Brock University

Department of Computer Science

Probabilistic granule analysis

Ivo Düntsch & Günther Gediga

Technical Report # CS-08-04
May 2008

Brock University
Department of Computer Science
St. Catharines, Ontario
Canada L2S 3A1
www.cosc.brocku.ca

Probabilistic granule analysis

Ivo Düntsch* & Günther Gediga

Department of Computer Science,
Brock University,
St. Catharines, Ontario, Canada, L2S 3A1
{duentsch,gediga}@brocku.ca

Abstract. We present a semi–parametric approach to evaluate the reliability of rules obtained from a rough set information system by replacing strict determinacy by predicting a random variable which is a mixture of latent probabilities obtained from repeated measurements of the decision variable. It is demonstrated that the algorithm may be successfully used for unsupervised learning.

1 Introduction

A simple and widely used form of data operationalization is the

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

relationship, where each object is described by its values with respect to properties chosen from a defined set Ω of features, and which is usually represented as a data table.

Rough set data analysis (RSDA), introduced in the early 1980s [1] uses the simple observation that each occurring feature vector determines a unique sets of objects – namely, all those objects which have these features – to construct rule systems on the basis of the granularity given by observed data; furthermore, feature reduction – a major issue in data analysis – can be achieved within these systems.

Although RSDA uses a only few parameters which need simple statistical *estimation* procedures, its results should be controlled using statistical *testing* procedures, in particular, when the method is used for modeling and prediction of events. If the claim of RSDA to be a fully fledged instrument for data analysis and prediction is to hold, the following issues must be addressed:

1. Significance of rules,
2. Model selection in case of competing rules,
3. Unreliability of measurements.

* Ivo Düntsch gratefully acknowledges support from the Natural Sciences and Engineering Research Council of Canada.

In earlier work, we have developed a procedure to determine the statistical significance of rough set rules based on randomization methods, and a method of model selection which combines the principle of indifference with the maximum entropy principle [2, 3]. The results support the view that a rule based method of data analysis does not, in principle, perform worse than traditional numerical methods, even on continuous data. Indeed, the direct comparison of linear discriminant analysis with RSDA based procedures by [4] on the Iris data [5] shows that the classification capability of non–parametric RSDA is as good as the parametric statistical method.

Traditionally, RSDA has concentrated on finding deterministic rules for the description of dependencies among attributes based on the *nominal scale assumption*: Once a deterministic rule has been found from a data set, it is tacitly assumed to hold without any error. Thus, in some sense, the theory is driven by the empirical data. However, if a measurement error is assumed to be an immeasurable part of the data, the pure RSDA approach may produce inaccurate results. On the one hand, even deterministic rules may be due to chance, and thus may not be reproducible; on the other hand, indeterministic information may be due to inaccurate measurement or the idiosyncrasies of a particular data set, thus possibly masking a theoretically deterministic situation.

In order to capture the uncertainty arising from measurement errors in a statistically sound way, we have proposed some 10 years ago the concept of *probabilistic information systems* [6], which may be viewed as an extension of the variable precision system of [7]. In the present contribution we take the opportunity to re–iterate this approach and extend it using well known procedures of classical test theory of psychometrics.

2 Definitions and notation

We assume familiarity with the basic notions of RSDA and will just briefly recall the necessary concepts. A *decision system* is a tuple $\mathscr{I} = \langle U, y, V_y, \Omega, (V_x)_{x\in\Omega}\rangle$, where

1. $U = \{a_1, \ldots, a_N\}$ is a finite set of objects.
2. $\Omega = \{x_1, \ldots, x_T\}$ is a finite set of mappings $x : U \to V_x$. Each x_i is called an *(independent) attribute*.
3. y is a mapping from U to V_y, called the *decision attribute*.
4. The functional dependency $\Omega \Rightarrow y$ holds, i.e.

$$\text{If } x(a) = x(b) \text{ for all } x \in \Omega, \text{ then } y(a) = y(b).$$

   This condition guarantees that the system is consistent.

If $\emptyset \neq X \subseteq \Omega$, we interpret X as a mapping $U \to \prod_{x\in X} V_x$ which assigns to each object $a \in U$ its feature vector $X(a) = x^X(a)$ with respect to the attributes in X; we will call $X(a)$ an *X–granule*; if $X = \Omega$, we will simply speak of a *granule*.

Each X – granule $X(a)$ can be understood as a piece of information about a set of objects in U given by the features in X, namely all those $b \in U$ for which $X(b) = X(a)$. The

equivalence relation on U induced by this condition is denoted by ψ_X, i.e. for $a,b \in U$,

$$a \equiv_{\psi_X} b \Longleftrightarrow X(a) = X(b). \tag{2.1}$$

Objects which are in the same class – and which are said to *belong to the same granule* – cannot be distinguished with the knowledge given by X.

Similarly, we define ψ_y on U by

$$a \equiv_{\psi_y} b \text{ iff } y(a) = y(b),$$

which gives us our target classification.

Suppose that $\emptyset \neq X \subseteq \Omega$. If a class M of ψ_X is contained totally within a class L of ψ_y, then $X(a)$ determines $y(b)$ for all $a,b \in M$. Such an M is called a *deterministic class of* ψ_X, and

$$\text{If } a,b \in M, \text{ then } y(a) = y(b) \tag{2.2}$$

is called a *deterministic X – rule*. Otherwise, M intersects exactly the classes $L_1,\ldots,L_k$ of ψ_y with associated values $l_1,\ldots,l_k$ in V_y, and we call

$$\text{If } a \in M, \text{ then } y(a) = l_1 \text{ or } \ldots \text{ or } y(a) = l_k \tag{2.3}$$

an *indeterministic X – rule*. The collection of all X – rules is denoted by $X \to y$, and – with some abuse of language – will sometimes be called a rule (of the information system).

The statistic

$$\gamma(X \to y) = \frac{|\bigcup\{M : M \text{ is a deterministic class of } X\}|}{|U|} \tag{2.4}$$

is called the *approximation quality of X* (with respect to y); it is the main indicator for the quality of feature reduction in RSDA [8]. It may be worthy of mention that this γ is only one of a whole family of such indicators, each of which may serve as useful approximation quality [9].

For our further discussion, we fix the following parameters:

- $U = \{a_1,\ldots,a_N\}$ is the set of objects.
- $\Omega = \{x_1,\ldots,x_T\}$ is the set of attributes.
- $G = \{g_1,\ldots,g_M\}$ is the set of granules. and T_i is the class of ψ_Ω associated with g_i, and $\nu(g_i) := |T_i|$.
- y is the decision attribute, $V_y = \{r_1,\ldots,r_D\}$ its set of values, and M_j is the class of ψ_y associated with r_j.
- For all $1 \leq i \leq M$, and $1 \leq j \leq D$, $\xi(i,j) := |T_i \cap M_j|$.

Table 1. A decision system

g_i	Ω		$y=r_1$	$y=r_2$	$\nu(g_i)$
	x_1	x_2	$\xi(i,1)$	$\xi(i,2)$	
g_1	0	1	5	1	6
g_2	1	0	2	8	10
	Σ		7	9	16

3 Probabilistic decision systems

In (deterministic) rule based systems a rule is either true or false, and a condition which holds for almost all cases will not contribute to the RSDA approximation quality. In the context of RSDA various remedies have been proposed which, instead of predicting hard decision values or intervals, regard the decision attribute as a random variable. For example, in standard rough set inclusion, deterministic rules for an indiscernibility class S and a decision class M are replaced by conditional probabilities which in the simplest case take the form

$$p(M|S) = \frac{|M \cap S|}{|M|}, \tag{3.1}$$

These considerations lead to probabilistic decision systems, sometimes called *Bayesian rough set models*, as structures of the from $\langle \mathscr{I}, Y \rangle$, where $\mathscr{I}$ is a classical RSDA information system, and $Y : G \times V_y \to [0,1]$ is a random variable; such structures have recently been an object of investigation, see e.g. [10–12]. Probabilistic rules have the form $x \to Y_j(x)$ which are pairs $\langle x, Y_j(x) \rangle$ where $x \in G$, and $Y_j(x)$ is the probability that x belongs to the decision class associated with r_j. Rough membership functions may be used to produce probabilistic decision systems such as the one shown in Table 1. There, we have $|U| = 16$, $\Omega = \{x_1, x_2\}$, and $V_y = \{r_1, r_2\}$, and both independent attributes are binary. Note that – up to indiscernibility – there are two granules, g_1, g_2. The rule system provided by the rough inclusion of (3.1) is obtained as

$$\langle 0,1 \rangle \to \{\langle 1, \tfrac{5}{6} \rangle, \langle 2, \tfrac{1}{6} \rangle\},$$
$$\langle 1,0 \rangle \to \{\langle 1, \tfrac{2}{10} \rangle, \langle 2, \tfrac{8}{10} \rangle\}.$$

Statistics such as rough inclusion are to some extent useful, however in principle they are subject to the same restrictions that the original problem poses, namely, that possible errors are not modeled within the system. In this sense, the problems persists, albeit with different, yet still "hard", boundaries for rule accuracy.

Computing the a–posteriori probability Y that a data element is assignable to a certain class requires distributional assumptions about the a priori distributions; estimation of priors is an inherent problem of Bayesian analysis. In most applications, however, it is not possible to observe the a priori distributions, and

> "A statistical problem is how to accurate are 'estimations' ... with regards to the *true* regions" [12].

If the observed rules are stable, then they should be the same for a different population. However, rules obtained from a second instance of a decision system may look quite different from the original one, even if the underlying structure is unchanged.

The well known test–retest paradigm of psychometrics offers a solution to the problem by using a distributional family such as a mixture of normal distributions or a mixture of triangle distributions, and a parameter fitting procedure given a learning data set. Since the true classification variable Y is principally unknown, we suppose that it is a mixture

$$Y = \sum_{1 \leq r \leq R} \omega_r Y_r, \tag{3.2}$$

of i.i.d. realizations Y_r based on an index R of unknown size and with unknown weights ω_r^i, for which $\sum_r \omega_r^i = 1$. It is safe to regard the Y_r as repeated measurements of the decision variable. In this way, the effects of an immeasurable measurement error are controlled and thus, the reliability of the rules can be tested in a statistically sound way.

The tasks now are

1. To estimate the best number R of replicas.
2. To estimate the parameters ω_r for each $1 \leq r \leq R$.

If we use the granules g_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 g_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 the class value of more than one granule; this will be indicated by a function

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

which maps the (set of indices of) the granules onto a set of (indices of) mixture components of Y.

In any estimation procedure, numerous models are produced, and one needs to decide which of these offers the best description of the data. Two standard procedures for model selection based on the size of the empirical data set and the number of parameters are the *Akaike Information Criterion* AIC [13] and Schwarz's *Bayesian Information Criterion* BIC [14]

$$\begin{aligned} AIC &= 2 \cdot (P - \ln(L(\max))) \\ BIC &= 2 \cdot \left(\frac{\ln(K)}{2} \cdot P - \ln(L(\max)) \right). \end{aligned}$$

Here, $L(\max)$ is the maximum likelihood of the data which may be obtained by optimizing the relevant binomial distribution by hill–climbing methods such as the EM

algorithm [15]. The lower AIC (and BIC respectively), the better the model. AIC and BIC are similar, but the penalty for parameters is higher in BIC then in AIC.

An algorithm to find the most appropriate model in our context using the AIC was first described in [6]. It 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

Table 2. Rule finding algorithm

$R := 0,\ \Delta(AIC) = 1.$
while $R \lesssim M$ and $\Delta(AIC) \gtrsim 0$ **do**
  $R := R + 1$
  Compute the best mapping $g : \{1,\ldots,M\} \to \{1,\ldots,R\}$ in terms
  of the product of the maximum likelihood of the Y replicas.
  Compute the number of parameters.
  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

the best mapping g is a combinatorial optimization 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

$$\hat{\pi}(i_k) = \frac{\sum_{t=1}^{s} \hat{\pi}_t(i_k)}{s}, \tag{3.3}$$

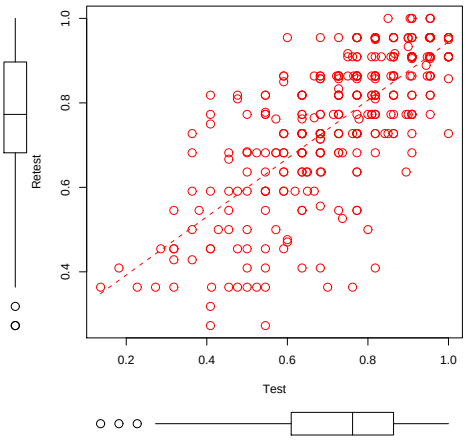
from which the likelihood can be found (Table 2).

4 Unsupervised learning and semi–parametric distribution estimates

In [6] we have shown that the AIC search algorithm may be used as a procedure for unsupervised learning. We have exemplified the procedure with Fisher's iris data [5] resulting in a classification quality of 85% which is quite acceptable for an unsupervised learning procedure. In the analysis, we have assumed that the attributes measure the same variable up to some scaling constants and that therefore the z – transformed attributes may be used as a basis for the analysis. Upon closer inspection, it turns out

the estimation of the mixture distributions is not a pure non–parametric procedure, because the standardization to z–values is, of course, a form of parametrization before the clustering procedure has started: The assumption "the attributes measure the same variables up to a some scaling constants" generates new variables which are assumed to be comparable on a standard scale.

Fig. 1. Test-Retest situation



To adjust the situation we may use the measureables to transform the data as suggested by the classical test theory of psychometrics: A test X may be retested or be used in a parallel form X' to estimate the reliability of the test, and the z–transformations of test and retest should be used to computed the reliability. Our original approach shows that a two–group representation combined with a non–parametric mixture of the distribution of the test values can be performed, and that there are no extra costs in terms of additional assumptions or parameters; in other words, it's simply for free. If the test–retest-paradigm is enhanced by further retesting, or if the test can be split additionally (e.g. by summing up odd and even items within the test to form test-values), it is easy to estimate more latent classes and their distribution estimates.

We shall illustrate the procedure with a typical example. An intelligence test applied to 331 subjects was tested and retested two weeks later using a parallel form of the test items (same solving principle, but different layout). Figure 1 shows the mean item solving probabilities of the subjects.

This procedure is routine part of the standardization of a psychometric test. Furthermore, test and retest are assumed to be identical in their expectation and variance. If these assumptions hold, the assumptions for searching the best-AIC-mapping to a decision attribute with two values ("solvers" and "non–solvers") holds as well.

Applying the algorithm we observe a clear cut optimum with two groups. Group 1 consists of 52,6% of the subsects showing a joint test–retest distribution given in Figure

2. This group of subjects shows a high probability to solve the test items (“solvers”). The group is rather homogeneous, because the correlation of test and retest value is very low.

Fig. 2. Test–Retest distribution group 1

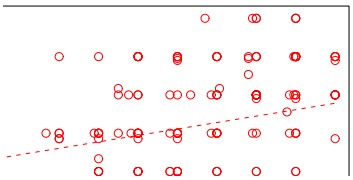
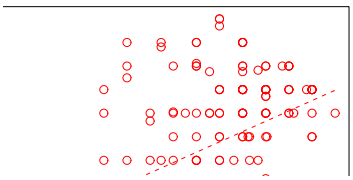


Fig. 3. Test–Retest distribution group 2

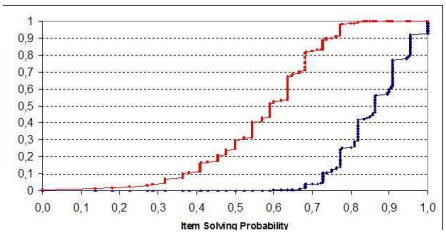


Group 2 consists of 47,4% of the subjects showing a joint test–retest distribution given in figure 3. This group has a much lower probability to solve the test items than the subjects in group 1 (“non–solvers”). Because the test–retest correlation is substantial, we have to argue that this group is not the final representation; owing to the restriction of

only two measurements, the best-AIC-mapping cannot squeeze out more groups from the data.

The cumulative distributions of the test values in the groups can now be used to classify the subjects (Figure 4).

Fig. 4. Cumulative distributions of test values



One can see, for example, that a subject showing a score of 0.6 is very likely to be a member of group 2, whereas a subject showing a score of 0.9 is member of group 1.

5 Conclusion and outlook

We have proposed a mixture model which enables traditional RSDA to handle possible measurement errors in the decision variable. The method makes only mild distributional assumptions which makes it well suited for the non–invasive approach of RSDA. In future work, we will extend the approach to predict unseen cases from partially known information and investigate estimations of semi–parametric mixture distributions and re–classification of latent groups in the context of RSDA. We will also apply our approach to estimate the reliability of data discretization procedures.

References

1. Pawlak, Z.: Rough sets. Internat. J. Comput. Inform. Sci. **11** (1982) 341–356
2. Düntsch, I., Gediga, G.: Statistical evaluation of rough set dependency analysis. International Journal of Human–Computer Studies **46** (1997) 589–604
3. Düntsch, I., Gediga, G.: Uncertainty measures of rough set prediction. Artificial Intelligence **106** (1998) 77–107
4. Browne, C., Düntsch, I., Gediga, G.: IRIS revisited: A comparison of discriminant and enhanced rough set data analysis. In Polkowski, L., Skowron, A., eds.: Rough sets in knowledge discovery, Vol. 2, Physica–Verlag (1998) 345–368
5. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7** (1936) 179–188

6. Gediga, G., Düntsch, I.: Statistical tools for rule based data analysis. In Komorowski, J., Düntsch, I., Skowron, A., eds.: Workshop on Synthesis of Intelligent Agent Systems from Experimental Data, ECAI'98. (1998)
7. Ziarko, W.: Variable precision rough set model. Journal of Computer and System Sciences **46** (1993)
8. Pawlak, Z.: Rough sets: Theoretical aspects of reasoning about data. Volume 9 of System Theory, Knowledge Engineering and Problem Solving. Kluwer, Dordrecht (1991)
9. Gediga, G., Düntsch, I.: Rough approximation quality revisited. Artificial Intelligence **132** (2001) 219–234
10. Pawlak, Z., Skowron, A.: Rough sets: Some extensions. Information Sciences **177** (2007) 28–40
11. Pawlak, Z.: A rough set view on Bayes' theorem. International Journal of Intelligent Systems **18** (2003) 487
12. Slezak, D., Ziarko, W.: The investigation of the Bayesian rough set model. International Journal of Approximate Reasoning **40** (2005) 81–91
13. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In Petrov, B.N., Cáski, F., eds.: Second International Symposium on Information Theory, Budapest, Akademiai Kaidó (1973) 267–281 Reprinted in *Breakthroughs in Statistics*, eds Kotz, S. & Johnson, N. L. (1992), volume I, pp. 599–624. New York: Springer.
14. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics **6** (1978) 461–464
15. Redner, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the EM algorithm. SIAM Review **26** (1984) 195–236