Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 287–294

ISBN 978-83-60810-51-4

Weighted λ precision models in rough set data analysis

Ivo Düntsch
Department of Computer Science
Brock University
St. Catharines, Ontario, L2S 3A1, Canada,
duentsch@brocku.ca

Günther Gediga
Fachbereich Psychologie
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, based on the well known PRE index λ of Goodman and Kruskal. Using weighted (parametric) λ models we show how expert knowledge can be integrated without losing the monotonic property of the index. Based on a weighted λ index we present a polynomial algorithm to determine an approximately optimal set of predicting attributes. Finally, we exhibit a connection to Bayesian analysis.**

I. Introduction

Rough set data analysis (RSDA) was introduced by Z. Pawlak in the early 1980s and have since become an established tool in information analysis and decision making. Lack of space allows us only to define some of the concepts we require, and we invite the reader to consult [1] for details.

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 [2] is shown in Table I.

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 $x, y \in U$,

$$x \equiv_Q y \iff (\forall a \in Q)[f_a(x) = f_a(y)]. \tag{I.1}$$

Ivo Düntsch gratefully acknowledges support by the Natural Sciences and Engineering Research Council of Canada and by the Bulgarian National Fund of Science, contract number DID02/32/2009.

TABLE I
A decision system from [2]

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	I	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					

The tenet of rough sets is that, 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 I, the classes of θ_Ω are

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

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:

$$\gamma(Q, d) = \frac{|\bigcup \mathrm{Pos}(Q, d)|}{|U|}. \tag{I.3}$$

$\gamma(Q, d)$ 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, $\gamma(\Omega, d) = \frac{1}{21}$.

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 [3]. Therefore, it would be advantageous to have a mechanism which allows some errors in order to result in a more stable prediction success. Such mechanism should be in the parameter – free spirit of the rough set model.

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

II. The variable precision model

A well established model which is less strict in terms of classification errors is the *variable precision rough set model* (VP – model) [2] 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 implies that more than 50% of the elements 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

$$X \overset{\beta}{\subseteq} Y \Longleftrightarrow c(X,Y) \leq \beta. \tag{II.1}$$

Given a family of nonempty subsets $\mathcal{X} = \{X_1, ..., X_k\}$ of U and $Y \subseteq U$, the *lower approximation* $\underline{Y}_\beta$ *of* Y *given* $\mathcal{X}$ *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,

$$\underline{Y}_\beta = \bigcup \{X \in \mathcal{X} : c(X,Y) \leq \beta\} \tag{II.2}$$

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

$$\gamma(Q,d,\beta) = \frac{|\operatorname{Pos}(Q,d,\beta)|}{|U|}, \tag{II.3}$$

where $\operatorname{Pos}(Q,d,\beta)$ 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 ([2], 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 is 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 [4]. As a counterexample, consider the information system shown in Table II which adds an additional independent attribute e to the system of Table I. Setting $P = \{a,b,c\}$ and $Q = \{a,b,c,e\}$, we observe

TABLE II
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						

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 (I.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.

III. 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, $\mathcal{X}$ with classes X_j, $1 \leq j \leq J$, respectively, $\mathcal{Y}$ 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 [5]). As an example, the information system of Table I is depicted as a contingency array in Table III.

TABLE III
Contingency table of the decision system of Table I

	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, and, 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 $|U|$. To conform 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.

IV. 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 [3], [6], [7]. 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,

$$\text{PRE} = 1 - \frac{\epsilon_1}{\epsilon_0}.$$

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 [8]. When no other information is given, it is reasonable to guess a decision category with highest frequency (such as Y_1 in Table III). 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. The PRE of knowing X instead of guessing is given by

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

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 $\mathcal{X}$ 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 [9], and a slightly modified approach in [10] – known as the 1R learning procedure – was shown to be a quite effective tool as well [11].

V. 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 a maximum per column j, γ counts this maximum only if $n_{ij} = n_{\bullet j}$, i.e if exactly one entry in column j is nonzero. We observe that $|U|$ is a suitable denominator for γ, since by our assumption θ_d has more than one class. In this situation, $n_{i\bullet} \neq |U|$ for any class Y_i, and therefore all observations have to be considered as "error".

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

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

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 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 $\mathcal{X}$ with respect to w by

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

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

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

The *w-boundary* now is the set

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

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

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

Various other indices may be defined: Let $\mathcal{X}$ be the partition associated with θ_Q and Y_i be a decision class. In a slight different meaning to machine learning, we will use the terms (Rough-)sensitivity and (Rough-)specificity for the results of our analysis:

1) The *Rough-sensitivity* of $\mathcal{X}$ with respect to Y_i (ratio of conditional positive to mutual positive results)

$$\alpha_w(\mathcal{X}, Y_i) = \frac{|\mathrm{Low}_w(Y_i)|}{|\mathrm{Upp}(Y_i)|},$$

2) The *Rough-specificity* of $\mathcal{X}$ with respect to Y_i (ratio of classified errors to mutual errors)

$$\zeta_w(\mathcal{X}, Y_i) = \frac{|\mathrm{Err}_w(Y_i)|}{|\mathrm{Bnd}_w(Y_i)|}$$

If $\mathcal{Y}$ is the partition induced by the decision attribute, we consider

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

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

2) The *Rough-specificity* of the partition $\mathcal{X}$ with respect to the partition $\mathcal{Y}$

$$\zeta_w(\mathcal{X}, \mathcal{Y}) = \frac{\sum_{Y_i \in \mathcal{Y}} |\mathrm{Err}_w(\mathcal{X}, Y_i)|}{\sum_{Y_i \in \mathcal{Y}} |\mathrm{Bnd}_w(\mathcal{X}, Y_i)|}$$

If $\mathcal{X}$ and $\mathcal{Y}$ are understood, we will just write γ_w and ζ_w. The Rough-sensitivity tells us about the approximation of the set or partition, whereas the Rough-specificity is an index which expresses the relative error of the classification procedure. Both indices are bounded by 0 and 1, and there is a partial monotone relationship: The higher the Rough-specificity the higher the Rough-sensitivity.

VI. Some weighting schemes

The Rough-sensitivity index γ_w 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 $w(U) = 0$.

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 $\lambda_{\mathrm{eq}} = \gamma$ if $C = 0$, and $\lambda_{\mathrm{eq}} = \lambda$ if $C = \infty$, i.e. if $\lambda_{\mathrm{eq}} \equiv 1$. Furthermore, if $C \leq \max_{j=1}^{J}(n_{\bullet j} - \max_{i=1}^{I} n_{ij})$, then the denominator of $\lambda(w_{\mathrm{eq}})$ 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 $\gamma_{w_{\mathrm{eq}}}$:

Proposition VI.1. *Let $Q_a = Q \cup \{a\}$ and $\mathcal{X}_a$ be its associated partition. Then, $\gamma_{w_{\mathrm{eq}}}^{C}(\mathcal{X}, \mathcal{Y}) \leq \gamma_{w_{\mathrm{eq}}}^{C}(\mathcal{X}_a, \mathcal{Y})$.*

Proof: We assume w.l.o.g. that a takes only the two values $0, 1$ (see e.g. [12] 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 $n_{ij}^0 = |X_i \cap Y_j \cap Z_0|$ and $n_{ij}^1 = |X_i \cap Y_j \cap Z_1|$ with respective columns $j0$ and $j1$, and sums $n_{\bullet j}^0$ and $n_{\bullet j}^1$. Then, $n_{ij}^0 + n_{ij}^1 = n_{ij}$, $n_{\bullet j}^0 + n_{\bullet j}^1 = n_{\bullet j}$, and $\max_{i=1}^{I} n_{ij}^0 + \max_{i=1}^{I} n_{ij}^1 \geq \max_{i=1}^{I} n_{ij}$ by the triangle inequality. Thus, if $n_{\bullet j} - \max_{i=1}^{I} n_{ij} \leq C$, then

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

Similarly, $n_{\bullet j}^1 - \max_{i=1}^{I} n_{ij}^1 \leq C$. Therefore, if $w_{\mathrm{eq}}(j) = 1$, then $w_{\mathrm{eq}}(j0) = w_{\mathrm{eq}}(j1) = 1$.

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. ∎

Note that ζ need not be monotonically increasing if an error class changes to a deterministic class when adding a new independent attribute. To prevent such behavior one may require that any deterministic class has to consist of more than C elements. Hence, using

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

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.

VII. 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 IV
Weighted contingency table of the decision system of Table I 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 IV and applying the bounds $E = 0, 0.2, 0.3, 0.6$ to compute $w_{\mathrm{eq}}^{E}(j)$, we observe the approximation qualities shown in Table V. We see that λ increases here as well

TABLE V
λ given various bounds

E	Formula (IV.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

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.

VIII. 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 $\gamma_w - \zeta_w$ 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 $\mathcal{X}$ 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 [13]:

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

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 **O** for the chosen cutpoint C
4) Choose a new attribute which shows the maximum in **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 $\mathcal{O}(J^2)$ and it will find a partition $\mathcal{X}$ which shows a good approximation of Y with an error less than C.

Applying the algorithm to the decision system given in Table I 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
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, since 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, since it is maximal in terms of **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 **O** does not increase.

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

IX. 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[1] . 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

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

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 (IX.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 $1 \leq i \leq I$.

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

$$s_C(X = X_j|Y = Y_i) = \frac{s_C(i, j)}{\sum_{k=1}^{I} s_C(k, j)}. \tag{IX.2}$$

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 (IX.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 (IX.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,

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

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.

X. 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.

[1]For other views of Bayes' Theorem and its connection to rough sets see e.g. [14–16].

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

Case number	a	b	c	d	e
1	1	0	0	0	0
2	1	1	0	0	0
3	1	1	1	0	0
4	1	1	1	1	0
5	1	1	1	1	1

produces a linear order as a concept lattice [17]. Now consider the following table with one erroneous observation:

Case number	a	b	c	d	e
1	1	0	0	0	0
2	1	1	0	0	0
3	1	1	1	0	0
4	1	1	1	1	0
5	**0**	1	1	1	1

This data structure results in a concept lattice consisting of $|U| - 2$ more nodes than the simple order structure. 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] I. Düntsch and G. Gediga, *Rough set data analysis: A road to non-invasive knowledge discovery*. Bangor: Methodos Publishers (UK), 2000. [Online]. Available: http://www.cosc.brocku.ca/~duentsch/archive/nida.pdf

[2] W. Ziarko, "Variable Precision Rough Set Model," *Journal of Computer and System Sciences*, vol. 46, 1993.

[3] G. Gediga and I. Düntsch, "Rough approximation quality revisited," *Artificial Intelligence*, vol. 132, pp. 219–234, 2001. [Online]. Available: http://www.cosc.brocku.ca/~duentsch/archive/gamma.pdf

[4] M. Beynon, "Reducts within the Variable Precision Rough Sets Model: A further investigation," *European Journal of Operatiobal Research*, vol. 134, pp. 592–605, 2001.

[5] J. M. Zytkow, "Granularity refined by knowledge: Contingency tables and rough sets as tools of discovery," in *Data Mining and Knowledge Discovery: Theory, Tools, and Technology II, SPIE*, B.Dasarathy, Ed., 2000.

[6] D. Hildebrand, J. Laing, and H. Rosenthal, "Prediction logic and quasi-independence in empirical evaluation of formal theory," *Journal of the Mathematical Sociology*, vol. 3, pp. 197–209, 1974.

[7] ——, *Prediction analysis of cross classification*. New York: Wiley, 1977.

[8] L. A. Goodman and W. H. Kruskal, "Measures of association for cross classification," *Journal of the American Statistical Association*, vol. 49, pp. 732–764, 1954.

[9] S. Wu and P. A. Flach, "Feature selection with labelled and unlabelled data," in *ECML/PKDD'02 workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*, M. Bohanec, B. Kasek, N. Lavrac, and D. Mladenic, Eds. University of Helsinki, August 2002, pp. 156–167.

[10] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63–90, 1993.

[11] C. G. Nevill-Manning, G. Holmes, and I. H. Witten, "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*, ser. ANNES '95. Washington, DC, USA: IEEE Computer Society, 1995, pp. 239–. [Online]. Available: http://dl.acm.org/citation.cfm?id=525883.786125

[12] I. Düntsch and G. Gediga, "Simple Data Filtering in Rough Set Systems," *International Journal of Approximate Reasoning*, vol. 18, no. 1–2, pp. 93–106, 1998. [Online]. Available: http://www.cosc.brocku.ca/~duentsch/archive/rgfilt.pdf

[13] C.-B. Chen and L.-Y. Wang, "Rough set based clustering with refinement using Shannon's entropy theory," *Computers and Mathematics with Applications*, vol. 52, pp. 1563 – 1576, 2006.

[14] Z. Pawlak, "A rough set view on Bayes' theorem," *International Journal of Intelligent Systems*, vol. 18, p. 487, May 2003.

[15] D. Slezak, "Rough sets and Bayes factor," in *Transactions on Rough Sets*, ser. Lecture Notes in Computer Science, J. F. Peters and A. Skowron, Eds., vol. 3400. Springer, 2005, pp. 202–229.

[16] Y. Yao, "Probabilistic rough set approximations," *Int. J. Approx. Reasoning*, vol. 49, no. 2, pp. 255–271, 2008.

[17] R. Wille, "Restructuring lattice theory: An approach based on hierarchies of concepts," in *Ordered sets*, ser. NATO Advanced Studies Institute, I. Rival, Ed. Dordrecht: Reidel, 1982, vol. 83, pp. 445–470.