Rough Sets and Bayes Factor

Dominik Ślęzak

Department of Computer Science, University of Regina,
Regina, SK, S4S 0A2 Canada
slezak@uregina.ca

Abstract. We present a novel approach to understanding the concepts of the theory of rough sets in terms of the inverse probabilities derivable from data. It is related to the Bayes factor known from the Bayesian hypothesis testing methods. The proposed Rough Bayesian model (RB) does not require information about the prior and posterior probabilities in case they are not provided in a confirmable way. We discuss RB with respect to its correspondence to the original Rough Set model (RS) introduced by Pawlak and Variable Precision Rough Set model (VPRS) introduced by Ziarko. We pay a special attention on RB's capability to deal with multi-decision problems. We also propose a method for distributed data storage relevant to computational needs of our approach.

Keywords: Rough Sets, Probabilities, Bayes Factor.

1 Introduction

The theory of *rough sets*, introduced by Pawlak in 1982 (see [10] for references), is a methodology of dealing with uncertainty in data. The idea is to approximate the *target concepts* (*events*, *decisions*) using the classes of *indiscernible* objects (in case of qualitative data – the sets of records with the same values for the features under consideration). Every concept X is assigned the *positive*, *negative*, and *boundary regions* of data, where X is *certain*, *impossible*, and *possible but not certain*, according to the data based information.

The above principle of rough sets has been extended in various ways to deal with practical challenges. Several extensions have been proposed as related to the data based probabilities. The first one, *Variable Precision Rough Set* (*VPRS*) model [24] proposed by Ziarko, softens the requirements for certainty and impossibility using the grades of the *posterior probabilities*. Pawlak [11, 12] begins research on the connections between rough sets and *Bayesian reasoning*, in terms of operations on the *posterior*, *prior*, and *inverse probabilities*. In general, one can observe a natural correspondence between the fundamental notions of rough sets and statistics, where a hypothesis (target concept X_1) can be verified *positively*, *negatively* (in favor of the *null hypothesis*, that is a *complement* concept X_0), or *undecided*, under the given evidence [4, 15].

Decision rules resulting from the rough set algorithms can be analyzed both with respect to the data derived posterior probabilities (certainty, accuracy)

J.F. Peters and A. Skowron (Eds.): Transactions on Rough Sets III, LNCS 3400, pp. 202–229, 2005.

and the inverse probabilities (coverage), like in machine learning methods [8, 9]. However, only the posterior probabilities decide about membership of particular cases to the positive/negative/boundary regions – the inverse probabilities are usually used just as the optimization parameters, once the posterior probabilities are good enough (cf. [16, 21]). Several rough set approaches to evaluation of "goodness" of the posterior probabilities were developed, like, for example, the above-mentioned parameter-controlled grades in VPRS. In [25] it is proposed to relate those grades to the prior probabilities of the target concepts. This is, actually, an implicit attempt to relate the rough set approximations with the Bayesian hypothesis testing, where comparison of the posterior and prior probabilities is crucial [1, 20].

In [18] a simplified probabilistic rough set model is introduced, where a given new object, supported by the evidence E, is in the positive region of X_1, if and only if the posterior probability $Pr(X_1|E)$ is greater than the prior probability $Pr(X_1)$. It is equivalent to inequality $Pr(E|X_1) > Pr(E|X_0)$, which means that the observed evidence is more probable assuming hypothesis X_1 than its complement X_0 (cf. [19]). This is the first step towards handling rough sets in terms of the inverse probabilities. Its continuation [17] points at relevance to the *Bayes factor* [4, 6, 15, 20], which takes the form of the following ratio:

$$B_0^1 = \frac{Pr(E|X_1)}{Pr(E|X_0)} \tag{1}$$

The Bayes factor is a well known example of comparative analysis of the inverse probabilities, widely studied not only by philosophers and statisticians but also within the domains of machine learning and data mining. Such analysis is especially important with regards to the rule confirmation and interestingness measures (cf. [5, 7]), considered also in the context of the rough set based decision rules [3, 12, 21]. In this paper we develop the foundations for *Rough Bayesian* (*RB*) model, which defines the rough-set-like positive/negative/boundary regions in terms of the Bayes factor. In this way, the inverse probabilities, used so far in the analysis of the decision rules obtained from the rough set model, become to be more directly involved in the specification of this model itself.

Operating with B_0^1 provides two major advantages, similar to those related to its usage in Bayesian reasoning and probabilistic data mining methods. Firstly, the posterior probabilities are not always derivable directly from data, in a reliable way (see e.g. Example 3 in Subsection 2.2). In such cases, information is naturally represented by means of the inverse probabilities corresponding to the observed evidence conditioned by the states we want to verify, predict, or approximate. Within the domain of statistical science, there is a discussion whether (and in which cases) the inverse probabilities can be combined with the prior probabilities using the *Bayes rule*. If it is allowed, then the proposed RB model can be rewritten in terms of the posterior probabilities and starts to work similarly as VPRS. However, such translation is impossible in case we can neither estimate the prior probabilities from data nor define them using background knowledge. Then, the data based inverse probabilities remain the only basis for constructing the rough-set-like models.

The second advantage of basing a rough set model on the Bayes factor is that the inverse probabilities provide clearer ability of comparing likelihoods of concepts. In the probabilistic rough set extensions proposed so far, the posterior probabilities $Pr(X_1|E)$ are compared to constant parameters [24, 25] or to the prior probabilities $Pr(X_1)$ of *the same* target concepts [18, 19]. A direct comparison of probabilities like $Pr(X_1|E)$ and $Pr(X_0|E)$ would not have too much sense, especially when the prior probabilities of X_1 and X_0 differ significantly. Comparison of the inverse probabilities $Pr(E|X_1)$ and $Pr(E|X_0)$ is more natural, as corresponding to relationship between the ratios of the posterior and prior probabilities for different concepts:

$$\frac{Pr(E|X_1)}{Pr(E|X_0)} = \frac{Pr(X_1|E)/Pr(X_1)}{Pr(X_0|E)/Pr(X_0)} \tag{2}$$

It shows that the analysis of the Bayes factor is equivalent to comparison of the ratios of the gain in probabilistic belief for X_1 and X_0 under the evidence E (cf. [18]). Therefore, the RB model can be more data sensitive than the approaches based on the posterior probabilities, especially for the problems with more than two target concepts to be approximated. RB is well comparable to Bayesian hypothesis testing methods, where B_0^1 is regarded as a summary of the evidence for X_1 against X_0 provided by the data, and also as the ratio of the posterior and prior odds. Finally, RB may turn out to be applicable to the problems where the prior probabilities are dynamically changing, remain unknown, or simply undefinable. Although we do not discuss such situations, we refer to the reader's experience and claim that it may be really the case for real-life data sets.

The article is organized as follows: Section 2 presents non-parametric probabilities derivable from data, with their basic intuitions and relations. It also contains basic information about the way of applying the Bayes factor in decision making. Section 3 presents the original rough set approach in terms of the posterior and, what is novel, the inverse data based probabilities. Then it focuses on foundations of the VPRS model and corresponding extensions of rough sets. Section 4 introduces the Rough Bayesian approach related to the Bayes factors calculated for the pairs of decision classes. The proposed model is compared with VPRS, both for the cases of two and more target concepts. In particular, it requires extending the original formulation of VPRS onto the multi-target framework, which seems to be a challenging task itself. Section 5 includes a short note on an alternative, distributed way of representing the data for the needs of the Rough Bayesian model. Section 6 summarizes the article and discusses directions for further research.

2 Data and Probabilities

2.1 Data Representation

In [10] it was proposed to represent the data as an *information system* $\mathbb{A} = (U, A)$, where U denotes the *universe* of *objects* and each *attribute* $a \in A$ is identified with function $a : U \rightarrow V_a$, for V_a denoting the set of values of a.

U	a_1	a_2	a_3	a_4	a_5	d
u_1	1	1	0	1	2	0
u_2	0	0	0	2	2	0
u_3	2	2	2	1	1	1
u_4	0	1	2	2	2	1
u_5	2	1	1	0	2	0
u_6	2	2	2	1	1	1
u_7	0	1	2	2	2	0
u_8	2	2	2	1	1	1
u_9	2	2	2	1	1	1
u_{10}	0	0	0	2	2	0

U	a_1	a_2	a_3	a_4	a_5	d
u_{11}	1	2	0	0	2	0
u_{12}	1	1	0	1	2	1
u_{13}	0	1	2	2	2	1
u_{14}	2	1	1	0	2	0
u_{15}	2	2	2	1	1	0
u_{16}	1	1	0	1	2	1
u_{17}	1	1	0	1	2	0
u_{18}	2	1	1	0	2	0
u_{19}	2	2	2	1	1	1
u_{20}	2	2	2	1	1	1

Fig. 1. Decision system $\mathbb{A} = (U, A \cup \{d\})$, $U = \{u_1, \ldots, u_{20}\}$, $A = \{a_1, \ldots, a_5\}$. Decision d induces classes $X_0 = \{u_1, u_2, u_5, u_7, u_{10}, u_{11}, u_{14}, u_{15}, u_{17}, u_{18}\}$ and $X_1 = \{u_3, u_4, u_6, u_8, u_9, u_{12}, u_{13}, u_{16}, u_{19}, u_{20}\}$.

Each subset $B \subseteq A$ induces a partition over U with classes defined by grouping together the objects having identical values of B. We obtain the partition space U/B, called the B-*indiscernibility relation* $IND_{\mathbb{A}}(B)$, where elements $E \in U/B$ are called the B-*indiscernibility classes* of objects.

Information provided by $\mathbb{A} = (U, A)$ can be applied to approximate the *target events* $X \subseteq U$ by means of the elements of U/B, $B \subseteq A$. We can express such targets using a distinguished attribute $d \notin A$. Given $V_d = \{0, \ldots, r-1\}$, we define the sets $X_k = \{u \in U : d(u) = k\}$. We refer to such extended information system $\mathbb{A} = (U, A \cup \{d\})$ as to a *decision system*, where d is called the *decision attribute*, and the sets X_k are referred to as the *decision classes*.

Elements of U/B correspond to B-*information vectors* $w \in V_B$ – collections of *descriptors* (a, v), $a \in B$, $v \in V_a$. They are obtained using B-*information function* $B : U \to V_B$ where $B(u) = \{(a, a(u)) : a \in B\}$.

Example 1. Consider $\mathbb{A} = (U, A \cup \{d\})$ in Fig. 1 and $B = \{a_1, a_3\}$. B-information vector $\{(a_1, 2), (a_3, 2)\}$ corresponds to conjunction of conditions $a_1 = 2$ and $a_3 = 2$, which is satisfied by the elements of $E = \{u_3, u_6, u_8, u_9, u_{15}, u_{19}, u_{20}\}$. In other words, $B(u_i) = \{(a_1, 2), (a_3, 2)\}$ holds for $i = 3, 6, 8, 9, 15, 19, 20$. □

2.2 Types of Probabilities

Let us assume that events X_k are labelled with the prior probabilities $Pr(X_k)$, $\sum_{l=0}^{r-1} Pr(X_l) = 1$, $r = |V_d|$. It is reasonable to claim that any X_k is likely to occur and that its occurrence is not certain – otherwise, we would not consider such an event as worth dealing with. The same can be assumed about indiscernibility classes $E \in U/B$, $B \subseteq A$, in terms of probabilities of their occurrence in data $\mathbb{A} = (U, A \cup \{d\})$. We can express such requirements as follows:

$$0 < Pr(X_k) < 1 \text{ and } 0 < Pr(E) < 1 \tag{3}$$

Let us also assume that each class E is labelled with the posterior probabilities $Pr(X_k|E)$, $\sum_{l=0}^{r-1} Pr(X_l|E) = 1$, which express beliefs that X_k will occur under the evidence corresponding to E.

Remark 1. We can reconsider probabilities in terms of the attribute-value conditions. For instance, if $k = 1$ and $E \in U/B$ groups the objects satisfying conditions $a_1 = 2$ and $a_3 = 2$, then we can write $Pr(d = 1)$ instead of $Pr(X_1)$, and $Pr(d = 1|a_1 = 2, a_3 = 2)$ instead of $Pr(X_1|E)$. □

In machine learning and data mining [5, 8, 9], the posterior probabilities correspond to the certainty (accuracy, precision) factors. One can also compare prior and posterior knowledge to see whether a new evidence (satisfaction of conditions) increases or decreases the belief in a given event (membership to a given decision class). This is, actually, the key idea of Bayesian reasoning [1, 20], recently applied also to rough sets [18, 19]. The easiest way of the data based prior and posterior probability estimation is the following:

$$Pr(X_k|E) = \frac{|X_k \cap E|}{|E|} \quad \text{and} \quad Pr(X_k) = \frac{|X_k|}{|U|} \tag{4}$$

Example 2. In case of Fig. 1, we get $Pr(d = 1|a_1 = 2, a_3 = 2) = 6/7$, which estimates our belief that objects satisfying $a_1 = 2$ and $a_3 = 2$ belong to X_1. It seems to increase the belief in X_1 with respect to $Pr(d = 1) = 1/2$. □

One can also use the inverse probabilities $Pr(E|X_k)$, $\sum_{E \in U/B} Pr(E|X_k) = 1$, which express a likelihood of the evidence E under the assumption about X_k [20]. The posterior probabilities are then derivable by using the Bayes rule. For instance, in case of $\mathbb{A} = (U, A \cup \{d\})$ with two decision classes, we have:

$$Pr(X_1|E) = \frac{Pr(E|X_1)Pr(X_1)}{Pr(E|X_0)Pr(X_0) + Pr(E|X_1)Pr(X_1)} \tag{5}$$

Remark 2. If we use estimations $Pr(E|X_k) = |X_k \cap E|/|X_k|$, then (5) provides the same value of $Pr(X_k|E)$ as (4). For instance, $Pr(d = 1|a_1 = 2, a_3 = 2) = (3/5 \cdot 1/2)/(1/10 \cdot 1/2 + 3/5 \cdot 1/2) = 6/7$. □

In some cases estimation (4) can provide us with invalid values of probabilities. According to the Bayesian principles, it is then desirable to combine the inverse probability estimates with the priors expressing background knowledge, not necessarily derivable from data. We can see it in the following short study:

Example 3. Let us suppose that X_1 corresponds to a rare but important target event like, e.g., some medical pathology. We are going to collect the cases supporting this event very accurately. However, we are not going to collect information about all the "healthy" cases as X_0. In the medical data sets we can rather expect the 50:50 proportion between positive and negative examples. It does not mean, however, that $Pr(X_1)$ should be estimated as $1/2$. It is questionable whether the posterior probabilities $Pr(X_1|E)$ should be derived from such data using estimation with $|E|$ in denominator – it is simply difficult to accept that $|E|$ is calculated as the non-weighted sum of $|E \cap X_0|$ and $|E \cap X_1|$. □

2.3 Bayes Factor

Example 3 shows that in some situations the posterior probabilities are not derivable from data in a credible way. In this paper, we do not claim that this is a frequent or infrequent situation and we do not focus on any specific real-life data examples. We simply show that it is still possible to derive valuable knowledge basing only on the inverse probabilities, in case somebody cannot trust or simply does not know priors and posteriors.

Our idea to refer to the inverse probabilities originates from the notion of *Bayes factor*, which compares the probabilities of the observed evidence E (indiscernibility class or, equivalently, conjunction of conditional descriptors) under the assumption concerning a hypothesis X_k (decision class) [4, 6, 15, 20]. In case of systems with two decision classes, the Bayes factor takes the form of B_0^1 defined by equation (1). It refers to the posterior and prior probabilities, as provided by equation (2). However, we can restrict to the inverse probabilities, if we do not know enough about the priors and posteriors occurring in (2).

The Bayes factor can be expressed in various ways, depending on the data type [15]. In case of decision table real valued conditional attributes, it would be defined as the ratio of probabilistic densities. For symbolic data, in case of more than two decision classes, we can consider pairwise ratios

$$B_l^k = \frac{Pr(E|X_k)}{Pr(E|X_l)} \tag{6}$$

for $l \neq k$, or ratios of the form

$$B_{\not k}^k = \frac{Pr(E|X_k)}{Pr(E|\neg X_k)} \qquad \text{where } \neg X_k = \bigcup_{l \neq k} X_l \tag{7}$$

In [6], it is reported that twice of the logarithm of B_0^1 is on the same scale as the deviance test statistics for the model comparisons. The value of B_0^1 is then used to express a degree of belief in hypothesis X_1 with respect to X_0, as shown in Fig. 2. Actually, the scale presented in Fig. 2 is quite widely used by statisticians while referring to the Bayes factors. We can reconsider this way of hypothesis verification by using the significance threshold $\varepsilon_1^0 \geq 0$ in the following criterion:

$$X_1 \text{ is verified } \varepsilon_1^0\text{-positively, if and only if } Pr(E|X_0) \leq \varepsilon_1^0 Pr(E|X_1) \tag{8}$$

For lower values of $\varepsilon_1^0 \geq 0$, the positive hypothesis verification under the evidence $E \in U/B$ requires more significant advantage of $Pr(E|X_1)$ over $Pr(E|X_0)$. Actually, it is reasonable to assume that $\varepsilon_1^0 \in [0, 1)$. This is because for $\varepsilon_1^0 = 1$, we simply cannot decide between X_1 and X_0 (cf. [18]) and for $\varepsilon_1^0 > 1$ one should rather consider X_0 instead of X_1 (by switching X_0 with X_1 and using possibly different $\varepsilon_0^1 \in [0, 1)$ in (8)). Another special case, $\varepsilon_1^0 = 0$, corresponds to *infinitely strong* evidence for hypothesis X_1, yielding $Pr(E|X_0) = 0$. This is the reason why we prefer to write $Pr(E|X_0) \leq \varepsilon_1^0 Pr(E|X_1)$ instead of $B_0^1 \geq 1/\varepsilon_1^0$ in (8).

The B_0^1 ranges proposed in [6]	Corresponding $2\log B_0^1$ ranges	Corresponding ε_1^0 ranges based on (8)	Evidence for X_1 described as in [6]
less than 1	less than 0	more than 1	negative (support X_0)
1 to 3	0 to 2	0.3 to 1	barely worth mentioning
3 to 12	2 to 5	0.1 to 0.3	positive
12 to 150	5 to 10	0.01 to 0.1	strong
more than 150	more than 10	less than 0.01	very strong

Fig. 2. The Bayes factor significance scale proposed in [6], with the corresponding ranges for $\varepsilon_1^0 \geq 0$ based on criterion (8). The values in the third column are rounded to better express the idea of working with inequality $Pr(E|X_0) \leq \varepsilon_1^0 Pr(E|X_1)$.

3 Rough Sets

3.1 Original Model in Terms of Probabilities

In Subsection 2.1 we mentioned that decision systems can be applied to approximation of the target events by means of indiscernibility classes. A method of such data based approximation was proposed in [10], as the theory of *rough sets*. Given $\mathbb{A} = (U, A \cup \{d\})$, $B \subseteq A$, and $X_k \subseteq U$, one can express the main idea of rough sets in the following way: The *B-positive*, *B-negative*, and *B-boundary rough set regions* (abbreviated as *RS-regions*) are defined as

$$\begin{aligned} \mathcal{POS}_B(X_k) &= \bigcup\{E \in U/B : Pr(X_k|E) = 1\} \\ \mathcal{NEG}_B(X_k) &= \bigcup\{E \in U/B : Pr(X_k|E) = 0\} \\ \mathcal{BND}_B(X_k) &= \bigcup\{E \in U/B : Pr(X_k|E) \in (0,1)\} \end{aligned} \tag{9}$$

$\mathcal{POS}_B(X_k)$ is an area of the universe where the occurrence of X_k is *certain*. $\mathcal{NEG}_B(X_k)$ covers an area where the occurrence of X_k is *impossible*. Finally, $\mathcal{BND}_B(X_k)$ defines an area where the occurrence of X_k is *possible* but *uncertain*. The boundary area typically covers large portion of the universe, if not all. If $\mathcal{BND}_B(X_k) = \emptyset$, then X_k is *B-definable*. Otherwise, X_k is a *B-rough set*.

Example 4. For $\mathbb{A} = (U, A \cup \{d\})$ from Fig. 1 and $B = \{a_1, a_3\}$, we obtain

$$\begin{aligned} \mathcal{POS}_B(X_1) &= \emptyset \\ \mathcal{NEG}_B(X_1) &= \{u_2, u_5, u_{10}, u_{14}, u_{18}\} \\ \mathcal{BND}_B(X_1) &= \{u_1, u_3, u_4, u_6, u_7, u_8, u_9, u_{11}, u_{12}, , u_{13}, u_{15}, u_{16}, u_{17}, u_{19}, u_{20}\} \end{aligned}$$

As we can see, X_1 is a B-rough set in this case. □

The following basic result emphasizes the decision-making background behind rough sets. To be sure (enough) about X_k we must be convinced (enough) against any alternative possibility X_l, $l \neq k$. This is a feature we would like to keep in mind while discussing extensions of the original rough set model, especially when the word "*enough*" becomes to have a formal mathematical meaning.

Proposition 1. *Let $\mathbb{A} = (U, A \cup \{d\})$, $X_k \subseteq U$, and $B \subseteq A$ be given. We have equality*

$$\mathcal{POS}_B(X_k) = \bigcap_{l:l \neq k} \mathcal{NEG}_B(X_l) \tag{10}$$

Proof. This is because $Pr(X_k|E) = 1$ holds, if and only if for every X_l, $l \neq k$, there is $Pr(X_l|E) = 0$. □

The RS-regions can be interpreted also by means of the inverse probabilities, which was not discussed so far in the literature. We formulate it as a theorem to emphasize its intuitive importance, although the proof itself is trivial. In particular, this result will guide us towards drawing a connection between rough sets and the Bayes factor based testing described in Subsection 2.3.

Theorem 1. *Let $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. Let the postulate (3) be satisfied. Consider the k-th decision class $X_k \subseteq U$. For any $E \in U/B$ we obtain the following characteristics:*

$$\begin{array}{ll} E \subseteq \mathcal{POS}_B(X_k) & \Leftrightarrow \forall_{l:\, l\neq k} Pr(E|X_l) = 0 \\ E \subseteq \mathcal{NEG}_B(X_k) & \Leftrightarrow Pr(E|X_k) = 0 \\ E \subseteq \mathcal{BND}_B(X_k) & \Leftrightarrow Pr(E|X_k) > 0 \wedge \exists_{l:\, l\neq k} Pr(E|X_l) > 0 \end{array} \tag{11}$$

Proof. Beginning with the positive region, we have

$$\forall_{l:\, l\neq k} Pr(E|X_l) = 0 \Leftrightarrow \forall_{l:\, l\neq k} Pr(X_l|E) = \frac{Pr(E|X_l)Pr(X_l)}{Pr(E)} = 0$$

Since $\sum_{l=0}^{r-1} Pr(X_l|E) = 1$, the above is equivalent to $Pr(X_k|E) = 1$. For the negative region we have

$$Pr(E|X_k) = 0 \Leftrightarrow Pr(X_k|E) = \frac{Pr(E|X_k)Pr(X_k)}{Pr(E)} = 0$$

Finally, for the boundary region, we can see that

$$\begin{array}{l} Pr(E|X_k) > 0 \Leftrightarrow Pr(X_k|E) > 0 \\ \exists_{l:\, l\neq k} Pr(E|X_l) > 0 \Leftrightarrow \exists_{l:\, l\neq k} Pr(X_l|E) > 0 \Leftrightarrow Pr(X_k|E) < 1 \end{array}$$

All the above equivalences follow from the postulate (3), the Bayes rule, and the fact that probability distributions sum up to 1. □

Remark 3. The formula for $\mathcal{POS}_B(X_k)$ can be also rewritten as follows:

$$E \subseteq \mathcal{POS}_B(X_k) \Leftrightarrow Pr(E|X_k) > 0 \wedge \forall_{l:\, l\neq k} Pr(E|X_l) = 0 \tag{12}$$

$Pr(E|X_k) > 0$ is redundant since conditions (3) and equalities $Pr(E|X_l) = 0$, $l \neq k$, force it anyway. However, the above form including $Pr(E|X_k) > 0$ seems to be more intuitive. □

Theorem 1 enables us to think about the rough set regions as follows (please note that interpretation of the positive region is based on characteristics (12)):

1. Object u belongs to $\mathcal{POS}_B(X_k)$, if and only if the vector $B(u) \in V_B$ is *likely* to occur under the assumption that u supports the event X_k and *unlikely* to occur under the assumption that it supports any alternative event X_l, $l \neq k$.

2. Object u belongs to $\mathcal{NEG}_B(X_k)$, if and only if the vector $B(u) \in V_B$ is *unlikely* to occur under the assumption that u supports the event X_k.
3. Object u belongs to $\mathcal{BND}_B(X_k)$, if and only if the vector $B(u) \in V_B$ is *likely* to occur under the assumption that u supports X_k but this is also the case for some alternative events X_l, $l \neq k$.

As a conclusion, the rough set model can be formulated without using the prior and posterior probabilities. It means that in case of rough sets we do not need any kind of background knowledge even if the only probabilities reasonably represented in data are the inverse ones. The rough set regions are not influenced by the changes of the prior probabilities. We do not even need the existence of those probabilities – postulate (3) could be then read as a requirement that every decision class under consideration is supported by some objects and that some alternative decisions are supported as well.

3.2 Variable Precision Rough Set Model

Although presented by means of probabilities in the previous subsection, the rough set regions were originally defined using simple set theoretic notions, namely inclusion (for positive regions) and empty intersection (for negative regions). Probabilities then occurred in various works [2, 13, 18, 21–24] to enable the initial rough set model to deal more flexibly with the indiscernibility classes *almost included* and *almost excluded* from the target events. In other words, one can use the probabilities to soften the requirements for certainty and impossibility in the rough set model. It provides better applicability to practical problems, where even a slight increase or decrease of probabilities can be as important as expecting them to equal 1 or 0.

The first method using non-0-1 posterior probabilities in rough sets is the *Variable Precision Rough Set* (*VPRS*) model [24]. It is based on parameter-controlled grades of the posterior probabilities in defining the approximation regions. The most general asymmetric VPRS model definition relies on the values of the *lower* and *upper limit certainty thresholds* α and β[1]. To deal with systems with many decision classes, we will understand α and β as vectors

$$\alpha = (\alpha_0, \ldots, \alpha_{r-1}) \text{ and } \beta = (\beta_0, \ldots, \beta_{r-1}) \tag{13}$$

where α_k and β_k refer to decision classes X_k, $k = 0, \ldots, r-1$, $r = |V_d|$. Let system $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. The *VPRS-regions* are defined as follows:

$$\begin{aligned} \mathcal{POS}_B^{\beta}(X_k) &= \bigcup\{E \in U/B : Pr(X_k|E) \geq \beta_k\} \\ \mathcal{NEG}_B^{\alpha}(X_k) &= \bigcup\{E \in U/B : Pr(X_k|E) \leq \alpha_k\} \\ \mathcal{BND}_B^{\alpha,\beta}(X_k) &= \bigcup\{E \in U/B : Pr(X_k|E) \in (\alpha_k, \beta_k)\} \end{aligned} \tag{14}$$

[1] Originally, the notation l, u was proposed for the lower and upper certainty thresholds. We use α, β instead to avoid coincidence with notation for decision classes, where l may occur as the index, and with notation for elements of the universe, often denoted by $u \in U$.

The β-positive region $\mathcal{POS}_B^{\beta}(X_k)$ is defined by the upper limit parameter β_k, which reflects the *least acceptable* degree of $Pr(X_k|E)$. Intuitively, β_k represents the desired level of improved prediction accuracy when predicting the event X_k based on the information that event E occurred. The α-negative region $\mathcal{NEG}_B^{\alpha}(X_k)$ is controlled by the lower limit parameter α_k. It is an area where the occurrence of the set X_k is *significantly*, as expressed in terms of α_k, less likely than *usually*. Finally, the (α, β)-boundary region $\mathcal{BND}_B^{\alpha,\beta}(X_k)$ is a "gray" area where there is no sufficient bias towards neither X_k nor its complement.

As proposed in [25], we suggest the following inequalities to be satisfied while choosing the VPRS parameters for particular decision systems:

$$0 \leq \alpha_k < Pr(X_k) < \beta_k \leq 1 \tag{15}$$

The reason lays in interpretation of the VPRS-regions. In case of $\mathcal{POS}_B^{\beta}(X_k)$, the improvement of prediction accuracy is possible only if $\beta_k > Pr(X_k)$. In case of $\mathcal{NEG}_B^{\alpha}(X_k)$, the word "*usually*" should be understood as the prior probability $Pr(X_k)$. Therefore, we should be sure to choose $\alpha_k < Pr(X_k)$ to obtain practically meaningful results.

Another explanation of (15) is that without it we could obtain $E \in U/B$ contained in negative or positive VPRS-regions of *all* decision classes in the same time. This is obviously an extremely unwanted situation since we should not be allowed to verify negatively all hypotheses in the same time. We could be uncertain about all the decision classes, which would correspond to the boundary regions equal to U for all decision classes, but definitely not negatively (positively) convinced about all of them.

Remark 4. The above explanation of postulate (15) should be followed by recalling the meaning of Proposition 1 in the previous subsection. Here, it should be connected with the following *duality* property of the VPRS regions [24, 25]: For $\mathbb{A} = (U, A \cup \{d\})$, $V_d = \{0, 1\}$, let us consider the limits satisfying equalities

$$\alpha_0 + \beta_1 = \alpha_1 + \beta_0 = 1 \tag{16}$$

Then we have the following identities:

$$\mathcal{POS}_B^{\beta}(X_0) = \mathcal{NEG}_B^{\alpha}(X_1) \text{ and } \mathcal{POS}_B^{\beta}(X_1) = \mathcal{NEG}_B^{\alpha}(X_0) \tag{17}$$

Further, equations (16) can be satisfied consistently with (15). This is because $0 \leq \alpha_0 < Pr(X_0) < \beta_0 \leq 1$ is equivalent to $1 \geq \beta_1 > Pr(X_1) > \alpha_1 \geq 0$. □

Identities (17) are important for understanding the nature of rough-set-like decision-making and its correspondence to the statistical hypothesis testing. It would be desirable to extend them onto the case of more than two decision classes, although it is not obvious how to approach it.

3.3 Further Towards the Inverse Probabilities

In the context of machine learning, the VPRS model's ability to flexibly control approximation regions' definitions allows for efficient capturing probabilistic relations existing in data. However, as we discussed before, the estimates of the posterior probabilities are not always reliable. Below we rewrite VPRS in terms of the inverse probabilities, just like we did in case of the original RS-regions.

Proposition 2. *Let $\mathbb{A} = (U, A \cup \{d\})$, $V_d = \{0, 1\}$, and $B \subseteq A$ be given. Consider parameters $\alpha = (\alpha_0, \alpha_1)$, $\beta = (\beta_0, \beta_1)$ such that conditions (15) and (16) are satisfied. Then we have inequalities*

$$\begin{array}{l} \mathcal{POS}_B^\beta(X_0) = \mathcal{NEG}_B^\alpha(X_1) = \bigcup\{E \in U/B : Pr(E|X_1) \leq \varepsilon_0^1 Pr(E|X_0)\} \\ \mathcal{POS}_B^\beta(X_1) = \mathcal{NEG}_B^\alpha(X_0) = \bigcup\{E \in U/B : Pr(E|X_0) \leq \varepsilon_1^0 Pr(E|X_1)\} \end{array} \tag{18}$$

where coefficients $\varepsilon_0^1, \varepsilon_1^0$ defined as

$$\varepsilon_0^1 = \frac{\alpha_1 Pr(X_0)}{\beta_0 Pr(X_1)} \quad and \quad \varepsilon_1^0 = \frac{\alpha_0 Pr(X_1)}{\beta_1 Pr(X_0)} \tag{19}$$

belong to the interval $[0, 1)$.

Proof. Consider α_0 and β_1 such that $\alpha_0 + \beta_1 = 1$ (the case of α_{01} and β_0 can be shown analogously). We want to prove

$$Pr(X_1|E) \geq \beta_1 \Leftrightarrow \alpha_0 \geq Pr(X_0|E) \Leftrightarrow Pr(E|X_0) \leq \frac{\alpha_0 Pr(X_1)}{\beta_1 Pr(X_0)} Pr(E|X_1) \tag{20}$$

We know that two first above inequalities are equivalent. By combining them together, we obtain the third equivalent inequality (its equivalence to both $Pr(X_1|E) \geq \beta_1$ and $Pr(X_0|E) \leq \alpha_0$ can be easily shown by contradiction):

$$Pr(X_1|E) \geq \beta_1 \Leftrightarrow \alpha_0 \geq Pr(X_0|E) \Leftrightarrow \alpha_0 Pr(X_1|E) \geq \beta_1 Pr(X_0|E) \tag{21}$$

It is enough to apply identity (2) to realize that the third inequalities in (20) and (21) are actually the same ones. It remains to show that $\varepsilon_0^1, \varepsilon_1^0 \in [0, 1)$. It follows from the assumption (15). For instance, we have inequality $\varepsilon_0^1 < 1$ because $\alpha_1 < Pr(X_1)$ and $Pr(X_0) < \beta_0$. □

The above correspondence can be used to draw a connection between VPRS and the statistical reasoning models. It is possible to refer inequality $Pr(E|X_0) \leq \varepsilon_1^0 Pr(E|X_1)$, rewritten as

$$B_0^1 \geq \frac{\beta_1 Pr(X_0)}{\alpha_0 Pr(X_1)} \tag{22}$$

to the Bayes factor based statistical principles discussed e.g. in [20]. However, the remaining problem is that we need to use $Pr(X_0)$ and $Pr(X_1)$ explicitly in (22), which is often too questionable from practical point of view.

In [18], another version of VPRS is considered. The idea is to detect *any* decrease/increase of belief in decision classes. The rough set region definitions proposed in [18] look as follows:

$$\begin{array}{ll} \mathcal{POS}_B^*(X_1) & = \bigcup\{E \in U/B : Pr(X_1|E) > Pr(X_1)\} \\ \mathcal{NEG}_B^*(X_1) & = \bigcup\{E \in U/B : Pr(X_1|E) < Pr(X_1)\} \\ \mathcal{BND}_B^*(X_1) & = \bigcup\{E \in U/B : Pr(X_1|E) = Pr(X_1)\} \end{array} \tag{23}$$

which can be equivalently expressed as follows (cf. [19]):

$$\begin{aligned}\mathcal{POS}^*_B(X_1) &= \bigcup\{E \in U/B : Pr(E|X_1) > Pr(E|X_0)\}\\ \mathcal{NEG}^*_B(X_1) &= \bigcup\{E \in U/B : Pr(E|X_1) < Pr(E|X_0)\}\\ \mathcal{BND}^*_B(X_1) &= \bigcup\{E \in U/B : Pr(E|X_1) = Pr(E|X_0)\}\end{aligned} \quad (24)$$

This simple interpretation resembles the VPRS characteristics provided by Proposition 2, for ε_0^1 and ε_1^0 tending to 1. It also corresponds to the limit $\varepsilon_1^0 \rightarrow 1$ applied to inequality $Pr(E|X_0) \leq \varepsilon Pr(E|X_1)$ in the Bayes factor criterion (8). We could say that according to the scale illustrated by Fig. 2 in Subsection 2.3 the region $\mathcal{POS}^*_B(X_1)$ gathers any, even *barely worth mentioning but still positive*, evidence for X_1. It is completely opposite to the original rough set model. Indeed $\mathcal{POS}_B(X_1)$ gathers, according to Theorem 1, only *infinitely strong* evidence for X_1. Let us summarize it as follows:

1. Object u belongs to $\mathcal{POS}_B(X_1)$ (to $\mathcal{POS}^*_B(X_1)$), if and only if X_1 can be *positively* verified under the evidence of $B(u)$ at the maximal (minimal) level of statistical significance, expressed by (8) for $\varepsilon_1^0 = 0$ ($\varepsilon_1^0 \rightarrow 1$).
2. Object u belongs to $\mathcal{NEG}_B(X_1)$ (to $\mathcal{NEG}^*_B(X_1)$), if and only if X_1 can be *negatively* verified under the evidence of $B(u)$ at the maximal (minimal) level of significance (we replace X_0 and X_1 and use ε_0^1 instead of ε_1^0 in (8)).
3. Object u belongs to $\mathcal{BND}_B(X_1)$ (to $\mathcal{BND}^*_B(X_1)$), if and only if it is not sufficient to verify X_1 *neither positively nor negatively* at the maximal (minimal) level of significance under the evidence of $B(u)$.

As a result, we obtain two models – the original rough set model and its modification proposed in [18] – which refer to the Bayes factor scale in two marginal ways. They also correspond to special cases of VPRS, as it is rewritable by means of the inverse probabilities following Proposition 2. They both do not need to base on the prior or posterior probabilities, according to characteristics (11) and (24). From this perspective, the main challenge of this article is to fill the gap between these two opposite cases of involving the Bayes factor based methodology into the theory of rough sets. An additional challenge is to extend the whole framework to be able to deal with more than two target events, as it was stated by Theorem 1 in case of the original RS-regions.

4 Rough Bayesian Model

4.1 RB for Two Decision Classes

After recalling basic methods for extracting probabilities from data and the VPRS-like extensions of rough sets, we are ready to introduce a novel extension based entirely on the inverse probabilities and the Bayes factor. To prepare the background, let us still restrict to systems with two decision classes. Using statistical terminology, we interpret classes X_1 and X_0 as corresponding to the *positive* and *negative verification* of some *hypothesis*.

Let us refer to the above interpretation of the RS-regions originating from substitution of $\varepsilon_1^0 = 0$ to the criterion (8). By using positive values of ε_1^0, we can soften the requirements for the positive/negative verification. In this way

$\varepsilon_1^0 \in [0,1)$ plays a role of a *degree of the significance approximation*. We propose the following model related to this degree. We will refer to this model as to the *Rough Bayesian* model because of its relationship to Bayes factor (cf. [17]).

Definition 1. *Let* $\mathbb{A} = (U, A \cup \{d\})$*,* $V_d = \{0,1\}$*, and* $B \subseteq A$ *be given. For any parameters* $\varepsilon = (\varepsilon_0^1, \varepsilon_1^0)$*,* $\varepsilon_0^1, \varepsilon_1^0 \in [0,1)$*, we define the* B-positive*,* B-negative*, and* B-boundary rough Bayesian regions *(abbreviated as* RB-regions*) as follows (the regions for* X_0 *are defined analogously):*

$$\begin{aligned} \mathcal{BAYPOS}_B^{\varepsilon}(X_1) &= \bigcup\{E \in U/B : Pr(E|X_0) \le \varepsilon_1^0 Pr(E|X_1)\} \\ \mathcal{BAYNEG}_B^{\varepsilon}(X_1) &= \bigcup\{E \in U/B : Pr(E|X_1) \le \varepsilon_0^1 Pr(E|X_0)\} \\ \mathcal{BAYBND}_B^{\varepsilon}(X_1) &= \bigcup\{E \in U/B : Pr(E|X_0) > \varepsilon_1^0 Pr(E|X_1) \wedge \\ &\qquad\qquad\qquad\quad Pr(E|X_1) > \varepsilon_0^1 Pr(E|X_0)\} \end{aligned} \tag{25}$$

Remark 5. The choice of ε_0^1 and ε_1^0 is a challenge comparable to the case of other parameter-controlled models, e.g. VPRS based on the threshold vectors α and β. It is allowed to put $\varepsilon_0^1 = \varepsilon_1^0$ and use a common notation $\varepsilon \in [0,1)$ for both coefficients. It obviously simplifies (but does not solve) the problem of parameter tuning. Further discussion with that respect is beyond the scope of this particular article and should be continued in the nearest future. □

Remark 6. As in Subsection 2.3, we prefer not to use the Bayes factor ratio explicitly because of the special case of zero probabilities. However, if we omit this case, we can rewrite the RB positive/negative/boundary regions using inequalities $B_0^1 \ge 1/\varepsilon_1^0$, $B_1^0 \ge 1/\varepsilon_0^1$, and $\max\{B_1^0\varepsilon_0^1, B_0^1\varepsilon_1^0\} < 1$, respectively, where $B_1^0 = Pr(E|X_0)/Pr(E|X_1)$ and $B_1^0 = Pr(E|X_1)/Pr(E|X_0)$. □

Proposition 3. *For* $\varepsilon = (0,0)$*, the RB-regions are identical with the RS-regions.*

Proof. Derivable directly from Theorem 1[2]. □

Below we provide possibly simplest way of understanding the RB-regions:

1. Object u belongs to $\mathcal{BAYPOS}_B^{\varepsilon}(X_1)$, if and only if $B(u)$ is *significantly* (up to ε_1^0) *more likely* to occur under X_1 than under alternative hypothesis X_0.
2. Object u belongs to $\mathcal{BAYNEG}_B^{\varepsilon}(X_1)$, if and only if the alternative hypothesis X_0 makes $B(u)$ *significantly more likely* (up to ε_0^1) than X_1 does.
3. Object u belongs to $\mathcal{BAYBND}_B^{\varepsilon}(X_1)$, if and only if it is not *significantly more likely* under X_1 than under X_0 but also X_0 does not make $B(u)$ *significantly more likely* than X_1 does.

Another interpretation refers to identity (2). It shows that by using condition (8) we actually require that the increase of belief in X_0 given E, expressed by $Pr(X_0|E)/Pr(X_0)$, should be ε*-negligibly small* with respect to the increase of belief in X_1, that is that $Pr(X_0|E)/Pr(X_0) \le \varepsilon_1^0 Pr(X_1|E)/Pr(X_1)$. According

[2] Although we refer here to the special case of two decision classes, the reader can verify that this proposition is also true for more general case discussed in the next subsection.

to yet another, strictly Bayesian interpretation, we are in $\mathcal{BAYPOS}_B^{\varepsilon}(X_1)$, if and only if the *posterior odds* $Pr(X_1|E)/Pr(X_0|E)$ are ε_1^0-significantly greater than the *prior odds* $Pr(X_1)/Pr(X_0)$. Identity (2) also shows that we do not need neither $Pr(X_k|E)$ nor $Pr(X_k)$ while comparing the above changes in terms of the belief gains and/or the prior and posterior odds.

The Rough Bayesian model enables us to test the target events directly against each other. For ε_1^0 tending to 1, we can replace $Pr(E|X_0) \leq \varepsilon_1^0 Pr(E|X_1)$ by $Pr(E|X_0) < Pr(E|X_1)$, as considered in Subsection 3.3. Also, across the whole range of $\varepsilon_1^0 \in [0,1)$, we obtain the following characteristics, complementary to Proposition 2:

Proposition 4. *Let* $\varepsilon = (\varepsilon_0^1, \varepsilon_1^0)$, $\varepsilon_0^1, \varepsilon_1^0 \in [0,1)$, *and* $\mathbb{A} = (U, A \cup \{d\})$ *with* $V_d = \{0,1\}$ *be given. The RB-regions are identical with the VPRS-regions for the following parameters:*

$$\alpha_0^{\varepsilon} = \frac{\varepsilon_1^0 Pr(X_0)}{\varepsilon_1^0 Pr(X_0) + Pr(X_1)} \quad \text{and} \quad \beta_0^{\varepsilon} = \frac{Pr(X_0)}{Pr(X_0) + \varepsilon_0^1 Pr(X_1)}$$

$$\alpha_1^{\varepsilon} = \frac{\varepsilon_0^1 Pr(X_1)}{\varepsilon_0^1 Pr(X_1) + Pr(X_0)} \quad \text{and} \quad \beta_1^{\varepsilon} = \frac{Pr(X_1)}{Pr(X_1) + \varepsilon_1^0 Pr(X_0)}$$

Proof. Let $B \subseteq A$ and $E \in U/B$ be given. We have to show the following:

$$\begin{array}{l} Pr(E|X_0) \leq \varepsilon_1^0 Pr(E|X_1) \Leftrightarrow Pr(X_1|E) \geq \beta_1^{\varepsilon} \Leftrightarrow Pr(X_0|E) \leq \alpha_0^{\varepsilon} \\ Pr(E|X_1) \leq \varepsilon_0^1 Pr(E|X_0) \Leftrightarrow Pr(X_1|E) \leq \alpha_1^{\varepsilon} \Leftrightarrow Pr(X_0|E) \geq \beta_0^{\varepsilon} \end{array} \tag{26}$$

Let us show, for example (the rest is analogous), that

$$Pr(E|X_0) \leq \varepsilon_1^0 Pr(E|X_1) \Leftrightarrow Pr(X_1|E) \geq \frac{Pr(X_1)}{Pr(X_1) + \varepsilon_1^0 Pr(X_0)} \tag{27}$$

Using the Bayes rule we rewrite the right above inequality as follows:

$$\frac{Pr(E|X_1)Pr(X_1)}{Pr(E|X_1)Pr(X_1) + Pr(E|X_0)Pr(X_0)} \geq$$

$$\geq \frac{Pr(E|X_1)Pr(X_1)}{Pr(E|X_1)Pr(X_1) + \varepsilon_1^0 Pr(E|X_1)Pr(X_0)}$$

The only difference is now between the term $Pr(E|X_0)$ at the left side and the term $\varepsilon_1^0 Pr(E|X_1)$ at the right side. Hence, (27) becomes clear. □

Example 5. Let us consider the data table from Fig. 1, for $B = \{a_1, a_3\}$ and $\varepsilon_0^1 = \varepsilon_1^0 = 1/5$. We obtain the following RB-regions:

$$\begin{array}{ll} \mathcal{BAYPOS}_B^{1/5}(X_1) &= \{u_3, u_6, u_8, u_9, u_{15}, u_{19}, u_{20}\} \\ \mathcal{BAYNEG}_B^{1/5}(X_1) &= \{u_2, u_5, u_{10}, u_{14}, u_{18}\} \\ \mathcal{BAYBND}_B^{1/5}(X_1) &= \{u_1, u_4, u_7, u_{11}, u_{12}, u_{13}, u_{16}, u_{17}\} \end{array}$$

In comparison to the original RS-regions, the case of $a_1 = 2$ and $a_3 = 2$ starts to support the B-positive RB-region of X_1. If we can assume that $Pr(X_1) = 1/2$, as derivable from the considered data table, then we obtain analogous result in terms of the VPRS-regions for

$$\alpha_1^{1/5} = \frac{1/5 \cdot 1/2}{1/5 \cdot 1/2 + (1 - 1/2)} = \frac{1}{6} \text{ and } \beta_1^{1/5} = \frac{1/2}{1/2 + 1/5(1 - 1/2)} = \frac{5}{6}$$

In particular, for $E = \{u \in U : a_1(u) = 2 \wedge a_3(u) = 2\}$, we get $Pr(X_1|E) = 6/7$ which is more than $\beta_1^{1/5} = 5/6$. □

In this way, the Rough Bayesian model refers to the VPRS idea of handling the posterior probabilities. We can see that coefficients $\alpha_k^\varepsilon, \beta_k^\varepsilon$, $k = 0, 1$, satisfy assumption (15). For instance we have

$$0 \leq \frac{\varepsilon_1^0 Pr(X_0)}{\varepsilon_1^0 Pr(X_0) + Pr(X_1)} < Pr(X_0) < \frac{Pr(X_0)}{Pr(X_0) + \varepsilon_0^1 Pr(X_1)} \leq 1$$

where inequalities hold, if and only if $\varepsilon_0^1, \varepsilon_1^0 \in [0, 1)$. The property (17) is satisfied as well, e.g.:

$$\frac{\varepsilon_1^0 Pr(X_0)}{\varepsilon_1^0 Pr(X_0) + Pr(X_1)} + \frac{Pr(X_1)}{Pr(X_1) + \varepsilon_1^0 Pr(X_0)} = 1$$

We can summarize the obtained results as follows:

Theorem 2. *Let $\mathbb{A} = (U, A \cup \{d\})$, $V_d = \{0, 1\}$, and $B \subseteq A$ be given. The VPRS and RB models are equivalent in the following sense:*

1. *For any $\alpha = (\alpha_0, \alpha_1)$ and $\beta = (\beta_0, \beta_1)$ satisfying (15) and (16), there exists $\varepsilon(\alpha, \beta) \in [0, 1) \times [0, 1)$ such that for $k = 0, 1$ we have*

$$\begin{aligned} \mathcal{BAYPOS}_B^{\varepsilon(\alpha,\beta)}(X_k) &= \mathcal{POS}_B^{\beta}(X_k) \\ \mathcal{BAYNEG}_B^{\varepsilon(\alpha,\beta)}(X_k) &= \mathcal{NEG}_B^{\alpha}(X_k) \\ \mathcal{BAYBND}_B^{\varepsilon(\alpha,\beta)}(X_k) &= \mathcal{BND}_B^{\alpha,\beta}(X_k) \end{aligned}$$

2. *For any $\varepsilon \in [0, 1) \times [0, 1)$, there exist $\alpha(\varepsilon)$ and $\beta(\varepsilon)$ satisfying (15) and (16) such that for $k = 0, 1$ we have*

$$\begin{aligned} \mathcal{POS}_B^{\beta(\varepsilon)}(X_k) &= \mathcal{BAYPOS}_B^{\varepsilon}(X_k) \\ \mathcal{NEG}_B^{\alpha(\varepsilon)}(X_k) &= \mathcal{BAYNEG}_B^{\varepsilon}(X_k) \\ \mathcal{BND}_B^{\alpha(\varepsilon),\beta(\varepsilon)}(X_k) &= \mathcal{BAYBND}_B^{\varepsilon}(X_k) \end{aligned}$$

Proof. Derivable directly from Propositions 2 and 4. □

It is important to remember that Theorem 2 holds only for $V_d = \{0, 1\}$. We will address more general case in the next subsections. For now, given $V_d = \{0, 1\}$, let us note that the advantage of the RB model with respect to VPRS is that any change of $Pr(X_1)$ results in automatic change of the lower and upper VPRS thresholds. It can be illustrated as follows:

Example 6. Let us continue the previous example but for $Pr(X_1) = 1/1000$, as if X_1 corresponded to a rare medical pathology discussed in Example 3. There is no sense to keep the upper limit for $Pr(X_1|E)$ equal to 5/6, so the VPRS parameters should be changed. However, there is no change required if we rely on the RB-regions. With the same $\varepsilon_0^1 = \varepsilon_1^0 = 1/5$ we simply get different interpretation in terms of the posterior probabilities. Namely, we recalculate the VPRS degrees as

$$\alpha_1^{1/5} = \frac{1/5 \cdot 1/1000}{1/5 \cdot 1/1000 + (1 - 1/1000)} \approxeq \frac{1}{5000} \quad \text{and} \quad \beta_1^{1/5} \approxeq \frac{1}{200}$$

One can see that this time $\beta_1^{1/5}$ would have nothing in common with previously calculated $Pr(X_1|E) = 6/7$. However, using standard estimation $Pr(X_1|E) = |X_k \cap E|/|E|$ is not reasonable in this situation. We should rather use the Bayes rule leading to the following result:

$$Pr(d = 1|a_1 = 2, a_3 = 2) = \frac{3/5 \cdot 1/1000}{1/10 \cdot (1 - 1/1000) + 3/5 \cdot 1/1000} = \frac{2}{335}$$

The posterior probability becomes then to be referrable to $\beta_1^{1/5}$. □

As a result, we obtain a convenient method of defining the rough-set-like regions based on the inverse probabilities, which – if necessary – can be translated onto the parameters related to more commonly used posterior probabilities. However, the Rough Bayesian model can be applied also when such translation is impossible, that is when the prior probabilities are unknown or even undefinable. The RB-regions have excellent statistical interpretation following from their connections with the Bayes factor. Actually, we obtain a kind of *variable significance rough set model*, as it is parameterized by the significance thresholds $\varepsilon = (\varepsilon_0^1, \varepsilon_1^0)$. The choice of ε refers to the choice of significance levels illustrated by Fig. 2, Subsection 2.3. We can draw a direct connection between the RB-regions and particular states of the statistical verification process. We can also base on statistical apparatus while tuning $\varepsilon = (\varepsilon_0^1, \varepsilon_1^0)$, with two important special cases – the original rough set model for $\varepsilon_0^1 = \varepsilon_1^0 = 0$ and the model introduced in [18] for $\varepsilon_0^1, \varepsilon_1^0$ tending to 1.

4.2 RB for More Decision Classes

The way of comparing the inverse probabilities in Definition 1 has a natural extension onto the case of more decision classes. Below we reconsider the Rough Bayesian model for such a situation. Please note that the regions from Definition 1 are the special cases of the following ones.

Definition 2. *Let $\mathbb{A} = (U, A \cup \{d\})$, $V_d = \{0, \ldots, r-1\}$, and $B \subseteq A$ be given. Consider matrix*

$$\varepsilon = \begin{bmatrix} * & \varepsilon_1^0 & \ldots & \varepsilon_{r-1}^0 \\ \varepsilon_0^1 & * & & \vdots \\ \vdots & & * & \varepsilon_{r-1}^{r-2} \\ \varepsilon_0^{r-1} & \ldots & \varepsilon_{r-2}^{r-1} & * \end{bmatrix} \tag{28}$$

U	a_1	a_2	a_3	a_4	a_5	d
u_1	1	1	0	1	2	0
u_2	0	0	0	2	2	2
u_3	2	2	2	1	1	1
u_4	0	1	2	2	2	1
u_5	2	1	1	0	2	0
u_6	2	2	2	1	1	1
u_7	0	1	2	2	2	2
u_8	2	2	2	1	1	1
u_9	2	2	2	1	1	1
u_{10}	0	0	0	2	2	0

U	a_1	a_2	a_3	a_4	a_5	d
u_{11}	1	2	0	0	2	0
u_{12}	1	1	0	1	2	1
u_{13}	0	1	2	2	2	1
u_{14}	2	1	1	0	2	0
u_{15}	2	2	2	1	1	2
u_{16}	1	1	0	1	2	1
u_{17}	1	1	0	1	2	2
u_{18}	2	1	1	0	2	2
u_{19}	2	2	2	1	1	1
u_{20}	2	2	2	1	1	1

Fig. 3. $\mathbb{A} = (U, A \cup \{d\})$, $U = \{u_1, \ldots, u_{20}\}$, $A = \{a_1, \ldots, a_5\}$. Decision classes: $X_0 = \{u_1, u_5, u_{10}, u_{11}, u_{15}\}$, $X_1 = \{u_3, u_4, u_6, u_8, u_9, u_{12}, u_{13}, u_{16}, u_{19}, u_{20}\}$, and $X_2 = \{u_2, u_7, u_{14}, u_{17}, u_{18}\}$.

where $\varepsilon_k^l \in [0,1)$*, for* $k \neq l$*. We define the* B-positive*,* B-negative*, and* B-boundary RB-regions *as follows:*

$$\begin{array}{ll} \mathcal{BAYPOS}_B^{\varepsilon}(X_k) &= \bigcup\{E \in U/B : \forall_{l:\, l\neq k} Pr(E|X_l) \leq \varepsilon_k^l Pr(E|X_k)\} \\ \mathcal{BAYNEG}_B^{\varepsilon}(X_k) &= \bigcup\{E \in U/B : \exists_{l:\, l\neq k} Pr(E|X_k) \leq \varepsilon_l^k Pr(E|X_l)\} \\ \mathcal{BAYBND}_B^{\varepsilon}(X_k) &= \bigcup\{E \in U/B : \exists_{l:\, l\neq k} Pr(E|X_l) > \varepsilon_k^l Pr(E|X_k) \wedge \\ & \qquad\qquad\qquad\quad \forall_{l:\, l\neq k} Pr(E|X_k) > \varepsilon_l^k Pr(E|X_l)\} \end{array} \tag{29}$$

Remark 7. As in Remark 6, we could use respectively conditions $\min_{l:l\neq k} B_l^k \geq 1/\varepsilon_k^l$, $\max_{l:l\neq k} B_k^l \geq 1/\varepsilon_l^k$, and $\max\left\{\min_{l:l\neq k} B_l^k \varepsilon_k^l, \max_{l:l\neq k} B_k^l \varepsilon_l^k\right\} < 1$, where the ratios B_l^k are defined by (6). The only special case to address would correspond to the zero inverse probabilities. □

Let us generalize the previous interpretation of the RB-regions as follows:

1. Object u belongs to $\mathcal{BAYPOS}_B^{\varepsilon}(X_k)$, if and only if $B(u)$ is *significantly more likely* to occur under X_k than under any other hypothesis X_l, $l \neq k$.
2. Object u belongs to $\mathcal{BAYNEG}_B^{\varepsilon}(X_k)$, if and only if there is an alternative hypothesis X_l, which makes $B(u)$ *significantly more likely* than X_k does.
3. Object u belongs to $\mathcal{BAYBND}_B^{\varepsilon}(X_k)$, if and only if $B(u)$ is not *significantly more likely* under X_k than under all other X_l but there is also no alternative hypothesis, which makes $B(u)$ *significantly more likely* than X_k does.

Remark 8. As in case of two decision classes, we can consider a simplified model with $\varepsilon_k^l = \varepsilon$, for every $k, l = 0, \ldots, r-1$, $k \neq l$. Appropriate tuning of many different parameters in the matrix (28) could be difficult technically, especially for large $r = |V_d|$. The examples below show that the multi-decision RB model has a significant expressive power even for one unified $\varepsilon \in [0,1)$. Therefore, we are going to put a special emphasis on this case in the future applications. □

Example 7. Fig. 3 illustrates decision system $\mathbb{A} = (U, A \cup \{d\})$ with $V_d = \{0, 1, 2\}$. Actually, it results from splitting the objects supporting X_0 in Fig. 1

onto two 5-object parts, now corresponding to decision classes X_0 and X_2. For $B = \{a_1, a_3\}$, we have five B-indiscernibility classes. We list them below, with the corresponding inverse probabilities.

U/B	Conditions	$P(E_i\|X_0)$	$P(E_i\|X_1)$	$P(E_i\|X_2)$
E_1	$a_1 = 1,\ a_3 = 0$	2/5	1/5	1/5
E_2	$a_1 = 0,\ a_3 = 0$	1/5	0	1/5
E_3	$a_1 = 2,\ a_3 = 2$	0	3/5	1/5
E_4	$a_1 = 0,\ a_3 = 2$	0	1/5	1/5
E_5	$a_1 = 2,\ a_3 = 1$	2/5	0	1/5

Let us start with the RS-regions. We obtain the following characteristics:

Decisions	$\mathcal{POS}_B(X_k)$	$\mathcal{NEG}_B(X_k)$	$\mathcal{BND}_B(X_k)$
X_0	$\emptyset$	$E_3 \cup E_4$	$E_1 \cup E_2 \cup E_5$
X_1	$\emptyset$	$E_2 \cup E_5$	$E_1 \cup E_3 \cup E_4$
X_2	$\emptyset$	$\emptyset$	U

Now, let us consider $\varepsilon = 1/3$. We obtain the following RB-regions:

Decisions	$\mathcal{BAYPOS}_B^{1/3}(X_k)$	$\mathcal{BAYNEG}_B^{1/3}(X_k)$	$\mathcal{BAYBND}_B^{1/3}(X_k)$
X_0	$\emptyset$	$E_3 \cup E_4$	$E_1 \cup E_2 \cup E_5$
X_1	E_3	$E_2 \cup E_5$	$E_1 \cup E_4$
X_2	$\emptyset$	E_3	$U \setminus E_3$

While comparing to the RS-regions, we can see that:

1. $\mathcal{BAYPOS}_B^{1/3}(X_1)$ and $\mathcal{BAYNEG}_B^{1/3}(X_2)$ start to contain E_3.
2. $\mathcal{BAYBND}_B^{1/3}(X_1)$ and $\mathcal{BAYBND}_B^{1/3}(X_2)$ do not contain E_3 any more.

This is because we have both $P(E_3|X_0) \leq 1/3 * P(E_3|X_1)$ and $P(E_3|X_2) \leq 1/3 * P(E_3|X_1)$. It means that E_3 at least three times more likely given hypothesis X_1 than given X_0 and X_2. According to the scale proposed in [4] and presented in Subsection 2.3, we could say that E_3 is a *positive* evidence for X_1. □

As a conclusion for this part, we refer to Proposition 1 formulated for the original rough set model as an important decision-making property. We can see that the Rough Bayesian model keeps this property well enough to disallow intersections between the positive and negative RB-regions of different decision classes. We will go back to this topic in the next subsection, while discussing the VPRS model for more than two decision classes.

Proposition 5. *Let $\mathbb{A} = (U, A \cup \{d\})$, $X_k \subseteq U$, and $B \subseteq A$ be given. Consider matrix ε given by (28) for $\varepsilon_k^l \in [0, 1)$, $k \neq l$. We have the following inclusion:*

$$\mathcal{BAYPOS}_B^{\varepsilon}(X_k) \subseteq \bigcap_{l:l\neq k} \mathcal{BAYNEG}_B^{\varepsilon}(X_l) \qquad (30)$$

Moreover, if inequalities

$$\varepsilon_k^l \geq \varepsilon_k^m \varepsilon_m^l \qquad (31)$$

hold for every mutually different k, l, m, then the equality holds in (30).

Proof. Assume $E \subseteq \mathcal{BAYPOS}_B^{\varepsilon}(X_k)$ for a given $k = 0, \ldots, r-1$. Consider any X_m, $m \neq k$. We have $E \subseteq \mathcal{BAYNEG}_B^{\varepsilon}(X_m)$ because there exists $l \neq m$ such that $Pr(E|X_m) \leq \varepsilon_l^m Pr(E|X_l)$. Namely, we can choose $l = k$.

Now, let us assume that $E \subseteq \mathcal{BAYNEG}_B^{\varepsilon}(X_l)$ for every X_l, $l \neq k$. Then, for any X_{l0}, $l0 \neq k$ there must exist X_{l1}, $l1 \neq l0$, such that $Pr(E|X_{l0}) \leq \varepsilon_{l1}^{l0} Pr(E|X_{l1})$. There are two possibilities: If $l1 = k$, then we reach the goal – we wanted to show that $Pr(E|X_{l0}) \leq \varepsilon_k^{l0} Pr(E|X_k)$ for any X_{l0}, $l0 \neq k$. If $l1 \neq k$, then we continue with $l1$. Since $l1 \neq k$, there must exist X_{l2}, $l2 \neq l1$, such that $Pr(E|X_{l1}) \leq \varepsilon_{l2}^{l1} Pr(E|X_{l2})$. Given inequalities (31), we get

$$Pr(E|X_{l0}) \leq \varepsilon_{l1}^{l0} Pr(E|X_{l1}) \leq \varepsilon_{l1}^{l0} \varepsilon_{l2}^{l1} Pr(E|X_{l2}) \leq \varepsilon_{l2}^{l0} Pr(E|X_{l2})$$

Therefore, we can apply the same procedure to every next $l2$ as we did with $l1$ above. At each next step we must select a brand new decision class – this is because the ε-matrix takes the values within $[0, 1)$. Since the number of decisions is finite, we must eventually reach the moment when a new $l2$ equals k. □

Corollary 1. *Let $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. Consider the RB model with unified parameter $\varepsilon \in [0, 1)$, that is $\varepsilon_k^l = \varepsilon$ for every $k, l = 0, \ldots, r-1$, $k \neq l$. Then we have always $\mathcal{BAYPOS}_B^{\varepsilon}(X_k) = \bigcap_{l: l \neq k} \mathcal{BAYNEG}_B^{\varepsilon}(X_l)$.*

Proof. Directly from Proposition 5. □

4.3 VPRS for More Decision Classes

The question is whether the Rough Bayesian model is still rewritable in terms of the posterior probabilities, similarly to the case of two decision classes described by Theorem 2. Let us first discuss requirements for a posterior probability based rough set model in such a case. In Subsection 3.2, we used the parameter vectors $\alpha = (\alpha_0, \ldots, \alpha_{r-1})$ and $\beta = (\beta_0, \ldots, \beta_{r-1})$ satisfying condition (15). One would believe that if a unique X_k is supported strongly enough, then the remaining classes cannot be supported in a comparable degree. However, this is the case only for two complementary decision classes. If $|V_d| > 2$, then there might be two different classes X_k and X_l, $k \neq l$, satisfying inequalities $Pr(X_k|E) \geq \beta_k$ and $Pr(X_l|E) \geq \beta_l$. It would lead to supporting two decisions in the same time, which is an unwanted situation.

Example 8. Consider the decision system illustrated in Fig. 3. Please note that $Pr(X_0) = Pr(X_2) = 1/4$ and $Pr(X_1) = 1/2$. Let us choose parameters $\alpha = (1/10, 1/4, 1/10)$ and $\beta = (13/20, 3/4, 13/20)$. One can see that inequalities (15) are then satisfied. For $B = \{a_1, a_3\}$, we have five B-indiscernibility classes, as in Example 7. Their corresponding posterior probabilities look as follows:

U/B	Conditions	$P(X_0\|E_i)$	$P(X_1\|E_i)$	$P(X_2\|E_i)$
E_1	$a_1 = 1,\ a_3 = 0$	2/5	2/5	1/5
E_2	$a_1 = 0,\ a_3 = 0$	1/2	0	1/2
E_3	$a_1 = 2,\ a_3 = 2$	0	6/7	1/7
E_4	$a_1 = 0,\ a_3 = 2$	0	2/3	1/3
E_5	$a_1 = 2,\ a_3 = 1$	2/3	0	1/3

We obtain the following characteristics, if we keep using conditions (14):

Decisions	$\mathcal{POS}_B^{\beta}(X_k)$	$\mathcal{NEG}_B^{\alpha}(X_k)$	$\mathcal{BND}_B^{\alpha,\beta}(X_k)$
X_0	E_5	$E_3 \cup E_4$	$E_1 \cup E_2$
X_1	E_3	$E_2 \cup E_5$	$E_1 \cup E_4$
X_2	$\emptyset$	$\emptyset$	U

Luckily enough, we do not obtain non-empty intersections between positive regions of different decision classes. However, there is another problem visible: E_5 (E_3) is contained in the positive region of X_0 (X_1) but it is not in the negative region of X_2. It is a lack of a crucial property of the rough-set-like regions, specially emphasized by Propositions 1 and 5. □

Obviously, one could say that α and β in the above example are chosen artificially to yield the described non-empty intersection situation. However, even if this is a case, it leaves us with the problem how to improve the requirements for the VPRS parameters to avoid such situations. We suggest embedding the property analogous to those considered for the RS and RB models directly into the definition. In this way, we can also simplify the VPRS notation by forgetting about the upper grades β. This is a reason why we refer to the following model as to the *simplified* VPRS model.

Definition 3. *Let* $\mathbb{A} = (U, A \cup \{d\})$ *and* $B \subseteq A$ *be given. Consider vector* $\alpha = (\alpha_0, \ldots, \alpha_{r-1})$ *such that inequalities*

$$0 \leq \alpha_k < Pr(X_k) \tag{32}$$

are satisfied for every $k = 0, \ldots, r-1$. *The* simplified VPRS-regions *are defined as follows:*

$$\begin{array}{r}\mathcal{POS}_B^{\alpha}(X_k) = \bigcup\{E \in U/B : \forall_{l:\, l \neq k} Pr(X_l|E) \leq \alpha_l\} \\ \mathcal{NEG}_B^{\alpha}(X_k) = \bigcup\{E \in U/B : Pr(X_k|E) \leq \alpha_k\} \\ \mathcal{BND}_B^{\alpha}(X_k) = \bigcup\{E \in U/B : Pr(X_k|E) > \alpha_k \wedge \exists_{l:\, l \neq k} Pr(X_l|E) > \alpha_l\}\end{array} \tag{33}$$

Proposition 6. *Let* $\mathbb{A} = (U, A \cup \{d\})$, $X_k \subseteq U$, $B \subseteq A$, *and* $\alpha = (\alpha_0, \ldots, \alpha_{r-1})$ *be given. We have the following equality:*

$$\mathcal{POS}_B^{\alpha}(X_k) = \bigcap_{l: l \neq k} \mathcal{NEG}_B^{\alpha}(X_l) \tag{34}$$

Proof. Directly from (33). □

The form of (33) can be compared with the way we expressed the original RS-regions by (11). There, we defined the positive region by means of conditions for the negative regions of all other classes, exactly like for simplified VPRS above. Further, we can reformulate the meaning of Remark 3 as follows:

Proposition 7. *Let* $\mathbb{A} = (U, A \cup \{d\})$, $B \subseteq A$, $E \in U/B$, *and* $\alpha = (\alpha_0, ..., \alpha_{r-1})$ *satisfying (32) be given. Let us define vector* $\beta = (\beta_0, \ldots, \beta_{r-1})$ *in the following way, for every* $k = 0, \ldots, r-1$*:*

$$\beta_k = 1 - \sum_{l: l \neq k} \alpha_k \tag{35}$$

Then, for any X_k, we have the following equivalence:

$$E \subseteq \mathcal{POS}_B^{\alpha}(X_k) \Leftrightarrow Pr(X_k|E) \geq \beta_k \wedge \forall_{l:\, l \neq k} Pr(X_l|E) \leq \alpha_l\}$$

Moreover, vectors α and β satisfy together the assumption (15).

Proof. Directly based on the fact that $\sum_{l=0}^{r-1} Pr(X_l|E) = 1$. □

In this way one can see that by removing $\beta = (\beta_0, \ldots, \beta_{r-1})$ from the definition of VPRS we do not change its meaning. Vector β is fully recoverable from α using equations (35), which actually generalize postulate (16). In particular, for a special case of two decision classes, we obtain the following result.

Proposition 8. *Let $V_d = \{0, 1\}$ and let equalities (16) be satisfied. Then Definition 3 is equivalent to the original VPRS model.*

Proof. For two decision classes, given (16), conditions $Pr(X_k|E) \geq \beta_k$ and $\forall_{l:\, l \neq k} Pr(X_l|E) \leq \alpha_l$ are equivalent – there is only one different $l = 0, 1$ and one of equalities (17) must take place. Therefore, $\mathcal{POS}_B^{\alpha}(X_k)$ takes the same form as in (14). Negative regions are formulated directly in the same way as in (14). Hence, the boundary regions must be identical as well. □

Example 9. Let us go back to the three-decision case illustrated by Fig. 3 and consider parameters $\alpha = (1/10, 1/4, 1/10)$, as in Example 8. Let us notice that the vector $\beta = (13/20, 3/4, 13/20)$ from that example can be calculated from α using (35). Now, let us compare the previously obtained regions with the following ones:

Decisions	$\mathcal{POS}_B^{\alpha}(X_k)$	$\mathcal{NEG}_B^{\alpha}(X_k)$	$\mathcal{BND}_B^{\alpha}(X_k)$
X_0	$\emptyset$	$E_3 \cup E_4$	$E_1 \cup E_2 \cup E_5$
X_1	$\emptyset$	$E_2 \cup E_5$	$E_1 \cup E_3 \cup E_4$
X_2	$\emptyset$	$\emptyset$	U

Although, on the one hand, the crucial property (34) is now satisfied, we do not get any relaxation of conditions for the positive regions with respect to the original RS-regions analyzed for the same system in Example 7. The problem with Definition 3 seems to be that even a very good evidence for X_k can be ignored (put into boundary) because of just one other X_l, $l \neq k$, supported by E to a relatively (comparing to X_k) low degree. The RB-regions presented in Example 7 turn out to be intuitively more flexible with handling the data based probabilities. We try to justify it formally below. □

After introducing a reasonable extension (and simplification) of VPRS for the multi-decision case, we are ready to compare it – as an example of the posterior probability based methodology – to the Rough Bayesian model. Since it is an introductory study, we restrict ourselves to the simplest case of RB, namely to the unified ε-matrix (28), where $\varepsilon_k^l = \varepsilon$ for every $k, l = 0, \ldots, r-1$, $k \neq l$, for some $\varepsilon \in [0, 1)$. It refers to an interesting special case of simplified VPRS, where

$$\frac{\alpha_0(1 - Pr(X_0))}{(1 - \alpha_0)Pr(X_0)} = \ldots = \frac{\alpha_{r-1}(1 - Pr(X_{r-1}))}{(1 - \alpha_{r-1})Pr(X_{r-1})} \tag{36}$$

According to (36) the parameters for particular decision classes satisfy inequalities (32) in a proportional way. Its advantage corresponds to the problem of tuning vectors $\alpha = (\alpha_0, \ldots, \alpha_{r-1})$ for large values of $r = |V_d|$. Given (36) we can handle the whole α using a single parameter:

Proposition 9. *Let $\mathbb{A} = (U, A \cup \{d\})$, $B \subseteq A$, and $\alpha = (\alpha_0, \ldots, \alpha_{r-1})$ satisfying both (32) and (36) be given. There exists $\varepsilon \in [0,1)$, namely*

$$\varepsilon = \frac{\alpha_k(1 - Pr(X_k))}{(1 - \alpha_k)Pr(X_k)} \quad \textit{for arbitrary } k = 0, \ldots, r-1 \tag{37}$$

such that for every $k = 0, \ldots, r-1$ the value of α_k is derivable as

$$\alpha_k = \frac{\varepsilon Pr(X_k)}{\varepsilon Pr(X_k) + (1 - Pr(X_k))} \tag{38}$$

Proof. It is enough to substitute the right side of (37) as ε to the right side (38) and check that it indeed equals α_k. □

The following result shows that at the one-parameter level the Rough Bayesian model can be potentially more data sensitive than the simplified VPRS model. Obviously, similar comparison of more general cases is a desired direction for further research.

Theorem 3. *Let $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. Consider vector $\alpha = (\alpha_0, \ldots, \alpha_{r-1})$ satisfying (32) and (36). Consider $\varepsilon \in [0,1)$ given by (37) as the unified parameter for the RB model, that is $\varepsilon_k^l = \varepsilon$ for every $k, l = 0, \ldots, r-1$. Then we have the following inclusions, for every $k = 0, \ldots, r-1$:*

$$\begin{array}{l} \mathcal{POS}_B^{\alpha}(X_k) \subseteq \mathcal{BAYPOS}_B^{\varepsilon}(X_k) \\ \mathcal{NEG}_B^{\alpha}(X_k) \subseteq \mathcal{BAYNEG}_B^{\varepsilon}(X_k) \\ \mathcal{BND}_B^{\alpha}(X_k) \supseteq \mathcal{BAYBND}_B^{\varepsilon}(X_k) \end{array} \tag{39}$$

Proof. Using the same technique as in the proof of Proposition 4, we can show

$$Pr(X_k|E) \leq \alpha_k \Leftrightarrow \varepsilon Pr(E|\neg X_k) \geq Pr(E|X_k)$$

where $\neg X_k = \bigcup_{l \neq k} X_l$. Further, using a simple translation, we can observe that

$$Pr(E|\neg X_k) = \frac{\sum_{l \neq k} Pr(X_l) Pr(E|X_l)}{\sum_{l \neq k} Pr(X_l)}$$

Now, we are ready to show inclusions (39). Let us begin with the second one. Assume that a given $E \in U/B$ is not in $\mathcal{BAYNEG}_B^{\varepsilon}(X_k)$, that is

$$\forall_{l: l \neq k}\, \varepsilon Pr(E|X_l) < Pr(E|X_k)$$

Then we get $\varepsilon \sum_{l:l\neq k} Pr(X_l)Pr(E|X_l) < \sum_{l:l\neq k} Pr(X_l)Pr(E|X_k)$, further equivalent to

$$Pr(E|X_k) > \varepsilon \cdot \frac{\sum_{l:l\neq k} Pr(X_l)Pr(E|X_l)}{\sum_{l:l\neq k} Pr(X_l)}$$

Hence, E is outside $\mathcal{NEG}_B^{\alpha}(X_k)$ and the required inclusion is proved. To show the first inclusion in (39), assume $E \subseteq \mathcal{POS}_B^{\alpha}(X_k)$. According to (34), we then have $E \subseteq \mathcal{NEG}_B^{\alpha}(X_l)$, for every X_l, $l \neq k$. Using just proved inclusion, we get $E \subseteq \mathcal{BAYNEG}_B^{\alpha}(X_l)$. By Corollary 1 we then obtain $E \subseteq \mathcal{BAYPOS}_B^{\alpha}(X_k)$, what we wanted to prove. The third inclusion in (39) is now derivable directly from the other two ones. □

Example 10. Let us recall the decision system from Fig. 3, where $Pr(X_0) = Pr(X_2) = 1/4$ and $Pr(X_1) = 1/2$. It turns out that the parameters $\alpha = (1/10, 1/4, 1/10)$ considered in Example 8 are derivable using (38) for $\varepsilon = 1/3$. According to Theorem 3, the RB-regions presented in Example 7 for $\varepsilon = 1/3$ are referrable to the simplified VPRS-regions from Example 9. It is an illustration for (39) – one can see that we should expect strict inclusions in all those inclusions. The specific problem with putting E_3 to the positive simplified VPRS-region of X_1 is that it is blocked by too high value of $Pr(X_2|E_3)$ although this value seems to be much lower than $Pr(X_1|E_3)$. We should avoid comparing these two posterior probabilities directly because it would be unfair with respect to X_2 for its prior probability is twice lower than in case of X_1. However, direct comparison of the inverse probabilities $Pr(E_3|X_1)$ and $Pr(E_3|X_2)$ shows that we can follow X_1 since E_3 is three times more likely given X_1 than given X_2. □

An interesting feature of the Rough Bayesian model is that it can use a single parameter $\varepsilon \in [0, 1)$ to produces valuable results, as illustrated by the above example. On the other hand, asymmetric extensions of RB are possible, even for $r(r-1)$ different parameters ε_k^l corresponding to comparison of $\varepsilon_k^l Pr(E|X_k)$ and $Pr(E|X_l)$. Further research is needed to understand expressive power of such extensions and their relevance to the other rough set approaches.

5 Distributed Decision Systems

In the examples considered so far, we referred to decision systems gathering objects supporting all decision classes together. On the other hand, while dealing with the Bayes factors and the RB-regions, we calculate only the inverse probabilities, which do not require putting the whole data in a single table. We propose a data storage framework, where the objects supporting the target concepts corresponding to different decision classes are stored in separate data sets. It emphasizes that in some situations data supporting particular events are *uncombinable* and the only probability estimates we can use are of the inverse character, that is they are naturally conditioned by particular decisions.

Definition 4. *Let the set of r mutually exclusive target events be given. By a* distributed decision system $\mathcal{A}$ *we mean the collection of r information systems*

$$\mathcal{A} = \{\mathbb{A}_0 = (X_0, A), \ldots, \mathbb{A}_{r-1} = (X_{r-1}, A)\} \tag{40}$$

where X_k denotes the set of objects supporting the k-th event, $k = 0, \ldots, r-1$, and A is the set of attributes describing all the objects in $X_0, \ldots, X_{r-1}$.

Any information derivable from $\mathbb{A}_k$ is naturally conditioned by X_k, for $k = 0, \ldots, r-1$. Given B-information vector $w \in V_B$, $B \subseteq A$, we can set up

$$Pr_k(B = w) = \frac{|\{u \in X_k : B(u) = w\}|}{|X_k|} \tag{41}$$

as the probability that a given object will have the values described by w on B *conditioned* by its membership to X_k.

Example 11. Let us consider $\mathcal{A}$ consisting of two information systems illustrated in Fig. 4. For instance, $Pr_0(a_1 = 2, a_3 = 2) = 1/10$ and $Pr_1(a_1 = 2, a_3 = 2) = 3/5$ are estimates of probabilities that a given object will satisfy $a_1 = 2$ and $a_3 = 2$, if it supports the events X_0 and X_1, respectively.

One can see that if we use estimation

$$Pr(B = w|d = k) = Pr_k(B = w) \tag{42}$$

then the inverse probabilities derivable from Fig. 4 are identical with those derivable from Fig. 1. Actually, we created Fig. 1 artificially by doubling the objects from $\mathbb{A}_1$ and merging them with $\mathbb{A}_0$ from Fig. 4. Therefore, if we assume that due to our knowledge we should put $Pr(X_0) = Pr(X_1) = 1/2$, then systems illustrated by Figures 1 and 4 will provide the same posterior probabilities. □

X_0	a_1	a_2	a_3	a_4	a_5
u_1	1	2	0	0	2
u_2	1	1	0	1	2
u_3	0	0	0	2	2
u_4	2	1	1	0	2
u_5	2	1	1	0	2
u_6	2	2	2	1	1
u_7	0	1	2	2	2
u_8	1	1	0	1	2
u_9	2	1	1	0	2
u_{10}	0	0	0	2	2

X_1	a_1	a_2	a_3	a_4	a_5
o_1	2	2	2	1	1
o_2	0	1	2	2	2
o_3	1	1	0	1	2
o_4	2	2	2	1	1
o_5	2	2	2	1	1

Fig. 4. Distribute decision system $\mathcal{A} = \{\mathbb{A}_0 = (X_0, A), \mathbb{A}_1 = (X_1, A)\}$, where $A = \{a_1, \ldots, a_5\}$, and $X_0 = \{u_1, \ldots, u_{10}\}$, $X_1 = \{o_1, \ldots, o_5\}$.

Distributed decision systems do not provide a means for calculation of the posterior probabilities unless we know the priors of all decision classes. On the other hand, we get full flexibility with respect to the changes of the prior probabilities, which can be easily combined with the estimates (42). For instance, let us go back to the case study from the end of Subsection 2.2 and assume that the objects in $\mathbb{A}_1 = (X_1, A)$ are very carefully chosen cases of a rare medical pathology while the elements of X_0 describe a representative sample of human beings not suffering from this pathology. Let us put $Pr(X_1) = 1/1000$. Then, as in Example 6, we get $Pr(d = 1|a_1 = 2, a_3 = 2) = 2/335$. It shows how different posterior probabilities can be obtained from the same distributed decision system for various prior probability settings. Obviously, we could obtain identical results from appropriately created classical decision systems (like in case of the system in Fig. 1). However, such a way of data translation is unnecessary or even impossible, if the prior probabilities are not specified.

From technical point of view, it does not matter whether we keep the data in the form of distributed or merged decision system, unless we use estimations (4). However, we find Definition 4 as a clearer way to emphasize the nature of the data based probabilities that we can really believe in. Indeed, the inverse probabilities (42) are very often the only ones, which can be reasonably estimated from real-life data sets. This is because the process of the data acquisition is often performed in parallel for various decisions and, moreover, the experts can (and wish to) handle the issue of the information representativeness only at the level of separate decision classes. Following this argumentation, let us reconsider the original RS-regions for distributed data, without the need of merging them within one decision system.

Definition 5. *Let the system* $\mathcal{A} = \{\mathbb{A}_0 = (X_0, A), \ldots, \mathbb{A}_{r-1} = (X_{r-1}, A)\}$ *be given. For any* X_k *and* $B \subseteq A$*, we define the* B-positive*,* B-negative*, and* B-boundary distributed rough set regions *(abbreviated as* DRS-regions*) as follows:*

$$\begin{array}{ll} \mathcal{DPOS}_B(X_k) &= \{w \in V_B : \forall_{l:\, l \neq k} Pr_l(B = w) = 0\} \\ \mathcal{DNEG}_B(X_k) &= \{w \in V_B : Pr_k(B = w) = 0\} \\ \mathcal{DBND}_B(X_k) &= \{w \in V_B : Pr_k(B = w) > 0 \wedge \exists_{l:\, l \neq k} Pr_l(B = w) > 0\} \end{array} \tag{43}$$

The difference between (43) and (9) is that the distributed rough set regions are expressed in terms of B-information vectors, regarded as the conditions satisfiable by the objects. Besides, both definitions work similarly if they refer to the same inverse probabilities.

Example 12. The DRS-regions obtained for $B = \{a_1, a_3\}$ from Fig. 4 look as follows:

$$\begin{array}{ll} \mathcal{DPOS}_B(X_1) &= \emptyset \\ \mathcal{DNEG}_B(X_1) &= \{\{(a_1, 0), (a_3, 0)\}, \{(a_1, 2), (a_3, 1)\}\} \\ \mathcal{DBND}_B(X_1) &= \{\{(a_1, 0), (a_3, 2)\}, \{(a_1, 1), (a_3, 0)\}, \{(a_1, 2), (a_3, 2)\}\} \end{array} \tag{44}$$

One can see that the supports of the above B-information vectors within the decision system from Fig. 1 correspond to the RS-regions in Example 4. □

The rough set extensions referring in a non-trivial way to the posterior and prior probabilities, like e.g. VPRS, cannot be rewritten in terms of distributed decision systems. However, it is possible for the Rough Bayesian model. Actually, it emphasizes that RB does not need to assume *anything* about the prior and posterior probabilities. We believe that in this form our idea of combining rough sets with the Bayes factor based approaches is possibly closest to the practical applications.

Definition 6. *Let* $\mathcal{A} = \{\mathbb{A}_0 = (X_0, A), \ldots, \mathbb{A}_{r-1} = (X_{r-1}, A)\}$ *be given. Consider matrix* ε *given by (28) for* $\varepsilon_k^l \in [0, 1)$*,* $k \neq l$*. For any* $k = 0, \ldots, r-1$ *and* $B \subseteq A$*, we define the* B-positive*,* B-negative*, and* B-boundary distributed rough Bayesian regions *(abbreviated as* DRB-regions*) as follows:*

$$\begin{array}{rl} \mathcal{DBAYPOS}_B^{\varepsilon}(X_k) &= \{w \in V_B : \forall_{l:\,l\neq k} Pr_l(B=w) \leq \varepsilon_k^l Pr_k(B=w)\} \\ \mathcal{DBAYNEG}_B^{\varepsilon}(X_k) &= \{w \in V_B : \exists_{l:\,l\neq k} Pr_k(B=w) \leq \varepsilon_l^k Pr_l(B=w)\} \\ \mathcal{DBAYBND}_B^{\varepsilon}(X_k) &= \{w \in V_B : \exists_{l:\,l\neq k} Pr_l(B=w) > \varepsilon_k^l Pr_k(B=w) \\ & \qquad \wedge\ \forall_{l:\,l\neq k} Pr_k(B=w) > \varepsilon_l^k Pr_l(B=w)\} \end{array} \tag{45}$$

Example 13. Fig. 5 illustrates a distributed system for three target events. They result from splitting $\mathbb{A}_0 = (X_0, A)$ from Fig. 4 onto equally large $\mathbb{A}_0 = (X_0, A)$ and $\mathbb{A}_2 = (X_2, A)$, similarly as we did in the previous sections with our exemplary non-distributed decision system. Let us start with calculation of the regions introduced in Definition 5. As usual, consider $B = \{a_1, a_3\}$. The DRS-regions for X_1 do not change with respect to (44). The remaining regions look as follows:

$$\begin{array}{ll} \mathcal{DPOS}_B(X_0) = \emptyset & \mathcal{DPOS}_B(X_2) = \emptyset \\ \mathcal{DNEG}_B(X_0) = \{\{(a_1,0),(a_3,2)\},\{(a_1,2),(a_3,2)\}\} & \mathcal{DNEG}_B(X_2) = \emptyset \\ \mathcal{DBND}_B(X_0) = V_B \setminus \mathcal{DNEG}_B(X_0) & \mathcal{DBND}_B(X_2) = V_B \end{array} \tag{46}$$

X_0	a_1	a_2	a_3	a_4	a_5
u_1	1	2	0	0	2
u_2	1	1	0	1	2
u_3	0	0	0	2	2
u_4	2	1	1	0	2
u_5	2	1	1	0	2

X_1	a_1	a_2	a_3	a_4	a_5
o_1	2	2	2	1	1
o_2	0	1	2	2	2
o_3	1	1	0	1	2
o_4	2	2	2	1	1
o_5	2	2	2	1	1

X_2	a_1	a_2	a_3	a_4	a_5
e_1	2	2	2	1	1
e_2	0	1	2	2	2
e_3	1	1	0	1	2
e_4	2	1	1	0	2
e_5	0	0	0	2	2

Fig. 5. System $\mathcal{A} = \{\mathbb{A}_0 = (X_0, A),\ \mathbb{A}_1 = (X_1, A),\ \mathbb{A}_2 = (X_2, A)\}$, where $A = \{a_1, \ldots, a_5\}$, $X_0 = \{u_1, \ldots, u_5\}$, $X_1 = \{o_1, \ldots, o_5\}$, $X_2 = \{e_1, \ldots, e_5\}$.

The B-boundary DRS-region for X_2 corresponds to the whole V_B. It means that any so far recorded B-information vector is *likely* to occur for a given object under the assumption that that object supports X_2, as well as under the assumption that it supports X_0 and/or X_1. Now, consider the DRB-regions for $\varepsilon = 1/3$. We obtain the following changes with respect to the (44) and (46):

1. $\mathcal{DBAYPOS}_B^{1/3}(X_1)$, $\mathcal{DBAYNEG}_B^{1/3}(X_2)$ start to contain $\{(a_1, 2), (a_3, 2)\}$.
2. $\mathcal{DBAYBND}_B^{1/3}(X_k)$, $k = 1, 2$, do not contain $\{(a_1, 2), (a_3, 2)\}$ any more.

The obtained DRB-regions are comparable with the RB-regions obtained previously for the corresponding non-distributed decision system from Fig. 3. □

Introduction of distributed decision systems has rather a technical than theoretical impact. It illustrates possibility of handling decomposed data, which can be especially helpful in case of many decision classes with diversified or unknown prior probabilities. Distributed systems provide the exact type of information needed for extracting the RB-regions from data. Hence, we plan implementing the algorithms referring to the Rough Bayesian model mainly for such systems.

6 Final Remarks

We introduced the Rough Bayesian model – a parameterized extension of rough sets, based on the Bayes factor and the inverse probabilities. We compared it with other probabilistic extensions, particularly with the VPRS model relying on the data based posterior probabilities. We considered both the two-decision and multiple-decision cases, where the direct comparison of the inverse probabilities conditioned by decision classes turns out to be more flexible than handling their posterior probabilities. Finally, we proposed distributed decision systems as a new way of storing data, providing estimations for the Rough Bayesian regions.

We believe that the framework based on the Rough Bayesian model is well applicable to the practical data analysis problems, especially if we cannot rely on the prior/posterior probabilities derivable from data and/or background knowledge. The presented results are also helpful in establishing theoretical foundations for correspondence between the theory of rough sets and Bayesian reasoning. Several basic facts, like, e.g., the inverse probability based characteristics of the original rough set model, support an intuition behind this correspondence.

Acknowledgments

Supported by the research grant from the Natural Sciences and Engineering Research Council of Canada.

References

1. Box, G.E.P., Tiao, G.C.: Bayesian Inference in Statistical Analysis. Wiley (1992).
2. Duentsch, I., Gediga, G.: Uncertainty measures of rough set prediction. Artificial Intelligence 106 (1998) pp. 77–107.
3. Greco, S., Pawlak, Z., Słowiński, R.: Can Bayesian confirmation measures be useful for the rough set decision rules? Engineering Application of Artificial Intelligence, 17 (2004) pp. 345–361.
4. Good, I.J.: The Interface Between Statistics and Philosophy of Science. Statistical Science 3 (1988) pp. 386–397.
5. Hilderman, R.J., Hamilton, H.J.: Knowledge Discovery and Measures of Interest. Kluwer (2002).
6. Jeffreys, H.: Theory of Probability. Clarendon Press, Oxford (1961).
7. Kamber, M., Shingal, R.: Evaluating the interestingness of characteristic rules. In: Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), Portland, Oregon (1996) pp. 263–266.
8. Kloesgen, W., Żytkow, J.M. (eds): Handbook of Data Mining and Knowledge Discovery. Oxford University Press (2002).
9. Mitchell, T.: Machine Learning. Mc Graw Hill (1998).
10. Pawlak, Z.: Rough sets – Theoretical aspects of reasoning about data. Kluwer Academic Publishers (1991).
11. Pawlak, Z.: New Look on Bayes' Theorem – The Rough Set Outlook. In: Proc. of JSAI RSTGC'2001, pp. 1–8.

12. Pawlak, Z.: Decision Tables and Decision Spaces. In: Proc. of the 6th International Conference on Soft Computing and Distributed Processing (SCDP'2002). June 24-25, Rzeszów, Poland (2002).
13. Pawlak, Z., Skowron, A.: Rough membership functions. In: Advances in the Dempster Shafer Theory of Evidence. Wiley (1994) pp. 251–271.
14. Polkowski, L., Tsumoto, S., Lin, T.Y. (eds): Rough Set Methods and Applications. Physica Verlag (2000).
15. Raftery, A.E.: Hypothesis testing and model selection. In: W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (eds), Markov Chain Monte Carlo in Practice. Chapman and Hall, London (1996) pp. 163–187.
16. Ślęzak, D.: Approximate Entropy Reducts. Fundamenta Informaticae, 53/3-4 (2002) pp. 365–390.
17. Ślęzak, D.: The Rough Bayesian Model For Distributed Decision Systems. In Proc. of RSCTC'2004 (2004) pp. 384–393.
18. Ślęzak, D., Ziarko, W.: Bayesian Rough Set Model. In: Proc. of FDM'2002 (2002) pp. 131–135.
19. Ślęzak, D., Ziarko, W.: The Investigation of the Bayesian Rough Set Model. International Journal of Approximate Reasoning (2005) to appear.
20. Swinburne, R. (ed.): Bayes's Theorem. Proc. of the British Academy 113 (2003).
21. Tsumoto, S.: Accuracy and Coverage in Rough Set Rule Induction. In: Proc. of RSCTC'2002 (2002) pp. 373–380.
22. Wong, S.K.M., Ziarko, W.: Comparison of the probabilistic approximate classification and the fuzzy set model. International Journal for Fuzzy Sets and Systems, 21 (1986) pp. 357–362.
23. Yao, Y.Y.: Probabilistic approaches to rough sets. Expert Systems, 20/5 (2003) pp. 287–297.
24. Ziarko, W.: Variable precision rough sets model. Journal of Computer and Systems Sciences, 46/1 (1993) pp. 39–59.
25. Ziarko, W.: Set approximation quality measures in the variable precision rough set model. Soft Computing Systems, Management and Applications, IOS Press (2001) pp. 442–452.