Brock University

Department of Computer Science

PRE and variable precision models in rough set data analysis

Ivo Düntsch and Günther Gediga
Technical Report # CS-13-06
April 2013

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

PRE and variable precision models in rough set data analysis*

Ivo Düntsch** and Günther Gediga

[1] Brock University, St. Catharines, Ontario, Canada, L2S 3A1, duentsch@brocku.ca
[2] Department of Psychology, Institut IV, Universität Münster, Fliednerstr. 21, Münster, Germany, gediga@uni-muenster.de

Abstract. We present a parameter free and monotonic alternative to the parametric variable precision model of rough set data analysis. The proposed model is based on the well known PRE index λ of Goodman and Kruskal. Using a weighted λ model it is possible to define a two dimensional space based on (Rough) sensitivity and (Rough) specificity, for which the monotonicity of sensitivity in a chain of sets is a nice feature of the model. As specificity is often monotone as well, the results of a rough set analysis can be displayed like a receiver operation curve (ROC) in statistics. Another aspect deals with the precision of the prediction of categories – normally measured by an index α in classical rough set data analysis. We offer a statistical theory for α and a modification of α which fits the needs of our proposed model. Furthermore, we show how expert knowledge can be integrated without losing the monotonic property of the index. Based on a weighted λ, we present a polynomial algorithm to determine an approximately optimal set of predicting attributes. Finally, we exhibit a connection to Bayesian analysis. We present several simulation studies for the presented concepts. The current paper is an extended version of [4].

1 Introduction

Rough sets were introduced by Z. Pawlak in the early 1980s [13] and have since become an established tool in information analysis and decision making. Given a finite set U and an equivalence relation θ on U the idea behind rough sets is that we know the world only up to the equivalence classes of θ. This leads to the following definition: Suppose that $X \subseteq U$. Then, the *lower approximation of* X is the set $\mathrm{Low}(X) = \{x \in U : \theta(x) \subseteq U\}$, and the *upper approximation of* X is the set $\mathrm{Upp}(X) = \{x \in U : \theta(x) \cap X \neq \emptyset\}$. Here, $\theta(x)$ is the equivalence class of x, i.e. $\theta(x) = \{y \in U : x\theta y\}$. A *rough set* now

* Ordering of authors is alphabetical, and equal authorship is implied

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

is a pair $\langle \mathrm{Low}(X), \mathrm{Upp}(X) \rangle$ for each $X \subseteq U$. A subset X of U is called *definable*, if $\mathrm{Low}(X) = \mathrm{Upp}(X)$. In this case, X is a union of classes of θ.

Rough set data analysis (RSDA) is an important tool in reasoning with uncertain information. Its basic data type is as follows: A *decision system* in the sense of rough sets is a tuple $\langle U, \Omega, (D_a)_{a\in\Omega}, (f_a)_{a\in\Omega}, d, D_d, f_d \rangle$, where

- U, Ω, D_a, D_d are nonempty finite sets. U is the set of objects, Ω is the set of (independent) attributes, and D_a is the domain of attribute a. The decision attribute is d, and D_d is its domain.
- For each $a \in \Omega$, $f_a : U \to D_a$ is a mapping; furthermore $f_d : U \to D_d$ is a mapping, called the *decision function*.

Since all sets under consideration are finite, an information system can be visualized as a matrix where the columns are labeled by the attributes and the rows correspond to feature vectors. An example from [20] is shown in Table 1.

Table 1. A decision system from [20]

U	a	b	c	d	U	a	b	c	d
1	1	0	0	1	12	0	1	1	1
2	1	0	0	1	13	0	1	1	2
3	1	1	1	1	14	1	1	0	2
4	0	1	1	1	15	1	1	0	2
5	0	1	1	1	16	1	1	0	2
6	0	1	1	1	17	1	1	0	2
7	0	1	1	1	18	1	1	0	3
8	0	1	1	1	19	1	0	0	3
9	0	1	1	1	20	1	0	0	3
10	0	1	1	1	21	1	0	0	3
11	0	1	1	1					

There, $U = \{1, \ldots, 21\}$ and $\Omega = \{a, b, c\}$. Each nonempty set Q of attributes leads to an equivalence relation $\equiv_Q$ on U in the following way: For all $x, y \in U$,

$$x \equiv_Q y \iff (\forall a \in Q)[f_a(x) = f_a(y)]. \tag{1.1}$$

According to the philosophy of rough sets, given a set Q of attributes, the elements of the universe U can only be distinguished up to the classes of $\equiv_Q$. A similar assumption holds for the decision classes of θ_d.

To continue the example of Table 1, the classes of θ_Ω are

$$\begin{aligned} X_1 &= \{1, 2, 19, 20, 21\}, & X_2 &= \{3\}, \\ X_3 &= \{4, \ldots, 13\}, & X_4 &= \{14, \ldots, 18\}, \end{aligned} \tag{1.2}$$

and the decision classes are

$$Y_1 = \{1,\ldots,12\},\ Y_2 = \{13,\ldots,17\},\ Y_3 = \{18,\ldots,21\}.$$

A class X of θ_Q is called *deterministic (with respect to d)* if there is a class Y of θ_d such that $X \subseteq Y$. In this case, all members of X have the same decision value. The set of all deterministic classes is denoted by $\mathrm{Pos}(Q,d)$.

The basic statistic used in RSDA is as follows:

$$\gamma(Q,d) = \frac{|\bigcup \mathrm{Pos}(Q,d)|}{|U|}. \tag{1.3}$$

$\gamma(Q,d)$ is called the *approximation quality of Q with respect to d*. If $\gamma(Q,d) = 1$, then each element of U can be correctly classified with the granularity given by Q. In the example, the only deterministic class is $\{3\}$, and thus, $\gamma(\Omega,d) = \frac{1}{21}$.

An important property of γ is monotony: If $Q \subseteq Q'$ then, $\gamma(Q,d) \leq \gamma(Q',d)$. In other words, increasing the granularity does not reduce the quality of classification.

In the sequel we exclude trivial cases and suppose that θ_Q and θ_d have more than one class.

2 The variable precision model

One problem of decision making using γ is the assumption of error free measurements, i.e. that the attribute functions f_a are exact, and even one error may reduce the approximation quality dramatically [6]. Therefore, it would be advantageous to have a procedure which allows some errors in order to result in a more stable prediction success.

A well established model which is less strict in terms of classification errors is the *variable precision rough set model* (VP – model) [20] with the following basic constructions: Let U be a finite universe, $X,Y \subseteq U$, and first define

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

Clearly, $c(X,Y) = 0$ if and only if $X = 0$ or $X \subseteq Y$, and $c(X,Y) = 1$ if and only if $X \neq \emptyset$ and $X \cap Y = \emptyset$. The *majority requirement* of the VP – model says that more than 50% of the elements in X should be in Y; this can be specified by an additional parameter β which is interpreted as an admissible classification error, where $0 \leq \beta < 0.5$. The *majority inclusion relation* $\overset{\beta}{\subseteq}$ (with respect to β) is now defined as

$$X \overset{\beta}{\subseteq} Y \Longleftrightarrow c(X,Y) \leq \beta. \tag{2.1}$$

Given a family of nonempty subsets $\mathscr{X} = \{X_1, ..., X_k\}$ of U and $Y \subseteq U$, the *lower approximation* $\underline{Y}_\beta$ *of* Y *given* $\mathscr{X}$ *and* β is defined as the union of all those X_i, which are in relation $X_i \overset{\beta}{\subseteq} Y$, in other words,

$$\underline{Y}_\beta = \bigcup \{X \in \mathscr{X} : c(X,Y) \leq \beta\} \tag{2.2}$$

The classical approximation quality $\gamma(Q,d)$ is now replaced by a three-parametric version which includes the external parameter β, namely,

$$\gamma(Q,d,\beta) = \frac{|\mathrm{Pos}(Q,d,\beta)|}{|U|}, \tag{2.3}$$

where $\mathrm{Pos}(Q,d,\beta)$ is the union of those equivalence classes X of θ_Q for which $X \overset{\beta}{\subseteq} Y$ for some decision class Y. Note that $\gamma(Q,d,0) = \gamma(Q,d)$. Continuing the example from the original paper ([20], p. 55), we obtain

$$\begin{aligned}
\gamma(\Omega,d,0) &= \frac{|X_2|}{|U|} &&= 1/21 \\
\gamma(\Omega,d,0.1) &= \frac{|X_2 \cup X_3|}{|U|} &&= 11/21 \\
\gamma(\Omega,d,0.2) &= \frac{|X_2 \cup X_3 \cup X_4|}{|U|} &&= 16/21 \\
\gamma(\Omega,d,0.4) &= \frac{|X_2 \cup X_3 \cup X_4 \cup X_1|}{|U|} &&= 21/21
\end{aligned}$$

Although the approach shows some nice properties, we think that care must be taken in at least three situations:

1. If we have a closer look at $\gamma(\Omega,d,0.1)$, we observe that, according to the table, object 13 is classified as being in class in Y_2, but with $\beta = 0.1$ it is assigned to the lower bound of Y_1. Intuitively, this assignment can be supported when the classification of the dependent attribute is assumed to be erroneous, and therefore, the observation is "moved" to a more plausible equivalence class due to approximation of the predicting variables. However, this may be problematic: Assume the decision classes arise from a medical diagnosis - why should an automatic device overrule the given diagnosis? Furthermore, the class changes are dependent on the actual predicting attributes in use, which may be problematic as well. This is evident if we assume for a moment that we want to predict d with only one class $X = U$. If we set $\beta = \frac{9}{21} < 0.5$, we observe that $U \overset{\frac{9}{21}}{\subseteq} Y_1$, resulting in $\gamma(\{U\},d,\frac{9}{21}) = 1$.
2. Classical reduct search is based on the monotone relation

$$P \subseteq Q \quad \text{implies} \quad \gamma(P,d) \leq \gamma(Q,d).$$

Unfortunately, the generalized $\gamma(Q,d,\beta)$ is not necessarily monotone [1]. As a counterexample, consider the information system shown in Table 2 which adds an additional independent attribute e to the system of Table 1. Setting $P = \{a,b,c\}$

Table 2. An enhanced decision system

U	a	b	c	e	d	U	a	b	c	e	d
1	1	0	0	0	1	12	0	1	1	1	1
2	1	0	0	0	1	13	0	1	1	1	2
3	1	1	1	0	1	14	1	1	0	0	2
4	0	1	1	0	1	15	1	1	0	0	2
5	0	1	1	0	1	16	1	1	0	0	2
6	0	1	1	0	1	17	1	1	0	0	2
7	0	1	1	0	1	18	1	1	0	0	3
8	0	1	1	0	1	19	1	0	0	0	3
9	0	1	1	1	1	20	1	0	0	0	3
10	0	1	1	1	1	21	1	0	0	0	3
11	0	1	1	1	1						

and $Q = \{a,b,c,e\}$, we observe that Q generates five classes for prediction. The three classes X_1, X_2, and X_4 are identical to those of the first example – given in (1.2) –, here given by P, but Q splits the class X_3 into the new classes $X_{3,0} = \{4...8\}$ and $X_{3,1} = \{9...13\}$. We now have

$$\gamma(Q,d,0.1) = \frac{|X_2 \cup X_{3,0}|}{|U|} = \frac{6}{21} < \gamma(P,d,0.1) = \frac{11}{21}.$$

The reason for this behavior is that $c(X_{3,1},Y) > 0.1$.

3. A third – perhaps minor – problem is the choice of $|U|$ as the denominator in $\gamma(Q,d,\beta)$. Using $|U|$ makes sense, when a no-knowledge-model cannot predict anything of d, and therefore any prediction success of Ω can be attributed to the predicting variables in Ω. But, as we have shown in the current section, there are situations in which a simple guessing model serves as a "perfect" model in terms of approximation quality.

Simulation 1 We conducted a simulation study based on an information system $\mathscr{I}$ with binary attributes A_1,A_2,A_3, and a decision attribute d with classes C_1,C_2,C_3 whose relative frequency is 0.6, 0.35, 0.05, respectively; there are 300 objects. Initially, for each $x \in U$ and $1 \leq i \leq 3$ we set

$$f_{A_i}(x) = \begin{cases} 1, & \text{if } f_d(x) \in C_i, \\ 0, & \text{otherwise.} \end{cases}$$

The information system $\mathscr{I}$ is assumed to be error free, i.e its reliability is 100 %. For each simulation, a certain percentage p of attribute values is changed to their opposite value, i.e. $f'_{A_i}(x) = 1 - f_{A_i}$ to obtain a different reliability $1 - p$. The expectation values of the prediction success for various values of β and reliabilities are shown in Table 3, based on 1000 simulations each. We observe that the prediction values are quite low for

Table 3. Simulation for the variable precision model

	Reliability		
β	.95	.90	0.85
0.00	0.5740	0.1553	0.0341
0.05	0.7315	0.4945	0.3754
0.10	0.7284	0.5045	0.3774
0.15	0.7356	0.4821	0.3734
0.20	0.7349	0.5146	0.3838
0.25	0.7349	0.4855	0.3731

$\beta = 0$ (no error) and are maximal already for small values of β. □

3 Contingency Tables and information systems

In this and the following sections we describe a formal connection of statistical and rough set data analysis. First of all, we need data structures which can be used for both types of analysis. It is helpful to observe that rough set data analysis is concept free because of its nominal scale assumption; in other words, only cardinalities of classes and intersection of classes are recorded. As $Q \subseteq \Omega$ and d induce partitions on U, say, $\mathscr{X}$ with classes X_j, $1 \leq j \leq J$, respectively, $\mathscr{Y}$ with classes Y_i, $1 \leq i \leq I$, it is straightforward to cross–classify the classes and list the cardinalities of the intersections $Y_i \cap X_j$ in a contingency table (see also [21]). As an example, the information system of Table 1 is shown as a contingency array in Table 4.

Table 4. Contingency table of the decision system of Table 1

	X_1	X_2	X_3	X_4	$n_{i\bullet}$
Y_1	2	**1**	**9**	0	**12**
Y_2	0	0	1	**4**	5
Y_3	**3**	0	0	1	4
$n_{\bullet j}$	5	1	10	5	21

The actual frequency of the occurrence, i.e. the cardinality of $Y_i \cap X_j$, is denoted by n_{ij} and the row and column sums by $n_{i\bullet}$ and $n_{\bullet j}$ respectively. The maximum of each column is shown in bold.

If a column X_j consists of only one non-zero entry, the corresponding set X_j is a deterministic class, i.e. it is totally contained in a decision class. In terms of classical rough set analysis, any column X_j which has at least two non-zero entries is not deterministic. The approximation quality $\gamma(Q,d)$ can now easily be derived by adding the frequencies n_{ij} in the columns with exactly one non-zero entry and dividing the sum by $|U|$. In the example we see that X_2 is the only column with exactly one nonzero entry, and $\gamma = \frac{1}{21}$. To be consistent with statistical notation, we will frequently speak of the classes of θ_Q as categories of the variable X and of the classes of θ_d as categories of the variable Y.

4 PRE measures and the Goodman-Kruskal λ

Statistical measures of prediction success – such as R^2 in multiple regression or η^2 in the analysis of variance – are often based on the comparison of the prediction success of a chosen model with the success of a simple zero model. In categorical data analysis the idea behind the *Proportional Reduction of Errors* (PRE) approach is to count the number of errors, i.e. events which should not be observed in terms of an assumed theory, and to compare the result with an "expected number of errors" , given a zero ("baseline") model [6, 8, 9]. If the number of expected errors is not zero, then

$$\text{PRE} = 1 - \frac{\text{number of observed errors}}{\text{number of expected errors}}$$

More formally, starting with a measure of error ε_0, the relative success of the model is defined by its proportional reduction of error in comparison to the baseline model,

$$\text{PRE} = 1 - \frac{\varepsilon_1}{\varepsilon_0}.$$

A very simple strategy in the analysis of categorical data is betting on the highest frequency; this strategy is normally used as the zero model benchmark ("baseline accuracy") in machine learning.

A simple modification which fits the contingency table was proposed by Goodman and Kruskal in the 1950s [7]. When no other information is given, it is reasonable to guess a decision category with highest frequency (such as Y_1 in Table 4). If the categories of X and the distribution of Y in each X_j are known, it makes sense to guess within each X_j some Y_i which shows the highest frequency, see also [10]. The PRE of knowing $\mathscr{X}$ instead of guessing is given by

$$\lambda = 1 - \frac{n - \sum_{j=1}^{J} \max_{i=1}^{I} n_{ij}}{n - \max_{i=1}^{I} n_{i\bullet}}. \tag{4.1}$$

Here, $n = |U|$. Note that $n - \max_{i=1}^{I} n_{i\bullet} \neq 0$, since we have assumed that θ_d has at least two classes. For our example we obtain

$$\lambda = 1 - \frac{21 - (3+1+9+4)}{21-12} = 1 - \frac{5}{9} = 0.444$$

We conclude that knowing $\mathscr{X}$ reduces the error of the pure guessing procedure by 44.4% in comparison to the baseline accuracy.

The λ-index is one of the most effective methods in ID3 [17], and a slightly modified approach in [10] – known as the *1R learning procedure* – was shown to be a quite effective tool as well [11].

5 Weighted λ

If we compare the set of classes $C(\beta)$ of θ_Q used to determine $\mathrm{Pos}(Q,d,\beta)$ in the VP-model, and the set of classes C used in the computation of λ , we observe that $C(\beta) \subseteq C$ for any value of $0 \leq \beta < 0.5$. The proof is simple: For every j more than 50% of the observations must be collected in one n_{ij}, and so these frequencies are the maximal frequency in column j.

The connection of λ and the approximation quality γ is straightforward: Whereas λ counts the maximum of each column j, γ counts this maximum only in the deterministic case if $n_{ij} = n_{\bullet j}$, i.e if exactly one entry in column j is nonzero.

Assume that we want to predict the decision attribute by one class only. In case that there is one attribute value of the decision attribute for which $n_{i\bullet} = |U| = n$ holds, we result in a situation in which the expected error is 0. Since this situation is of no interest for prediction, we should exclude it – the decision attribute is deterministic itself.

In all other cases, the decision attribute is indeterministic in the sense that there is no deterministic rule for prediction given no attributes; hence, in this case, the expected error is n. We observe that $|U| = n$ is a suitable denominator for γ.

As γ is a special case by filtering maximal categories by an additional condition, we define a *weighted* λ by

$$\lambda(w) = 1 - \frac{n - \sum_{j=1}^{J} (\max_{i=1}^{I} n_{ij}) \cdot w(j)}{n - (\max_{i=1}^{I} n_{i\bullet}) \cdot w(U)}. \tag{5.1}$$

where $w : \{1,\ldots,J\} \cup \{U\} \to [0,1]$ is a function weighting the maxima of the columns of the contingency table. In the cases we consider, w will be an indicator function taking its values from $\{0,1\}$.

Now we set

$$X_j \subseteq_w Y_i \Longleftrightarrow n_{ij} = \max_{k=1}^{I} n_{kj} \text{ and } w(j) > 0,$$

and define the lower approximation of Y_i by $\mathscr{X}$ with respect to w by

$$\mathrm{Low}_w(\mathscr{X}, Y_i) = Y_i \cap \bigcup \{X_j : X_j \subseteq_w Y_i\}.$$

Observe that $\mathrm{Low}_w(Y_i) \subseteq Y_i$, unlike in the lower approximation of the VP – model. For the upper approximation we choose the "classical" definition

$$\mathrm{Upp}(\mathscr{X}, Y_i) = \bigcup \{X_j : X_j \cap Y_i \neq \emptyset\}.$$

The *w-boundary* now is the set

$$\mathrm{Bnd}_w(\mathscr{X}, Y_i) = \mathrm{Upp}(\mathscr{X}, Y_i) \setminus \mathrm{Low}_w(\mathscr{X}, Y_i).$$

Unlike in the VP – model, elements of non–deterministic classes are not re–classified with respect to the decision attribute but are left in the boundary region.

We can now specify the error of the lower bound classification by

$$\mathrm{Err}_w(\mathscr{X}, Y_i) = \bigcup \{X_j \setminus Y_i : X_j \subseteq_w Y_i\}.$$

If we assume that errors are proportional to the number of entries in the contingency table – but independent of the joint distribution – it makes sense to count the absolute error $c_j = n_{\bullet j} - \max_{i=1}^{I} n_{ij}$ for every column j and compare it to some cutpoint C. This leads to the following definition:

$$w_{\mathrm{eq}}^{C}(j) = \begin{cases} 1, & \text{if } n_{\bullet j} - \max_{i=1}^{I} n_{ij} \leq C, \\ 0, & \text{otherwise} \end{cases}$$

and

$$w_{\mathrm{eq}}^{C}(U) = \begin{cases} 1, & \text{if } n - \max_{i=1}^{I} n_{i\bullet} \leq C, \\ 0, & \text{otherwise.} \end{cases}$$

respectively.

It is easy to see that $\lambda_{\text{eq}} = \gamma$ if $C = 0$, and $\lambda_{\text{eq}} = \lambda$ if $C = \infty$, i.e. if $\lambda_{\text{eq}} \equiv 1$. Furthermore, if $C \leq \max_{j=1}^{J}(n_{\bullet j} - \max_{i=1}^{I} n_{ij})$, then the denominator of $\lambda(w_{\text{eq}})$ is $|U|$.

In classical rough set theory, adding an independent attribute while keeping the same decision attribute will not decrease the approximation quality γ. The same holds for $\gamma_{w_{\text{eq}}}$:

Proposition 1. *Let $Q_a = Q \cup \{a\}$ and $\mathscr{X}_a$ be its associated partition. Then, $\gamma^C_{w_{\text{eq}}}(\mathscr{X}, \mathscr{Y}) \leq \gamma^C_{w_{\text{eq}}}(\mathscr{X}_a, \mathscr{Y})$.*

Proof. We assume w.l.o.g. that a takes only the two values $0, 1$ (see e.g. [5] for the binarization of attributes). Let Z_0, Z_1 be the classes of θ_a. The classes of θ_{Q_a} are the non–empty elements of $\{X_i \cap Z_0 : 1 \leq 1 \leq I\} \cup \{X_i \cap Z_1 : 1 \leq 1 \leq I\}$. Each n_{ij} is split into $n^0_{ij} = |X_i \cap Y_j \cap Z_0|$ and $n^1_{ij} = |X_i \cap Y_j \cap Z_1|$ with respective columns $j0$ and $j1$, and sums $n^0_{\bullet j}$ and $n^1_{\bullet j}$. Then, $n^0_{ij} + n^1_{ij} = n_{ij}$, $n^0_{\bullet j} + n^1_{\bullet j} = n_{\bullet j}$, and $\max_{i=1}^{I} n^0_{ij} + \max_{i=1}^{I} n^1_{ij} \geq \max_{i=1}^{I} n_{ij}$ by the triangle inequality. Thus, if $n_{\bullet j} - \max_{i=1}^{I} n_{ij} \leq C$, then

$$\begin{aligned} n^0_{\bullet j} - \max_{i=1}^{I} n^0_{ij} &\leq n^0_{\bullet j} - \max_{i=1}^{I} n^0_{ij} + n^1_{\bullet j} - \max_{i=1}^{I} n^1_{ij} \\ &= n^0_{\bullet j} + n^1_{\bullet j} - (\max_{i=1}^{I} n^0_{ij} + \max_{i=1}^{I} n^1_{ij}) \\ &= n_{\bullet j} - (\max_{i=1}^{I} n^0_{ij} + \max_{i=1}^{I} n^1_{ij}) \\ &\leq n_{\bullet j} - \max_{i=1}^{I} n_{ij} \\ &\leq C. \end{aligned}$$

Similarly,

$$n^1_{\bullet j} - \max_{i=1}^{I} n^1_{ij} \leq C.$$

Therefore, if $w_{\text{eq}}(j) = 1$, then $w_{\text{eq}}(j0) = w_{\text{eq}}(j1) = 1$.

Again by the triangle inequality, the sum of errors in the two $j0$ and $j1$ columns is no more than the error in the original column j. As the overall error is simply the sum of the errors per column, the proof is complete. □

Simulation 2 In order to compare the new model with the variable precision model, we assume the same setup as in Simulation 1, but use the equal weights λ PRE model instead. The results can be seen in Table 5. We note that the prediction quality increases with the cutpoint C, and from a certain C is larger than the asymptotic value of the variable precision model. The value depends on the chosen reliability. As a rule, the percentage of successful predictions is higher than that of the variable precision model. However, choosing $C = 1$ will not increase the prediction quality compared to the variable precision model when the reliability is low.

Table 5. Simulation for the λ model

	Reliability		
C	0.95	0.90	0.85
0	0.5740	0.1553	0.0341
1	0.7985	0.4798	0.1339
2	0.8508	0.5916	0.2813
3	0.8615	0.6707	0.4603

Table 6. Variable β-values to mimic the λ model in the example given a sample size of $n = 300$

Group sizes	C			
	0	1	2	3
0.60	0.000	0.006	0.011	0.017
0.35	0.000	0.010	0.019	0.029
0.05	0.000	0.067	0.133	0.200

We may consider the λ PRE model as a "variable precision model" when we assume that the β-boundaries vary in dependence of the group sizes. Table 6 shows the dependencies. □

6 Rough–sensitivity and Rough–specificity

Various other indices may be defined: Let $\mathscr{X}$ be the partition associated with θ_Q and Y_i be a decision class. In a slightly different meaning than in machine learning, we will use the terms *Rough-sensitivity* and *Rough-specificity* for the results of our analysis: If $\mathscr{Y}$ is the partition induced by the decision attribute, we consider

1. The *Rough-sensitivity* of the partition $\mathscr{X}$ with respect to the partition $\mathscr{Y}$

$$\gamma_w(\mathscr{X},\mathscr{Y}) = \frac{\sum_{Y_i\in\mathscr{Y}} |\operatorname{Low}_w(\mathscr{X},Y_i)|}{|U|}$$

2. The *Rough-specificity* of the partition $\mathscr{X}$ with respect to the partition $\mathscr{Y}$ is based on $\zeta_w(\mathscr{X},\mathscr{Y})$, which is defined by

$$\zeta_w(\mathscr{X},\mathscr{Y}) = \begin{cases} \frac{\sum_{Y_i\in\mathscr{Y}} |\operatorname{Err}_w(\mathscr{X},Y_i)|}{\sum_{Y_i\in\mathscr{Y}} |\operatorname{Bnd}_w(\mathscr{X},Y_i)|}, & \text{if } \sum_{Y_i\in\mathscr{Y}} |\operatorname{Bnd}_w(\mathscr{X},Y_i)| > 0 \\ 1, & \text{otherwise.} \end{cases}$$

   The *Rough-specificity* is defined by $1-\zeta_w(\mathscr{X},\mathscr{Y})$.

If $\mathscr{X}$ and $\mathscr{Y}$ are understood, we will just write γ_w and ζ_w or just ζ. The Rough-sensitivity tells us about the approximation of the set or partition, whereas ζ is an index which expresses the relative error of the classification procedure. Both indices are bounded by 0 and 1, and in most cases monotonically related (a counter example is discussed below). Rough-sensitivity reflects the relative precision of deterministic rules, which are true up to some specified error. It captures the rough set approximation quality γ in case w is defined as

$$w(j) = \begin{cases} 1, & \text{if } n_{\bullet j} = \max_{i=1}^{I} n_{ij} \\ 0, & \text{otherwise,} \end{cases}$$

and $w(U) = 0$.

Rough-specificity is a new concept: Whereas errors are addressed to the lower bound in the classic variable precision model, in our model an error is an instance of the boundary – it addresses those elements which are errors of the prediction rules in contrast to indeterministic elements, which cannot be predicted by prediction rules. The value of ζ tells us the relative magnitude of the "hard boundary" within the boundary. In other words, $1-\zeta$ (the Rough-specificity) is the relative number of elements of the boundary which may become deterministic, if we consider more attributes.

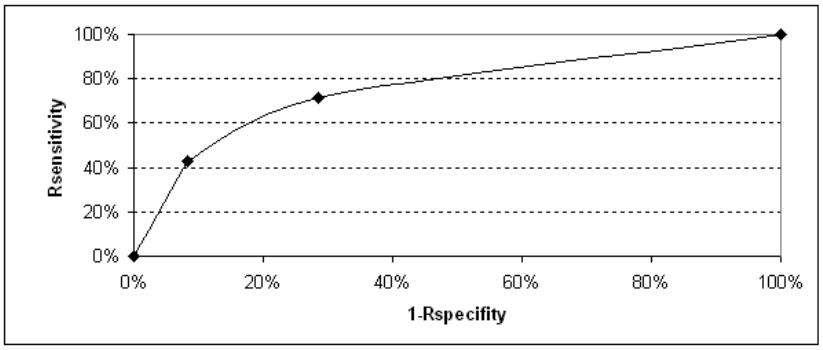
Using the data of Table 1 and the chain $(\emptyset, \{a\}, \{a,b\}, \{a,b,c\}, \{a,b,c,d\})$ of attributes and $C = 1$ we obtain the results shown in Table 7.

Table 7. Rough–Sensitivity and Rough–Specificity given the data of Table 1

Attribute Sets	{}	{a}	{a,b}	{a,b,c}	{a,b,c,d}
Lower Bound	{}	{4–12}	{4–12}	{3–12,14–17}	U
Error	{}	{13}	{13}	{13,18}	{}
Upper Bound	U	{1,2,3,13–21}	{1,2,3,13–21}	{1,2,13,18–21}	{}
Sensitivity	0.000	0.429	0.429	0.714	1,000
$\zeta = 1 -$ Specificity	0.000	0.083	0.083	0.286	1,000
Difference	0.000	0.345	0.345	0.429	1.000

A diagram of our results – which we may call a *rough receiver operation curve* (Rough–ROC) is depicted in Figure 1. Apart from the boundary values $0,1$, we find that the sensitivity γ_w is much higher than ζ_w. We call the difference $\gamma_w - \zeta_w$ the *Rough-Youden-index* (RY). In ROC analysis – the statistical counterpart to analyze sensitivity and specificity – the Youden index is a good heuristic to capture a good cut–point for prediction [2,19]. In rough set data analysis, the largest RY within a chain of attributes tells us a promising set for prediction – in case of the example the set $\{a,b,c\}$ seems to be good choice for prediction.

Fig. 1. A Rough–sensitivity / ζ = 1-Rough–specificity diagram



Note that if the decision attribute contains more than 2 classes, ζ – unlike γ – need not be monotonically increasing in case an error class changes to a deterministic class when adding a new independent attribute. As long as we do not split a deterministic class by adopting a further attribute, any new granule will not decrease the lower bound, and will not increase the number of elements in the boundary; therefore, the error will not decrease. Hence, ζ will not decrease, when we add a further attribute for prediction, and the deterministic classes are unchanged.

Table 8. A non–monotonic ζ

	X	X_0	X_1
Y_0	0	0	0
Y_1	3	3	0
Y_2	2	0	2

As an example of a non monotonic ζ, consider an information system with a granule $G = \langle\langle X, Y_0, 0\rangle, \langle X, Y_1, 3\rangle, \langle X, Y_2, 2\rangle\rangle$, see Table 8. G is deterministic, if we choose $C = 2$. Let n_e be the number of errors outside this granule, and $n_b > 0$ be the number of elements of the boundary outside this granule; then, $n_e < n_b$ and $\zeta = \frac{n_e+2}{n_b+2}$.

Suppose that a new attribute splits exactly G into G_0 and G_1 according to Table 8. Then, we obtain $\zeta = \frac{n_e}{n_b}$ as two elements are moved from the boundary to the lower bound. Since $n_b > 0$ and $n_e < n_b$, it follows that $\frac{n_e}{n_b} < \frac{n_e+2}{n_b+2}$.

There are various ways how to deal with the non-monotonicity of ζ:

1. Use a rule in the algorithm that prevent the split of deterministic classes.

2. Require that any deterministic class has to consist of more than C elements. Hence, using

$$\tilde{w}^C_{\mathrm{eq}}(j) = \begin{cases} 1, & \text{if } n_{\bullet j} - \max_{i=1}^{I} n_{ij} \leq C \text{ and } \max_{i=1}^{I} n_{ij} > C, \\ 0, & \text{otherwise} \end{cases}$$

is a weighting function for which ζ is monotone when classes are split. It is straightforward to show that γ is monotone as well when using $\tilde{w}$ as the weight function.
3. We leave everything unchanged. If ζ decreases when adding an attribute, we assume that this behavior is due to spurious deterministic rules, and consider this as another stopping rule for adopting more attributes.

7 Precision and its confidence bounds

In order to avoid technical problems, we assume that $\mathrm{Low}(Y_i) > 0$ for each class Y_i of the decision attribute, so that there is at least one rule which predicts membership in Y_i.

In classical rough set data analysis the accuracy of approximation α of Y_i is defined as

$$\alpha(Y_i) = \frac{|\mathrm{Low}(Y_i)|}{|\mathrm{Upp}(Y_i)|} = \frac{|\mathrm{Low}(Y_i)|}{|\mathrm{Low}(Y_i)| + |\mathrm{Bnd}(Y_i)|} = \frac{1}{1 + |\mathrm{Bnd}(Y_i)|/|\mathrm{Low}(Y_i)|}.$$

To approximate the standard error of the accuracy we use the Delta method. Broadly speaking, in a first step we linearize the fractions by taking the logarithms, secondly, we approximate the logarithm by the first order Taylor expansion, see e.g. [12] for details.

Letting $\pi_1 = |\mathrm{Low}(Y_i)|/|U|$ and $\pi = |\mathrm{Bnd}(Y_i)|/|U|$, we obtain

$$Var(\ln(\alpha(Y_i))) = \pi_1/|U| \cdot \left(\frac{\pi}{\pi_1^2 + \pi_1 \pi}\right)^2 + \pi/|U| \cdot \left(\frac{1}{\pi_1 + \pi}\right)^2$$

Example 1. Suppose that in a company with 500 employees there are 80 employees who are involved in accidents per year (Y_1). Of these, 70 can be predicted correctly, while of the other 420 cases (Y_0) 300 can be predicted correctly. As the decision attribute consists of two categories only, the number of subjects in the boundary is determined by $500 - 70 - 300 = 130$.

For the category Y_1 ("had accidents") we obtain the precision

$$\alpha(Y_1) = \frac{70}{70 + 130} = 0.35$$

The standard error of $\ln(\alpha(Y_1))$ is given by $\mathrm{SE}(\ln \alpha(Y_1))) = 0.120$ resulting in a 95% confidence interval $[0.277, 0.442]$ for $\alpha(Y_1)$.

For the category Y_0 ("no accidents") the precision is given by

$$\alpha(Y_0) = \frac{300}{300+130} = 0.698$$

The standard error of $\ln(\alpha(Y_0))$ is $\mathrm{SE}(\ln\alpha(Y_0))) = 0.103$ resulting in a 95% confidence interval $[0.570, 0.854]$ for $\alpha(Y_0)$. As both confidence intervals do not intersect, we conclude, that the precision of Y_0 is higher than the precision of Y_1. □

Assume now that there are some errors in the prediction. It makes no sense to count the errors for the prediction of the other categories as possible indeterministic rules for the category under study. Therefore we eliminate the errors from the other categories from the boundary by

$$|\,\mathrm{Bnd}_{\mathrm{corrected};Y_i}\,| = |\,\mathrm{Bnd}(Y_i)| - \sum_{k\neq i} |\,\mathrm{Err}(Y_k)|$$

and use the corrected boundary instead in the computation of α_c (a corrected α).

Example 2. We use the data from the preceding example, but assume additionally, that there were 30 errors to predict Y_1 and 100 errors to predict Y_0 due to application of our PRE model.

$$\alpha_c(Y_1) = \frac{70}{70+(130-100)} = 0.7 \qquad (95\%CI = [0.485, 1.000])$$
$$\alpha_c(Y_0) = \frac{300}{300+(130-30)} = 0.75 \qquad (95\%CI = [0.601, 0.936]).$$

Obviously, the estimated precision of Y_1 is enhanced dramatically. Note that $\alpha_c(Y_1)$ and $\alpha_c(Y_0)$ cannot be improved, as the boundary consists of error elements only. $\alpha_c(Y_1) = 0.7$ means in this case that 70% of the rules are deterministic and lead to the correct result Y_1, but 30% of the rules for Y_1 cannot be described in this way. □

In the variable precision model the error is moved to the lower bound. It is interesting to see how α_c looks like in this case. We select β large enough that the same errors occur as in our example using the PRE model. In this case we observe

Example 3.

$$\alpha_c(Y_1;\mathrm{VPRM}) = \frac{70+30}{70+30+(130-100-30)} = 1$$

and

$$\alpha_c(Y_0;\mathrm{VPRM}) = \frac{300+100}{300+100+(130-100-30)} = 1$$

In case of the VPRM, the α_c values signal a "perfect" precision of the model. □

8 Using additional expert knowledge

Weights given by experts or a priori probabilities of the outcomes Y_i ($1 \leq i \leq I$) are one of the simplest assumptions of additional knowledge which can be applied to a given situation: We let π_i ($1 \leq i \leq I$) be weights of the outcomes and w.l.o.g. we assume that $\sum_i \pi_i = 1$. Now, we obtain a weighted contingency table simply by defining $n^*_{ij} = n_{ij} \cdot \pi_i$ and use n^*_{ij} instead of n_{ij} of the original table.

Table 9. Weighted contingency table of the decision system of Table 1 using $\pi = \langle 0.5, 0.3, 0.2 \rangle$

	X_1	X_2	X_3	X_4	$n^*_{i\bullet}$
Y_1	**1**	**0.5**	**4.5**	0	**6**
Y_2	0	0	0.3	**1.2**	1.5
Y_3	0.6	0	0	0.2	0.8
$n^*_{\bullet j}$	1.6	.5	4.8	1.4	8.3

Using Table 9 and applying the bounds $E = 0, 0.2, 0.3, 0.6$ to compute $w^E_{\text{eq}}(j)$, we observe the approximation qualities shown in Table 10. We see that λ increases here as

Table 10. λ given various bounds

E	Formula (4.1)	Weighted λ
0.0	$1 - \frac{8.3-0.5}{8.3}$	0.060
0.2	$1 - \frac{8.3-0.5-1.2}{8.3}$	0.250
0.3	$1 - \frac{8.3-0.5-1.2-4.5}{8.3}$	0.747
0.6	$1 - \frac{8.3-0.5-1.2-4.5-1}{8.3}$	0.867

well as in case of the unweighted λ, but if we consider the weighted λ, the approximation qualities differ from those in the unweighted case. Furthermore, even the (approximate) deterministic class may change, if the weights differ largely: Note, that in case $E = 0.6$ we choose class X_1 as the (approximate) deterministic class, whereas X_3 would be chosen, if we use equal (or no) weights.

The algorithm given below and the monotonicity of λ given a split (or using an additional attribute) stay valid in case of introducing weights for the decision category as in the unweighted case. This holds because we have changed the entries of the table only – the structure of the table remains unchanged.

9 A simple decision tree algorithm based on rough sets

In order to find an algorithm for optimization, not only the Rough-sensitivity but also the Rough-specificity must be taken into account, and we have to find a function which reflects the status of the partitions in a suitable way. Numerical experiments show that neither the difference $\gamma_w - \zeta_w$ (the RY-index) nor the odds $\frac{\gamma_w}{\zeta_w}$ are appropriate for the evaluation of the partitions. The reason for this seems to be that the amount of deterministic classification, which is a function of $|U| \cdot \gamma_w$, as well as the amount of the probabilistic part of ζ_w are not taken into account.

Therefore we define an objective function based on entropy measures, which computes the fitness of the partition $\mathscr{X}$ on the basis of the difference of the coding complexity of the approximate deterministic and indeterministic classes, which is an instance of a mutual entropy [3]:

$$\mathbf{O}(\mathscr{Y}|\mathscr{X}) = -\gamma_w \ln(|U| \cdot \gamma_w) + \zeta_w \ln(\sum_{Y_i \in \mathscr{Y}} |\operatorname{Bnd}_w(\mathscr{X}, Y_i)| \cdot \zeta_w)$$

The algorithm proceeds as follows:

1. Set a cutpoint C for the algorithm.
2. Start with $Q = \emptyset$.
3. Add any attribute from $\Omega \setminus Q$ to Q. Compute $\mathbf{O}$ for the chosen cutpoint C
4. Choose a new attribute which shows the maximum in $\mathbf{O}$.
5. If the new maximum is less than or equal to the maximum of the preceding step, then stop.
   Otherwise add the new attribute to Q and proceed with step 2.

The time complexity of the algorithm is bounded by $\mathscr{O}(J^2)$ and it will find a partition $\mathscr{X}$ which shows a good approximation of Y with an error less than C.

Applying the algorithm to the decision system given in Table 1 and using $C = 1$ (we allow 1 error per column), results in the following steps:

`Step 1` $C = 1$
`Step 2.0` $Q = \emptyset$
`Step 3.0.a` Test attribute a

	X_1 (a=0)	$X_2(a=1)$
Y_1	**9**	3
Y_2	1	4
Y_3	0	4
$n_{\bullet j}$	10	11
$\mathbf{O}$	0.942	

`Step 3.0.b` Test attribute b

	X_1 (b=0)	$X_2(b=1)$
Y_1	2	10
Y_2	0	5
Y_3	3	1
$n_{\bullet j}$	6	16
O	0.000	

`Step 3.0.c` Test attribute c

	X_1 (b=0)	$X_2(b=1)$
Y_1	2	**10**
Y_2	4	1
Y_3	4	0
$n_{\bullet j}$	10	11
O	1.096	

`Step 4.0` Choose attribute c, because it is maximal in terms of **O**.

`Step 5.0` Iterate step 2.1

`Step 2.1` $Q = \{c\}$.

`Step 3.1.a` Test attribute a.

	X_1 $(c=0,a=1)$	X_2 $(c=1,a=0)$	X_3 $(c=1,a=1)$
Y_1	2	**9**	**1**
Y_2	4	1	0
Y_3	4	0	0
$n_{\bullet j}$	10	10	1
O	1.096		

`Step 3.1.b` Test attribute b

	X_1 $(c=0,b=0)$	X_2 $(c=0,b=1)$	X_3 $(c=1,b=1)$
Y_1	2	0	**10**
Y_2	0	**4**	1
Y_3	3	1	0
$n_{\bullet j}$	5	5	11
O	1.561		

`Step 4.1` Choose attribute b, because it is maximal in terms of $\mathbf{O}$.
`Step 5.2` Iterate step 2.2
`Step 2.2` $Q = \{b, c\}$.
`Step 3.2.a` Test attribute a.

	X_1	X_2	X_3	X_4
Y_1	2	**1**	**9**	0
Y_2	0	0	1	**4**
Y_3	3	0	0	1
$n_{\bullet j}$	5	1	10	5
O	1.561			

`Step 4.2` Stop, because $\mathbf{O}$ does not increase.

The attributes.$Q = \{b, c\}$ show the best behavior in terms of $\mathbf{O}$.

10 Bayesian considerations

As we introduced weights for the decision attribute, and since the weights may be interpreted as prior probabilities, it is worthwhile to find a connection to Bayesian posterior probabilities[3] . Choose some cutpoint C; we shall define a two dimensional strength function $s_C(i, j)$ $(1 \leq i \leq I, 1 \leq j \leq J)$, which reflects the knowledge given in column X_i to predict the category Y_j. As we use approximate deterministic classes as basis of our knowledge, the strength function is dependent on C as well.

First consider the case that the column X_j satisfies the condition

$$n_{\bullet j} - \max_{i=1}^{I} n_{ij} \leq C. \tag{10.1}$$

In that case there is one class with frequency $\max_{i=1}^{I} n_{ij}$ which is interpreted as the approximate deterministic class; all other frequencies are assumed as error. In this case we define $s_C(i, j) := \frac{n(i,j)}{n}$. This is simply the joint relative frequency $p(i, j)$ of the occurrence of $Y = Y_i$ and $X = X_j$. If the column X_j does not fulfill condition (10.1), we conclude that X_j cannot be used for approximation.

In this case no entry of column X_j contains (approximate) rough information about the decision attribute. Therefore we define $s_C(i, j) := 0$ for $1 \leq i \leq I$.

[3] For other views of Bayes' Theorem and its connection to rough sets see e.g. [14, 15, 18].

Now we define a conditional strength $s_C(X = X_j|Y = Y_i)$: If there is a least one $1 \leq j \leq J$ with $s_C(i,j) > 0$, then there is at least one (approximate) deterministic class X_j, which predicts Y_i. In this case we set

$$s_C(X = X_j|Y = Y_i) = \frac{s_C(i,j)}{\sum_{k=1}^{I} s_C(k,j)}. \tag{10.2}$$

Obviously, $s_C(X = X_j|Y = Y_i)$ reflects the relative strength of a rule predicting $Y = Y_i$.

If there is no (approximate) deterministic attribute $X = X_j$, which predicts $Y = Y_i$, the fraction $s_C(X = X_j|Y = Y_i)$ of (10.2) is undefined, since its denominator is 0. In this case – as we do not know the result –, we use $\underline{s}_C(X = X_j|Y = Y_i) = 0$ as the lower bound, and $\overline{s}_C(X = j|Y = Y_i) = 1$ as the upper bound.

Now we are able to define lower and upper posterior strength values by setting

$$\overline{s}_C(Y = Y_i|X = X_j) = \frac{\underline{s}_C(X = X_j|Y = Y_i)\pi_i}{\sum_r \underline{s}_C(X = X_j|Y = Y_r)\pi_r}$$

and

$$\underline{s}_C(Y = Y_i|X = X_j) = \frac{\underline{s}_C(X = X_j|Y = Y_i)\pi_i}{\sum_r \overline{s}_C(X = X_j|Y = Y_r)\pi_r}$$

If $C \geq n$, i.e. if the cutpoint is not less than the number of objects, then (10.1) is true for every X_j, and we observe that $s_C(Y = Y_i|X = X_j) = \frac{n(i,j)}{n} = p(i,j)$ for any i, j. Hence,

$$\overline{s}_n(Y = Y_i|X = X_j) = \underline{s}_n(Y = Y_i|X = X_j) = p(Y = Y_i|X = X_j)$$

and we result in the ordinary posterior probability of $Y = Y_i$ given $X = X_j$. Note, that although $\overline{s}_C \geq \underline{s}_C$ holds, the probability estimators $p(Y = Y_i|X = X_j)$ may be greater than $\overline{s}_C$ or smaller than $\underline{s}_C$. This is due the fact that the strength tables for different cutpoints C may looks quite different.

11 Summary and outlook

Whereas the variable precision model uses a parameter β to relax the strict inclusion requirement of the classical rough set model and to compute an approximation quality, a parameter free λ model based on proportional reduction of errors can be adapted to the rough set approach to data analysis. This index has the additional property that it is monotone in terms of attributes, i.e. if our knowledge of the world increases, so does the approximation quality. Weighted λ measures can be used to include expert or other context knowledge into the model, and an algorithm was given which approximates optimal sets of independent attributes and that is polynomial in the number of attributes. In the final section we showed how to explain Bayesian reasoning into this model. In

future work we shall compare our algorithm with other machine learning procedures and extend our approach to unsupervised learning.

Furthermore, we would like to point out that the approach can be characterized as a task to "generate deterministic structures which allow C errors within a substructure", and that this approach can be generalized for other structures as well. For example, finding deterministic orders of objects may be quite unsatisfactory, because given a linear order and adding one error could result in a much larger deterministic structure.

As an example note that the data table

Object	Attr 1	Attr 2	Attr 3	Attr 4	Attr 5
Obj 1	1	0	0	0	0
Obj 2	1	1	0	0	0
Obj 3	1	1	1	0	0
Obj 4	1	1	1	1	0
Obj 5	1	1	1	1	1

produces a linear order as a concept lattice [16]. Now consider the following table with one erroneous observation:

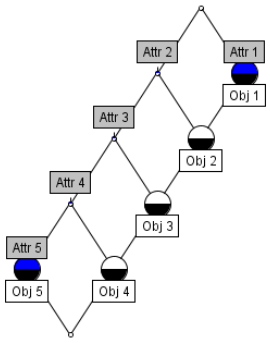
Object	Attr 1	Attr 2	Attr 3	Attr 4	Attr 5
Obj 1	1	0	0	0	0
Obj 2	1	1	0	0	0
Obj 3	1	1	1	0	0
Obj 4	1	1	1	1	0
Obj 5	**0**	1	1	1	1

This system results in a concept lattice consisting of $|U|-2$ more nodes than the simple order structure, see Figure 2. Hence, leaving out some erroneous observation may lead to a smaller, stronger and mutually more stable structure. We will investigate this in future work.

References

1. Beynon, M.: Reducts within the Variable Precision Rough Sets Model: A further investigation. European Journal of Operatiobal Research 134, 592–605 (2001)
2. Böhning, D., Böhning, W., Holling, H.: Revisiting Youden's index as a useful measure of the misclassification error in meta–analysis of diagnostic studies. Statistical Methods in Medical Research 17, 543–554 (2008)
3. Chen, C.B., Wang, L.Y.: Rough set based clustering with refinement using Shannon's entropy theory. Computers and Mathematics with Applications 52, 1563 – 1576 (2006)

Fig. 2. Concept lattice resulting from one error



4. Düntsch, I., Gediga, G.: Weighted λ precision models in rough set data analysis. In: Proceedings of the Federated Conference on Computer Science and Information Systems, Wrocław, Poland. pp. 309–316. IEEE (2012)
5. Düntsch, I., Gediga, G.: Simple Data Filtering in Rough Set Systems. International Journal of Approximate Reasoning 18(1–2), 93–106 (1998), `http://www.cosc.brocku.ca/~duentsch/archive/rgfilt.pdf`
6. Gediga, G., Düntsch, I.: Rough approximation quality revisited. Artificial Intelligence 132, 219–234 (2001), `http://www.cosc.brocku.ca/~duentsch/archive/gamma.pdf`
7. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classification. Journal of the American Statistical Association 49, 732–764 (1954)
8. Hildebrand, D., Laing, J., Rosenthal, H.: Prediction logic and quasi-independence in empirical evaluation of formal theory. The Journal of Mathematical Sociology 3, 197–209 (1974)
9. Hildebrand, D., Laing, J., Rosenthal, H.: Prediction analysis of cross classification. Wiley, New York (1977)
10. Holte, R.C.: Very simple classification rules perform well on most commonly used datasets. Machine Learning 11, 63–90 (1993)
11. Nevill-Manning, C.G., Holmes, G., Witten, I.H.: The development of Holte's 1R classifier. In: Proceedings of the 2nd New Zealand Two-Stream International Conference on Artificial Neural Networks and Expert Systems. pp. 239–246. ANNES '95, IEEE Computer Society, Washington, DC, USA (1995), `http://dl.acm.org/citation.cfm?id=525883.786125`
12. Oehlert, G.: A note on the Delta method. American Statistician 46, 27–29 (1992)
13. Pawlak, Z.: Rough Sets. Internat. J. Comput. Inform. Sci. 11, 341–356 (1982)
14. Pawlak, Z.: A rough set view on Bayes' theorem. International Journal of Intelligent Systems 18, 487–498 (May 2003)
15. Slezak, D.: Rough sets and Bayes factor. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets. Lecture Notes in Computer Science, vol. 3400, pp. 202–229. Springer (2005)

16. Wille, R.: Restructuring lattice theory: An approach based on hierarchies of concepts. In: Rival, I. (ed.) Ordered sets, NATO Advanced Studies Institute, vol. 83, pp. 445–470. Reidel, Dordrecht (1982)
17. Wu, S., Flach, P.A.: Feature selection with labelled and unlabelled data. In: Bohanec, M., Kasek, B., Lavrac, N., Mladenic, D. (eds.) ECML/PKDD'02 workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning. pp. 156–167. University of Helsinki (August 2002)
18. Yao, Y.: Probabilistic rough set approximations. Int. J. Approx. Reasoning 49(2), 255–271 (2008)
19. Youden, W.: Index for rating diagnostic tests. Cancer 3, 32–35 (1950)
20. Ziarko, W.: Variable precision rough set model. Journal of Computer and System Sciences 46, 39–59 (1993)
21. Zytkow, J.M.: Granularity refined by knowledge: Contingency tables and rough sets as tools of discovery. In: B.Dasarathy (ed.) Proc. SPIE 4057, Data Mining and Knowledge Discovery: Theory, Tools, and Technology II. pp. 82–91 (2000)