t and u are incomparable with respect to $\subseteq$, (overlap)

$tDu \iff t \cap u = \emptyset.$ (disjoint)

Note that O is a partial overlap, since we exclude the inclusions. The relations of (9.9.13) partition $M \times M$. Such "intersection tables" have been considered in qualitative spatial reasoning, for example, in [39, 40] for the interior I and boundary B of sets in a topological space.

Set $H = I \cup B$. Given x, y in U, there are nine ways of relating an element of $\{I(x), B(x), H(x)\}$ with an element of $\{I(y), B(y), H(y)\}$, and we denote these possibilities by row headings

(9.9.14) $$II,\ IB,\ IH,\ BI,\ BB,\ BH,\ HI,\ HB,\ HH.$$

We can now construct a relational table by indicating below each heading which of the relations of (9.9.13) holds. Of course, not all configurations are possible, since we have to observe the constraints

(9.9.15) $$H = I \cup B \text{ and } I \cap B = \emptyset,$$

which imply other conditions. If, for example, one of the entries is $=$, then the additional constraints are listed in Table 9.1. There, for example, the entry D in the cell $\langle I(x) = I(y), BI\rangle$ means that $I(x) = I(y)$ implies $B(x) \cap I(y) = \emptyset$.

Using the relationdefined above, we can now express quite general relationships amnong the objects in U based on their behaviour with respect to the attribute relations. Suppose that

Table 9.1: Equality constraints

	II	IB	IH	BI	BB	BH	HI	HB	HH
$I(x) = I(y)$	$=$	D	$\subsetneq$	D			$\supsetneq$		
$I(x) = B(y)$	D	$=$	$\subsetneq$		D			$\supsetneq$	
$I(x) = H(y)$	$\supsetneq$	$\supsetneq$	$=$	D	D	D	$\supsetneq$	$\supsetneq$	$\supsetneq$
$B(x) = I(y)$	D			$=$	D	$\subsetneq$	$\supsetneq$		
$B(x) = B(y)$		D		D	$=$	$\subsetneq$		$\supsetneq$	
$B(x) = H(y)$	D	D	D	$\supsetneq$	$\supsetneq$	$=$	$\supsetneq$	$\supsetneq$	$\supsetneq$
$H(x) = I(y)$	$\subsetneq$	D	$\subsetneq$	$\subsetneq$	D	$\subsetneq$	$=$	D	$\subsetneq$
$H(x) = B(y)$	D	$\subsetneq$	$\subsetneq$	D	$\subsetneq$	$\subsetneq$	D	$=$	$\subsetneq$
$H(x) = H(y)$			$\subsetneq$			$\subsetneq$	$\supsetneq$	$\supsetneq$	$=$

$R, S \in \{I, B, H\}$, and that Q is one of the relations of (9.9.13). A relation T on U is called an *elementary information relation* if it has the form

(9.9.16) $$xTy \iff \langle R(x), S(y)\rangle \in Q.$$

Any $\cup, \cap$ – combination of elementary information relations is called an *a-information relation*. ; this generalises the information relations of [85]. This can be further generalised by considering more than one attribute, but we shall not consider this here. A proof theory for relational attribute systems has been presented in [38].

As an example we will consider the following situation: A procedure often employed in psychological research is *expert-based categorisation*: A collection of N items a_i – such as statements, behaviour sequences etc – are presented to an expert, who is asked to assign each one to exactly one of n categories C_i. If two experts solve this task, then these categories can be cross-classified in a table as follows:

Category:	C_1	C_2	...	C_n
No. of agreements:	k_1	k_2	...	k_n

One problem of this procedure is that experts often cannot or will not assign the items to a unique category, since statements or behavioural sequences can often be interpreted in more than one way, so that there could be more than one category to which they could be assigned. By having to assign an item to exactly one category, this information is suppressed, and, in case the experts ratings differ significantly, it cannot be said whether the experts strongly disagree, or whether the categories are not sufficiently discriminating.

In order to surmount this problem, one can offer the experts a choice among the following alternatives:

(9.9.17) Each item is assigned to a unique category, as described above.

This is the standard procedure. But there are less restrictive possibilities:

(9.9.18) Each item is assigned to a main category and zero or more lesser categories.

(9.9.19) Each item is assigned to one or more categories “aequo loco”.

We can express these situations with our RAS operationalisation as follows: Let $U = \{E_1, \ldots E_t\}$ be the set of experts, and for each item a_i, $1 \leq i \leq N$, let $V_{a_i} = \{C_1, \ldots, C_n\}$ be the set of possible categories. The relations which we consider are I_{a_i} and B_{a_i}; their meaning is given by

- $\langle E, C\rangle \in I_{a_i}$ means that expert E classifies item a_i as certainly belonging to category C.
- $\langle E, C\rangle \in B_{a_i}$ means that expert E classifies item a_i as possibly belonging to category C.

The constraints corresponing to the conditions (9.9.17) – (9.9.19) can be described by

(9.9.20) I_{a_i} is a function, and $B_{a_i} = \emptyset$ for all a_i.

(9.9.21) I_{a_i} is a function, and $I_{a_i} \cap B_{a_i} = \emptyset$ for all a_i.

(9.9.22) I_{a_i} is total and $B_{a_i} = \emptyset$ for all a_i.

All these constraints can be expressed as relational equations.

Various statistics can now be employed to gauge the quality of (dis-)agreement, and one can explore the areas of possible reconciliation. We invite the reader to consult [37] for more details and examples.

9.2 Non-invasive test theory

The aim of an assessment is to find out which skills a person has in a given area. An operationalisation of this domain of interest is often a set of problems, and a test containing these problems can be regarded as an empirical system. A scaling would map the test results to some linear scale of grades. These grades, however, are so far removed from the domain of interest – the skills a person has –, that alternative methods of “measurement” need to be employed, if sensible information is to be gained. The theory of *knowledge structures* has been developed by Doignon and Falmagne [20, 21] to handle structural dependencies among sets of problems. Typically, it provided rules such as

> If person x can solve problem p, then (s)he can also solve problem q.

The theory of knowledge structures has been extensively used for automated assessment. One drawback of the theory was its assumption that the ability to solve problems could be equated to the possession of appropriate skills, which, however, are not part of the theory. Subsequently, Doignon [19] and Düntsch and Gediga [27] independently developed similar skill theories; in other words, they studied the operationalisation

$$\text{Set of skills } \mapsto \text{ Set of problems.}$$

In the sequel, we will follow the approach given in [27, 46].

Suppose that S is a finite set of skills, Q a finite set of dichotomous problems, and U a finite set of subjects[1]. For each $x \in U$ we let $s(x)$ be the set of problems (s)he is able to solve. The collection

$$\mathcal{K}_s = \{s(x) : x \in U\}$$

is called the *empirical knowledge structure* (EKS) with respect to Q and U; the elements of K_s are called *empirical knowledge states*.

Observe that s determines a relation $T_s \subseteq U \times Q$ such that

$$xTq \iff x \text{ solves } t. \tag{9.9.23}$$

With this in mind, we sometimes will call a triple $\langle U, Q, T\rangle$ an EKS, where $T \subseteq U \times Q$. This gives us the freedom to intepret the knowledge states in other settings; for example, xTq could mean "Person x has symptom q". We remark in passing that such a structure is the same as a context in the sense of [125].

A *problem function* is a mapping $2^S \to q^Q$, such that

$$\delta(\emptyset) = \emptyset,\ \delta(S) = Q. \tag{9.9.24}$$

$$\delta \text{ is monotone with respect to } \subseteq. \tag{9.9.25}$$

We interpret $\delta(X)$ as the set of problems which can be solved with the skills in X. The set

$$\mathcal{K}_\delta = \{K(X) : X \subseteq S\}$$

is called the *skill knowledge structure* (SKS) with respect to S and Q. Its elements are called *skill* or *theoretical knowledge states*.

Associated with δ is a relation $\Gamma_\delta \subseteq Q \times (2^S \setminus \{\emptyset\})$ such that

$$q\Gamma_\delta X \iff q \text{ can be solved with the skills in } X, \text{ but not with any proper subset of } X.$$

[1] We follow, not without hesitation, the nomenclature of the psychological literature.

It can be shown that, given our assumptions on δ,

(9.9.26) $$\mathrm{dom}(\Gamma_\delta) = Q,$$

(9.9.27) $$q\Gamma_\delta X \Rightarrow X \neq \emptyset,$$

(9.9.28) $q\Gamma_\delta X$ and $q\Gamma_\delta Y \Rightarrow X = Y$ or X, Y are incomparable with respect to $\subseteq$.

Conversely, it can be shown that for every relation Γ satisfying (9.9.26) – (9.9.28), there is exactly one problem function δ such that $\Gamma_\delta = \Gamma$. A relation with these properties is called a *skill relation*. We can interpret $q\Gamma X$ as "X is minimally necessary to solve q", and call X a *strategy for* q. In what follows, we suppose that a problem function δ is given, along with its skill relation Γ. We also let $\sigma(x)$ be the true skill state of x, i.e. the set of all skills from S which x possesses.

The aim of a test should not be only to observe which problems a student has solved, but rather, which skills the student has. Due to possibly different solving strategies and random influences, the true skill state of a subject t is usually not directly observable, even if $s(t)$ is a knowledge state which is consistent with the theory. Therefore, we have to look for ways to approximate $\sigma(t)$. To acquaint the reader with the proposed concepts, we first look at some examples:

Example 1. Suppose that $Q = \{m, p, q, r\}$, $S = \{a, b, c, d\}$, and

$$\Gamma(m) = \{\{a\}, \{b\}\}, \Gamma(p) = \{\{c\}, \{d\}\}, \Gamma(q) = \{\{a, c\}\}, \Gamma(r) = \{\{a, d\}\}.$$

The associated knowledge structure $\mathcal{K}_\delta$ is given in Table 9.2.

Table 9.2: Knowledge structure, Example 1

Skills	$\emptyset$	a	b	c	d	ab	ac	ad
Problems	$\emptyset$	m	m	p	p	m	mpq	mpr
Skills	bc	bd	cd	abc	abd	acd	bcd	S
Problems	mp	mp	p	mpq	mpr	Q	mp	Q

Now, let t solve exactly m and p. Then, $s(t)$ is a knowledge state, and $\delta(X) = s(t)$ if and only if X is one of the sets

$$\{b, c\},\ \{b, d\}, \{b, c, d\}$$

Since we do not know which of these sets led to $s(t)$, it follows that t certainly has skill b, which is contained in all these sets, and possibly the skills b, c, d. If we had $\delta(\{b, c, d\}) = \{m, p, r\}$, then we would still have the same set $\{b, c, d\}$ of possible skills with the additional information that t cannot have both skills c, d, since $r \notin s(t)$. □

Example 2. Whereas Example 1 shows that true states within a knowledge structure not necessarily imply a unique set of skills for each problem, ambiguity will come into play, even if every problem can be solved by exactly one minimal skill set.

Let

$$\begin{aligned}
\Gamma(p_1) &= \{\{s_1\}\},\\
\Gamma(p_2) &= \{\{s_1, s_2\}\},\\
\Gamma(p_3) &= \{\{s_1, s_2, s_3\}\},\\
\Gamma(p_4) &= \{\{s_1, s_2, s_3, s_4\}\},\\
\Gamma(p_5) &= \{\{s_1, s_2, s_3, s_4, s_5\}\},\\
\Gamma(p_6) &= \{\{s_1, s_2, s_3, s_4, s_5, s_6\}\},
\end{aligned}$$

The states of $\mathcal{K}_\delta$ are

$$\emptyset, \{p_1\}, \{p_1, p_2\}, \{p_1, p_2, p_3\}, \ldots \{p_1, p_2, p_3, p_4, p_5, p_6\}.$$

Suppose now that $s(t) = \{p_1, p_2, p_3, p_6\}$. The problem arises that $s(t)$ is **not** a knowledge state in $\mathcal{K}_\delta$; however, since in this case $\mathcal{K}_\delta$ is actually closed under $\cup$ and $\cap$, we can find the largest state $\underline{s(t)} = \{p_1, p_2, p_3\}$ below $s(t)$ and the smallest state $\overline{s(t)} = \{p_1, p_2, p_3, p_4, p_5, p_6\}$ above $s(t)$.

The subject solves the first three problems consistent with the skill theory. Due to the structure of the skill assignment, we conclude that the body of certain knowledge of the subject is $\{s_1, s_2, s_3\}$. Although the subject did not solve problems p_4 and p_5, the subject solves p_6. Though this is inconsistent with the theory, we cannot preclude that t indeed has the skills necessary to solve p_6. Hence, with some justification, we can say that an upper bound for $\sigma(t)$ is S. The interpretation of the set $\{s_4, s_5, s_6\}$ contained in the upper bound but not in the lower bound is that the skills in this set are those which the subject possibly has; in other words, the set describes the "border" of the subject's knowledge. □

Example 3. Let $Q = \{p, q, r, m, n\}$, $S = \{a, b, c, d\}$, and

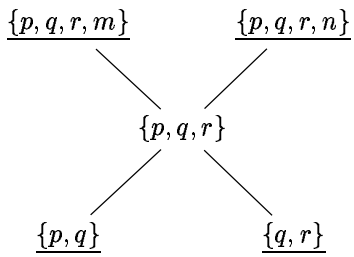
$$\begin{aligned}
\Gamma(p) &= \{\{a\}, \{b\}\},\\
\Gamma(q) &= \{\{a\}, \{b\}, \{c\}, \{d\}\},\\
\Gamma(r) &= \{\{c\}, \{d\}\},\\
\Gamma(m) &= \{\{a, c\}, \{a, d\}\},\\
\Gamma(n) &= \{\{b, c\}, \{b, d\}\},
\end{aligned}$$

The associated knowledge structure $\mathcal{K}_\delta$ is given in Table 9.3.

Table 9.3: Knowledge structure, Example 3

Skills	$\emptyset$	a	b	c	d	ab	ac	ad
Problems	$\emptyset$	pq	pq	qr	qr	pq	$pqrm$	$pqrm$
Skills	bc	bd	cd	abc	abd	acd	bcd	S
Problems	$pqrn$	$pqrn$	qr	Q	Q	Q	Q	Q

Figure 9.1: Example 3



Suppose that $s(t) = \{p, q, r\}$. The states in $\mathcal{K}_\delta \setminus \{\emptyset, Q\}$ are underlined in Figure 9.1. In particular, $s(t) \notin \mathcal{K}_\delta$. Since the true states $\{p, q\}$ and $\{q, r\}$ are subsets of $s(t)$, we can assume, however, that t possesses any skill which belongs to all skill sets X with $\delta(X) = \{p, q\}$, as well as any skill which belongs to all skill sets Y with $\delta(Y) = \{q, r\}$. The set of all these skills determines the lower bound of $\sigma(t)$, which, in this case is $\emptyset$.

If we consider the set of skills which t possibly has, we see that any of the sets $\{a, b\}$, $\{b, c\}$, $\{b, d\}$ is consistent with $\{p, q, r\}$, and therefore, we conclude that no skill can be definitely excluded from $\sigma(t)$. □

These examples show that, in general, it is not possible to give a necessary and sufficient condition relating $s(t)$ and $\sigma(t)$. Furthermore, it is not clear what the connection between $\sigma(t)$ and $s(t)$ should be. One extreme would be that there is no connection at all: Each $q \in s(t)$ is the result of a lucky guess, and each $p \in S \setminus \sigma(t)$ is the result of a careless error. In order to build a sensible structural theory, we clearly need to limit the scope of random influences, and therefore, in the sequel we shall assume two conditions. First, let

$$\underline{s(t)} = \{T \in \mathcal{K}_\delta : T \subseteq s(t) \text{ and } T \text{ is maximal with this property}\},$$

Observe that $s(t) \in \mathcal{K}_\delta$ implies $\underline{s(t)} = s(t)$. Now,

(9.9.29) $$s \in \sigma(t) \Rightarrow (\exists q \in s(t))(\exists X \subseteq S)[q \Gamma X \text{ and } s \in X].$$

(9.9.30) $$(\exists T \in \underline{s(t)})(\forall q \in T)(\forall X \subseteq S)[q \Gamma X \Rightarrow s \in X] \Rightarrow s \in \sigma(t).$$

Condition (9.9.29) relates skills to solved problems: If t has a skill s, then this should be witnessed by some problem which t is able to solve, and some strategy of which requires s. Note that we do not take into account latent skills and thus, careless errors. Condition (9.9.30) says that if $s(t)$ contains a maximal knowledge state T according to δ, then we can assume that t possesses each skill which is contained in each strategy for T. This limits the possibility of lucky guesses in $s(t)$.

This leads to the following definitions:

1. The *upper approximation of* $\sigma(t)$ is the set

   (9.9.31) $$\overline{\sigma(t)} = \{s \in S : (\exists q \in s(t))(\exists X \subseteq S)[q\Gamma X \text{ and } s \in X]\}.$$

2. The *lower approximation of* $\sigma(t)$ is the set

   (9.9.32) $$\underline{\sigma(t)} = \{s \in S : (\exists T \in \underline{s(t)})(\forall q \in T)(\forall X \subseteq S)[q\Gamma X \Rightarrow s \in X]\}.$$

3. The *area of uncertainty* is the set

   (9.9.33) $$\partial(t) = \overline{\sigma(t)} \setminus \underline{\sigma(t)}.$$

It is not hard to check that these definitions give us the upper and lower bounds heuristically obtained in Example 1 – 3. The conditions (9.9.29) and (9.9.30) now immediately imply

Proposition 9.1. $\underline{\sigma(t)} \subseteq \sigma(t) \subseteq \overline{\sigma(t)}$.

The interpretation of (9.9.31) as the upper bound of skills which t possibly has implies the assumption that

t certainly does not have a skill s if and only if $(\forall q \in s(t))(\forall X \subseteq S)[q\Gamma X \Rightarrow s \notin X]$.

9.2.1 Application: Guttman scaling revisited

One of the first investigators to realise that different scaling techniques are needed for qualitative attributes and quantitative ones was L. Guttman [51, 52]:

> "Guttman implicitly challenged the justification of applying correlation analysis to attributes, that is, manifest classifications for which only a few categories are available" [60].

He proposed a mathematical scale model in which one assigns numbers to individuals and problems such that t solves q if and only if the number assigned to t is greater than the number assigned to q. In the more formal terms of [24], an EKS $\langle U, Q, T\rangle$ is *Guttman scalable* if and only if there are functions $f : U \to \mathbb{N},\ g : Q \to \mathbb{N}$ such that

$$xTq \Longleftrightarrow f(x) \gtrsim g(q). \tag{9.9.34}$$

In one of the first applications of Guttman's scaling technique, Suchman [119] investigated physical reactions to dangers of battle experienced by soldiers who have been under fire. He showed that subjects and experienced symptons form an almost perfect Guttman scale. Presence of symptoms in decreasing frequency was as follows:

X_1	Violent pounding of the heart	84%
X_2	Sinking feeling of the stomach	73%
X_3	Feeling sick at the stomach	57%
X_4	Shaking or trembling all over	52%
X_5	Feeling of stiffness	50%
X_6	Feeling of weakness or feeling faint	42%
X_7	Vomiting	35%
X_8	Loosing control of the bowels	21%
X_9	Urinating in pants	9%

A skill theory consistent with Guttman scaling must assume that the items become increasingly "difficult" with decreasing manifest frequency. Therefore, a strategy for a "harder" item must contain all the skills necessary for "easier" items and at least one additional skill. In its simplest form, we obtain the following skill relation:

$$\begin{aligned}
\Gamma_G(X_1) &= \{\{A_1\}\},\\
\Gamma_G(X_2) &= \{\{A_1, A_2\}\},\\
\Gamma_G(X_3) &= \{\{A_1, A_2, A_3\}\},\\
\Gamma_G(X_4) &= \{\{A_1, A_2, A_3, A_4\}\},\\
\Gamma_G(X_5) &= \{\{A_1, A_2, A_3, A_4, A_5\}\},\\
\Gamma_G(X_6) &= \{\{A_1, A_2, A_3, A_4, A_5, A_6\}\},\\
\Gamma_G(X_7) &= \{\{A_1, A_2, A_3, A_4, A_5, A_6, A_7\}\},\\
\Gamma_G(X_8) &= \{\{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8\}\},\\
\Gamma_G(X_9) &= \{\{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9\}\}.
\end{aligned}$$

It is straightforward to see that the corresponding $\mathcal{K}_{\delta_G}$ has 10 states.

Looking at the symptoms (and not the observed data), we observe the following three categories:

- Slight somatic symptoms X_1, X_2.
- Medium to severe symptons without excretion X_3, X_4, X_5, X_6.
- Excretion X_7, X_9, X_8.

Assuming that the corresponding populations can be well separated, and that occurence of all symptoms in one group implies occurence of at least one sympton of each lower group, we can construct the following skill relation:

$$\begin{aligned}
\Gamma_3(X_1) &= \{\{B_1\}\},\\
\Gamma_3(X_2) &= \{\{B_2\}\},\\
\Gamma_3(X_3) &= \{\{B_1, B_2, B_3\}\},\\
\Gamma_3(X_4) &= \{\{B_1, B_2, B_4\}\},\\
\Gamma_3(X_5) &= \{\{B_1, B_2, B_5\}\},\\
\Gamma_3(X_6) &= \{\{B_1, B_2, B_6\}\},\\
\Gamma_3(X_7) &= \{\{B_1, B_2, B_3, B_4, B_5, B_6, B_7\}\},\\
\Gamma_3(X_8) &= \{\{B_1, B_2, B_3, B_4, B_5, B_6, B_8\}\},\\
\Gamma_3(X_9) &= \{\{B_1, B_2, B_3, B_4, B_5, B_6, B_9\}\}.
\end{aligned}$$

$\mathcal{K}_{\delta_3}$ has the 27 states which have one of the forms

$$\begin{aligned}
P, &\qquad P \subsetneq \{X_1, X_2\},\\
\{X_1, X_2\} \cup P, &\qquad P \subsetneq \{X_3, X_4, X_5, X_6\},\\
\{X_1, X_2, X_3, X_4, X_5, X_6\} \cup P, &\qquad P \subseteq \{X_7, X_8, X_9\}.
\end{aligned}$$

We have shown elsewhere [46] that the skill theory ending up in $\mathcal{K}_{\delta_3}$ is more suitable to the empirical data published in [119, p. 140] than the original Guttman scale version. Table 9.4 on the next page demonstrates that the set difference of lower and upper approximation of skill sets is much smaller in the new scale version than in the original, which supports the statistical findings in [46].

Table 9.4: Skill estimation in the soldier data

Pattern (X9 .. X1)	Freq.	27 states lower	27 states upper	10 states lower	10 states upper
000000000	7	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
000000010	1	$\{B_2\}$	$\{B_2\}$	$\emptyset$	$\{A_1, A_2\}$
000000001	7	$\{B_1\}$	$\{B_1\}$	$\{A_1\}$	$\{A_1\}$
000100000	1	$\emptyset$	$\{B_1, B_2, B_6\}$	$\emptyset$	${}^1_A \setminus \{A_7, A_8, A_9\}$
001000000	1	$\emptyset$	${}^1_B \setminus \{B_8, B_9\}$	$\emptyset$	${}^1_A \setminus \{A_8, A_9\}$
000001010	1	$\{B_1, B_2, B_4\}$	$\{B_1, B_2, B_4\}$	$\emptyset$	$\{A_1, A_2, A_3, A_4\}$
000010010	2	$\{B_2\}$	$\{B_1, B_2, B_5\}$	$\emptyset$	$\{A_1, A_2, A_3, A_4, A_5\}$
000000011	7	$\{B_1, B_2\}$	$\{B_1, B_2\}$	$\{A_1, A_2\}$	$\{A_1, A_2\}$
010000001	1	$\{B_1\}$	${}^1_B \setminus \{B_7, B_9\}$	$\{A_1\}$	${}^1_A \setminus \{A_9\}$
001000001	1	$\{B_1\}$	${}^1_B \setminus \{B_8, B_9\}$	$\{A_1\}$	${}^1_A \setminus \{A_8, A_9\}$
000010001	1	$\{B_1\}$	$\{B_1, B_2, B_5\}$	$\{A_1\}$	$\{A_1, A_2, A_3, A_4, A_5\}$
000000101	1	$\{B_1\}$	$\{B_1, B_2, B_3\}$	$\{A_1\}$	$\{A_1, A_2, A_3\}$
000001001	1	$\{B_1\}$	$\{B_1, B_2, B_4\}$	$\{A_1\}$	$\{A_1, A_2, A_3, A_4\}$
000100100	1	$\emptyset$	$\{B_1, B_2, B_6\}$	$\emptyset$	${}^1_A \setminus \{A_7, A_8, A_9\}$
000001011	3	$\{B_1, B_2, B_4\}$	$\{B_1, B_2, B_4\}$	$\{A_1, A_2\}$	$\{A_1, A_2, A_3, A_4\}$
000000111	2	$\{B_1, B_2, B_3\}$	$\{B_1, B_2, B_3\}$	$\{A_1, A_2, A_3\}$	$\{A_1, A_2, A_3\}$
000100011	1	$\{B_1, B_2, B_6\}$	$\{B_1, B_2, B_6\}$	$\{A_1, A_2\}$	${}^1_A \setminus \{A_7, A_8, A_9\}$
001010110	1	$\{B_2\}$	${}^1_B \setminus \{B_8, B_9\}$	$\emptyset$	${}^1_A \setminus \{A_8, A_9\}$
000011110	1	$\{B_2\}$	$\{B_1, B_2, B_3, B_4, B_5\}$	$\emptyset$	$\{A_1, A_2, A_3, A_4, A_5\}$
000010111	2	$\{B_1, B_2, B_3, B_5\}$	$\{B_1, B_2, B_3, B_5\}$	$\{A_1, A_2, A_3\}$	$\{A_1, A_2, A_3, A_4, A_5\}$
000100111	3	$\{B_1, B_2, B_3, B_6\}$	$\{B_1, B_2, B_3, B_6\}$	$\{A_1, A_2, A_3\}$	${}^1_A \setminus \{A_7, A_8, A_9\}$
000001111	2	$\{B_1, B_2, B_3, B_4\}$	$\{B_1, B_2, B_3, B_4\}$	$\{A_1, A_2, A_3, A_4\}$	$\{A_1, A_2, A_3, A_4\}$
000101011	1	$\{B_1, B_2, B_3, B_6\}$	$\{B_1, B_2, B_3, B_6\}$	$\{A_1, A_2\}$	${}^1_A \setminus \{A_7, A_8, A_9\}$
000011011	3	$\{B_1, B_2, B_4, B_5\}$	$\{B_1, B_2, B_4, B_5\}$	$\{A_1, A_2\}$	$\{A_1, A_2, A_3, A_4, A_5\}$
001000111	1	$\{B_1, B_2, B_3\}$	${}^1_B \setminus \{B_8, B_9\}$	$\{A_1, A_2, A_3\}$	${}^1_A \setminus \{A_8, A_9\}$
001010011	1	$\{B_1, B_2, B_5\}$	${}^1_B \setminus \{B_8, B_9\}$	$\{A_1, A_2\}$	${}^1_A \setminus \{A_8, A_9\}$
010011001	1	$\{B_1\}$	${}^1_B \setminus \{B_7, B_9\}$	$\{A_1\}$	${}^1_A \setminus \{A_9\}$
001110101	1	$\{B_1\}$	${}^1_B \setminus \{B_8, B_9\}$	$\{A_1\}$	${}^1_A \setminus \{A_9\}$
001100111	1	$\{B_1, B_2, B_3, B_6\}$	${}^1_B \setminus \{B_8, B_9\}\{A_1, A_2, A_3\}$	${}^1_A \setminus \{A_8, A_9\}$	
001011101	1	$\{B_1\}$	${}^1_B \setminus \{B_8, B_9\}$	$\{A_1\}$	${}^1_A \setminus \{A_8, A_9\}$
000111011	1	$\{B_1, B_2, B_4, B_5, B_6\}$	$\{B_1, B_2, B_4, B_5, B_6\}$	$\{A_1, A_2\}$	${}^1_A \setminus \{A_7, A_8, A_9\}$
010011101	1	$\{B_1\}$	${}^1_B \setminus \{B_7, B_9\}$	$\{A_1\}$	${}^1_A \setminus \{A_9\}$
001010111	1	$\{B_1, B_2, B_3, B_5\}$	${}^1_B \setminus \{B_8, B_9\}$	$\{A_1, A_2, A_3\}$	${}^1_A \setminus \{A_8, A_9\}$
000011111	2	$\{B_1, B_2, B_3, B_4, B_5\}$	$\{B_1, B_2, B_3, B_4, B_5\}$	$\{A_1, A_2, A_3, A_4, A_5\}$	$\{A_1, A_2, A_3, A_4, A_5\}$
001001111	2	$\{B_1, B_2, B_3, B_4\}$	${}^1_B \setminus \{B_8, B_9\}$	$\{A_1, A_2, A_3, A_4\}$	${}^1_A \setminus \{A_8, A_9\}$
000101111	1	$\{B_1, B_2, B_3, B_4, B_6\}$	$\{B_1, B_2, B_3, B_4, B_6\}$	$\{A_1, A_2, A_3, A_4\}$	${}^1_A \setminus \{A_7, A_8, A_9\}$
110000111	1	$\{B_1, B_2, B_3\}$	${}^1_B \setminus \{B_7\}$	$\{A_1, A_2, A_3\}$	${}^1_A$
011000111	1	$\{B_1, B_2, B_3\}$	${}^1_B \setminus \{B_9\}$	$\{A_1, A_2, A_3\}$	${}^1_A \setminus \{A_9\}$
001111101	1	$\{B_1\}$	${}^1_B \setminus \{B_8, B_9\}$	$\{A_1\}$	${}^1_A \setminus \{A_8, A_9\}$
001111011	1	$\{B_1, B_2, B_4, B_5, B_6\}$	${}^1_B \setminus \{B_8, B_9\}$	$\{A_1, A_2\}$	${}^1_A \setminus \{A_8, A_9\}$
001101111	1	$\{B_1, B_2, B_3, B_4, B_6\}$	${}^1_B \setminus \{B_8, B_9\}$	$\{A_1, A_2, A_3, A_4\}$	${}^1_A \setminus \{A_8, A_9\}$
001011111	1	$\{B_1, B_2, B_3, B_4, B_5\}$	${}^1_B \setminus \{B_8, B_9\}$	$\{A_1, A_2, A_3, A_4, A_5\}$	${}^1_A \setminus \{A_8, A_9\}$
000111111	6	${}^1_B \setminus \{B_7, B_8, B_9\}$	${}^1_B \setminus \{B_7, B_8, B_9\}$	${}^1_A \setminus \{A_7, A_8, A_9\}$	${}^1_A \setminus \{A_7, A_8, A_9\}$
011110111	1	${}^1_B \setminus \{B_4, B_9\}$	${}^1_B \setminus \{B_9\}$	$\{A_1, A_2, A_3\}$	${}^1_A \setminus \{A_9\}$
001111111	5	${}^1_B \setminus \{B_8, B_9\}$	${}^1_B \setminus \{B_8, B_9\}$	${}^1_A \setminus \{A_8, A_9\}$	${}^1_A \setminus \{A_8, A_9\}$
100111111	1	${}^1_B \setminus \{B_7, B_8\}$	${}^1_B \setminus \{B_7, B_8\}$	${}^1_A \setminus \{A_7, A_8, A_9\}$	${}^1_A$
010111111	1	${}^1_B \setminus \{B_7, B_9\}$	${}^1_B \setminus \{B_7, B_9\}$	${}^1_A \setminus \{A_7, A_8, A_9\}$	${}^1_A$
110111111	1	${}^1_B \setminus \{B_7\}$	${}^1_B \setminus \{B_7\}$	${}^1_A \setminus \{A_7, A_8, A_9\}$	${}^1_A$
011111111	7	${}^1_B \setminus \{B_9\}$	${}^1_B \setminus \{B_9\}$	${}^1_A \setminus \{A_9\}$	${}^1_A$
111111111	6	${}^1_B$	${}^1_B$	${}^1_A$	${}^1_A$

Chapter 10

Epilogue

In this book, we have given an introduction to and an overwiew of non-invasive data analysis based on the paradigm of RSDA. We hope to have shown that there are effective and mathematically well-founded tools for non-invasive data analysis, and that the gain in "model certainty" may outweigh the precise, but often uncertain, results of those methods which are based on strong model assumptions.

At the end, we would like to stress another argument for using non-invasive data analysis techniques, which follows from one of the most expensive negative results in the research on data analysis: The STATLOG project [72] had the aim to find the best algorithm(s) for supervised learning. Its main result was that most algorithms within the competition have about the same effectiveness. Because it is obvious that we will not find a "best" data analysis method when following this path of maximal effectiveness, we need other criteria for choosing a good data analysis strategy. We think that the criterion of "non-invasiveness" is a good choice: A good data analysis algorithm should be not a black, but a white box; the user should be aware of all assumptions and their consequences or, at least, most of them. This book makes a first step in this direction, and our procedures – e.g. the rough-entropy based searching (Section 6.2) or the lattice based classificatory algorithms ([124]) – are as good as the established fine tuned methods for supervised learning.

The considerations of this book show that many situations can be described with the simple trichotomy

Inside - Boundary - Outside.

and a very worthwhile research task is the investigation of systems with boundaries. There are already many flourishing areas which are concerned with such investigation, for example, qualitative spatial reasoning [13, 61], visual perception [47], or test theory [46].

An example for the three-part conceptualisation of the domain of interest is the “preparedness theory” of Seligman and Schwartz [108] in the field of learning psychology. They distinguish among being

- Prepared - for pure basic instinct reactions,
- Unprepared - for actions which an individual is able to learn,
- Contraprepared - for actions which an individual is not able to learn.

One cause for the low impact of this theory may have been the unavailability of a methodology which could deal with such boundary systems.

Most algorithms used in RSDA are NP-hard, and therefore, good heuristics are needed to approximate optimal solutions, as well as software which implements the procedures. We will endeavour to list RSDA resources in general and current software development in particular at

`www.methodos.co.uk/noninv/resources.html`

Bibliography

[1] Acock, A. (1997). Working with missing data. *Family Science Review*, 10:76–102.

[2] Adams, D. (1979). *The Hitch Hiker's Guide to the Galaxy*. Pan Books, London.

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

[4] Arbuckle, J. (1996). *Amos Users Guide: Version 3.6*. SmallWaters Corp., Chicago.

[5] Bazan, J. (1998). A comparison of dynamic and non-dynamic rough set methods fro extracting laws from decision tables. In [93], pages 321–365.

[6] Bazan, J., Nguyen, H., Nguyen, T., Skowron, A., and Stepaniuk, J. (1998). Synthesis of decision rules for object classification. In [93], pages 23–57.

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

[8] Bentler, P. (1996). *EQS: Structural Equations Program Manual*. BMDP Statistical Software, Los Angeles.

[9] Bjorvand, A. T. and Komorowski, J. (1997). Practical applications of genetic algorithms for efficient reduct computation. In [120], pages 601–606.

[10] Browne, C., Düntsch, I., and Gediga, G. (1998). IRIS revisited: A comparison of discriminant and enhanced rough set data analysis. In [94], pages 345–368.

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

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

[13] Cohn, A. G. and Gotts, N. M. (1996). The 'egg-yolk' representation of regions with

indeterminate boundaries. In Burrough, P. and Frank, A. M., editors, *Proc. of the GISDATA Specialist Meeting on Geographical Objects with Undetermined Boundaries*, pages 171–187. Francis Taylor.

[14] Comer, S. (1993). On connections between information systems, rough sets, and algebraic logic. In Rauszer, C., editor, *Algebraic Methods in Logic and Computer Science*, volume 28 of *Banach Center Publications*, pages 117–124. Polish Academy of Science, Warszawa.

[15] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society (B)*, 39:1–38.

[16] Devlin, K. (1997). *Goodbye, Descartes*. Wiley.

[17] Diday, E. (1987)). Introduction a l'approche symbolique en analyse des donnees. In *Actes des journees symboliques-numeriques pour l'apprentissage de connaissances a partir des donnes*. Paris.

[18] Diday, E. and Roy, L. (1988). Generating rules by symbolic data analysis and application to soil feature recognition. In *Actes des 8emes Journees Internationales: Les systemes experts et leurs applications*. Avignon.

[19] Doignon, J.-P. (1994). Knowledge spaces and skill assignments. In Fischer, G. and Laming, D., editors, *Contributions to Mathematical Psychology, Psychometrics, and Methodology*. Springer, Berlin.

[20] Doignon, J.-P. and Falmagne, J.-C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, 23:283–303.

[21] Doignon, J. P. and Falmagne, J. C. (1999). *Knowledge spaces*. Springer-Verlag, Berlin.

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

[23] Dubois, D. and Prade, H. (1992). Putting rough sets and fuzzy sets together. In [116], pages 203–232.

[24] Ducamp, A. and Falmagne, J. (1969). Composite measurement. *J. Math. Psych.*, 6:359–390.

[25] Düntsch, I. (1997). A logic for rough sets. *Theoretical Computer Science*, 179(1-2):427–436.

[26] Düntsch, I. and Gediga, G. (1995a). Rough set dependency analysis in evaluation studies: An application in the study of repeated heart attacks. *Informatics Research Reports*, 10:25–30.

[27] Düntsch, I. and Gediga, G. (1995b). Skills and knowledge structures. *British J. Math. Statist. Psych.*, 48:9–27.

[28] Düntsch, I. and Gediga, G. (1997a). Algebraic aspects of attribute dependencies in information systems. *Fundamenta Informaticae*, 29:119–133.

[29] Düntsch, I. and Gediga, G. (1997b). The rough set engine GROBIAN. In [120], pages 613–618.

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

[31] Düntsch, I. and Gediga, G. (1998a). Knowledge structures and their applications in CALL systems. In Jager, S., Nerbonne, J., and van Essen, A., editors, *Language teaching and language technology*, pages 177–186. Swets & Zeitlinger.

[32] Düntsch, I. and Gediga, G. (1998b). Simple data filtering in rough set systems. *International Journal of Approximate Reasoning*, 18(1–2):93–106.

[33] Düntsch, I. and Gediga, G. (1998c). Uncertainty measures of rough set prediction. *Artificial Intelligence*, 106(1):77–107.

[34] Düntsch, I. and Gediga, G. (2000a). *A concise introduction into the foundations of building intelligent artifacts*. Methodos Publishers, Bangor. To appear.

[35] Düntsch, I. and Gediga, G. (2000b). ROUGHIAN – Rough Information Analysis. *International Journal of Intelligent Systems*. To appear.

[36] Düntsch, I. and Gediga, G. (2000c). *Sets, relations, functions*, volume 1 of *Methodos Primers*. Methodos Publishers, Bangor.

[37] Düntsch, I., Gediga, G., and Orłowska, E. (1999). Relational attribute systems. Submitted for publication.

[38] Düntsch, I., McCaull, W., and Orłowska, E. (2000). Structures with many-valued information and their relational proof theory. In *Proc. of 30th IEEE International Symposium on Multiple-Valued Logic*. To appear.

[39] Egenhofer, M. (1994). Deriving the composition of binary topological relations. *Journal of Visual Languages and Computing*, 5:133–149.

[40] Egenhofer, M. and Franzosa, R. (1991). Point–set topological spatial relations. *International Journal of Geographic Information Systems*, 5(2):161–174.

[41] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *Artificial Intelligence Magazine*, 17:37–54.

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

[43] Gediga, G. (1998). *Skalierung*. Lit Verlag, Münster.

[44] Gediga, G. and Düntsch, I. (1999a). Maximum consistency of incomplete data via non-invasive imputation. Submitted for publication.

[45] Gediga, G. and Düntsch, I. (1999b). Probabilistic granule analysis. Draft paper.

[46] Gediga, G., Düntsch, I., and Ievers, M. (2000). Skill set analysis in knowledge struc-

tures. Preprint.

[47] Gibson, J. J. (1986). *The ecological approach to visual perception*. Lawrence Erlbaum, Hillsdale, 2 edition.

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

[49] Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. (1997). Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery*, 1:11–28.

[50] Graham, J. W., Hofer, S. M., and Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. In Collins, L. M. and Seitz, L., editors, *Advances in Data Analysis for Prevention Intervention Research*, Washington. NIDA Research Monograph. Series 142.

[51] Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9:139–150.

[52] Guttman, L. (1950). The basis for scalogram analysis. In [118], pages 60–90.

[53] Hand, D. J. (1994). Deconstructing statistical questions. *J. Roy. Statist. Soc. Ser. A*, 157:317–356.

[54] Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63–91.

[55] Huebener, J. (1972). Complementation in the lattice of regular topologies. *Pacific J. Math.*, 41:139–149.

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

[57] Jaynes, E. T. (1996). Probability Theory: The Logic of Science. Fragmentary edition of March 1996, `http://www.math.albany.edu:8008/JaynesBook.html`.

[58] Krusińska, E., Babic, A., Słowiński, R., and Stefanowski, J. (1992). Comparison of the rough sets approach and probabilistic data analysis techniques on a common set of medical data. In [116], pages 251–265.

[59] Kryszkiewicz, M. (1998). Properties of incomplete information systems in the framework of rough sets. In [93], pages 422–450.

[60] Lazarsfeld, P. F. (1968). *Latent structure analysis*. Houghton Mifflin, Boston.

[61] Lehmann, F. and Cohn, A. G. (1994). The EGG/YOLK reliability hierarchy: Semantic data integration using sorts with prototypes. In *Proc. Conf. on Information Knowledge Management*, pages 272–279. ACM Press.

[62] Li, M. and Vitányi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications*. Graduate Texts in Computer Science. Springer-Verlag, New York, 2 edition.

[63] Liau, C.-J. (1996). An algebraic formalization of the relationship between evidential structures and data tables. *Fundamenta Informaticae*, 27:57–76.

[64] Lin, T. Y. (1992). Topological and fuzzy rough sets. In [116], pages 287–304.

[65] Lin, T. Y. and Cercone, N., editors (1997). *Rough sets and data mining*, Dordrecht. Kluwer.

[66] Lin, T. Y., Liu, Q., and Yao, Y. Y. (1994). Logic systems for approximate reasoning: via rough sets and topology. In Raś, Z. W. and Zemankova, M., editors, *Methodologies for intelligent systems*, pages 64–74. Springer, Berlin.

[67] Lipski, W. (1976). Informational systems with incomplete information. In Michaelson, S. and Milner, R., editors, *Third International Colloquium on Automata, Languages and Programming*, pages 120–130, University of Edinburgh. Edinburgh University Press.

[68] Lipski, W. (1979). On semantic issues connected with incomplete information data bases. *ACM Trans. Database Systems*, 4(3):262–296.

[69] Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[70] Manly, B. F. J. (1997). *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. Chapman and Hall, London.

[71] Meng, X. L. (1995). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 10:538–573.

[72] Michie, D., Spiegelhalter, D. J., and Taylor, C. C. (1994). *Machine learning, neural and statistical classification*. Ellis Horwood.

[73] Muggleton, S., editor (1992). *Inductive Logic Programming*. Academic Press.

[74] Neyman, J. and Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society (A)*, 231:289–337.

[75] Nguyen, H. (1998a). From optimal hyperplanes to optimal decision trees. *Fundamenta Informaticae*, 34:145–174.

[76] Nguyen, H. S. (1998b). From optimal hyperplanes to otimal decision trees. *Fundamenta Informaticae*, 34:145–174.

[77] Nguyen, H. S. and Nguyen, S. H. (1996). Some efficient algorithms for rough set methods. In *Proc. of IPMU'96*, pages 1451–1456.

[78] Nguyen, H. S. and Nguyen, S. H. (1998a). Discretization methods in data mining. In [93], pages 451–482.

[79] Nguyen, H. S. and Nguyen, S. H. (1998b). Pattern extraction from data. *Fundamenta Informaticae*, 34:129–144.

[80] Nguyen, H. S., Nguyen, S. H., and Skowron, A. (1996). Searching for features defined by hyperplanes. In Ras, Z. W. and Michalewicz, M., editors, *ISMIS-96, Ninth International Symposium on Methodologies for Intelligent Systems*, volume 1079 of *Lecture Notes in Artificial Intelligence*, pages 366–375, Berlin. Springer–Verlag.

[81] Nguyen, H. S. and Skowron, A. (1996). Quantization of real value attributes: Rough set

and Boolean reasoning approach. *Bulletin of International Rough Set Society*, 1(1):5–16.

[82] Nguyen, S. H., Skowron, A., and Synak, P. (1998). Discovery of data patterns with applications to decomposition and classification problems. In [94], pages 55–97.

[83] Novotný, M. (1997). Dependence spaces of information systems. In [86], pages 193–246.

[84] Orłowska, E. (1984). Modal logics in the theory of information systems. *Zeitschr. f. Math. Logik und Grundlagen der Math.*, 30:213–222.

[85] Orłowska, E. (1995). Information algebras. In *Proceedings of AMAST 95*, volume 639 of *Lecture Notes in Computer Science*. Springer–Verlag.

[86] Orłowska, E., editor (1997). *Incomplete Information – Rough Set Analysis*. Physica – Verlag, Heidelberg.

[87] Orłowska, E. and Pawlak, Z. (1987). Representation of nondeterministic information. *Theoretical Computer Science*, 29:27–39.

[88] Pagliani, P. (1997). Rough sets theory and logic-algebraic structures. In [86], pages 109–190.

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

[90] Pawlak, Z. and Skowron, A. (1994). Rough membership functions. In Zadeh, L. and Kacprzyk, J., editors, *Fuzzy logic for the management of uncertainty*, pages 251–271, New York. Wiley.

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

[92] Peters, J. F. (1998). Time and clock information systems: Concepts and roughly fuzzy Petri net models. In [94], pages 385–417.

[93] Polkowski, L. and Skowron, A., editors (1998a). *Rough sets in knowledge discovery, Vol. 1*. Physica–Verlag, Heidelberg.

[94] Polkowski, L. and Skowron, A., editors (1998b). *Rough sets in knowledge discovery, Vol. 2*. Physica–Verlag, Heidelberg.

[95] Pomykala, J. and Pomykala, J. A. (1988). The Stone algebra of rough sets. *Bull. Polish Acad. Sci. Math.*, 36:495–508.

[96] Prediger, S. (1997). Symbolic objects in formal concept analysis. Preprint 1923, Fachbereich Mathematik, Technische Hohschule Darmstadt.

[97] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1:81–106.

[98] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo.

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

[100] Rauszer, C. (1991). Reducts in information systems. *Fundamenta Informaticae*, 15:1–

12.

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

[102] Rissanen, J. (1985). Minimum – description – length principle. In Kotz, S. and Johnson, N. L., editors, *Encyclopedia of Statistical Sciences*, pages 523–527, New York. Wiley.

[103] Rogner, J., Bartram, M., Hardinghaus, W., Lehr, D., and Wirth, A. (1994). Depressiv getönte Krankheitsbewältigung bei Herzinfarktpatienten – Zusammenhänge mit dem längerfristigen Krankheitsverlauf und Veränderbarkeit durch eine Gruppentherapie auf indirekt–suggestiver Grundlage. In Schüßler, G. and Leibing, E., editors, *Coping. Verlaufs– und Therapiestudien chronischer Krankheit*, pages 95–109. Hogrefe, Göttingen.

[104] Rubin, D. B. (1987). *Multiple Imputations for Nonresponse in Surveys*. Wiley, New York.

[105] Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91:473–489.

[106] Schafer, J. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall.

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

[108] Seligman, M. E. P. and Schwartz, G. E. (1972). *Biological Boundaries of Learning*. Appelton-Century-Crofts, New York.

[109] Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.

[110] Shannon, C. E. and Weaver, W. (1963). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

[111] Skowron, A. (1990). The rough sets theory and evidence theory. *Fundamenta Informaticae*, 13:245–262.

[112] Skowron, A. (1995). Extracting laws from decision tables - a rough set approach. *Computational Intelligence*, 11:371–388.

[113] Skowron, A. and Grzymała-Busse, J. W. (1993). From rough set theory to evidence theory. In Yager, R., Fedrizzi, M., and Kasprzyk, J., editors, *Advances in the Dempster–Shafer Theory of Evidence*, pages 193–236. Wiley, New York.

[114] Skowron, A. and Polkowski, L. (1997). Synthesis of decision systems from data tables. In [65], pages 259–299.

[115] Skowron, A. and Rauszer, C. (1992). The discernibility matrices and functions in information systems. In [116], pages 331–362.

[116] Słowiński, R., editor (1992). *Intelligent decision support: Handbook of applications and advances of rough set theory*, volume 11 of *System Theory, Knowledge Engineering and Problem Solving*. Kluwer, Dordrecht.

[117] Słowinski, R. and Stefanowski, J. (1996). Rough set reasoning about uncertain data.

Fundamenta Informaticae, 27:229–243.

[118] Stouffer, S., Guttman, L., Suchman, E., Lazarsfeld, P., Star, S., and Clausen, J., editors (1950). *Measurement and Prediction*. Princeton University Press.

[119] Suchman, E. (1950). The utility of scalogram analysis. In [118], pages 122–171.

[120] Sydow, A., editor (1997). *Proc. 15th IMACS World Congress*, volume 4, Berlin. Wissenschaft und Technik Verlag.

[121] Tanaka, M., Ishibuchi, H., and Shigenaga, T. (1992). Fuzzy inference system based on rough sets and its application to medical diagnosis. In [116], pages 111–118.

[122] Tsumoto, S. and Tanaka, H. (1996). A common algebraic framework for empirical learning methods based on rough sets and matroid theory. *Fundamenta Informaticae*, 27(7):273–288.

[123] Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

[124] Wang, H., Düntsch, I., and Gediga, G. (2000). Classificatory filtering in decision systems. *International Journal of Approximate Reasoning*, 23:111–136.

[125] Wille, R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In Rival, I., editor, *Ordered sets*, volume 83 of *NATO Advanced Studies Institute*, pages 445–470. Reidel, Dordrecht.

[126] Wong, S. K. M., Ziarko, W., and Ye, R. L. (1986). Comparison of rough–set and statistical methods in inductive learning. *International Journal of Man-Machine Studies*, 24:53–72.

[127] Wroblewski, J. (1995). Finding minimal reducts using genetic algorithms. ICS Research Report 16, Warsaw University of Technology.

[128] Yao, Y. (1997). Combination of rough and fuzzy sets based on α-level sets. In [65], pages 301–321.

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

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

Citation index

Index