Rough set data analysis in the KDD process

Ivo Düntsch[*]
Faculty of Informatics
University of Ulster
Newtownabbey, N.Ireland
I.Duentsch@ulst.ac.uk

Günther Gediga[*]
FB Psychologie / Methodenlehre
Universität Osnabrück
Osnabrück, Germany
Guenther@Gediga.de

Hung Son Nguyen[*]
Institute of Mathematics
Warsaw University
Warsaw, Poland
son@mimuw.edu.pl

Abstract

In this paper, we position rough set data analysis (RSDA)in the KDD process, and emphasise its difference to other KDD methods, especially with respect to model assumptions. As an example we describe a method of non-invasive imputation.

1 Introduction

According to the widely accepted description of [5], 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.

Rough set data analysis (RSDA) was introduced by [14] as a method of rule-based data analysis. Since then, it has become increasingly popular with over 1100 publications until 1998 [15]. In this paper, we position RSDA in the various steps of the KDD process, and emphasise its difference to other KDD methods.

2 Data models and model assumptions

As the basis for our presentation we choose the data model of [7]: The centre of the modelling process is the researcher who chooses

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 *operationalisation*. The operationalisation mapping is often called *representation* in KDD, and the empirical model a *domain model*.
3. A *numerical system*[1] $\mathcal{M}$, 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.

[*] The ordering of authors is alphabetical, and equal authorship is implied.

[1] The term "numerical system" is historical and, besides traditional numerical systems, it comprises other systems in modern scaling theory, such as networks or knowledge spaces

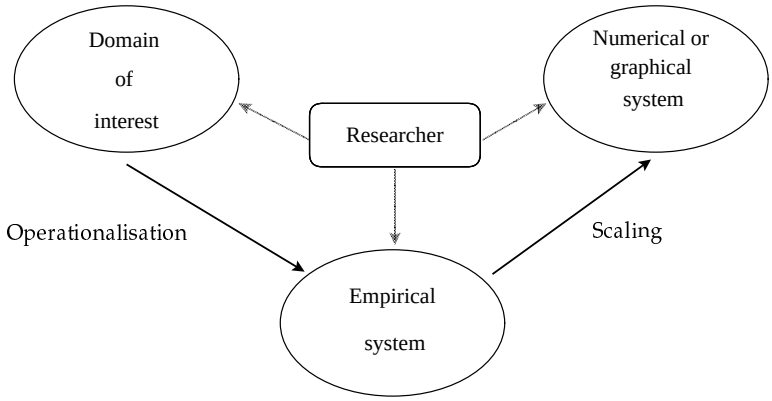

Figure 1: Data modelling

The choice of each of the parts of the model is a pragmatic decision by researchers, how they want to model 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. As a simple 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 are the test scores achieved by the students.

The operationalisation is to a large part subjective, and thus the first source of uncertainty ("model selection bias"): One question is whether the elements and relations of the empirical model $\mathcal{E}$ are representative for the objects of the domain $\mathcal{D}$ and the relations among them, another whether the choice of attributes covers the relevant aspects of $\mathcal{D}$. Operationalisation and an empirical model – along with the assumptions that go with them – are necessary in any type of data analysis, while a numerical system is not, as we shall show below.

All statistical and most KDD methods make external model assumptions, and thus reside on the level of the numerical model; 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 loglinear models; and
> 3. a class of models for mixed model and categorical data ..." [16].

It may be argued that this is particular to the "hard" statistical methods, and that a "soft" approach does not use numerical systems. However,

> *The principal components of soft computing are fuzzy logic, neural network theory, and probabilistic reasoning* [18],

and all of these "soft" methods require "hard" parameters outside the observed phenomena – membership degrees, prior probabilities, parameters for differential equations.

Model assumptions are not always spelled out "in toto", and thus, it is not clear to an observer on what basis and with which justification a particular method 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 are only a few number of features. Furthermore, the influence of the model assumptions on the results is not always taken into account when these are interpreted.

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 [10] 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.

3 Rough set data analysis

An instance of KDD, which is based on minimal model assumptions, and which admits ignorance when no conclusion can be drawn from the data at hand is rough set data analysis (RSDA) [14] which draws all its information from the given data. In other words, RSDA remains at the level of the empirical system; more formally, the numerical and the empirical system coincide, and the scaling is the identity function.

Operationalisation in RSDA is based on the viewpoint that objects are known (only) up to their description by attribute vectors: An *information system* $\mathcal{I}$ consists of a set U of objects, and a set Ω of attributes; the latter are functions $a : U \to V_a$ which assign to each object x a value $a(x)$ in the set V_a of values which x can take under a. If $\emptyset \neq Q \subseteq \Omega$, we denote the feature vector of x with respect to the attributes in Q by $\vec{x}^Q$.

This operationalisation by Object $\to$ Attribute data tables assumes the "nominal scale restriction" which postulates that each object has exactly one value of each attribute at a given time, and that the observation of this value is without error.

KDD 1 and KDD 2 are not part of RSDA, which assumes that enough care has been taken in these steps so that the operationalisation of data is sufficiently accurate to be a sound basis for analysis. Traditional RSDA takes subjectivity at this stage for granted as a fact of life, and starts with an empirical system given by such an operationalisation. An extended approach which includes parts of the operationalisation into the RSDA process is presented in [4].

Step KDD 3 consists of several mechanisms to solve problems with the data structure at hand: Missing data treatment is one of the issues which was not part of the classical RSDA; we will show below how this can be done, at least initially, on the level of the empirical model.

Noise reduction in the sense of a statistical KDD-procedure does not apply to RSDA, because RSDA has no concept of noise. Nevertheless, reducing complexity by removing dependency within the data set is a procedure which reduces "noise" as well. Therefore, it is sometimes used as a KDD 3-procedure within a statistical KDD-process. Indeed, RSDA can be regarded as a pre-processing device to recognise the potentially important variables for the construction of a neural network or multiple regression in order to reduce the problem of multi-co-linearity.

Data reduction (KDD 4) is a major feature of RSDA. Each $Q \subseteq \Omega$ determines an equivalence relation θ_Q on U by setting

$$x \equiv_{\theta_Q} y \iff (\forall a \in Q) a(x) = a(y).$$

The finest equivalence obtained in this way is θ_Ω. If $Q \subseteq \Omega$ and

$$\theta_Q = \theta_\Omega, \tag{3.1}$$

then the attributes in Q are sufficient to describe the classification induced by Ω, and thus, one can project Ω to Q. Note that only information by the data is used for attribute reduction. A set Q of attributes which is minimal with respect to (3.1) is called a *reduct of* $\mathcal{I}$.

Because of its model assumptions, RSDA can only use classification in KDD 5: The data table is enhanced by a decision attribute d with corresponding

equivalence relation θ_d, and the relationships between the various θ_Q and θ_d are considered. If $\theta_Q \subseteq \theta_d$, then the attributes in Q completely determine the decision attribute. Having said this, we may mention that one can also use procedures based on the non-invasive RSDA reasoning for unsupervised learning [1].

As mentioned in the previous Section, at least the hypothesis and model selection in the statistical sense of KDD 6 should have been taken care of after KDD 2. In RSDA there is no numerical system different from the operationalisation of the observed data, and there are no outside parameters to be chosen, nor is there a statistical model to be fitted. Within the practice of RSDA, however, there are numerous attempts to find an optimal RSDA model. In principle, there are three strands:

1. The classical main approach concentrates on finding (near) reducts and short rules via Boolean reasoning to explain the decision variable, a brief introduction to which can be found in [12]. Evaluation is done by simple probability measures such as approximation quality, rough membership and rough inclusion which are obtained from within the data to estimate and optimise the explanation and prediction quality of various attribute sets. These measures are conditional on the choice of attribute sets.

2. A second strand integrates the complexity of rules and the estimation of prediction error into a common unconditional measure by employing various entropy functions [3]. These methods are substantially different from the reduct-based ones, and the fine structure of their relation to the traditional methods and their consequences still needs to be fully understood. In this respect, RSDA has the same problems as other methods of data analysis.

3. Finally, hybrid approaches such as the integration of RSDA and fuzzy techniques (which, strictly speaking, lie outside the RSDA philosophy) are being used more widely [13].

Given a decision system $\langle \mathcal{I}, d \rangle$, the data mining step (KDD 7) in all RSDA variations consists of searching for deterministic and indeterministic rules offered by the data: Suppose that $\emptyset \neq Q \subseteq \Omega$, and that d is a decision attribute. Each class X of θ_Q gives us a rule in the following way: First, note that X is associated with a unique feature vector $\langle t_q \rangle_{q \in Q}$. Suppose that X intersects exactly the classes $Y_i, i \leq k$, of θ_d; each class Y_i corresponds to a unique value m_i of the decision attribute. We now have the rule

$$(\forall x \in U) \left[\bigwedge_{q \in Q} x^q = t_q \Rightarrow x^d = m_0 \vee \ldots \vee x^d = m_k \right].$$

If in the implication above we have $k = 0$, i.e. if X is contained in a class of θ_d, we call X (and the rule) d – *deterministic* (with respect to Q). We write $Q \to d$ for the collection of all these rules, and, with some abuse of language, call $Q \to d$ itself a rule. If each class of θ_Q is deterministic, then the classification according to d is completely determined by Q; in this case we call d *dependent on* Q. Observe that $Q \to d$ may contain indeterministic rules (expressed by a disjunction on the right hand side), contrary to the assertion in [5] that "a logical model is always deterministic".

KDD 8 is of special importance for RSDA. Even though, as a symbolic method, RSDA uses a only few parameters which need simple statistical estimation procedures, its results must be controlled using statistical testing procedures, in particular, when they are used for modelling and prediction of events. The problem here, of course, is not to allow subjective model assumptions to creep in through the back door, when testing procedures are applied. In earlier papers, two of us have described how RSDA can be supplemented by non-invasive procedures for testing rule significance by randomisation methods [1], data compression by classificatory filtering [2], and model selection using a suitable entropy and the principle of minimum description length [3].

As for KDD 9, experience has shown, that the non-invasive approach can be successfully used for KDD, at least as an indicator, before harder methods are applied, and we invite the reader to consult [15] for a recent overview of applications of RSDA.

4 Non-invasive imputation

Most KDD applications contain a large amount of missing data which have to be taken care of in KDD 3. As an example for a non-invasive procedure in the spirit of RSDA, we present a method of data imputation, which is described more fully in [6]. Other recent work in the RSDA context includes [9, 17, 8, 11], and for the statistical methods we recommend the excellent book [16].

In contrast to statistical imputation procedures, RSDA analysis offers no straightforward way to define loss functions or a likelihood function; these are based on statistical pre-assumptions, which are not given in rule based data analysis. 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, which means 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 an attribute vector $\vec{x}$ by presenting a list of possible values drawn from the set of all vectors $\vec{y}$ which do not contradict $\vec{x}$, i.e. they have the same entries wherever both are defined.

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

$$\mathrm{rel}_Q(x) = \{a \in Q : a(x) \text{ is defined}\}$$

be the set of *Q-relevant attributes for x*. Let $\vec{x}_Q$ be the feature vector of x with respect to the attributes in Q, i.e.

$$\vec{x}_Q = \langle a(x) : a \in Q \rangle.$$

Here, we assume that the attributes in Ω, and thus in Q, are suitably indexed. Each $\vec{x}_Q$ is called a *Q-granule*; if $Q = \Omega$ or Q is understood, we just speak of *granules*. Observe that

$$\vec{x}_Q = \vec{x}_{\mathrm{rel}_Q(x)}.$$

For each $\emptyset \neq Q \subseteq \Omega$ we define a relation $Q_{\mathcal{I}}$ on U by

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

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-indiscernible, 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.

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.

Let us 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 $a \in Q$ those entries which appear as entries $a(y)$ in the granules induced by a $y \in U$ which is consistent to x.

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

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 $\langle x, a \rangle$.

We 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) $\mathrm{rel}_\Omega(x)$. At some stage, there is a k such that $Q_{\mathcal{I}_k}(x) = Q_{\mathcal{I}_{k+1}}(x)$ for all $x \in U$, see [6] for a proof. 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.

As an example, consider the Table 1. After the first step we note that we cannot make the table complete, since there is no a_1-value for x_2 . 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.

Table 1: Imputation I

	a_1	a_2	a_3	a_4
x_1	5.0	4.0	3.0	2.0
x_2	?	3.0	2.0	1.0
x_3	2.0	?	4.0	5.0
x_4	?	2.0	?	3.0
x_5	2.0	2.0	?	?
x_6	5.0	4.0	3.0	2.0
x_7	3.0	2.0	1.0	1.0
x_8	3.0	2.0	5.0	?

Table 2: Imputation II

	a_1	a_2	a_3	a_4
x_1	5.0	4.0	3.0	2.0
x_2	?	3.0	2.0	1.0
x_3	2.0	{2.0}	4.0	5.0
x_4	$\{a_1(x_8)\}$	2.0	{5.0}	3.0
x_5	2.0	2.0	{4.0}	$\{a_4(x_3)\}$
x_6	5.0	4.0	3.0	2.0
x_7	3.0	2.0	1.0	1.0
x_8	3.0	2.0	5.0	{3.0}

Table 3: Imputation III

	a_1	a_2	a_3	a_4
x_1	5.0	4.0	3.0	2.0
x_2	?	3.0	2.0	1.0
x_3	2.0	{2.0}	4.0	5.0
x_4	{3.0}	2.0	{5.0}	3.0
x_5	2.0	2.0	{4.0}	{5.0}
x_6	5.0	4.0	3.0	2.0
x_7	3.0	2.0	1.0	1.0
x_8	3.0	2.0	5.0	{3.0}

A simulation study which shows the scope of the applicability of this method has been performed in [6]. 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.

References

[1] Ivo Düntsch and Günther Gediga, *Statistical evaluation of rough set dependency analysis*, International Journal of Human–Computer Studies **46** (1997), 589–604.

[2] _____, *Simple data filtering in rough set systems*, International Journal of Approximate Reasoning **18** (1998), no. 1–2, 93–106.

[3] _____, *Uncertainty measures of rough set prediction*, Artificial Intelligence **106** (1998), no. 1, 77–107.

[4] Ivo Düntsch, Günther Gediga, and Ewa Orłowska, *Relational attribute systems*, Submitted for publication, 1999.

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

[6] Günther Gediga and Ivo Düntsch, *Maximum consistency of incomplete data via non-invasive imputation*, Submitted for publication, 1999.

[7] G. Gigerenzer, *Messung und Modellbildung in der Psychologie*, Birkhäuser, Basel, 1981.

[8] Salvatore Greco, Benedetto Matarazzo, and Roman Słowinski, *Handling missing values in rough set analysis of multi-attribute and multi-criteria decision problems*, In Zong et al. [19], pp. 146–157.

[9] Jerzy W. Grzymala-Busse, Witold J. Grzymala-Busse, and Linda K. Goodwin, *A closest fit approach to missing attribute values in preterm birth data*, In Zong et al. [19], pp. 405–413.

[10] David J. Hand, *Deconstructing statistical questions*, J. Roy. Statist. Soc. Ser. A **157** (1994), 317–356.

[11] M. Kryszkiewicz, *Properties of incomplete information systems in the framework of rough sets*, Rough sets in knowledge discovery, Vol. 1 (Lech Polkowski and Andrzej Skowron, eds.), Physica–Verlag, Heidelberg, 1998, pp. 422–450.

[12] H.S. Nguyen and A. Skowron, *Boolean reasoning for feature extraction problems*, Proc. of the 10th International Symposium on Methodologies for Intelligent Systems (ISMIS'97) (Berlin) (Z.W. Ras and A. Skowron, eds.), Lecture Notes in Artificial Intelligence, vol. 1325, Springer–Verlag, 1997, pp. 117–126.

[13] S.K. Pal and A. Skowron (eds.), *Rough fuzzy hybridization*, Springer–Verlag, 1999.

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

[15] Lech Polkowski and Andrzej Skowron (eds.), *Rough sets in knowledge discovery, vol. 2*, Physica–Verlag, Heidelberg, 1998.

[16] J.L. Schafer, *Analysis of incomplete multivariate data*, Chapman & Hall, 1997.

[17] J. Stefanowski and A. Tsoukiàs, *On the extension of rough sets under incomplete information*, In Zong et al. [19], pp. 73–81.

[18] Lotfi A. Zadeh, *What is BISC?*, `http://http.cs.berkeley.edu/projects/Bisc/bisc.memo.html`, University of California, 1994.

[19] N. Zong, A. Skowron, and S. Ohsuga (eds.), *New directions in rough sets, data mining, and granular soft computing*, Lecture Notes in Artificial Intelligence, vol. 1711, Berlin, Springer–Verlag, 1999.