The Rough Set Engine GROBIAN

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
gg@Luce.Psycho.Uni-Osnabrueck.DE

and

Institut für semantische Informationsverarbeitung
Universität Osnabrück

Abstract

Rough set analysis is a non–numerical method of data analysis. Among other purposes, it can be used rule extraction, classification and to recognize dependencies of attributes. The paper describes a software system called GROBIAN[1] which performs rough set data analysis.

In addition to the "traditional" procedures of rough set analysis such as reduct analysis and rule generation, we have implemented statistical methods to evaluate the prediction quality of such analysis, and procedures to investigate the change the granularity within single attributes.

1 A brief introduction to the rough set model

The rough set model for data analysis has been developed by Z. Pawlak and his co–workers since the early 1980s (5). The original idea behind the model is the assumption that – in many situations – objects within a given population can only be distinguished up to a set of features. This may be due to subjective causes such as vagueness or to objective ones such as measuring errors or insufficient knowledge. In these situations, sets can only be roughly described via upper and lower approximations which are induced by the classes of an equivalence relation on the population.

The main thrust of applications of rough set analysis, however, is knowledge discovery in databases. Many of examples of real – life applications of rough set analysis can be found in (9).

Knowledge representation is done with information systems which are a tabular form of an OBJECT → ATTRIBUTE VALUE relationship. If knowledge of objects is represented by attributes and their values, it

[*] Equal authorship implied

[1] An acronym for the German expression "**Grob**mengen **I**nformations – **An**alysator". An adequate English translation is ROUGHIAN – **Rough**set **I**nformation **An**alyzer.

is important to find the relationships among the attributes. Knowing dependencies simplifies the original information system, reduces computational overhead and, at times, may indicate causal relationship.
If P is a set of attributes, and d is an attribute, we write $P \to d$ to express the dependency (also called a *rule*):

> *Whenever two objects agree on all attributes in P, then they agree on d.*

A *reduct* with respect to a (dependent) attribute d is a set Q of attributes, minimal with respect to the property that $Q \to d$. The intersection of all reducts is called the *core*. Determination of reducts and the core are at the heart of rough set data analysis.
Attributes within the core are necessary for the representation of the decision attribute, whereas an empty core signals a high substitution rate among the attributes.
Since it is based on equivalence relations, rough set analysis needs only internal information and does not rely on additional model assumptions as fuzzy set methods or probabilistic models do. In other words, instead of using external numbers or other additional parameters, rough set analysis utilizes only the structure of the given data and its inherent metrics. A widely used metric is the *approximation quality* of a partition with respect to another one; this is, roughly speaking, the ratio of the number of all correctly classified elements to the total number of objects. The *approximation quality of an attribute* is measured by the drop of the approximation quality when the attribute is removed from the set of (independent) attributes under consideration.
A more detailed overview of the rough set method can be found in (1) and a comprehensive presentation is the monograph (6).
We have used GROBIAN to (re–) analyze the results of two applications of rough set analysis: The duodenal ulcer data of (7), and the earthquake data of (11). They search for premonitory factors of earthquakes by emphasizing gas geochemistry.

2 GROBIAN: A short program description

GROBIAN uses a few functions of the RSL-library (RSL; 4), a collection of C-coded routines which cover a broad range of problems in rough set analysis. We have enhanced the library by several new functions, and have ported it to C^{++} classes which resulted in a more flexible, more stable and more transparent coding of the functions. The design of GROBIAN allows any procedure implemented in the RSL to be transformed easily into a WIN3.x/WIN95 user interface. In addition, we have implemented new theoretical results such as granularity analysis (2) and statistical validation of dependencies (3), both of which are described below.
After startup, GROBIAN needs a file containing the **R**ough **I**nformation **S**ystem (RIS) structure. The RIS file is in ASCII format, and contains all information needed to perform the data analysis There are several ways to produce such a file via the file menu.Whereas in the standard C-code one needs to look after the functions (e.g. permute) to know which information system is really in use, the classes in the C^{++}-code immediately show what is going on. Another advantage of using C^{++}-classes is the easiness to implement checks whether a function is applicable at all.

3 Major tasks of rough set analysis

3.1 Coding and filtering

Any data analysis starts with so called ”raw data”. These are unfiltered measurements of attributes within the domain under investigation. Rough set analysis – like other types of analysis – needs a preprocessing step whose result is ”data” suitable for further analysis. GROBIAN provides two basic preprocessing steps:

- Data conversion to convert attribute intervals of groups of attributes into coded equivalence classes and
- Data ranges to define suitable ranges of (groups of) attributes for further analysis.

The result of both procedures can be stored temporarily to perform an experimental filtering, or permanently to define a new information system where the filtered data is considered to be the new raw data. The ”View data” option of GROBIAN presents the raw data as well as their filtered counterpart, which enables the researcher to control the results of the filter procedures. Because equivalence relations are the only data type allowed in rough set analysis, processing starts with the construction of suitable equivalence classes induced by an attribute by identifying objects which have the same value with respect to this attribute.

3.2 Finding reducts and core

GROBIAN's search menu provides the items ”Reducts” and ”Core”. As an example we demonstrate how GROBIAN handles the analysis of the HSV data ((8)). There is only one dependent variable to define the classification, namely, the ”Visick_coding” attribute which signifies the healing success; all other attributes are viewed as independent variables. GROBIAN searches all reducts which describe the data up to a predefined ”minimal approximation quality”. If the ”minimal approximation quality” equals 1, the prediction of the dependent variable(s) must be perfect, while lower values indicate that we are willing to allow an error margin in the prediction. GROBIAN enables the researcher to perform reduct and core analysis within his information system using a systematic decrease of the minimal approximation quality. Because the core is defined as the intersection of all reducts of the information system, a reduct analysis can always determine the core as well. The analysis of the HSCA data shows that there is no perfect prediction of ”Visick_coding”, or in other words, $C(1.00) = \emptyset$. If we use a minimal approximation quality of 0.91, the corresponding core is empty $C(0.91) = \{\emptyset\}$. As a starting point for further analysis, we analyze the core $C(0.94) = \{3, 4, 5, 6, 8, 10\}$ and, because a core need not be a good candidate for a reduct with high approximation quality, reducts ”close to” this core such as the reduct $\{2, 3, 4, 5, 8, 9\}$ suggested by (7).

Since the first step of Rough Set Analysis is always the determination of the core, and since the computation of the core can be done more efficiently than using the intersection of all reducts, GROBIAN offers an extra entry for the core computation.

The core $C(1.00)$ of the earthquake information system is empty, which indicates that the data filtering process is not complete. The next section shows how one can proceed in such a situation.

3.3 Analysis of the empty core situation

An analysis of the earthquake data with the rough set approach discovers that there is an empty core for the dependent attribute ”Seismic activity”. One possibility to solve this problem is to eliminate some of the attribute values, and check whether the new system has a core. Because rough set analysis gives no hint which values should be eliminated, we have suggested a strategy how values of the (non-decision) attributes can be identified without losing information with respect to the decision attributes (2). The result of this analysis is a data driven filter procedure which produces additional coding rules within the attributes. We call the result of this procedure a *rough filter*. The GROBIAN analysis discovers that the granularity of some of the attributes is too high, and that some variables can be recoded without loss of information. GROBIAN proposes the following transformations:

- Radon11 should by filtered by $\{1, 2\} \to \{1\}$ and $\{3, 4, 5\} \to \{2\}$,
- Radon21 should by handled by the same filter procedure,
- Radon32 should by filtered by $\{1, 5\} \to \{1\}$ and $\{2, 3, 4\} \to \{2\}$,
- Radon62 should by filtered by $\{1, 4\} \to \{1\}$ and $\{2, 3, 5\} \to \{2\}$,
- Atmospheric pressure should by filtered by $\{1, 2, 3\} \to \{1\}$ and $\{4, 5\} \to \{2\}$.

Although rough filtering aims to lower the exchangeability of rules, there is no guarantee that the core of the filtered information system is not empty as well (as in the case of our example). Nevertheless, the researcher gains further insight into the structure of the data and may obtain ideas how to perform the coding and filtering procedures which may result in a more suitable information system.

3.4 Statistical analysis of a reduct

A high approximation quality is not a guarantee that the result of a rough set analysis is valid, and attention needs to be paid to the underlying statistical assumptions of the model. If, for example, rough set analysis discovers a rule of the form $P \to d$ in a database, and if the rule is based on only a few observations, its usefulness for prediction is arguable, and it needs to be statistically validated.

In (3) we have developed two simple procedures, both based on randomization techniques, which evaluate the validity of prediction based on the approximation quality of attributes of rough set dependency analysis. These are incorporated into the GROBIAN engine.

If a rule is based on only one observation, we call the result a *deterministic casual dependency*. Let $\gamma_{\mathcal{R}}$ be the probability distribution of all γ-values; given the same dependency structure, but randomized assignments of cases, the probability $p(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}})$ should be low to validate the dependency. If $p(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}}) > 0.05$, we call the result of the analysis a *casual dependency*. A similar idea leads to the definition of *conditional casual attributes* within rough set dependency analysis.

We have used GROBIAN to apply these considerations to the the duodenal ulcer data of (8).

The attribute ”Visick_coding” – which signifies the healing success – determines a partition of the set of patients into four classes. (8) shows that the attribute set R, consisting of ”Duration of disease” (**2**), ”Basic volume of gastric juice” (**5**), ”Complication” (**3**), ”Stimulated HCL concentration” (**8**), ”Basic

HCL concentration" (**4**), "Stimulated volume of gastric juice" (**4**) suffices to predict membership in a "Visick_coding" class. Based on the decline of the approximation quality they suggest that the attribute sets

$$A \overset{\text{def}}{=} \{3,4,5,8,9\}, \qquad B \overset{\text{def}}{=} \{2,3,5,8,9\}, \qquad C \overset{\text{def}}{=} \{2,3,4,5,9\}$$

are candidates for future research. Using the statistical analysis of the dependency of the decision attribute $\{11\}$ and the independent attributes $\{2,3,4,5,8,9\}$ the overall prediction success is found to be $\gamma = 0.795$. It is not casual, because the randomization analysis shows that $\hat{p}(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}}) = 0.013$. Furthermore, the attributes within R (Table 1) are checked using the technique of determining the conditional casualness. The underlined attribute in col. 1 is the attribute under study. The astonishing result: All attributes are conditional casual within R. This means that there are always only a few of the 122 observations which can be predicted additionally by introducing the attribute under study into the set. If we doubled all observations and analyzed the resulting set of 244 objects, no attribute would be conditional casual. Thus, one can argue that the number of observations in the duodenal ulcer information system is too small to predict the influences of the attributes within R. Six further cases are analyzed by leaving out one of the independent attributes. The results of the randomization tests based on 1000 simulations for each are given in Table 2. Col. 1 shows the attributes under consideration, col. 2 the observed approximation quality γ of this set, col. 3 the estimated position of γ in the distribution of the random matching assumption, and col. 4 the estimated 5% cutpoint in the distribution of γ assuming random matching. We observe that the prediction success of the attribute set $\{2,3,4,5,8,9\}$ is satisfactory. The proposed attribute sets B and C are not casual, whereas the proposed attribute set A is casual. Furthermore, the interesting attribute set indexed by $*$ was overlooked in the original study.

Table 1: GROBIAN re–analysis of the duodenal ulcer data, I

Attributes	decline of γ_{obs}	overall γ_{obs}	$\hat{p}(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}})$	$\gamma_{\mathcal{R}}(\alpha = 5\%)$	interpretation
$\underline{2}$,3,4,5,8,9	0.590	0.795	0.182	0.828	cond. casual
2,$\underline{3}$,4,5,8,9	0.516	0.795	0.099	0.811	cond. casual
2,3,$\underline{4}$,5,8,9	0.680	0.795	0.394	0.844	cond. casual
2,3,4,$\underline{5}$,8,9	0.549	0.795	0.107	0.811	cond. casual
2,3,4,5,$\underline{8}$,9	0.631	0.795	0.127	0.811	cond. casual
2,3,4,5,8,$\underline{9}$	0.648	0.795	0.310	0.844	cond. casual

Table 2: GROBIAN re–analysis of the duodenal ulcer data, II

Attributes	γ_{obs}	$\hat{p}(\gamma_{\mathcal{R}} \geq \gamma_{\text{obs}})$	$\gamma_{\mathcal{R}}(\alpha = 5\%)$	interpretation
2,3,4,5,8,9	0.795	0.013	0.770	not casual
·,3,4,5,8,9 (A)	0.590	0.153	0.623	casual
2,·,4,5,8,9	0.516	0.199	0.557	casual
2,3,·,5,8,9 (B)	0.680	0.018	0.656	not casual
2,3,4,·,8,9	0.549	0.084	0.556	casual
2,3,4,5,·,9 (*)	0.631	0.008	0.590	not casual
2,3,4,5,8,· (C)	0.648	0.011	0.607	not casual

3.5 Other options in GROBIAN

The *Analysis* menu shows two entries which we have not yet discussed: *Rule* and *Assignment*. Both types of analysis are coded in the Rough Set Library.

The *Rule* dialog computes a system of rules based upon a reduct of a dependency structure. The computation of the rules is quite resource demanding. Therefore, the RSL offers 5 different strategies (very fast, fast, middle, normal, the best) to compute a rule system.

The *Assignment* dialog find out how new observations should be classified given a fixed rule system within a dependency structure. If there is not a rule for a new observation, the best possible rule is determined and a change of the rule system can be invoked. For both types of analysis we refer to (10) for further details.

References

[1] Ivo Düntsch and Günther Gediga, *The rough set model for data analysis – introduction and overview*, Preprint, `http://www.infj.ulst.ac.uk/~cccz23/papers/roughmod.html`, August 1996.

[2] ______, *Simple data filtering in rough set systems*, International Journal of Approximate Reasoning (1997), to appear, `http://www.infj.ulst.ac.uk/~cccz23/papers/bininf.html`.

[3] ______, *Statistical evaluation of rough set dependency analysis*, International Journal of Human–Computer Studies (1997), To appear, `http://www.infj.ulst.ac.uk/~cccz23/papers/rougheva.html`.

[4] M. Gwaryś and J. Sienkiewicz, *Rough set library, Version 2.0*, User manual, Warsaw University of Technology, 1993.

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

[6] ______, *Rough sets: Theoretical aspects of reasoning about data*, Kluwer, Dordrecht, 1991.

[7] Krysztof Słowiński and Roman Słowiński, *Sensitivity analysis of rough classification*, Internat. J. Man–Mach. Stud. **32** (1990), 693–705.

[8] Krzysztof Słowiński, *Rough classification of HSV patients*, In *Intelligent decision support: Handbook of applications and advances of rough set theory* (9), pp. 77–94.

[9] Roman Słowiński, *Intelligent decision support: Handbook of applications and advances of rough set theory*, System Theory, Knowledge Engineering and Problem Solving, vol. 11, Kluwer, Dordrecht, 1992.

[10] Roman Słowiński and Jerzy Stefanowski, *'ROUGHDAS' and 'ROUGHCLASS' software implementations of the rough sets approach*, In *Intelligent decision support: Handbook of applications and advances of rough set theory* (9), pp. 445–456.

[11] Jacques Teghem and J.-M. Charlet, *Use of "rough sets" method to draw premonitory factors for earthquakes by emphasing gas geochemistry: the case of a low seismic activity context in Belgium*, In *Intelligent decision support: Handbook of applications and advances of rough set theory* (9), pp. 165–179.