Rough approximation quality revisited

Günther Gediga*
Institut für Evaluation und Marktanalysen
Brinkstr. 19
49143 Jeggen, Germany
gediga@eval-institut.de

Ivo Düntsch*
School of Information and Software Engineering
University of Ulster
Newtownabbey, BT 37 0QB, N.Ireland
I.Duentsch@ulst.ac.uk

Abstract

In rough set theory, the approximation quality γ is the traditional measure to evaluate the classification success of attributes in terms of a numerical evaluation of the dependency properties generated by these attributes. In this paper we re-interpret the classical γ in terms of MZ and PRE measures, and exhibit infinitely many possibilities to define γ-like measures which are meaningful in situations different from the classical one.

Key words: Rough sets, approximation quality, PRE measures, empirical evaluation

1 Introduction

One of the strengths of rough set theory is the fact that all its parameters are obtained from the given data:

> "The numerical value of imprecision is not pre-assumed, as it is in probability theory or fuzzy sets – but is calculated on the basis of approximations which are the fundamental concepts used to express imprecision of knowledge". [9]

Approximations are measured simply be counting those observations which support the theoretic assumptions of rough sets and are afterward normalised by the number of all observations. Although this seems to be straightforward, two questions remain:

- Given two partitions, what is an *observation* in the problem of approximating set (or a partition) by a partition?
- Is there a good reason to use the number of all observations as a normalisation factor? Are there possibly other meaningful normalisation factors?

We show in this paper that the answers to both questions lead to the result that γ as the classical rough set evaluation of approximation is one possible instance – but certainly not the only one. Thus, if γ is used as an approximation quality, this is a (conscious or unconscious) choice made by the researcher, and not a necessity given by the data.

*Equal authorship is implied.

2 Pawlak's approximation quality

Throughout this paper, we suppose that U is a finite nonempty set. If $X \subseteq U$, we denote the relative number $\frac{|X|}{|U|}$ of elements of X with respect to U by $p_U(X)$, or just by $p(X)$, if U is understood.

Let us recall a few facts about partitions and equivalence relations. Suppose that $\mathcal{P}$ is a partition of U. If $x \in U$, we let $\mathcal{P}(x)$ be the class of $\mathcal{P}$ containing x, and $\theta_{\mathcal{P}}$ be the equivalence relation associated with $\mathcal{P}$, i.e.

$$x \theta_{\mathcal{P}} y \iff \mathcal{P}(x) = \mathcal{P}(y). \tag{2.1}$$

We say that $\mathcal{P}$ *is finer than a partition* $\mathcal{R}$, and write $\mathcal{P} \preceq \mathcal{R}$, if $\theta_{\mathcal{P}} \subseteq \theta_{\mathcal{R}}$, i.e. if every class of of $\mathcal{R}$ is a union of classes of $\mathcal{P}$. The *identity partition* is the partition containing only singleton sets. It is the finest partition on any nonempty set.

Rough set data analysis (RSDA) [8] is based on the conviction that knowledge about the world is available only up to a certain granularity, and that granularity can be expressed mathematically by partitions and their associated equivalence relations.

If $Y \subseteq U$ and $\mathcal{P}$ is a partition of U, then the *lower approximation* (of Y by $\mathcal{P}$) is defined as

$$\underline{Y}_{\mathcal{P}} = \bigcup \{X \in \mathcal{P} : X \subseteq Y\}, \tag{2.2}$$

and the *upper approximation* by

$$\overline{Y}^{\mathcal{P}} = \bigcup \{X \in \mathcal{P} : X \cap Y \neq \emptyset\}. \tag{2.3}$$

A pair of the form $\langle \underline{Y}, \overline{Y} \rangle$ is called a *rough set*. It is easily seen that the upper approximation is expressible using set complement and lower approximation by

$$\overline{Y}^{\mathcal{P}} = U \setminus \underline{(-Y)}_{\mathcal{P}}. \tag{2.4}$$

The *area of uncertainty* or *boundary region* is defined as

$$\partial_{\mathcal{P}}(Y) = \overline{Y}^{\mathcal{P}} \setminus \underline{Y}_{\mathcal{P}}. \tag{2.5}$$

If $\emptyset \neq U' \subseteq U$, we let $\mathcal{P} \restriction U'$ be the restriction of $\mathcal{P}$ to U', i.e.

$$\mathcal{P} \restriction U' = \{X \cap U' : X \in \mathcal{P}\} \setminus \{\emptyset\}. \tag{2.6}$$

Clearly, $\mathcal{P} \restriction U'$ partitions U', and we have

Lemma 2.1. *Let $Y \subseteq U' \subseteq U$. Then, $\underline{Y}_{\mathcal{P}} = \underline{Y}_{\mathcal{P} \restriction U'}$.*

Proof. Suppose $Y \subseteq U' \subseteq U$. Then,

$$\begin{aligned} x \in \underline{Y}_{\mathcal{P}} &\iff \mathcal{P}(x) \subseteq Y \\ &\iff \mathcal{P}(x) \cap U' \subseteq Y, && \text{since } Y \subseteq U', \\ &\iff \mathcal{P} \restriction U'(x) \subseteq Y, \\ &\iff x \in \underline{Y}_{\mathcal{P} \restriction U'}, \end{aligned}$$

which proves the claim. □

As a numerical measure of imprecision, Pawlak [8, 9] recommends for $Y \neq \emptyset$ the ratio

$$\alpha(\mathcal{P}, Y) = \frac{|\underline{Y}_{\mathcal{P}}|}{|\overline{Y}^{\mathcal{P}}|} \tag{2.7}$$

called the *accuracy measure* of Y by $\mathcal{P}$. It expresses the degree of completeness of our knowledge about Y, given the granularity of $\mathcal{P}$. This measure not only depends on the approximation of Y, because by (2.4) it depends on the approximation of $-Y$ as well:

$$\alpha(\mathcal{P}, Y) = \frac{|\underline{Y}_{\mathcal{P}}|}{|\overline{Y}^{\mathcal{P}}|} = \frac{|\underline{Y}_{\mathcal{P}}|}{|U| \setminus |\underline{(-Y)}_{\mathcal{P}}|}. \tag{2.8}$$

This is not surprising, and, indeed, a necessity of the rough set view that the world (and hence, the complement of Y) is known only up to the granularity given by the classes of $\mathcal{P}$. As a consequence, it is worth noting that $\alpha(\mathcal{P}, Y)$ can be used in all three steps of modelling – learning, testing and applying a model –, because the rough set $\langle \underline{Y}, \overline{Y} \rangle$ is properly defined with the knowledge of the set Y in the learning and testing stage, and without knowing Y in the application stage. This property of $\alpha(\mathcal{P}, Y)$ is seldom observed in rule generating procedures.

Suppose that two views of the world are given by the partitions $\mathcal{P}$ and $\mathcal{R}$ of the universe U, with associated equivalence relations $\theta_{\mathcal{P}}$ and $\theta_{\mathcal{R}}$. We assume that a class of a partition corresponds to a property of its members, and with some abuse of language we identify the name of a class with the name of the property it signifies. The question arises how well one partition can be expressed by the other. If a class X of $\mathcal{P}$ is a subset of a class Y of $\mathcal{R}$, then we can be sure that any element of U having property X also has property Y. In this case, X is called *deterministic with respect to* $\mathcal{R}$, or just *deterministic*, if $\mathcal{R}$ is understood. On the other hand, if X intersects the $\mathcal{R}$-classes $Y_1, \ldots, Y_k$, then we can only say that each element of X has one of the properties $Y_1, \ldots, Y_k$.

An often applied measure for this situation is the *quality of approximation of* $\mathcal{R}$ *by* $\mathcal{P}$, also called the *degree of dependency*. It is defined by

$$\gamma(\mathcal{P}, \mathcal{R}) = \frac{\sum\{|\underline{Y}_{\mathcal{P}}| : Y \in \mathcal{R}\}}{|U|}, \tag{2.9}$$

and evaluates the deterministic part of the rough set description of $\mathcal{R}$ by counting those elements which can be re-classified to blocks of $\mathcal{R}$ with the knowledge given by $\mathcal{P}$ (see Pawlak [9, p.22], Komorowski et al. [6, p.17], Pawlak [10, p. 52]).

Since each class of $\mathcal{P}$ contained in a class of $\mathcal{R}$ corresponds to a deterministic rule (and vice versa), we see that γ is also the relative number of elements of U which can be described by deterministic rules.

3 Re-interpretation of the Pawlak approximation quality

A simple statistic for the precision of (deterministic) approximation of Y given $\mathcal{P}$ which is not affected by the approximation of $-Y$ is

$$\pi(\mathcal{P}, Y) = \frac{|\underline{Y}_{\mathcal{P}}|}{|Y|}. \tag{3.1}$$

Table 1: α and π

U	1	2	3	4	5	6	7	8	$\alpha(\mathcal{P}, Y)$	$\pi(\mathcal{P}, Y)$	$\pi(\mathcal{P}, -Y)$
$\mathcal{P}_1$	x	x	x	y	y	y	y	y	0.375	0.750	0.000
$\mathcal{P}_2$	a	a	b	b	b	c	c	c	0.400	0.500	0.750
Y	*	*	*	*							

This is just the relative number of elements in Y which can be approximated by $\mathcal{P}$; clearly, $\pi(\mathcal{P}, Y) \geq \alpha(\mathcal{P}, Y)$. It is important to point out, that $\pi(\mathcal{P}, Y)$ requires complete knowledge of Y, whereas α does not, since the latter uses only the rough set $\langle \underline{Y}, \overline{Y} \rangle$. Unlike $\alpha(\mathcal{P}, Y)$, the precision measure $\pi(\mathcal{P}, Y)$ can only be applied while learning from data or testing with data. Obviously, $\pi(\mathcal{P}, Y)$ cannot be used in an application step such as prediction. Since this is no drawback for a descriptive measure, we will use $\pi(\mathcal{P}, Y)$ in the sequel.

In the sequel we shall require a monotony property, the simple proof of which is left to the reader:

Lemma 3.1. *If $\mathcal{P}_1 \preceq \mathcal{P}_2$, then $\pi(\mathcal{P}_1, Y) \geq \pi(\mathcal{P}_2, Y)$.* □

The following example demonstrates how α and π differ: Consider $U = \{1, 2, \ldots, 8\}, Y = \{1, 2, 3, 4\}$, and two partitions $\mathcal{P}_1,\ \mathcal{P}_2$ of U shown in Table 1. Since $\mathcal{P}_2$ is more structured in $-Y$ than $\mathcal{P}_1$, and α "knows" only rough sets, we have $\alpha(\mathcal{P}_1, Y) \lneq \alpha(\mathcal{P}_2, Y)$. On the other hand, if we are interested in the precision of the approximation of Y by the classes of $\mathcal{P}$, then we expect $\mathcal{P}_1$ to deliver the better result. For this goal, clearly π is the better index.

The γ statistic is an aggregate measure of the sets of a partition approximated by another partition, and therefore, accommodates both points of view – that of "knowing the world up to $\mathcal{P}$" and that of "approximating Y by $\mathcal{P}$". Indeed, γ turns out to be a weighted average of the π as well as of the α statistics: In the first case, for each class Y of $\mathcal{R}$, the quality of approximation of $\mathcal{P}$ with respect to Y is weighted by the cardinality of Y relative to the number of elements in U, and we obtain

$$\gamma(\mathcal{P}, \mathcal{R}) = \sum_{Y \in \mathcal{R}} \frac{|Y|}{|U|} \cdot \pi(\mathcal{P}, Y) = \sum_{Y \in \mathcal{R}} p(Y) \cdot \pi(\mathcal{P}, Y). \tag{3.2}$$

Therefore $\gamma(\mathcal{P}, \mathcal{R})$ is the mean precision of the approximation of $\mathcal{R}$ by $\mathcal{P}$. Using α as a basis, we have

$$\gamma(\mathcal{P}, \mathcal{R}) = \sum_{Y \in \mathcal{R}} \frac{|\overline{Y}|}{|U|} \cdot \alpha(\mathcal{P}, Y) = \sum_{Y \in \mathcal{R}} p(\overline{Y}) \cdot \alpha(\mathcal{P}, Y), \tag{3.3}$$

Thus, $\gamma(\mathcal{P}, \mathcal{R})$ can also be regarded as the weighted mean of the accuracies of approximation of the sets $Y \in \mathcal{R}$ by $\mathcal{P}$.

Yao [12] connects rough set approximation with a classic distance measure based on sets, called *Marczewski-Steinhaus metric* [7] (MZ), which is defined by

$$MZ(X, Y) = \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y|}.$$

The indices above can be redefined using MZ as

$$\alpha(\mathcal{P}, X) = 1 - MZ(\underline{X}_{\mathcal{P}}, \overline{X}^{\mathcal{P}})$$
$$\pi(\mathcal{P}, X) = 1 - MZ(\underline{X}_{\mathcal{P}}, X)$$
$$\gamma(\mathcal{P}, \mathcal{R}) = 1 - MZ(\bigcup_{X \in \mathcal{R}} \underline{X}_{\mathcal{P}}, \bigcup_{X \in \mathcal{R}} \overline{X}^{\mathcal{P}}) = 1 - MZ(\bigcup_{X \in \mathcal{R}} \underline{X}_{\mathcal{P}}, U)$$

This is a mathematically elegant reinterpretion, and MZ itself has the nice properties of a metric. Nevertheless, it has the disadvantage that it is not clear which model assumptions are needed to state that the proposed fraction is as a meaningful expression.

In the work of Hildebrand et al. [3, 4], the idea of empirical evaluation of a theory (in terms of a formal logic) was formulated by introducing a system of measures called *proportional reduction of errors* (PRE) measures. They have shown that most of the commonly used descriptive statistics have a PRE interpretation, and that this interpretation brings to the fore the important characteristics of the statistic. The idea behind the 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 suitable benchmark model:

$$\gamma_{\text{PRE}} = \begin{cases} 1 - \dfrac{\text{number of observed errors}}{\text{number of expected errors}}, & \text{if number of expected errors} \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

We use the convention that $1 - \frac{0}{0} = 0$, because if there is no error, the error reduction can only be 0.

If our theory says that "X and Y are the same sets", then every element in $(X \cup Y) \setminus (X \cap Y)$ can be regarded as an error for this statement. In this sense, we can interpret MZ as an "error fraction" by

$$MZ(X, Y) = \frac{|X \cup Y| - |X \cap Y|}{|X \cup Y| - |\emptyset|}.$$

where the denominator addresses the worst case $X \cap Y = \emptyset$ as a benchmark model. Using this interpretation for the re-interpreted measures we obtain the combination of errors and expected errors given in Table 2.

Table 2: Measures, approximation errors, and expected approximation errors

Measure	Approximation error	Expected approximation error
$\alpha(\mathcal{P}, Y)$	$\overline{Y}^{\mathcal{P}} \setminus \underline{Y}_{\mathcal{P}}$ = Boundary of Y	$\overline{Y}^{\mathcal{P}}$ = Upper bound of Y
$\pi(\mathcal{P}, Y)$	$Y \setminus \underline{Y}_{\mathcal{P}}$ = Indeterministic cases of Y	Y = All elements in Y
$\gamma(\mathcal{P}, \mathcal{R})$	$U \setminus \bigcup_{Y \in \mathcal{R}} \underline{Y}_{\mathcal{P}}$ = Indeterministic cases of U	U = All elements

As these measures are instances of the MZ metric, we see that they are PRE measures, based on the assumption of a maximal error rate for the benchmark model. In case of α, the elements outside the lower approximation but inside the upper approximation are considered errors, and these errors are compared with a maximum number of errors, which is assumed to be the number of elements in the upper approximation. This occurs exactly in the worst case, namely, when the lower approximation is empty. In case of π, the computation of the error rate assumes that the set X can been described by the data; up to this difference, the construction of the PRE-measure π is the same as the construction of α.

4 More PRE measures for rough sets

The idea of a PRE measure is to answer the question how much better a given model fits – in terms of percentage of error reduction – than a "straw man" benchmark model. We have shown, that rough set based parameters such as γ are PRE measures, when a worst-case benchmark error model is used; therefore, these indices are upper bounds for suitable PRE measures in that situation, and can be regarded as optimistic measures for error reduction. In this Section we will show that it is possible to result in more reasonable estimations for the parameters of interest.

We start with variables X, Y (which may be thought of as sets of attributes of an information system), which generate equivalence relations and partitions in the usual way: If $s, t \in U$, then

$$s\theta_X t \iff X(s) = X(t). \tag{4.1}$$

We denote by $\mathcal{P}_X$ the partition belonging to θ_X, and likewise, we define θ_Y and $\mathcal{P}_Y$. To avoid trivialities, we assume that $|\mathcal{P}_Y| \geq 2$.

According to (2.9), the approximation quality in classical rough set theory is perfect if $\gamma(\mathcal{P}_X, \mathcal{P}_Y) = 1$, i.e. if

$$(\forall K \in \mathcal{P}_X)(\exists L \in \mathcal{P}_Y) K \subseteq L. \tag{4.2}$$

We say that *an equivalence class $K \in \mathcal{P}_X$ counts towards an error*, if

$$(\forall L \in \mathcal{P}_Y) K \not\subseteq L \tag{4.3}$$

holds. The *error function* $\mathrm{ERR} : \mathcal{P}_X \to \{\top, \bot\}$ is now defined by

$$\mathrm{ERR}(K, \mathcal{P}_Y) = \begin{cases} \top, & \text{if (4.3) is true,} \\ \bot, & \text{otherwise.} \end{cases} \tag{4.4}$$

Because a PRE measure is defined with respect to an expectation value, the definition of a suitable random variable χ is necessary; therefore, we assume that $\mathcal{P}_X$ is drawn at random from a set $\chi(\mathcal{P})$ of partitions. On this basis, PRE measures in RSDA can now be defined by

$$\gamma_{\mathrm{PRE}}(\mathcal{P}_X, \mathcal{P}_Y) = 1 - \frac{\sum_{K \in \mathcal{P}_X \text{ and } \mathrm{ERR}(K, \mathcal{P}_Y) = \top} |K|}{\mathcal{E}_{\mathcal{P}_X \in \chi(\mathcal{P})} \left[\sum_{K \in \mathcal{P}_X \text{ and } \mathrm{ERR}(K, \mathcal{P}_Y) = \top} |K| \right]} \tag{4.5}$$

Observe that the "error classes" are weighted with the number of elements they contain. This is necessary, since it is enough for one element to "fail" to discard the whole class.

Obviously, there are different PRE measures which differ in computation of the expected value in the denominator of γ_{PRE} (= the normalisation parameter), or by the assumptions about the structure of $\chi(\mathcal{P})$.

If we regard the classical γ as a PRE measure, then

$$\begin{aligned} \gamma &= \frac{\sum_{K \in \mathcal{P}_X \text{ and } \mathrm{ERR}(K, \mathcal{P}_Y) = \bot} |K|}{|U|} \\ &= 1 - \frac{\sum_{K \in \mathcal{P}_X \text{ and } \mathrm{ERR}(K, \mathcal{P}_Y) = \top} |K|}{|U|}, \end{aligned}$$

which implies that

$$\mathcal{E}_{\mathcal{P}_X \in \chi(\mathcal{P})}\left[\sum_{K \in \mathcal{P}_X \text{ and } \mathrm{ERR}(K,\mathcal{P}_Y)=\top} |K|\right] = |U|. \tag{4.6}$$

This is only possible if $\gamma(\mathcal{P}_X, \mathcal{P}_Y) = 0$ for each $\mathcal{P}_X \in \chi(\mathcal{P})$. A suitable random model χ therefore consists of (any number of) partitions from which no deterministic rule can be derived with respect to the dependent variable. The problem with this result is that the admissible $\mathcal{P}_X$ partitions can only be determined after the data are drawn.

Since there is only one partition which contains no deterministic classes for any $\mathcal{P}_Y$ with at least two elements, namely, the one element partition $\mathcal{P}_X = \{U\}$, an a-priori model that describes γ as a PRE measure must be based on this partition, and therefore, $\chi(\mathcal{P})$ contains $\{U\}$ as its sole element. Therefore, neither the possible a-posteriori nor the a-priori benchmark model take into account the structure of the empirical $\mathcal{P}_X$ – which must of course be known if we want to perform a PRE description.

The description of approximation quality in terms of an error reduction measure offers some insights into the relevant statistical properties of the representation of the data. As shown above, the γ index is not very informative in this sense, because the baseline model for measuring the error reduction is a rather artificial one. A PRE measure with a standard baseline model can be a valuable supplement to the standard measure of approximation quality.

In order to estimate the expectation of errors based on random assignment, we will use randomisation procedures similar to those which we have proposed in [2] for the evaluation of the statistical significance of rough set rules. Randomisation procedures are particularly suitable to RSDA since they do not require outside information; in particular, it is not assumed that the data under discussion are a representative sample.

Suppose that Σ is the set of all permutations of U. If $\sigma \in \Sigma$ and $\mathcal{P}$ is a partition of U, we let

$$\mathcal{P}^\sigma = \{\sigma[K] : K \in \mathcal{P}\}, \tag{4.7}$$

where $\sigma[K] = \{\sigma(x) : x \in K\}$. Observe that $\mathcal{P}^\sigma$ preserves the class sizes of $\mathcal{P}$. We now assume the null hypothesis to be

H_0: "Objects are randomly assigned to classes".

The value

$$p(\gamma(\mathcal{P}_X, \mathcal{P}_Y)|H_0) := \frac{|\{\gamma(\mathcal{P}_X^\sigma, \mathcal{P}_Y) : \sigma \in \Sigma \text{ and } \gamma(\mathcal{P}_X^\sigma, \mathcal{P}_Y) \geq \gamma(\mathcal{P}_X, \mathcal{P}_Y)\}|}{|U|!} \tag{4.8}$$

measures the statistical significance of the observed approximation quality. If $p(\gamma(\mathcal{P}_X, \mathcal{P}_Y)|H_0)$ is low, traditionally below 5%, then the approximation quality is deemed significant, and the (statistical) hypothesis "The value $\gamma(\mathcal{P}_X, \mathcal{P}_Y)$ is due to chance" can be rejected.

With a similar random assignment procedure as a benchmark, we can estimate the expectation of errors using the randomised $\gamma(\mathcal{P}_X^\sigma, \mathcal{P}_Y)$ by

$$\mathcal{E}[\gamma(\mathcal{P}_X, \mathcal{P}_Y)] = \frac{1}{|U|!} \cdot \sum_{\sigma \in \Sigma} \gamma(\mathcal{P}_X^\sigma, \mathcal{P}_Y) \text{ and}$$
$$\mathcal{E}[\text{number of errors}] = |U| - |U| \cdot \mathcal{E}[\gamma(\mathcal{P}_X, \mathcal{P}_Y)].$$

Now,

$$\gamma_{\text{PRE}}(\mathcal{P}_X, \mathcal{P}_Y) = 1 - \frac{|U| - |U| \cdot \gamma(\mathcal{P}_X, \mathcal{P}_Y)}{|U| - |U| \cdot \mathcal{E}[\gamma(\mathcal{P}_X, \mathcal{P}_Y)]} \tag{4.9}$$

$$= \frac{\gamma(\mathcal{P}_X, \mathcal{P}_Y) - \mathcal{E}[\gamma(\mathcal{P}_X, \mathcal{P}_Y)]}{1 - \mathcal{E}[\gamma(\mathcal{P}_X, \mathcal{P}_Y)]} \tag{4.10}$$

is our desired PRE-measure of approximation quality, using $\frac{0}{0} \stackrel{\text{def}}{=} 0$ for the degenerate case.

The PRE measure has negative values, if the observed number of errors (approximation quality) is above (below) the number of error (approximation quality) which can be achieved by random. Furthermore, it is straightforward to see that γ_{PRE} is not monotone decreasing with $\supseteq$: If $\mathcal{P}_X$ consists only of singletons, $\gamma(\mathcal{P}_X, \mathcal{P}_Y) = \mathcal{E}[\gamma(\mathcal{P}_X, \mathcal{P}_Y)] = 1$ and therefore $\gamma_{\text{PRE}}(\mathcal{P}_X, \mathcal{P}_Y) = 0$, if $\mathcal{P}_Y$ consists of at least 2 classes. This maximal dissociation of classical γ and γ_{PRE} relies on the fact that the interpretation of the approximation of $\mathcal{P}_Y$ by singletons in $\mathcal{P}_X$ is totally different in the respective measures. Whereas in classical γ the singletons are the best case, because they are part of the lower approximation, the same singletons are part of the lower approximation for every randomised partition as well. This reduces the denominator of γ_{PRE}, and may lead to such dramatic differences in a situation with many singletons in $\mathcal{P}_X$.

The computation of the expectation of the distribution of γ_{PRE} without simulation is quite costly. However, a short hand correction can be done easily, if we use the fact that a singleton class in $\mathcal{P}_X^\sigma$ is never an error class for any σ. If s is the number of singletons of $\mathcal{P}_X$, then

$$|U| - s \geq \mathcal{E}[\text{number of errors}] = |U| \cdot (1 - \mathcal{E}[\gamma(\mathcal{P}_X, \mathcal{P}_Y)]. \tag{4.11}$$

Let $p_s = \frac{s}{|U|}$, and

$$\gamma^* = \frac{\gamma_1 - p_s}{1 - p_s} \tag{4.12}$$

Proposition 4.1. $\gamma \geq \gamma^* \geq \gamma_{PRE}$.

Proof. Note that γ_1, γ^* and γ_{PRE} are functions of γ of the form

$$f(z) = \begin{cases} \frac{\gamma - z}{1 - z}, & \text{if } z \neq 1, \\ 0, & \text{otherwise.} \end{cases} \tag{4.13}$$

where

$$z = \begin{cases} 0, & \text{for } \gamma, \\ p_s, & \text{for } \gamma^*, \\ \mathcal{E}[\gamma(\mathcal{P}_X, \mathcal{P}_Y)], & \text{for } \gamma_{\text{PRE}}. \end{cases} \tag{4.14}$$

If $p_s = 1$ or $\mathcal{E}[\gamma(\mathcal{P}_X, \mathcal{P}_Y)] = 1$, then $\gamma = 1$ and $\gamma^* = \gamma_{\text{PRE}} = 0$ by our previous remarks. Otherwise, looking at the derivative

$$\frac{df(z)}{dz} = \frac{\gamma - 1}{(1 - z)^2} \leq 0, \tag{4.15}$$

of $f(z)$, we see that $f(z)$ is monotonically decreasing for $0 \leq z \lneq 1$. By (4.11), $\mathcal{E}[\gamma(\mathcal{P}_X, \mathcal{P}_Y)] \geq p_s \geq 0$, whence the claim follows. □

The value of γ is therefore an upper bound for a PRE measure of approximation quality, and γ^* is a "quick-and-dirty" measure in between γ and γ_{PRE}.

Another interpretation makes γ^* even more interesting: In every rule system which is used for a descriptive purpose we can distinguish among two types of "trueness":

1. Conditional "trueness", which means that there is the possibility of counter-examples.
2. Tautologies, i.e. rules which are true in any model. These are exactly the ones which arise from singleton classes.

In terms of description, γ^* seems the optimal measure, because it is built by assuming that tautologies can by expected to be part in any benchmark model, and it provides additional and interesting information about the approximation quality.

5 Cube measures

Up to now, the paper has dealt with the variation of normalisation factors for expressing measures of interest for rough approximation and precision. In this section we will vary the theme "observation", because before counting the number of valid "observations", one needs to reflect on what the "observations" actually should be.

If $\mathcal{P}$ is a partition of U and $1 \leq k$, we let

$$\begin{aligned}\mathcal{P}^k &= \overbrace{\mathcal{P} \times \cdots \times \mathcal{P}}^{k-\text{times}}, \\ &= \{X_1 \times \cdots \times X_k : X_i \in \mathcal{P}\}, \\ &= \{\mathcal{P}(x_1) \times \cdots \times \mathcal{P}(x_k) : x_i \in U\}.\end{aligned}$$

Clearly, $\mathcal{P}^k$ is a partition of U^k, and its corresponding +equivalence is described by

$$\begin{aligned}\vec{x}\theta_{\mathcal{P}^k}\vec{y} &\iff \mathcal{P}^k(\vec{x}) = \mathcal{P}^k(\vec{y}) \\ &\iff (\forall 1 \leq i \leq k)\mathcal{P}(x_i) = \mathcal{P}(y_i).\end{aligned}$$

The *k-th cube relation generated by $\mathcal{P}$* is defined as

$$C^k_{\mathcal{P}} = \bigcup_{Y \in \mathcal{P}} Y^k. \tag{5.1}$$

Note that $C^1_{\mathcal{P}} = U$, $C^2_{\mathcal{P}} = \theta_{\mathcal{P}}$, and, in general, $C^k_{\mathcal{P}} \subseteq \bigcup \mathcal{P}^k$.

For later use, let us look at the approximation of a cube Y^k by the partition $\mathcal{P}^k$:

Lemma 5.1. *For each $Y \subseteq U$, $1 \leq k$,*

$$(\underline{Y}_{\mathcal{P}})^k = \underline{Y^k}_{\mathcal{P}^k} \text{ and } (\overline{Y}_{\mathcal{P}})^k = \overline{Y^k}^{\mathcal{P}^k}. \tag{5.2}$$

Proof.

$$\begin{aligned}\langle x_1,\ldots,x_k\rangle \in (\underline{Y}_{\mathcal{P}})^k &\iff (\forall 1\le i\le k)x_i \in \underline{Y}_{\mathcal{P}}\\ &\iff (\forall 1\le i\le k)\mathcal{P}(x_i)\subseteq Y\\ &\iff \mathcal{P}(x_1)\times\cdots\times\mathcal{P}(x_k)\subseteq Y^k\\ &\iff \langle x_1,\ldots,x_k\rangle \in \underline{Y^k}_{\mathcal{P}^k},\end{aligned}$$

$$\begin{aligned}\langle x_1,\ldots,x_k\rangle \in (\overline{Y}^{\mathcal{P}})^k &\iff (\forall 1\le i\le k)\mathcal{P}(x_i)\cap Y\neq\emptyset\\ &\iff \mathcal{P}(x_1)\times\cdots\times\mathcal{P}(x_k)\cap Y^k\neq\emptyset\\ &\iff \langle x_1,\ldots,x_k\rangle \in \overline{Y^k}^{\mathcal{P}^k}.\end{aligned}$$

□

Corollary 5.2. $\pi(\mathcal{P},Y)^k=\pi(\mathcal{P}^k,Y^k)$ *and* $\alpha(\mathcal{P},Y)^k=\alpha(\mathcal{P}^k,Y^k)$. □

Just like the classical γ, a cube-γ index can be defined in three equivalent ways:

$$\gamma_k(\mathcal{P},\mathcal{R}) = \sum_{Y\in\mathcal{R}}\frac{|\underline{Y}^k|}{|C^k_{\mathcal{R}}|} \tag{5.3}$$

$$= \sum_{Y\in\mathcal{R}}\frac{|Y^k|}{|C^k_{\mathcal{R}}|}\cdot\pi(\mathcal{P}^k,Y^k) \tag{5.4}$$

$$= \sum_{Y\in\mathcal{R}}\frac{|\overline{Y}^k|}{|C^k_{\mathcal{R}}|}\cdot\alpha(\mathcal{P}^k,Y^k). \tag{5.5}$$

Note that

$$\begin{aligned}\gamma_1(\mathcal{P},\mathcal{R}) &= \gamma(\mathcal{P},\mathcal{R}),\\ \gamma_2(\mathcal{P},\mathcal{R}) &= \sum_{Y\in\mathcal{R}}\frac{|Y^2|}{|\theta_{\mathcal{R}}|}\cdot\frac{|\underline{Y^2}_{\mathcal{P}^2}|}{|Y^2|}\\ &= \sum_{Y\in\mathcal{R}}\frac{|\underline{Y^2}_{\mathcal{P}^2}|}{|\theta_{\mathcal{R}}|}\end{aligned}$$

Furthermore, observing that $\{Y^k : Y\in\mathcal{R}\}$ partitions $C^k_{\mathcal{R}}$, and using Lemma 2.1, we obtain

$$\gamma_k(\mathcal{P},\mathcal{R}) = \sum_{Y\in\mathcal{R}}\frac{|Y^k|}{|C^k_{\mathcal{R}}|}\cdot\pi(\mathcal{P}^k\restriction C^k_{\mathcal{R}},Y^k) = \gamma(\mathcal{P}^k\restriction C^k_{\mathcal{R}},\{Y^k : Y\in\mathcal{R}\}). \tag{5.6}$$

With Corollary 5.2 and some basic arithmetic, we can rewrite (5.3) as

$$\gamma_k(\mathcal{P},\mathcal{R}) = \sum_{Y\in\mathcal{R}}\frac{|Y|^k}{\sum_{Y\in\mathcal{R}}|Y|^k}\cdot\pi(\mathcal{P},Y)^k = \sum_{Y\in\mathcal{R}}\frac{|\overline{Y}|^k}{\sum_{Y\in\mathcal{R}}|Y|^k}\cdot\alpha(\mathcal{P},Y)^k. \tag{5.7}$$

If we want to compare $\gamma_k(\mathcal{P}, \mathcal{R})$ for different values of k, we have to consider the different dimensions of the k-cubes. This can be done by using

$$g_k(\mathcal{P}, \mathcal{R}) = \sqrt[k]{\gamma_k(\mathcal{P}, \mathcal{R})}. \tag{5.8}$$

The value g_k is the length of an edge of a k-dimensional cube of size γ_k in one dimension, and therefore, g_k-values can be compared for different dimensions k. The g_k measures can be interpreted analogous to Minkowski norms: If $k = 1$, the statistic looks at one dimensional "cubes" with the consequence that the approximation is measured per element. The larger we choose k, the higher becomes the weight of the larger categories in $\mathcal{R}$ for the description of the approximation quality using $\mathcal{P}$.

Consider the example given in Table 3.

Table 3:

U	1	2	3	4	5	6	7	8	$\gamma_1(\mathcal{P}_i, d)$	$\gamma_2(\mathcal{P}_i, d)$	$g_2(\mathcal{P}_i, d)$
d	A	A	A	A	B	B	C	C			
$\mathcal{P}_1$	1	1	1	2	2	3	3	3	0.375	0.375	0.612
$\mathcal{P}_2$	1	1	2	2	2	3	2	5	0.400	0.250	0.500

The values of $\gamma_j(\mathcal{P}_i, d)$ show a dissociation: The value of γ_1 votes for $\mathcal{P}_2$, whereas γ_2 votes for $\mathcal{P}_2$. The difference of $\mathcal{P}_1$ and $\mathcal{P}_2$ is that the latter consistently approximates more elements, but $\mathcal{P}_1$ consistently approximates more pairs of elements of the equivalence relation θ_d. This is due to the fact that the approximation of P_1 is concentrated in one class (A) of θ_d.

But does it matter? The answer is that $\mathcal{P}_2$ is preferred over $\mathcal{P}_1$, if the application context requires rules of the form

$$(\forall x \in U)[f_q(x) = v_q \Rightarrow f_d(x) = v_d], \tag{5.9}$$

whereas $\mathcal{P}_1$ should be preferred to $\mathcal{P}_2$ if the application context considers rules such as

$$(\forall (x, y) \in U^2)[f_q(x) = f_q(y) \Rightarrow f_d(x) = f_d(y)]. \tag{5.10}$$

An index such as γ_k is useful for model selection only if it has the same monotony properties as γ_1. Our next result shows this to be the case:

Proposition 5.3. *1. If $\mathcal{P}_1 \preceq \mathcal{P}_2$, then*

(a) $\gamma_k(\mathcal{P}_1, d) \geq \gamma_k(\mathcal{P}_2, d)$,

(b) $g_k(\mathcal{P}_1, d) \geq g_k(\mathcal{P}_2, d)$.

2. If $\mathcal{P}$ is the identity partition, then $\gamma_k(\mathcal{P}, d) = g_k(\mathcal{P}, d) = 1$ for all $k \geq 1$.

Proof. 1. Both statements follow immediately from Lemma 3.1 and the definitions of γ_k and g_k.

2. Since every class of $\mathcal{P}$ is a singleton, the approximation is perfect for every $k \geq 1$.

□

Table 4: Contraception data

Country		q_1	q_2	q_3	q_4	d
(1)	Lesotho	3.9	4	73	0	6
(2)	Kenya	0.9	4	108	6	9
(3)	Peru	2.7	17	367	0	14
(4)	Sri Lanka	3.8	20	142	12	22
(5)	Indonesia	1.2	9	61	14	25
(6)	Thailand	2.1	8	142	20	36
(7)	Colombia	2.7	47	284	16	37
(8)	Malaysia	1.6	29	313	18	38
(9)	Guayana	6.1	20	318	0	42
(10)	Jamaica	6.9	8	593	23	44
(11)	Jordan	1.4	53	197	0	44
(12)	Panama	5.3	50	570	19	59
(13)	Costa Rica	4.7	18	464	21	59
(14)	Fiji	3.7	15	321	22	60
(15)	Korea	4.5	15	188	24	61

Table 5: Recoded data

Country		q_1	q_2	q_3	q_4	d
(1)	Lesotho	1	0	0	0	0
(2)	Kenya	0	0	0	0	0
(3)	Peru	1	1	2	0	0
(4)	Sri Lanka	1	1	0	1	0
(5)	Indonesia	0	0	0	1	0
(6)	Thailand	1	0	0	1	1
(7)	Colombia	1	2	1	1	1
(8)	Malaysia	0	1	2	1	1
(9)	Guayana	2	1	2	0	1
(10)	Jamaica	2	0	2	2	1
(11)	Jordan	0	2	1	0	1
(12)	Panama	2	2	2	1	2
(13)	Costa Rica	2	1	2	2	2
(14)	Fiji	1	1	2	2	2
(15)	Korea	2	1	1	2	2

Looking at the results obtained in this Section, we conclude that the type of approximation quality (and the "best set of attributes") depends on the context of the application – a boundary condition which cannot be defined by the data alone. The value $k = 1$, which is used in RSDA, is perhaps the simplest choice, but it is by far not the only one possible. One has to consider which γ_k is relevant for expressing the quality of approximation, and which k-cube relation is of interest. Thus, the researcher has to make a decision how the weights of the approximation of sets are to be chosen.

6 Example

To demonstrate our procedures, we will use a data set published in [1] shown in Table 4, with the data recoded as in Table 5. It is aimed to approximate the values of the countries in the attribute

- % ever practising contraception (d)

from the characteristics

- Average years of education (q_1),
- Percent urbanised (q_2),
- Gross national product per capita (q_3),
- Expenditures on family planning (q_4),

and to find those characteristics that are most valuable to approximate the dependent attribute d.

We use the notation PRE_k for $\gamma_{k_{\text{PRE}}}$, where $\gamma_{k_{\text{PRE}}}$ is defined in analogy to γ_{PRE} (4.5). The results for the various indices are shown in Table 6. The full set of variables results in the identity partition, and thus in a perfect approximation quality ($\gamma_1 = 1$). Since the random assignment preserves the cardinalities

Table 6: Analysis of the contraception data

Set	γ_1	$\mathcal{E}[\gamma_1]$	PRE_1	γ_1^*	$\mathcal{E}[\gamma_1 *]$	PRE_1^*	γ_2	$\mathcal{E}[\gamma_2]$	PRE_2	γ_3	$\mathcal{E}[\gamma_3]$	PRE_3
$\{q_1, q_2, q_3, q_4\}$	1	1	0	0	0	0	1	1	0	1	1	0
$\{q_2, q_3, q_4\}$	0.73	0.62	0.31	0.50	0.28	0.31	0.53	0.42	0.19	0.38	0.31	0.11
$\{q_1, q_3, q_4\}$	0.73	0.81	-0.43	0	0.30	-0.43	0.53	0.68	-0.48	0.38	0.59	-0.51
$\{q_1, q_2, q_4\}$	1	0.91	1	1	0.29	1	1	0.83	1	1	0.77	1
$\{q_1, q_2, q_3\}$	0.60	0.62	-0.06	0.25	0.29	-0.06	0.38	0.43	-0.09	0.24	0.32	-0.11
$\{q_3, q_4\}$	0.33	0.34	-0.01	0.17	0.17	-0.01	0.12	0.17	-0.06	0.04	0.10	-0.06
$\{q_2, q_4\}$	0.47	0.34	0.20	0.38	0.24	0.20	0.22	0.17	0.06	0.11	0.10	0.01
$\{q_2, q_3\}$	0.40	0.31	0.13	0.18	0.13	0.13	0.18	0.13	0.05	0.08	0.07	0.02
$\{q_1, q_4\}$	0.33	0.31	-0.03	0.17	0.19	-0.03	0.12	0.08	-0.08	0.04	0.11	-0.07
$\{q_1, q_3\}$	0.40	0.36	0.06	0.18	0.13	0.06	0.18	0.18	0.00	0.09	0.10	-0.01
$\{q_1, q_2\}$	0.47	0.44	0.05	0.20	0.16	0.05	0.27	0.24	0.04	0.18	0.15	0.04
$\{q_4\}$	0	0.00	0.00	0	0.00	0.00	0	0.00	0.00	0	0.00	0.00
$\{q_3\}$	0	0.01	-0.01	0	0.01	-0.01	0	0.01	-0.01	0	0.01	-0.01
$\{q_2\}$	0	0.01	-0.01	0	0.01	-0.01	0	0.00	0.00	0	0.00	0.00
$\{q_1\}$	0	0.01	-0.01	0	0.01	-0.01	0	0.00	0.00	0	0.00	0.00

of the partition classes, the expectation of γ_1 is 1 as well, and therefore, $\gamma_{\text{PRE}} = 0$. For the same reason, $\gamma_1^* = 0$.

Because $\{q_1, q_2, q_4\}$ is a reduct (i.e. a set of attributes minimal with respect to the property $\gamma_1 = 1$), RSDA prefers the set as the best attribute set for describing d. Most of the indices show that $\{q_1, q_2, q_4\}$ is indeed a good choice. However, the expectation values of the γ_k are quite high; this means that the resulting rules are based on only a small number of examples, and consequently the approximation is not significant ($\alpha = 0.29$).

Of some interest is the comparison of $\{q_2, q_3, q_4\}$ and $\{q_1, q_3, q_4\}$, because both have identical γ_k-values. The PRE interpretation offers a different view: $\{q_2, q_3, q_4\}$ results in a positive PRE-measure, whereas the PRE results of $\{q_1, q_3, q_4\}$ are negative. This means that a higher approximation quality can be achieved by using our benchmark of random assignment.

7 Discussion

Our starting point were the basic questions

- What are the *observations* that should be counted?
- What could be used as a meaningful normalisation factor?

We have shown that there are various reasonable choices for both problems, and that these choices lead to evaluations of approximation quality which are different from the standard γ_1 statistic. Differentiating among different types of basic information led to approximation measures which show characteristics similar to the Minkowski norm in metric data analysis. Although these measures exhibit a certain dissociation from γ_1 as shown by the examples, applications of different γ_k measures in a reduct search situation shows that the reducts do not differ from those constructed by γ_1. The example is typical for this observation: The rank order of the evaluation of the conditional attribute sets is very stable, given different γ_k, if γ_1 is moderately large.

One problem remains: If there are so many possible measures – what measure should be chosen? We think the question is justified, but hard to solve. It certainly is dependent on the context and the intentions of the researcher. There are some examples in the literature, such as the well established RAND index, originally used for the evaluation of cluster analysis results [5, 11]. This measure acts on the same domain as γ_2, and can be used to evaluate the equivalence of two partitions.

A different look at the normalisation factor resulted in dramatic changes of the evaluation of approximation quality. Assuming statistical independence of the decision attribute from the class assignments $\mathcal{P}_X$ of the conditional attributes, a γ-like PRE measure shows behaviour extremely different from γ_1, if $\mathcal{P}_X$ contains very small classes. Whereas γ_1 tends to be very high in this situation, γ_{PRE} tends to show low values. This is explained by the fact that small classes have a high chance to be in the lower approximation of a set. We could show additionally that, at any rate, γ_1 is an optimistic measure of approximation quality, because $\gamma_1 \geq \gamma_{\text{PRE}}$. We therefore recommend to use both statistics to describe approximation quality. However, since the computation of γ_{PRE} is very costly, we propose a "quick - and - dirty" measure γ_1^* for the evaluation of approximation quality, which can be computed with minimal additional effort. Because $\gamma_1 \geq \gamma_1^* \geq \gamma_{\text{PRE}}$, we result in a better upper approximation of the PRE approximation measure than can be achieved by γ_1. Because the computation of γ_1^* is not costly at all, it should be implemented as a routine procedure in rough set algorithms.

We have shown that the interpretation of γ_1 as a PRE measure leads to strange set combinations, which are not compatible to standard statistical assumptions such as independence. It is worthwhile to look at the problem from a different direction: Consider set combinations which can be defined by $\gamma_1 = 0$ and define a "rough independence" axiomatically, which implies $\gamma_1 = 0$ in these cases. But this is a different story –

Acknowledgement

We thank the anonymous referees for their constructive remarks which helped to improve the clarity of our thoughts and, hopefully, of the paper.

References

[1] Cliff, N. (1994). Predicting ordinal relations. *British J. Math. Statist. Psych.*, **47**, 127–150.

[2] Düntsch, I. & Gediga, G. (1997). Statistical evaluation of rough set dependency analysis. *International Journal of Human–Computer Studies*, **46**, 589–604.

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

[4] Hildebrand, D., Laing, J. & Rosenthal, H. (1977). Prediction analysis of cross classification. New York: Wiley.

[5] Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**, 193–218.

[6] Komorowski, J., Pawlak, Z., Polkowski, L. & Skowron, A. (1999). Rough sets: A tutorial. In S. Pal & A. Skowron (Eds.), *Rough Fuzzy Hybridization*, 3–98. Springer–Verlag.

[7] Marcewski, E. & Steinhaus, H. (1958). On a certain distance of sets and the corresponding distance of functions. *Colloquium Mathematicum*, **6**, 319–327.

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

[9] Pawlak, Z. (1991). Rough sets: Theoretical aspects of reasoning about data, vol. 9 of *System Theory, Knowledge Engineering and Problem Solving*. Dordrecht: Kluwer.

[10] Pawlak, Z. (1997). Rough set approach to knowledge-based decision support. *European Journal of Operational Research*, **99**, 48–57.

[11] Rand, M. W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–850.

[12] Yao, Y. (2001). Information granulation and rough set approximation. *International Journal of Intelligent Systems*, **16**, 87–104.