Stochastic Dominance-based Rough Set Model for Ordinal Classification

Wojciech Kotłowski
Institute of Computing Science
Poznań University of Technology
Poznań, 60-965, Poland
wkotlowski@cs.put.poznan.pl
Krzysztof Dembczyński
Institute of Computing Science
Poznań University of Technology
Poznań, 60-965, Poland
kdembczynski@cs.put.poznan.pl
Salvatore Greco
Faculty of Economics
University of Catania
95129 Catania, Italy
salgreco@unict.it
Roman Słowiński
Institute of Computing Science
Poznań University of Technology
Poznań, 60-965, Poland
Institute for Systems Research
Polish Academy of Sciences
Warsaw, 01-447, Poland
rslowinski@cs.put.poznan.pl

Abstract

In order to discover interesting patterns and dependencies in data, an approach based on rough set theory can be used. In particular, Dominance-based Rough Set Approach (DRSA) has been introduced to deal with the problem of ordinal classification with monotonicity constraints (also referred to as multicriteria classification in decision analysis). However, in real-life problems, in the presence of noise, the notions of rough approximations were found to be excessively restrictive. In this paper, we introduce a probabilistic model for ordinal classification problems with monotonicity constraints. Then, we generalized the notion of lower approximations to the stochastic case. We estimate the probabilities with the maximum likelihood method which leads to the isotonic regression problem for two-class (binary) case. The approach is easily generalized for the multi-class case. Finally, we show the equivalence of the variable consistency rough sets to the specific empirical risk-minimizing decision rule in the statistical decision theory.

Keywords: Dominance-based Rough Set Approach, Ordinal Classification, Monotonicity Constraints, Isotonic Regression, Maximum Likelihood Estimation, Variable Consistency Models, Statistical Decision Theory, Empirical Risk Minimization, Multiple Criteria Decision Analysis.

1 Introduction

We consider an *ordinal classification problem* that consists in assignment of objects to K ordered *classes* Cl_k, $k \in Y = \{1, \ldots, K\}$, such that if $k > k'$ then class Cl_k is higher than class $Cl_{k'}$. Objects are evaluated on a set of m *attributes* with ordered value sets. Here, without loss of generality we assume that value set of each attribute is a subset of $\mathbb{R}$ (even if the scale is purely ordinal, evaluation of attributes can be numbercoded) and the order relation is a linear order $\geq$, so that each object x_i is an m-dimensional vector $(x_{i1}, \ldots, x_{im})$. It is assumed that *monotonicity constraints* are present in the data: a higher evaluation of an object on an attribute, with other evaluations being fixed, should not decrease its assignment to the class. In order to construct a data model, one can induce it from a *training* set $U = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, consisting of n objects (denoted with x) already assigned to their classes (class indices denoted with $y \in Y$). We will also denote $X = \{x_1, \ldots, x_n\}$ and by class Cl_k we mean the subset of X consisting of objects x_i having class index $y_i = k$, $Cl_k = \{x_i \in X : y_i = k\}$.

Thus, ordinal classification problem with monotonicity constraints resembles typical classification problem considered in machine learning [8, 14], with two additional constraints. The first one is the assumption of the ordinal scale on each attribute and between class indices. The second constraint is the monotonicity property: the expected class index increases with increasing values on attributes. Such properties are commonly encountered in real-life applications, yet rarely taken into account.

In decision theory, a *multicriteria classification problem* is considered [11], which has exactly the form of ordinal classification problem with monotonicity constraints. Moreover, in many different domains monotone properties follows from the domain knowledge about the problem and should not be neglected. They have been recognized in applications such as bankruptcy risk prediction[9], breast cancer diagnosis [21], house pricing [19], credit rating [7], liver disorder diagnosis [22] and many others.

Despite the monotone nature of the data, it still may happen that in the training set U, there exists an object x_i not worse than another object x_j on all attributes, however, x_i is assigned to a worse class than x_j; such a situation violates the monotone properties of data, so we shall call objects x_i and x_j *inconsistent*. Rough set theory [16, 17, 18] has been adapted to deal with this kind of inconsistency and the resulting methodology has been called *Dominance-based Rough Set Approach* (DRSA) [10, 11]. In DRSA, the classical indiscernibility relation has been replaced by a dominance relation. Using the rough set approach to the analysis of multicriteria classification problem, we obtain lower and upper (rough) approximations of unions of classes. The difference between upper and lower approximations shows inconsistent objects with respect to the dominance principle. It can happen that due to the presence of noise, the data is so inconsistent, that too much information is lost, thus making the DRSA inference model not accurate. To cope with the problem of excessive inconsistency the *variable consistency* model within DRSA has been proposed (VC-DRSA) [12].

In this paper, we look at DRSA from a different point of view, identifying its connections with statistics and statistical decision theory. We start with the overview of the classical rough sets theory and show that the variable-precision model [24, 25] comes from the maximum likelihood estimation method. Then we briefly present main concepts of DRSA. Afterwards, the main part of the paper follows: we introduce the probabilistic model for a general class of ordinal classification problems with monotonicity constraints, and we generalize lower approximations to the stochastic case. Using the maximum likelihood method we show how the probabilities can be estimated in a nonparametric way. It leads to the statistical problem of isotonic regression, which is then solved by the optimal object reassignment problem. Finally, we explain the approach as being a solution to the problem of finding a decision minimizing the empirical risk [1].

We stress that the theory presented in this paper is related to the training set only. In order to classify well objects outside the training set, some generalizing classification function must be constructed. We do not consider this problem here. The aim of this paper is the analysis of inconsistencies in the dataset, handling and correcting them according to the probabilistic model assumption, which comes from exploring the monotonicity constraints. This analysis can be seen as a stochastic extension of DRSA. Therefore, the methodology presented here can be treated as a form of preprocessing and improving the data.

2 Maximum likelihood estimation in the classical variable precision rough set approach

We start with the classical rough set approach [16], which neither take into account any monotonicity constraints nor classes and attributes are ordered. It is based on the assumption that objects having the same description are indiscernible (similar) with respect to the available information. The *indiscernibility* relation thus generated induces a partition of the universe into blocks of indiscernible objects, called *granules* [16, 11]. The indiscernibility relation I is defined as:

$$I = \{(x_i, x_j) \in X \times X\colon\ x_{it} = x_{jt}\ \ \forall\, t = 1, \ldots, m\} \tag{1}$$

where x_{it} is the evaluation of object x_i on attribute t, as defined in previous section. The equivalence classes of I are called *granules.* The equivalence class for an object $x \in X$ is denoted $I(x)$. Any subset S of the universe may be expressed in terms of the granules either precisely (as a union of granules) or approximately only. In the latter case, the subset S may be characterized by two ordinary sets, called *lower* and *upper approximations.* Here, we always assume, that the approximated set S is some class $Cl_k, k \in Y$. The lower and upper approximations of class Cl_k are defined, respectively, by:

$$\underline{Cl}_k = \{x_i \in X\colon I(x_i) \subseteq Cl_k\} \tag{2}$$

$$\overline{Cl}_k = \{x \in X : I(x) \cap Cl_k \neq \emptyset\} \tag{3}$$

It follows from the definition, that $\underline{Cl}_k$ is the largest union of the granules included in Cl_k, while $\overline{Cl}_k$ is the smallest union of the granules containing Cl_k [11]. It holds, that $\underline{Cl}_k \subseteq Cl_k \subseteq \overline{Cl}_k$. Therefore, if an object $x \in X$ belongs to $\underline{Cl}_k$, it is certainly also an element of Cl_k, while if x belongs to $\overline{Cl}_k$, it may belong to the class Cl_k.

For application to the real-life data, a less restrictive definition were introduced under the name of *variable consistency rough sets* (VPRS) [24, 25]. The new definitions of approximations (where lower approximation is usually replaced by the term *positive region*, which will not be used here for simplicity) are expressed in the probabilistic terms in the following way. Let $\Pr(y = k|I(x))$ be a probability that an object x_i from granule $I(x)$ belongs to the class Cl_k. The probabilities are unknown, but are estimated by frequencies $\Pr(y = k|I(x)) = \frac{|Cl_k \cap I(x)|}{|I(x)|}$. Then, the lower approximation of class Cl_k is defined as:

$$\underline{Cl}_k = \{x \colon \Pr(y = k|I(x)) \geq u\} \tag{4}$$

so it is the sum of all granules, for which the probability of class Cl_k is at least equal to some threshold u. Similarly, the upper approximation of class Cl_k is defined as:

$$\overline{Cl}_k = \{x \colon \Pr(y = k|I(x)) \geq l\} \tag{5}$$

where l is usually set to $1 - u$ for the complementarity reasons.

It can be shown that frequencies used for estimating probabilities are the maximum likelihood (ML) estimators under assumption of common class probability distribution for every object within each granule. The sketch of the derivation is the following. Let us choose some granule $G = I(x)$. Let n_G be the number of objects in G, and for each class Cl_t, let n_G^k be the number of objects from this class in G. Then the class index y has a multinomial distribution when conditioned on granule G. Let us denote those probabilities $\Pr(y = k|G)$ by p_G^k.

Then the conditional probability of observing the $n_G^1, \ldots n_G^K$ objects in G, given $p_G^1, \ldots, p_G^K$ (conditional likelihood) is the following:

$$L(p; n_G|G) = \prod_{k=1}^{K} (p_G^k)^{n_G^k} \tag{6}$$

so that the log-likelihood is:

$$\mathcal{L}(p; n_G|G) = \ln L(n; p, G) = \sum_{k=1}^{K} n_G^k \ln p_G^k \tag{7}$$

The maximization of $\mathcal{L}(p; n_G|G)$ with additional constraint $\sum_{k=1}^{K} p_G^k = 1$ leads to the well-known fomula for ML estimators $\hat{p}_G^k$ in multinomial distribution:

$$\hat{p}_G^k = \frac{n_G^k}{n_G} \tag{8}$$

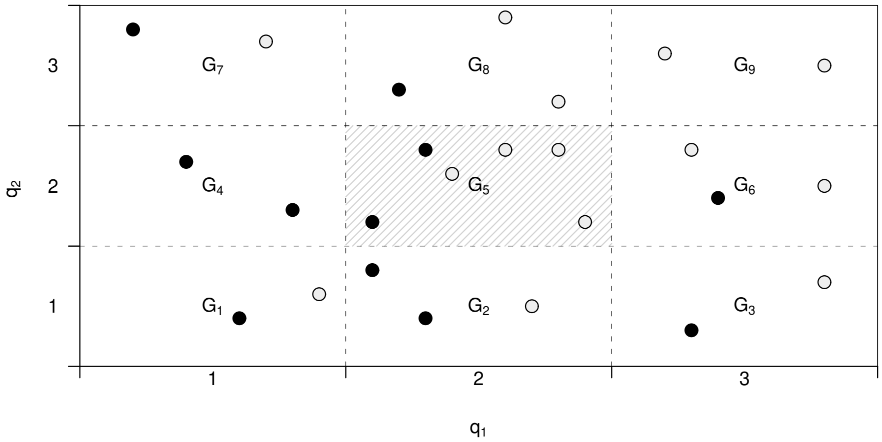

Figure 1: Example of two-class problem. Black points are objects from class 1, light points — from class 2. The value sets of two attributes q_1, q_2 are $\{1, 2, 3\}$. At granule G_5, $n^1_{G_5} = 2$, $n^2_{G_5} = 4$, so that $p^1_{G_5} = 1/3, p^2_{G_5} = 2/3$. We see, that for precision threshold $u = 2/3$ or lower, G_5 belongs to $\underline{Cl_2}$

which are exactly the frequencies used in VPRS. This observation will lead us in section 4 to the stochastic generalization of Dominance-based Rough Set Approach.

3 Dominance-based rough set approach (DRSA)

Within DRSA [10, 11], we define the *dominance* relation $\succeq$ as a binary relation on X in the following way: for any $x_i, x_j \in X$ we say that x_i *dominates* x_j, $x_i \succeq x_j$, if on every attribute, x_i has evaluation not worse than x_j, $x_{it} \geq x_{jt}$, for all $t = 1, \ldots, m$. The dominance relation $\succeq$ is a partial pre-order on X, i.e. it is reflexive and transitive. The *dominance principle* can be expressed as follows. For all $x_i, x_j \in X$ it holds:

$$x_i \succeq x_j \Longrightarrow y_i \geq y_j \tag{9}$$

The dominance principle follows from the monotone relationship between class indices and attributes. However, in many real-life applications the dominance principle is not satisfied, i.e. there exists at least one pair of objects violating (9). We say, that an object x_i is *inconsistent* if there exist another object x_j, such that x_i, x_j violates (9). Otherwise, we say that object x_i is *consistent*. We will sometimes also use the following expression: object x_i is consistent with x_j, if a pair x_i, x_j satisfies (9).

The rough approximations concern granules resulting from information carried out by the class indices. The approximation is made using granules resulting from information carried out by attributes. These granules are called *decision* and *condition* granules, respectively. This is due to the fact, that in rough set theory the class index for a given object is called *a decision value* and attributes are called *condition attributes.* The decision granules can be expressed by unions of classes:

$$Cl_k^{\geq} = \{x_i \in X : y_i \geq k\} \tag{10}$$

$$Cl_k^{\leq} = \{x_i \in X : y_i \leq k\}. \tag{11}$$

The condition granules are dominating and dominated sets defined, respectively, as:

$$D^{+}(x) = \{x_i \in X : x_i \succeq x\} \tag{12}$$

$$D^{-}(x) = \{x_i \in X : x \succeq x_i\}. \tag{13}$$

Let us remark that both decision and condition granules are cones in decision (Y) and condition (X) spaces, respectively. Using class unions instead of separate classes and dominating (dominated) sets instead of single objects is a general property of most of methodologies dealing with ordinal classification problem with the monotonicity constraints and follows directly from the monotone nature of the data.

Lower dominance-based approximations of $Cl_k^{\geq}$ and $Cl_k^{\leq}$ are defined as follows:

$$\underline{Cl_k^{\geq}} = \{x_i \in X : D^{+}(x_i) \subseteq Cl_k^{\geq}\}, \tag{14}$$

$$\underline{Cl_k^{\leq}} = \{x_i \in X : D^{-}(x_i) \subseteq Cl_k^{\leq}\}. \tag{15}$$

They reflect the objects which certainly belong to class $Cl_k^{\geq}$ (or $Cl_k^{\leq}$). This certainty comes from the fact, that object x_i belongs to the lower approximation of class union $Cl_k^{\geq}$ (respectively $Cl_k^{\leq}$), if no other object in the dataset X contradicts it, i.e. x_i is consistent with every other object outside of $Cl_k^{\geq}$ (respectively $Cl_k^{\leq}$). Otherwise, if there exists object outside of $Cl_k^{\geq}$, which dominates x_i, then due to the dominance principle (following from the monotonicity constraints) we cannot say that x_i should belong to $Cl_k^{\geq}$ with certainty.

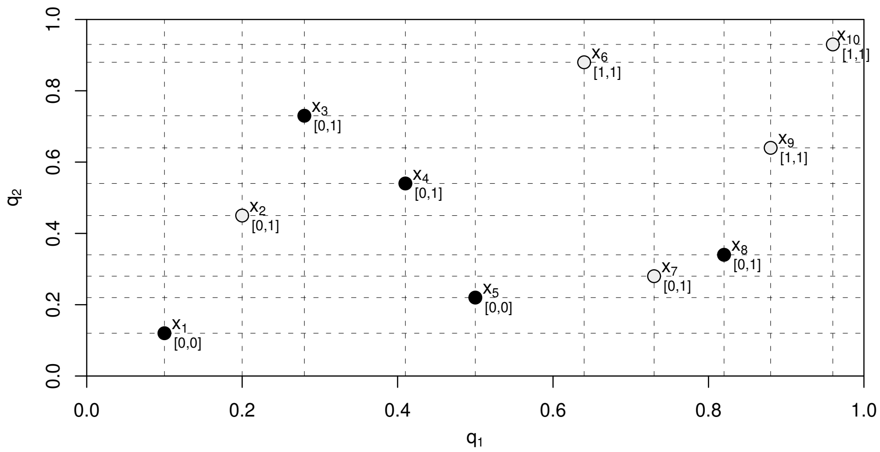

Figure 2: Example of two-class problem. Black points are objects from class 0, light points — from class 1 (indexing classes from 0 will be used later in this paper). We have trivially $\underline{Cl_0^{\geq}} = \underline{Cl_1^{\leq}} = X$, $\underline{Cl_0^{\leq}} = \{x_1, x_5\}, \underline{Cl_1^{\geq}} = \{x_6, x_9, x_6\}$. The generalized decisions for objects $x_1, \ldots, x_{10}$ were shown on the chart in brackets.

Notice, that for any $k \in Y$, we have $Cl_k^{\geq} \cup Cl_{k-1}^{\leq} = X$. It is not the case with the lower approximations. Therefore we define *the boundary (doubtful) region* [11] for class unions $Cl_k^{\geq}$ and $Cl_{k-1}^{\leq}$ as:

$$B_k = X \backslash (\underline{Cl_k^{\geq}} \cup \underline{Cl_{k-1}^{\leq}}) \tag{16}$$

which reflects the area which does not belong to lower approximations of class unions $Cl_k^{\geq}$ and $Cl_{k-1}^{\leq}$. Notice, that DRSA separates the analysis of inconsistencies into $K-1$ binary problems: for each $k = 2, \ldots, K$ we have lower approximations $\underline{Cl_k^{\geq}}$, $\underline{Cl_{k-1}^{\leq}}$ and boundary B_k which together form the whole set X. Such a separation will be also used in the stochastic extension of DRSA.

For the purpose of this paper, we will focus our attention on another concept from DRSA (as we shall shortly see, equivalent to the notion of approximations), the generalized decision [5]. Suppose, object $x_i \in \underline{Cl_k^{\geq}}$. Since the lower approximation of class union $Cl_k^{\geq}$ is a region, in which objects certainly belong to $Cl_k^{\geq}$, we can state, that class index of x_i should be at least k. Choosing the greatest k, for which $x_i \in \underline{Cl_k^{\geq}}$ holds (denoted by $l(x_i)$), we know that the class index of x_i must be at least $l(x_i)$ and we cannot give more precise statement, since we are not certain that the class index of x_i is at least $l(x_i)+1$, because $x_i \notin \underline{Cl_k^{\geq}}$.

On the other hand, if $x_i \in \underline{Cl}_k^{\leq}$, we known that the class index of x_i must be at most k. Similarly choosing the lowest k, for which $x_i \in \underline{Cl}_t^{\leq}$ (denoted by $u(x_i)$), we end up with the interval of classes $[l(x_i), u(x_i)]$, for which we known that object x_i may belong to. This interval is often denoted by $\delta(x_i)$, and is called *generalized decision* (we remind that class assignments are called decision values in rough set theory):

$$\delta(x_i) = [l(x_i), u(x_i)] \tag{17}$$

where:

$$\begin{aligned} l(x_i) &= \max\left\{k \colon x_i \in \underline{Cl}_k^{\geq}\right\} & (18)\\ u(x_i) &= \min\left\{k \colon x_i \in \underline{Cl}_k^{\leq}\right\} & (19) \end{aligned}$$

The generalized decision reflects an interval of decision classes to which an object may belong due to the inconsistencies with the dominance principle. Investigating the definitions of lower approximations (14)-(15) one can show, that generalized decision can be easily computed without reference to the lower approximation:

$$\begin{aligned} l(x_i) &= \min\{y_j \colon x_j \succeq x_i, x_j \in X\}, & (20)\\ u(x_i) &= \max\{y_j \colon x_i \succeq x_j, x_j \in X\}. & (21) \end{aligned}$$

Thus, $l(x_i)$ is the lowest class, to which objects dominating x_i belong; $u(x_i)$ is the highest class, to which objects dominated by x_i belong. Obviously, $l(x_i) \leq y_i \leq u(x_i)$ for every $x_i \in X$ and if $l(x_i) = u(x_i)$, then object x_i is consistent with respect to the dominance principle with every other object $x_j \in X$. Notice, that the wider the generalized decision, the less precise knowledge about the object we have.

Let us remark that the description with generalized decision is fully equivalent with description with rough approximations. Namely, dominance-based lower approximations may be expressed using generalized decision:

$$\begin{aligned} \underline{Cl}_k^{\geq} &= \{x_i \in X : l(x_i) \geq k\}, & (22)\\ \underline{Cl}_k^{\leq} &= \{x_i \in X : u(x_i) \leq k\}. & (23) \end{aligned}$$

Finally, notice that the definitions of lower approximations and generalized decisions for DRSA are very restrictive. Suppose, there exist one object dominating many other objects from dataset, but its class index is the lowest one (e.g. due to mistake). Then, many of the objects will we included into boundary regions and their generalized decision will be broaden. Therefore, relaxed definitions of lower approximations have been introduced under the name *variable consistency* DRSA (VC-DRSA) [12, 2], which allow object x_i to be incorporated into lower approximations, if a high fraction of objects dominating x_i (or being dominated by x_i) is consistent with x_i. The stochastic model introduced in the next section has also similar properties, therefore it can be regarded as a sort of VC-DRSA model.

4 Stochastic model of DRSA

In this section, we introduce new definitions of lower approximations for DRSA. The definitions will be based on the probabilistic model for the ordinal classification problems.

In section 2, we have made the assumption that in a single granule $I(x)$, each object $x \in G$ has the same conditional probability distribution, $\Pr(y|I(x))$. This is due to the property of indiscrenibility of objects within a granule. In case of DRSA, indiscernibility is replaced by a dominance relation, so that a different relation between the probabilities must hold. Namely, we conclude from the dominance principle that:

$$x_i \succeq x_j \Longrightarrow \Pr(y \geq k|x_i) \geq \Pr(y \geq k|x_j) \qquad \forall k \in Y, \;\; \forall x_i, x_j \in X \tag{24}$$

where $\Pr(y \geq k|x_i)$ is a probability (conditioned on x_i) of class index at least t. In other words, if object x_i dominates object x_j, probability distribution conditioned at point x_i *stochastically dominates* probability distribution conditioned at x_j. Equation (24) will be called *stochastic dominance principle*. It reflects the general property of a probability distribution in the problems with monotonicity constraints. Moreover, reversing our reasoning, we can give a statistical definition of the ordinal classification problem with monotonicity constraints: it is every classification problem with ordered value sets of attributes and classes with the probabilistic model for which (24) holds.

Having stated the probabilistic model, we introduce the stochastic DRSA by relaxing the definitions of lower approximation of classes:

$$\underline{Cl_k^{\geq}} = \{x_i \in X : \Pr(y \geq k|x_i) \geq \alpha\}, \tag{25}$$

$$\begin{aligned} \underline{Cl_k^{\leq}} &= \{x_i \in X : \Pr(y \leq k|x_i) \geq \alpha\} = \\ &= \{x_i \in X : \Pr(y \geq k+1|x_i) \leq 1-\alpha\}. \end{aligned} \tag{26}$$

where α is some fixed threshold. Thus, lower approximations of class unions are regions, for which objects are assigned to those class unions with high probabilities (at least α). For a given k and two unions $Cl_k^{\geq}$ and $Cl_{k-1}^{\leq}$, the boundary region (defined as before: $B_k = X \backslash (Cl_k^{\geq} \cup Cl_{k-1}^{\leq})$) is the region in which objects belong to any of those unions with probability in the range $(1-\alpha, \alpha)$. Two special cases are important. When $\alpha = 1$, lower approximations reflect the certain region for a given class union (contain only those objects, which surely belong to this class union) and, as we shall shortly see, the stochastic definition boils down to the classical definition of dominance-based lower approximations. When $\alpha \to \frac{1}{2}$, only objects for which $\Pr(y \leq k-1|x_i) = \Pr(y \geq k|x_i) = \frac{1}{2}$ are on the boundary, which correspond to the Bayes boundary between classes [8].

Assume for a moment that the probabilities are known so that we could obtain lower approximations for each class union. It may happen for some object x_i, that although it does not belong to class e.g. $Cl_k^{\geq}$, its class probability satisfies $\Pr(y \geq k|x_i) \geq \alpha$, so that it belongs to $\underline{Cl_k^{\geq}}$. The interpretation of this fact is that although the class index of x_i is less than k, $y_i < t$, such a situation is much less likely than the opposite, since with high probability x_i should have class index greater then k, so that we should change its class union to the more probable one. Therefore stochastic approximations lead to reassigning the objects.

To determine, what should be the range of classes to which an object x_i belongs with high probability, we must take the greatest class index k for which it still holds $x_i \in \underline{Cl_k^{\geq}}$ and the smallest class index k for which it still holds $x_i \in \underline{Cl_k^{\leq}}$. This is exactly the generalized decision defined in (18)-(19), but using the stochastic lower approximations (25)-(26) in the definition (however, equations (20)-(21) does not longer hold). To distinguish between the classical and stochastic definitions of generalized decision we will refer to the latter one as a *stochastic decision*. Concluding, stochastic decision reflects the classes, to which an object belong with high probability, therefore it can be regarded as a sort of confidence interval. In a special case $\alpha = 1$ those class intervals boil down to the generalized decision and cover the whole probability distribution conditioned at a given object x_i – it is certain, that the real class of x_i is inside the interval. On the other hand, such intervals may be too wide, so that we loose information about the objects. Therefore in real-life data, lower values of α are more appropriate.

However, the real probabilities are unknown in almost every case.

Therefore, the problem considered in the next few sections will be the nonparametric estimation of probabilities under stochastic dominance assumption, which is much harder task than in the VPRS case with indiscernibility relation. Since for each $k = 2, \dots, K$ we need to obtain two lower approximations $\underline{Cl}_k^{\geq}$ and $\underline{Cl}_{k-1}^{\leq}$, we must solve $K-1$ binary problems, where in each problem the "positive" class corresponds to the class union $Cl_k^{\geq}$ and the "negative" class – to the class union $Cl_{k-1}^{\leq}$. Therefore one need to estimate probabilities only for a binary-class problems. This will be considered in sections 5, 6 and 7. In sections 8 and 9 we show, that for a given α one can directly obtain stochastic lower approximations without estimating the probabilities. Finally, in section 10 we justify that splitting the procedure into $K-1$ binary problems showing that it does not lead to inconsistent results.

5 Binary-class probability estimation

In this section, we will restrict the analysis to two-class (binary) problem, so we assume $Y = \{0, 1\}$ (indices start with 0 for simplicity and 0 denote "negative" class, while 1 – "positive"). Notice, that $\underline{Cl}_0^{\geq}$ and $\underline{Cl}_1^{\leq}$ are trivial (they are equal to X), so that only $\underline{Cl}_1^{\geq}$ and $\underline{Cl}_0^{\leq}$ are used and will be denoted simply by $\underline{Cl}_1$ and $\underline{Cl}_0$, respectively. Finally notice, that in case of generalized decision, $l(x_i) = u(x_i) = 0$ for $x_i \in \underline{Cl}_0$, $l(x_i) = u(x_i) = 1$ for $x_i \in \underline{Cl}_1$ and $l(x_i) = 0, u(x_i) = 1$ for $x_i \in B$, where B denotes boundary region.

We denote $p_i^1 = \Pr(y \geq 1|x_i) = \Pr(y = 1|x_i)$ and $p_i^0 = \Pr(y \leq 0|x_i) = \Pr(y = 0|x_i)$. The stochastic approximations (25)-(26) now have the following form:

$$\underline{Cl}_k = \{x_i \in X : p_i^k \geq \alpha\}, \tag{27}$$

for $k \in \{0, 1\}$, where α is a chosen threshold. Notice, that for (27) to make sense, it must hold $\alpha \in (0.5, 1]$, since for any x_i, $p_i^0 + p_i^1 = 1$. Since we do not know probabilities p_i^k, we will use instead their ML estimators $\hat{p}_i^k$, where the nonparametric procedure of ML estimation will be used, based only on the stochastic dominance principle. The conditional likelihood function (probability of classes with X being fixed) is a product of binomial distributions and is given by:

$$L(p; y|X) = \prod_{i=1}^{n} (p_i^1)^{y_i} (p_i^0)^{1-y_i} \tag{28}$$

By using $p_i := p_i^1$ (since $p_i^0 = 1 - p_i$), the likelihood can be written as:

$$L(p; y|X) = \prod_{i=1}^{n} (p_i)^{y_i} (1 - p_i)^{1-y_i} \tag{29}$$

The log-likelihood is then:

$$\mathcal{L}(p; y|X) = \sum_{i=1}^{n} \left(y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)\right) \tag{30}$$

The stochastic dominance principle (24) in binary-class case simplifies to:

$$x_i \succeq x_j \Longrightarrow p_i \geq p_j \qquad \forall x_i, x_j \in X \tag{31}$$

To obtain probability estimators $\hat{p}_i$, we need to maximize (30) subject to constraints (31). This is exactly the problem of statistical inference under the order restriction [20].

At the moment, we can prove the following theorem, which strongly reduces the size of the problem:

Theorem 1 *Object $x_i \in X$ is consistent with respect to the dominance principle if and only if $\hat{p}_i = y_i$.*

Proof. We consider the case $y_i = 1$ (the case $y_i = 0$ is analogous). If x_i is consistent, then there is no other object x_j, such that $x_j \succeq x_i$ and $y_j = 0$ (otherwise, it would violate dominance principle and consistency of x_i as well). Thus, for every x_j, such that $x_j \succeq x_i$, $y_j = 1$ and y_j is also consistent (otherwise, due to transitivity of dominance, x_i wouldn't be consistent). Thus, we can set $\hat{p}_j = 1$ for x_j and $\hat{p}_i = 1$ for x_i, and these are the values that maximize the log-likelihood (30) for those objects while satisfying the constraints (31)

Now, suppose $\hat{p}_i = 1$ and assume the contrary that x_i is not consistent, i.e. there exists x_j, $x_j \succeq x_i$, but $y_j = 0$. Then, due to the monotonicity constraints (31), $\hat{p}_j \geq \hat{p}_i = 1$, so $\hat{p}_j = 1$, and the log-likelihood (30) equals to minus infinity, which is surely not the optimal solution to the maximization problem (since at least one feasible solution $\hat{p} \equiv \frac{1}{2}$ with a finite objective value exists). ■

We see, that only consistent objects have probability estimates equal to 1. Therefore, stochastic approximations with $\alpha = 1$ boil down to the classical DRSA lower approximations.

Using Theorem 1 we can set $\hat{p}_i = y_i$ for each consistent object $x_i \in X$ and optimize (30) only for inconsistent objects, which usually gives a large reduction of the problem size (number of variables). In the next section, we show that solving (30) boils down to the isotonic regression problem.

6 Isotonic regression

The problem of isotonic regression [20] appears naturally during the analysis of statistical inference when the order constraints are present. For the purpose of this paper we consider the simplified version of the problem. It is defined in the following way[20]. Let $X = \{x_1, \ldots, x_n\}$ be a finite set with some pre-order (reflexive and transitive) relation $\succeq \subseteq X \times X$. Suppose also that $y: X \to \mathbb{R}$ is some function on X, where $y(x_i)$ is shortly denoted y_i. Any function $p: X \to \mathbb{R}$ is called *isotonic*, if $p_i \geq p_j$ whenever $x_i \succeq x_j$ (where we again used the shorter notation p_i instead of $p(x_i)$). A function $y^*: X \to \mathbb{R}$ is an *isotonic regression* of y if it is the optimal solution to the problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} (y_i - p_i)^2 \\ \text{subject to} \quad & x_i \succeq x_j \Longrightarrow p_i \geq p_j \qquad \forall\, 1 \leq i, j \leq n \end{aligned} \tag{32}$$

so that it minimizes the squared error in the class of all isotonic functions p. In our case, the ordering relation $\succeq$ is the dominance relation, the set X and values of function y on X, i.e. $\{y_1, \ldots, y_n\}$ will have the same meaning as before.

Although squared error seems to be arbitrarily chosen, it can be shown that minimizing many other error functions yields to the same function y^* as in the case of (32). Clearly, we sketch below the assumptions and the content of the theorem, which leads to the so called *generalized* isotonic regression. Details can be found in [20].

Suppose that Φ is a convex function finite on an interval I containing the range of function y on X, i.e. $y(X) \subseteq I$ and Φ has value $+\infty$ elsewhere. Let ϕ be a nondecreasing function on I such that, for each $u \in I$, $\phi(u)$ is a *subgradient* of Φ, i.e. $\phi(u)$ is a number between the left derivative of Φ at u and the right derivative of Φ at u. For each $u, v \in I$ define the function $\Delta_\Phi(u, v)$ by:

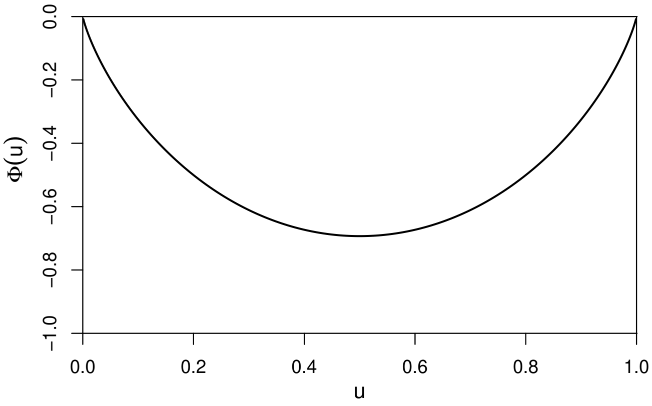

Figure 3: Function $\Phi(u) = u \ln u + (1-u)\ln(1-u)$.

$$\Delta_\Phi(u, v) = \Phi(u) - \Phi(v) - (u - v)\phi(v) \tag{33}$$

Then the following theorem holds:

Theorem 2 *[20] Let y^* be an isotonic regression of y on X, i.e. y^* solves (32). Then it holds:*

$$\sum_{x_i \in X} \Delta_\Phi(y_i, f(x_i)) \geq \sum_{x_i \in X} \Delta_\Phi(y_i, y^*(x_i)) + \sum_{x_i \in X} \Delta_\Phi(y^*(x_i), f(x_i)) \tag{34}$$

for any isotonic function f with the range in I, so that y^* minimizes

$$\sum_{x_i \in X} \Delta_\Phi(y_i, f(x_i)) \tag{35}$$

in the class of all isotonic functions f with range in I. The minimizing function is unique if Φ is strictly convex.

Theorem 2 states, that for any convex function Φ satisfying the assumptions, the isotonic regression function minimizes also the function Δ_Φ. Thus, Theorem 2 can be used to show that the isotonic regression provides a solution for a wide variety of restricted estimation problems in which the objective function do not look at all like least squares [20]. Here, this property will be used to solve the problem (30) under the order restrictions (31).

Let $I = [0, 1]$ and define Φ to be [20]:

$$\Phi(u) = \begin{cases} u \ln u + (1-u)\ln(1-u) & \text{for } u \in (0,1) \\ 0 & \text{for } u \in \{0,1\} \end{cases} \tag{36}$$

(see Fig. 3). One can shown that Φ is indeed convex on I. Then, the first derivative ϕ is given by:

$$\phi(u) = \begin{cases} -\infty & \text{for } u = 0 \\ \ln u - \ln(1-u) & \text{for } u \in (0,1) \\ +\infty & \text{for } u = 1 \end{cases} \tag{37}$$

Then $\Delta_\Phi(u, v)$ for $u, v \in (0,1)$ is given by:

$$\Delta_\Phi(u,v) = u \ln u + (1-u)\ln(1-u) - u \ln v - (1-u)\ln(1-v) \tag{38}$$

It is easy to check, that $\Delta_\Phi(u,v) = 0$ if $u = v = 1$ or $u = v = 0$, and that $\Delta_\Phi(u,v) = +\infty$ for $u = 0, v = 1$ or $u = 1, v = 0$. Now suppose that we want to minimize the function $\sum_{i=1}^n \Delta_\Phi(y_i, f(x_i))$ between all isotonic functions f in the range $I = [0,1]$. Then the first two term in (38) depend only on y_i, so they can be removed from the objective function, thus leading to the problem of minimizing:

$$-\sum_{i=1}^{n} \left(y_i \ln f(x_i) + (1-y_i)\ln(1-f(x_i))\right) \tag{39}$$

between all isotonic functions f in the range I. By denoting $p_i := f(x_i)$ and multiplying by -1 (for maximization) we end up with the problem of maximizing (30) subject to constraints (31).

To summarize, we can find solution to the problem (30) subject to (31) by solving the problem of isotonic regression (32).

Suppose A is a subset of X and $f\colon X \to \mathbb{R}$ is any function. We define $Av(f, A) = \frac{1}{|A|}\sum_{x_i \in A} f(x_i)$ to be an average of f on a set A. Now suppose y^* is the isotonic regression of y. By a *level set* of y^*, $[y^* = a]$ we mean the subset of X, on which y^* has constant value a, i.e. $[y^* = a] = \{x \in X\colon y^*(x) = a\}$. The following theorem holds:

Theorem 3 *[20] Suppose y^* is the isotonic regression of y. If a is any real number such that the level set $[y^* = a]$ is not empty, then $a = Av(y, [y^* = a])$.*

Theorem 3 states, that for a given x, $y^*(x)$ equals to the average of y over all the objects having the same value $y^*(x)$. In other words, if we

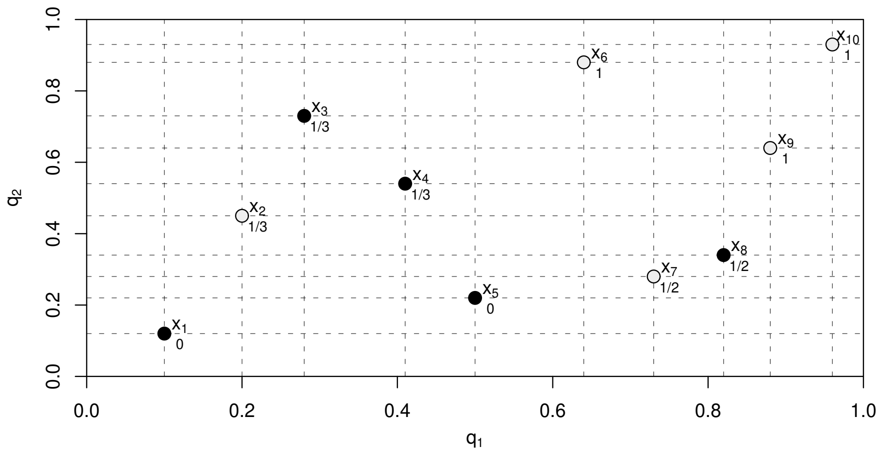

Figure 4: Example of isotonic regression problem. Black points are objects from class 0, light points — from class 1. For every x_i, the values of probabilities $\hat{p}_i$, where $\hat{p}$ is the optimal solution to the isotonic regression problem (32), are shown on the chart.

divide X into disjoint subsets such that for any subset all of the objects have the same value of $y^*(x)$ (so that those subsets are level sets), than $y^*(x)$ must be equal to the average value of y within this subset. Since there is a finite number of divisions of X into level sets, we conclude there are only finite number of values that y^* can possibly take. In our case, since $y_i \in \{0, 1\}$, all values of y^* must be of the form $\frac{r}{r+s}$, where r is the number of objects from class Cl_1 in the level set, while s is the number of objects from Cl_0.

7 Minimal reassignment problem

In this section we briefly describe the problem of minimal reassignment, introduced in [6]. We focus only on the binary problem. Comparing to [6], the notation for decision variables was unified with the notation used in this paper.

We define the reassignment of an object $x_i \in X$ as changing its class index y_i. Moreover, by minimal reassignment we mean reassigning the smallest possible number of objects to make the set X consistent (with respect to the dominance principle). One can see, that such a reassignment of objects corresponds to indicating and correcting possible errors in the dataset. We denote the minimal number of reassigned objects from X by R. To compute R, one can formulate a linear programming problem. Such problems were already considered in [4] (under the name *isotonic separation*, in the context of binary and multi-class classification) and also in [3] (in the context of boolean regression). In [6] the similar problem was formulated, but with a different aim.

Assume $y_i \in \{0, 1\}$. For each object $x_i \in X$ we introduce a binary variable d_i which is to be a new class index for x_i. The request that the class indices must be consistent with respect to the dominance principle implies:

$$x_i \succeq x_j \Longrightarrow d_i \geq d_j \qquad \forall 1 \leq i, j \leq n \tag{40}$$

Notice, that (40) has the form of the stochastic dominance principle (31). The reassignment of an object x_i takes place if $y_i \neq d_i$. Therefore, the number of reassigned objects (which is also the objective function for minimal reassignment problem) is given by:

$$R = \sum_{i=1}^{n} |y_i - d_i| = \sum_{i=1}^{n} (y_i(1 - d_i) + (1 - y_i)d_i) \tag{41}$$

when the last equality is due to the fact, that both $y_i, d_i \in \{0, 1\}$ for each i. Finally notice, that the matrix of constraints (40) is totally unimodular [4, 15, 6], so we can relax the integer condition for d_i reformulating it as $0 \leq d_i \leq 1$, and get a linear programming problem:

$$\begin{array}{lll} \text{minimize} & \displaystyle\sum_{i=1}^{n} (y_i(1 - d_i) + (1 - y_i)d_i) & \\ \text{subject to} & x_i \succeq x_j \Longrightarrow d_i \geq d_j & \forall 1 \leq i, j \leq n \\ & 0 \leq d_i \leq 1 & \forall 1 \leq i \leq n \end{array} \tag{42}$$

We will rewrite the problem (42) in a slightly different form:

$$\begin{array}{lll} \text{minimize} & \displaystyle\sum_{i=1}^{n} |y_i - d_i| & \\ \text{subject to} & x_i \succeq x_j \Longrightarrow d_i \geq d_j & \forall 1 \leq i, j \leq n \end{array} \tag{43}$$

where the last constraint $0 \leq d_i \leq 1$ has been dropped, because if there were any $d_i \geq 1$ (or $d_i \leq 0$) in any feasible solution, we could decrease their values down to 1 (or increase up to 0), obtaining a new feasible solution with smaller value of the objective function of (43).

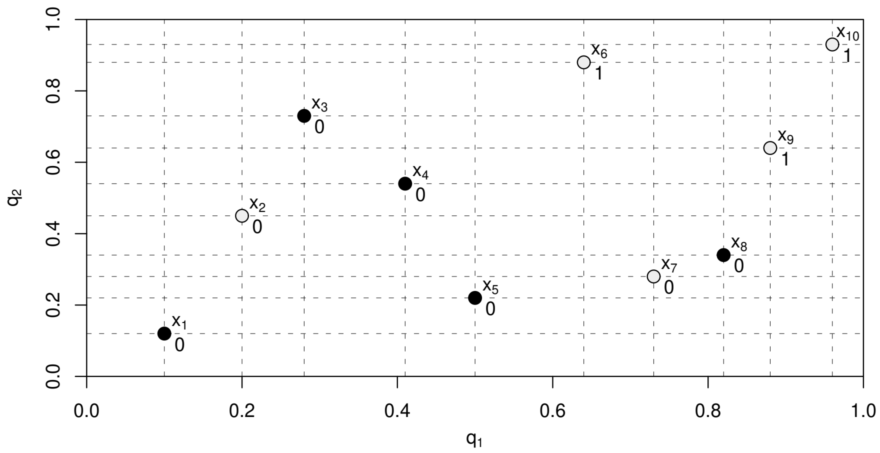

Figure 5: Example of minimal reassignment problem. Black points are objects from class 0, light points — from class 1. For every x_i, the new labels $\hat{d}_i$, where $\hat{d}$ is one (not unique!) of the optimal solutions to the minimal reassignment problem (43), are shown on the chart. There is one more optimal solution $\hat{d}'$, which differs from $\hat{d}$ only for objects x_7, x_8, namely $\hat{d}'_7 = 1, \hat{d}'_8 = 1$.

Comparing (43) with (32), we notice that, although both problems emerged in different context, they look very similar and the only difference is in the objective function. In (32) we minimize L_2-norm (sum of squares) between vectors y and p, while in (43) we minimize L_1-norm (sum of absolute values). In fact, both problems are closely connected, which will be shown in the next section.

8 Relationship between isotonic regression and minimal reassignment

To show the relationship between isotonic regression and minimal reassignment problems we consider the latter to be in more general form, allowing the cost of reassignment to be different for different classes. The *weighted* minimal reassignment problem is given by

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} w_{y_i} |y_i - d_i| \\ \text{subject to} \quad & x_i \succeq x_j \Longrightarrow d_i \geq d_j \qquad \forall 1 \leq i, j \leq n \end{aligned} \tag{44}$$

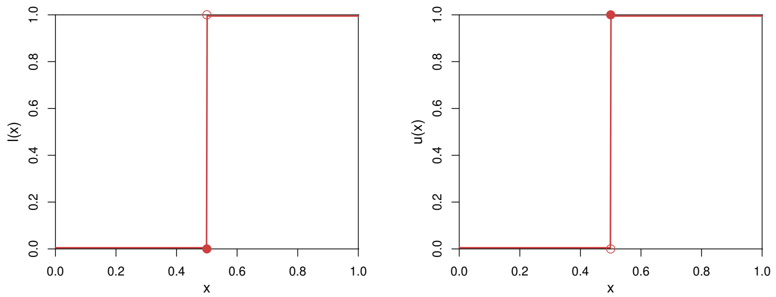

Figure 6: Functions $l(x)$ and $u(x)$ defined in (45)-(46), for value $\alpha = \frac{1}{2}$.

where w_{y_i} are arbitrary, positive weights associated with classes. The following results hold:

Theorem 4 *Suppose $\hat{p} = \{\hat{p}_1, \ldots, \hat{p}_n\}$ is an optimal solution to the problem of isotonic regression (32). Choose some value $\alpha \in [0,1]$ and define two functions:*

$$l(x) = \begin{cases} 0 & \text{if } x \leq \alpha \\ 1 & \text{if } x > \alpha \end{cases} \tag{45}$$

and

$$u(x) = \begin{cases} 0 & \text{if } x < \alpha \\ 1 & \text{if } x \geq \alpha \end{cases} \tag{46}$$

where $x \in \mathbb{R}$ (see Fig. 6). Then the solution $\hat{d}^l = \{\hat{d}^l_1, \ldots, \hat{d}^l_n\}$ given by $\hat{d}^l_i = l(\hat{p}_i)$ for each $i \in \{1, \ldots, n\}$ and the solution $\hat{d}^u = \{\hat{d}^u_1, \ldots, \hat{d}^u_n\}$ given by $\hat{d}^u_i = u(\hat{p}_i)$ for each $i \in \{1, \ldots, n\}$ are the optimal solutions to the problem of weighted minimal reassignment (44) with weights:

$$\begin{aligned} w_0 &= \alpha \\ w_1 &= 1 - \alpha \end{aligned} \tag{47}$$

Moreover, if $\hat{d} = \{\hat{d}_1, \ldots, \hat{d}_n\}$ is an optimal integer solution to the problem of weighted minimal reassignment with weights (47), it must hold $\hat{d}^l_i \leq \hat{d}_i \leq \hat{d}^u_i$, for all $i \in \{1, \ldots, n\}$. In particular, if $\hat{d}^l \equiv \hat{d}^u$, then the solution of the weighted minimal reassignment problem is unique.

Proof. Let us define a function $\Phi(u)$ on the interval $I = [0,1]$ in the following way:

$$\Phi(u) = \begin{cases} \alpha(u - \alpha) & \text{for } x \geq \alpha \\ (1-\alpha)(\alpha - u) & \text{for } x < \alpha \end{cases} \tag{48}$$

It is easy to check, that $\Phi(u)$ is a convex function, but not a strictly convex function. Φ has derivative $\phi(u) = \alpha - 1$ for $u \in [0, \alpha)$ and $\phi(u) = \alpha$ for $u \in (\alpha, 1]$. At point $u = \alpha$, $\Phi(u)$ is not differentiable, but each value in the range $[\alpha - 1, \alpha]$ is a subgradient of $\Phi(u)$.

First, suppose we set $\phi(\alpha) = \alpha - 1$. We remind, that:

$$\Delta_\Phi(u, v) = \Phi(u) - \Phi(v) - (u - v)\phi(v) \tag{49}$$

Now, assume $u \in \{0, 1\}$. To calculate $\Delta_\Phi(u, v)$, we need to consider four cases, depending what are the values of u and v:

1. $u = 0$, $v > \alpha$; then $\Phi(u) = \alpha(1 - \alpha)$, $\Phi(v) = \alpha(v - \alpha)$, $\phi(v) = \alpha$, so that $\Delta_\Phi(u, v) = \alpha$.

2. $u = 0$, $v \leq \alpha$; then $\Phi(u) = \alpha(1 - \alpha)$, $\Phi(v) = (1 - \alpha)(\alpha - v)$, $\phi(v) = \alpha - 1$, so that $\Delta_\Phi(u, v) = 0$.

3. $u = 1$, $v > \alpha$; then $\Phi(u) = \alpha(1 - \alpha)$, $\Phi(v) = \alpha(v - \alpha)$, $\phi(v) = \alpha$, so that $\Delta_\Phi(u, v) = 0$.

4. $u = 1$, $v \leq \alpha$; then $\Phi(u) = \alpha(1 - \alpha)$, $\Phi(v) = (1 - \alpha)(\alpha - v)$, $\phi(v) = \alpha - 1$ so that $\Delta_\Phi(u, v) = 1 - \alpha$.

Using definition (45) of function l, we can comprehensively write those results as:

$$\Delta_\Phi(u, v) = w_u |l(v) - u| \tag{50}$$

for $u \in \{0, 1\}$, where w_u are given by (47). Thus, according to Theorem 2, $\hat{p}$ is the optimal solution to the problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} w_{y_i} |l(p_i) - y_i| && (51) \\ \text{subject to} \quad & x_i \succeq x_j \Longrightarrow p_i \geq p_j \qquad \forall 1 \leq i, j \leq n && (52) \end{aligned}$$

Notice, that $\hat{d}^l = l(\hat{p})$ is also the optimal solution to the problem (51)-(52), because l is a nondecreasing function, so if $\hat{p}$ satisfies constraints (52), then so does $\hat{d}^l$. Moreover, $l(l(x)) = l(x)$, so the value of the objective function (51) is the same for both $\hat{p}$ and $\hat{d}^l$. But $\hat{d}^l$ is integer and, for integer solutions, problems (51)-(52) and (44) are the same, so $\hat{d}^l$ is a solution to the problem (44) with the lowest objective value among all the integer solutions to this problem. But, from the analysis of the unimodularity of constraints matrix of (44) we know that if $\hat{d}^l$ is the solution to (44) with the lowest objective value among the integer solutions, it is also the optimal solution, since there exists an optimal solution to (44), which is integer.

Now, setting $\phi(\alpha) = \alpha$, we repeat the above analysis, which leads to the function u instead of l and shows, that also $\hat{d}^u$ is the optimal solution to the problem (44).

We now prove the second part of the theorem. Assume $v \in \{0, 1\}$ and fix again $\phi(\alpha) = \alpha - 1$. To calculate $\Delta_\Phi(u, v)$, we consider again four cases, depending what are the values of u and v:

1. $u > \alpha$, $v = 0$; then $\Phi(u) = \alpha(u-\alpha)$, $\Phi(v) = \alpha(1-\alpha)$, $\phi(v) = \alpha-1$, so that $\Delta_\Phi(u, v) = u - \alpha > 0$.

2. $u \geq \alpha$, $v = 1$; then $\Phi(u) = \alpha(u - \alpha)$, $\Phi(v) = \alpha(1 - \alpha)$, $\phi(v) = \alpha$, so that $\Delta_\Phi(u, v) = 0$.

3. $u \leq \alpha$, $v = 0$; then $\Phi(u) = (1 - \alpha)(\alpha - u)$, $\Phi(v) = \alpha(1 - \alpha)$, $\phi(v) = \alpha - 1$, so that $\Delta_\Phi(u, v) = 0$.

4. $u < \alpha$, $v = 1$; then $\Phi(u) = (1 - \alpha)(\alpha - u)$, $\Phi(v) = \alpha(1 - \alpha)$, $\phi(v) = \alpha$, so that $\Delta_\Phi(u, v) = \alpha - u > 0$.

From Theorem 2 it follows that:

$$\sum_{i=1}^{n} \Delta_\Phi(y_i, f(x_i)) \geq \sum_{i=1}^{n} \Delta_\Phi(y_i, \hat{p}_i) + \sum_{i=1}^{n} \Delta_\Phi(\hat{p}_i, f(x_i)) \tag{53}$$

for any isotonic function f in the range $[0, 1]$. Notice that if the last term in (53) is nonzero, then f cannot be optimal to the problem (51)-(52) (since then $\hat{p}$ has strictly lower cost than f).

Suppose now that $\hat{d}$ is an optimal integer solution to the minimal reassignment problem (44). But then it is also the solution to the problem (51)-(52) with the lowest objective value between all the integer solutions (since both problems are exactly the same for integer solutions). Since $\hat{d}^l$ is optimal solution to the problem (51)-(52) and is integer (so that there exist integer solution which is optimal), $\hat{d}$ is also optimal solution to this problem. Then, however, the last term in (53) must be zero, so for each $i \in \{1, \ldots, n\}$ it must hold $\Delta_\Phi(\hat{p}_i, \hat{d}_i) = 0$ (since all those terms are nonnegative). As $\hat{d}$ is integer, it is clear from the above analysis of $\Delta_\Phi(u, v)$ for v being integer, that it may only happen, if the following conditions hold:

$$\hat{p}_i > \alpha \Longrightarrow \hat{d}_i = 1 \tag{54}$$

$$\hat{p}_i < \alpha \Longrightarrow \hat{d}_i = 0 \tag{55}$$

for all $i \in \{1, \ldots, n\}$. From the definitions of $\hat{d}^l$ and $\hat{d}^u$ it follows, that for $\hat{p}_i = \alpha$ it holds that $\hat{d}^l_i = 0$ and $\hat{d}^u_i = 1$, for $\hat{p}_i > \alpha$ it holds $\hat{d}^l_i = \hat{d}^u_i = 1$ and for $\hat{p}_i < \alpha$ it holds $\hat{d}^l_i = \hat{d}^u_i = 0$. From this and from (54)-(55) we conclude that:

$$\hat{d}^l_i \leq \hat{d}_i \leq \hat{d}^u_i \tag{56}$$

for all $i \in \{1, \ldots, n\}$, for any optimal integer solution $\hat{d}$ to problem (44). ■

Theorem 4 clearly states, that if the optimal value for a variable $\hat{p}_i$ in the isotonic regression problem (32) is greater (or smaller) than α, then the optimal value for the corresponding variable $\hat{d}_i$ in the weighted minimal reassignment problem (44) with weights (47) is 1 (or 0). In particular, for $\alpha = \frac{1}{2}$ we have $w_0 = w_1 = 1$, so we obtain reassignment problem (43). Comparing Figure 4 and Figure 5, one can see this correspondence (notice, that for objects x_7, x_8 we have $\hat{p}_7 = \hat{p}_8 = \frac{1}{2}$ and thus there are two optimal solutions for minimal reassignment problem — see description under Figure 5).

It also follows from Theorem 4, that if α cannot be taken by any $\hat{p}_i$ in the optimal solution $\hat{p}$ to the isotonic regression problem (32), the optimal solution to the weighted minimal reassignment problem (44) is unique. It follows from the Theorem 3, that $\hat{p}$ can take only finite number of values,which must be of the form $\frac{r}{r+s}$, where $r < n_1$ and $s < n_0$ are integer (n_0 and n_1 are numbers of objects from class, respectively, 0 and 1). Since it is preferred to have a unique solution to the reassignment problem, from now on, we always assume that α was chosen not to be of the form $\frac{r}{r+s}$ (in practice it can easily be done by choosing α to be some simple fraction, e.g. 2/3 and adding some small number ϵ). We call such value of α to be *proper*.

It is worth noticing that weighted minimal reassignment problem is easier to solve than isotonic regression. It is linear, so that one can use linear programming, also it can be transformed to the network flow problem [4] and solved in $O(n^3)$. In the next section, we show, that to obtain stochastic lower approximations, one does not need to solve isotonic regression problem and solves two reassignment problems instead. In other words, one does not need to estimate probabilities and can directly estimate stochastic lower approximations.

9 Summary of stochastic DRSA for binary-class problem

We begin with reminding the definitions of lower approximations of classes (for two-class problem) for threshold α:

$$\underline{Cl}_k = \{x_i \in X : p_i^k \geq \alpha\} \tag{57}$$

for $k \in \{0, 1\}$. The probabilities p^k are estimated using the ML approach and from the previous analysis it follows that the set of estimators $\hat{p}$ is the optimal solution to the isotonic regression problem.

As it was stated in the previous section we choose α to be proper, so that the definition (57) can be equivalently stated as:

$$\begin{aligned} \underline{Cl}_1 &= \{x_i \in X : \hat{p}_i > \alpha\} \\ \underline{Cl}_0 &= \{x_i \in X : 1 - \hat{p}_i > \alpha\} = \{x_i \in X : \hat{p}_i < 1 - \alpha\} \end{aligned} \tag{58}$$

where we replace the probabilities by their maximum likelihood estimators and we used "$>$" instead of "$\geq$", since proper values of α cannot be taken by any $\hat{p}_i$. It follows from Theorem 4, that to obtain $\underline{Cl}_0$ and $\underline{Cl}_1$ we does not need to solve isotonic regression. Instead we solve two weighted minimal reassignment problems (44), first one with weights $w_0 = \alpha$ and $w_1 = 1 - \alpha$, second one with $w_0 = 1 - \alpha$ and $w_1 = \alpha$. Then, objects with new class indices (optimal assignments) $\hat{d}_i = 1$ in the first problem form $\underline{Cl}_1$, while objects with new class indices $\hat{d}_i = 0$ in the second problem form $\underline{Cl}_0$. It is easy to show that the boundary between classes is composed of objects, for which new class indices are different in those two problems.

10 Extension to the multi-class case

Till now, we focused only on binary problems in case of DRSA. However, the theory should also be valid for more general problems, when the number of classes equals to an arbitrary number K.

The first idea is to use the multinomial probability distribution for each point x_i, $\{p_i^1, \ldots, p_i^K\}$. Then, using the maximum likelihood method, we obtain the problem of the following form. We maximize:

$$\mathcal{L}(p; y|X) = \ln L(p; y|X) = \sum_{i=1}^{n} \ln(p_i^{y_i}) \tag{59}$$

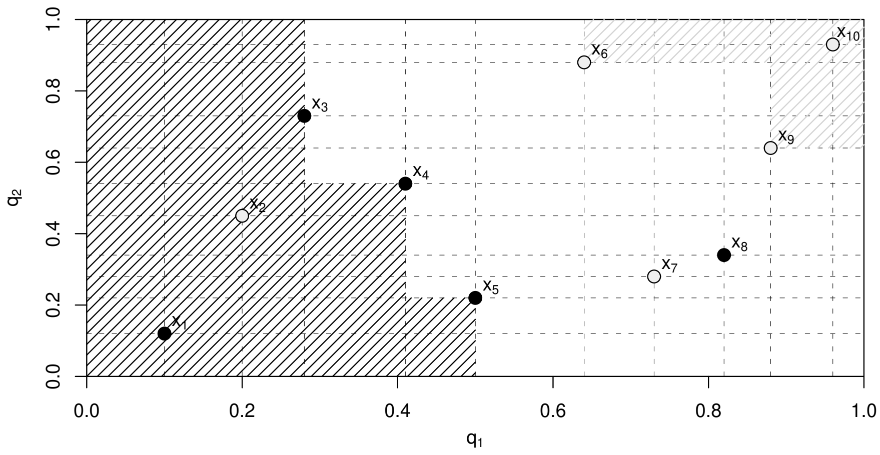

Figure 7: Black points are objects from class 0, light points — from class 1. Lower approximations for threshold $\alpha = 0.6$ were shown on the chart (dashed regions): $\underline{Cl}_0 = \{x_1, x_2, x_3, x_4, x_5\}$, $\underline{Cl}_1 = \{x_6, x_9, x_{10}\}$. Notice, that x_7, x_8 do not belong to any lower approximation, so they are at the boundary between classes.

which is the extension of (30), subject to the constraints:

$$x_i \succeq x_j \Longrightarrow p_i^k \geq p_j^k \qquad \forall k \in Y, \ \forall x_i, x_j \in X \tag{60}$$

Unfortunately, there is a serious problem already present in (59) – it has objective function, which is not strictly convex, so that the problem may not have a unique solution. It is usually the case, that at a certain point x_i there is only one object, i.e. it is not a common situation, that $x_i = x_j$ for some $i, j \in \{1, \dots, n\}$. Then, usually we have only one value y_i to estimate the full probability distribution $\{p_i^1, \dots, p_i^K\}$ at point x_i, from which the lack of strict convexity follows.

Here we propose a different approach, which always gives the unique solution and is based on the sequence of two-class (binary) problems, as was already noted in section 4. By using the unions of classes, DRSA is naturally incorporated to this procedure.

Suppose we have K-class problem. Suppose, we want to calculate the lower approximations of upward union for class k, $\underline{Cl}_k^{\geq}$, and the lower approximation of downward union for class $k-1$, $\underline{Cl}_{k-1}^{\leq}$. Then we form the "negative" class to be $Cl_0 = Cl_{k-1}^{\leq}$, and the "positive" class to be $Cl_1 = Cl_k^{\geq}$. Having obtained the binary problem, we can solve it and get the lower approximations $\underline{Cl}_{k-1}^{\leq}$ and $\underline{Cl}_k^{\geq}$. Repeating the process $K-1$

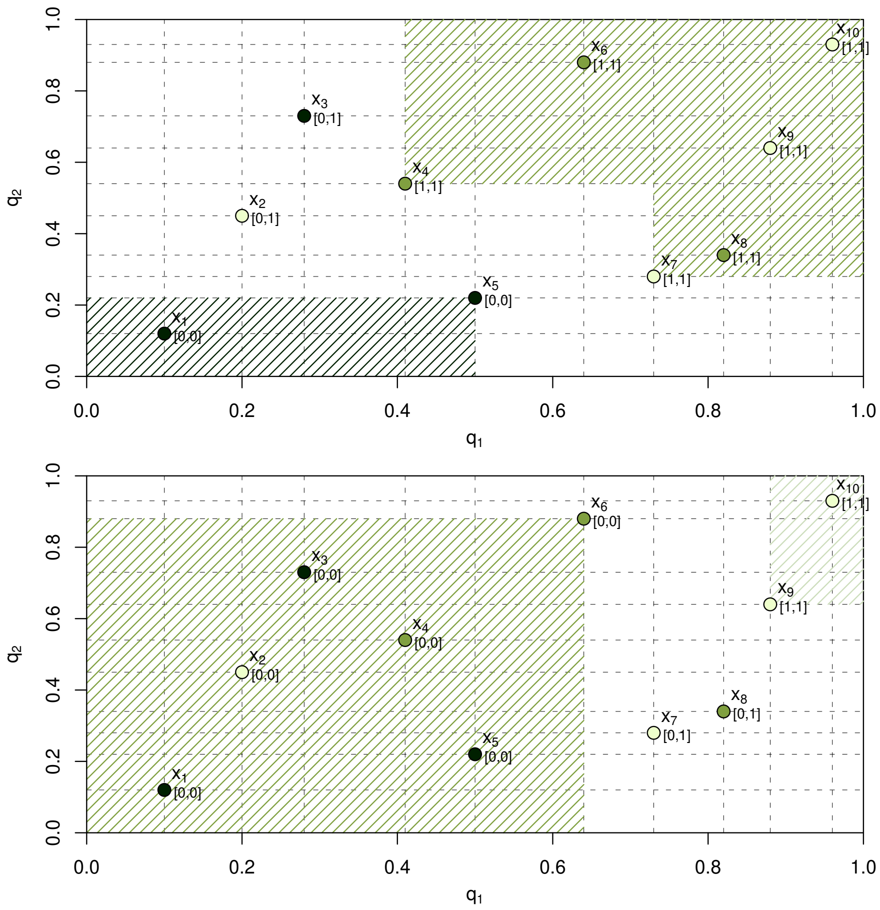

Figure 8: Example of three-class case. Black points are from class 1, green — from class 2, while light — from class 3. The threshold $\alpha = 0.6$. On the upper chart, the solution to the binary problem $Cl_1^{\leq}$ vs. $Cl_2^{\geq}$ is shown (in brackets assignments of new labels in two weighted minimal reassignment problems as described in Section 9). We see, that $\underline{Cl}_1^{\leq} = \{x_1, x_5\}, \underline{Cl}_2^{\geq} = \{x_4, x_6, x_7, x_8, x_9, x_{10}\}$. On the lower chart, the solution to the problem $Cl_2^{\leq}$ vs. $Cl_3^{\geq}$ is shown. Notice, that $\underline{Cl}_2^{\leq} = \{x_1, x_2, x_3, x_4, x_5, x_6\}, \underline{Cl}_3^{\geq} = \{x_9, x_{10}\}$.

times for $k = 2, \ldots, K$ we obtain the whole set of lower approximations for upward and downward unions.

Thus, we divide the problem into $K - 1$ binary problems. This procedure gives the unique solution, since each binary subprocedure gives the solution, which is unique. Notice, that for the procedure to be consistent, it must follow that for any $k' > k$, $\underline{Cl}_{k'}^{\geq} \subseteq \underline{Cl}_{k}^{\geq}$ and $\underline{Cl}_{k}^{\leq} \subseteq \underline{Cl}_{k'}^{\leq}$. In other words, the solution has to satisfy the property of inclusion that is one of the fundamental properties considered in rough set theory. Fortunately, the relation always hold. First we need to prove the following lemma:

Lemma 5 *Let $\hat{p}$ be the optimal solution to the isotonic regression problem (32) for class indices y. Suppose, we introduce a new vector of class indices y', such that $y'_i \geq y_i$ for all $i \in \{1, \ldots, n\}$. Then, $\hat{p}'$, the isotonic regression of y' (optimal solution to the isotonic regression problem for values y'), has the following property: $\hat{p}'_i \geq \hat{p}_i$, for all $i \in \{1, \ldots, n\}$.*

Proof. Assume the contrary, let $\hat{p}'$ be the isotonic regression of y', and there exists i, such that $\hat{p}'_i < \hat{p}_i$. Define two other solutions, $\hat{p}^+$ and $\hat{p}^-$ in the following way:

$$\hat{p}_i^+ = \max\{\hat{p}_i, \hat{p}'_i\}, \tag{61}$$

$$\hat{p}_i^- = \min\{\hat{p}_i, \hat{p}'_i\}. \tag{62}$$

Notice that $\hat{p}^+ \neq \hat{p}'$ and $\hat{p}^- \neq \hat{p}$, since for some i, $\hat{p}'_i < \hat{p}_i$. We show that $\hat{p}^+, \hat{p}^-$ are feasible solutions, i.e. they satisfy constraints of (32). Suppose $x_i \succeq x_j$. Then, since $\hat{p}, \hat{p}'$ are feasible, it follows that $\hat{p}_i \geq \hat{p}_j$ and $\hat{p}'_i \geq \hat{p}'_j$. But from definition of $\hat{p}_i^+$ we have, that $\hat{p}_i^+ \geq \hat{p}_i$ and $\hat{p}_i^+ \geq \hat{p}'_i$, so it also holds that $\hat{p}_i^+ \geq \hat{p}_j$ and $\hat{p}_i^+ \geq \hat{p}'_j$. Then, $\hat{p}_i^+ \geq \max\{\hat{p}_j, \hat{p}'_j\} = \hat{p}_j^+$.

Similarly, from the definition of $\hat{p}_j^-$ we have, that $\hat{p}_j^- \leq \hat{p}_j$ and $\hat{p}_j^- \leq \hat{p}'_j$, so it also holds that $\hat{p}_j^- \leq \hat{p}_i$ and $\hat{p}_j^- \leq \hat{p}'_i$. But then $\hat{p}_j^- \leq \min\{\hat{p}_i, \hat{p}'_i\} = \hat{p}_i^-$. Thus, both $\hat{p}^+, \hat{p}^-$ are feasible.

Let us denote the objective function of (32) as $F(y, p) = \sum_{i=1}^n (y_i - p_i)^2$. Then, we have:

$$\begin{aligned} F(y', \hat{p}^+) - F(y', \hat{p}') &= \sum_{i=1}^n \left(\hat{p}_i^{+2} - \hat{p}_i'^2 - 2y'_i\hat{p}_i^+ - 2y'_i\hat{p}'_i\right) = \\ &= \sum_{i=1}^n \left((\hat{p}_i^+ - \hat{p}'_i)(\hat{p}_i^+ + \hat{p}'_i) - 2y'_i(\hat{p}_i^+ - \hat{p}'_i)\right) \end{aligned} \tag{63}$$

Since from the definition (61) it holds that $\hat{p}_i^+ - \hat{p}'_i \geq 0$ and from the assumption of the theorem it holds $y'_i \geq y_i$, we have:

$$\sum_{i=1}^n 2y'_i(\hat{p}_i^+ - \hat{p}'_i) \geq \sum_{i=1}^n 2y_i(\hat{p}_i^+ - \hat{p}'_i) \tag{64}$$

so that:

$$F(y', \hat{p}^+) - F(y', \hat{p}') \leq \sum_{i=1}^{n} \big((\hat{p}_i^+ - \hat{p}_i')(\hat{p}_i^+ + \hat{p}_i') - 2y_i(\hat{p}_i^+ - \hat{p}_i')\big) \quad (65)$$

Moreover, from (61)-(62) it holds that $\hat{p}_i^+ + \hat{p}_i^- = \hat{p}_i' + \hat{p}_i$, so that:

$$\hat{p}_i^+ - \hat{p}_i' = \hat{p}_i - \hat{p}_i^- \quad (66)$$

and by adding $2\hat{p}_i'$ to both sides of (66):

$$\hat{p}_i^+ + \hat{p}_i' = 2(\hat{p}_i' - \hat{p}_i^-) + (\hat{p}_i + \hat{p}_i^-). \quad (67)$$

Putting (66)-(67) into (65), we finally obtain:

$$\begin{aligned}
&F(y', \hat{p}^+) - F(y', \hat{p}') \leq \\
&\leq \sum_{i=1}^{n} \big((2(\hat{p}_i' - \hat{p}_i^-) + (\hat{p}_i + \hat{p}_i^-))(\hat{p}_i - \hat{p}_i^-) - 2y_i(\hat{p}_i - \hat{p}_i^-)\big) \\
&= \sum_{i=1}^{n} \big(2(\hat{p}_i - \hat{p}_i^-)(\hat{p}_i' - \hat{p}_i^-) + (\hat{p}_i - \hat{p}_i^-)(\hat{p}_i + \hat{p}_i^-) - 2y_i(\hat{p}_i - \hat{p}_i^-)\big) \\
&= \sum_{i=1}^{n} \big(2(\hat{p}_i - \hat{p}_i^-)(\hat{p}_i' - \hat{p}_i^-) + \hat{p}_i^2 - 2y_i\hat{p}_i - \hat{p}_i^{-2} + 2y_i\hat{p}_i^-\big) \\
&= \sum_{i=1}^{n} 2(\hat{p}_i - \hat{p}_i^-)(\hat{p}_i' - \hat{p}_i^-) + F(y, \hat{p}) - F(y, \hat{p}^-) \\
&< \sum_{i=1}^{n} 2(\hat{p}_i - \hat{p}_i^-)(\hat{p}_i' - \hat{p}_i^-) \qquad (68)
\end{aligned}$$

where the last inequality is from the assumption that $\hat{p}$ is the isotonic regression of y, so it is the unique optimal solution for class indices y and $\hat{p} \neq \hat{p}^-$. In the last sum, however, for each i, either $\hat{p}_i = \hat{p}_i^-$ or $\hat{p}_i' = \hat{p}_i^-$, so the sum vanishes. Thus, we have:

$$F(y', \hat{p}^+) - F(y', \hat{p}') < 0 \quad (69)$$

which is a contradiction, since $\hat{p}'$ is the isotonic regression of y', it is the unique optimal solution for class indices y'. ∎

Now, we may state the following theorem:

Theorem 6 *For each $k = 2, \ldots, K$, let $\underline{Cl}_{k-1}^{\leq}$ and $\underline{Cl}_k^{\geq}$ be the sets obtained from solving two-class isotonic regression problem with threshold α for binary classes $Cl_0 = Cl_{k-1}^{\leq}$ and $Cl_1 = Cl_k^{\geq}$. Then, we have:*

$$k' \geq k \implies \underline{Cl}_{k-1}^{\leq} \subseteq \underline{Cl}_{k'-1}^{\leq} \tag{70}$$

$$k' \geq k \implies \underline{Cl}_{k'}^{\geq} \subseteq \underline{Cl}_k^{\geq} \tag{71}$$

Proof. Suppose we have solved the problem for some k. Denote $y_i = 1$ if $x_i \in Cl_k^{\geq}$ and $y_i = 0$ if $x_i \in Cl_{k-1}^{\leq}$. Suppose we have also solved the problem for some $k' \geq k$. Denote $y_i' = 1$ if $x_i \in Cl_{k'}^{\geq}$ and $y_i' = 0$ if $x_i \in Cl_{k'-1}^{\leq}$. Clearly, from the definition of $Cl_{k-1}^{\leq}, Cl_k^{\geq}$ it follows that $y_i \geq y_i'$ for each $i \in \{1, \ldots, n\}$. Then, according to Lemma 5, if $x_i \in \underline{Cl}_{k-1}^{\leq}$ (so that $\hat{p}_i < \alpha$), then also $x_i \in \underline{Cl}_{k'-1}^{\leq}$ (since then $\hat{p}_i' \leq \hat{p}_i < \alpha$). Analogously, if $x_i \in \underline{Cl}_{k'}^{\geq}$, then also $x_i \in \underline{Cl}_k^{\geq}$. This proves the theorem. ■

To summarize, in the previous sections we focused on estimating the stochastic lower approximations. Since the probabilities in the definitions (25)-(26) are unknown, we use their maximum likelihood estimates instead. We showed that we do not need to estimate those probabilities (which is hard), rather we directly calculate lower approximation for a given threshold α (which is easier). Now, having obtain stochastic approximations, we can assign to each object a stochastic decision, as it was described in section 4. In the next section we show, that such intervals have interesting decision-theoretic properties.

11 Decision-theoretical view of variable precision/consistency rough sets

In this section we will look at the problem of variable precision classical rough sets and stochastic DRSA from the point of view of statistical decision theory [1, 14]. A decision-theoretic approach has already been proposed in [23] (for VPRS) and in [13] (for DRSA). The theory presented here for VPRS is slightly different than in [23], while the decision-theoretic view for DRSA proposed in this section is completely novel.

Suppose, we seek for a function (classifier) $f(x)$, which, for a given input vector x, predicts value y as well as possible. To assess the goodness of prediction, the *loss function* $L(y, f(x))$ is introduced for penalizing the prediction error. The simplest loss function for binary classification problem, when $y, f(x) \in \{0, 1\}$, is *0-1 loss*, given by:

$$L_{0-1}(y, f(x)) = \begin{cases} 0 & \text{if } f(x) = y \\ 1 & \text{if } f(x) \neq y \end{cases} \tag{72}$$

However, more complicated loss functions are intensively used in machine learning [14]. Since x and y are the observed values of some random variables, the overall measure of the classifier $f(x)$ is the *expected loss* or *risk*, which is defined as a functional:

$$R(f) = E[L(y, f(x))] = \int L(y, f(x)) dP(y, x) \tag{73}$$

for some probability measure $P(y, x)$. Since $P(y, x)$ is unknown in almost all the cases, one usually minimize the *empirical risk*, which is the value of risk taken for the points from a training sample U:

$$R_e(f) = \sum_{i=1}^{n} L(y_i, f(x_i)) \tag{74}$$

Function f is usually chosen from some restricted family of functions. We now show that the rough set theory leads to the classification procedures, which are naturally suited for dealing with problems when the classifiers are allowed to abstain from giving answer in some cases.

Let us start with the classical theory of variable precision rough sets. We consider the multi-class problem and allow the classification function to give no answer, which is denoted as $f(x) =?$. The loss function suitable for the problem is the following:

$$L_c(y, f(x)) = \begin{cases} 0 & \text{if } f(x) = y \\ 1 & \text{if } f(x) \neq y \\ \beta & \text{if } f(x) =? \end{cases} \tag{75}$$

As we see, there is a penalty β for giving no answer. To be consistent with the classical rough set theory, we assume, that any function must be constant within each granule, i.e. for each $G = I(x)$ for some $x \in X$, we have:

$$x_i, x_j \in G \Longrightarrow f(x_i) = f(x_j) \qquad \forall x_i, x_j \in X \tag{76}$$

which is in fact the principle of indiscernibility. We now state:

Theorem 7 *The function f^* minimizing the empirical risk (74) with loss function (75) between all functions satisfying (76) is equivalent to the VPRS in the sense, that $f^*(G) = t$ if and only if granule G belongs to the lower approximation of class t with the precision threshold $u = 1-\beta$, otherwise $f^*(G) =?$.*

Proof. Since apart from (76) there are no other restrictions for possible functions f, we can analyze the value of f in each granule independently. Let us choose then some granule $G = I(x)$ for some $x \in X$. Let us also denote the number of objects in G as n_G, and for each class index $k \in Y$, let us denote n_G^k as the number of objects from class k in G. It is clear that the total loss of some function f in granule G is the following:

$$L(f(G)) = \begin{cases} n_G - n_G^k & \text{if } f(G) = k \\ \beta \cdot n_G & \text{if } f(G) =? \end{cases} \tag{77}$$

This follows from the fact that if $f(G) = k$, then for each $x_i \in G$ such that $y_i \neq k$, f suffers loss 1. On the other hand, if $f(G) =?$, for each $x_i \in G$, function f suffers loss β. It is obvious that the best strategy is to choose the majority class in G or give no answer, depending which loss is greater. The preferred strategy is choosing the majority class if for some t it holds $n_G - n_G^k \leq \beta n_G$ or:

$$\beta \geq 1 - \frac{n_G^k}{n_G} \tag{78}$$

Otherwise, if no k satisfies this relation, $f^*(G) =?$ is chosen. Comparing this results with section 2, one can see that the decision $f^*(G) = k$ is chosen if granule G belongs to the lower approximation of class k with the precision threshold $u = 1 - \beta$. Clearly, from (4) with probabilities estimated by (8) the above equality follows (we assume that $u > \frac{1}{2}$, so granule G may belong to the lower approximation of one class only). If there is no class for which G is in its lower approximation, the optimal function f^* abstain from answer. ∎

Concluding, the variable precision rough sets can be derived by considering the class of function constant in each granule and choosing the function f^*, which minimizes the empirical risk (74) for loss function (77) with parameter $\beta = 1-u$. For each granule G, if $G \subseteq \underline{Cl}_k$ for some $k \in Y$, then $f^*(x) = k$ for each $x \in G$. Otherwise $f^*(x) =?$ (abstains from answer). As we see, the classical rough set theory suits well for considering the problems, when the classification procedure is allowed to abstain from predictions for some x.

We now turn back to DRSA. The problem here is different, since now we assume, that to each point x, classification function f assigns the interval of classes, denoted as $[l(x), u(x)]$. The lower and upper ends of each interval are supposed to be consistent with the dominance principle:

$$
\begin{aligned}
x_i \succeq x_j &\implies l(x_i) \geq l(x_j) \qquad \forall x_i, x_j \in X \\
x_i \succeq x_j &\implies u(x_i) \geq u(x_j) \qquad \forall x_i, x_j \in X
\end{aligned} \tag{79}
$$

The loss function $L(y, f(x))$ is composed of two terms. First term is a penalty for the size of the interval (degree of imprecision) and equals to $\beta(u(x) - l(x))$. Second term measures the accuracy of the classification and is zero, if $y \in [l(x), u(x)]$, otherwise $f(x)$ suffers additional loss equal to distance of y from the closer interval end:

$$
L(y, f(x)) = \beta(u(x)-l(x)) + I(y \notin [l(x), u(x)]) \min\{|y-l(x)|, |y-u(x)|\} \tag{80}
$$

where $I(\cdot)$ is an indicator function. We now state the following theorem:

Theorem 8 *The function f^* minimizing the empirical risk (74) with loss function (80) between all interval functions satisfying (79) is equivalent to the stochastic DRSA with threshold $\alpha = 1 - \beta$ in the sense, that for each $x \in X$, $f^*(x) = [l^*(x), u^*(x)]$ is a stochastic decision defined by (18)-(19).*

Proof. We first show, how to find the function minimizing the empirical risk using the linear programming approach. Let $l_{ik}, u_{ik} \in \{0, 1\}$, be binary decision variables for each $i \in \{1, \ldots, n\}, k \in \{2, \ldots, K\}$. We code the ranges of interval $f(x_i)$ as $l(x_i) = 1 + \sum_{k=2}^{K} l_{ik}$ and $u(x_i) = 1 + \sum_{k=2}^{K} u_{ik}$. In order to provide the unique coding for each value of $l(x_i)$ and $u(x_i)$ and to ensure that $u(x_i) \geq l(x_i)$ the following properties are sufficient:

$$
u_{ik} \geq l_{ik} \qquad \forall i \in \{1, \ldots, n\}, k \in \{2, \ldots, K\} \tag{81}
$$

$$
l_{ik} \geq l_{ik'} \qquad \forall i \in \{1, \ldots, n\}, k < k' \tag{82}
$$

$$
u_{ik} \geq u_{ik'} \qquad \forall i \in \{1, \ldots, n\}, k < k' \tag{83}
$$

Moreover, for dominance principle (79) to hold, we must also have:

$$
x_i \succeq x_j \Longrightarrow l_{ik} \geq l_{ik} \qquad \forall i \in \{1, \ldots, n\}, k \in \{2, \ldots, K\} \tag{84}
$$

$$
x_i \succeq x_j \Longrightarrow u_{ik} \geq u_{ik} \qquad \forall i \in \{1, \ldots, n\}, k \in \{2, \ldots, K\} \tag{85}
$$

It is not hard to verify, that the loss function (80) for object x_i can be written as:

$$L_i = L(f(x_i), y_i) = \beta \sum_{k=2}^{K} (u_{ik} - l_{ik}) + \sum_{k=y_i+1}^{K} l_{ik} + \sum_{k=2}^{y_i} (1 - u_{ik}) \quad (86)$$

Denoting $y_{ik} = I(y_i \geq k)$, where $I(\cdot)$ is the indicator function, we have:

$$\begin{aligned} L_i &= (1-\beta)\sum_{k=2}^{K} l_{ik}(1-y_{ik}) - \beta \sum_{k=2}^{K} l_{ik} y_{ik} + \\ &\quad \beta \sum_{k=2}^{K} u_{ik}(1-y_{ik}) - (1-\beta)\sum_{k=2}^{K} u_{ik} y_{ik} + \sum_{k=2}^{K} y_{ik} \\ &= \sum_{k=2}^{K} w_{y_{ik}}^{II} |l_{ik} - y_{ik}| + \sum_{k=2}^{K} w_{y_{ik}}^{I} |u_{ik} - y_{ik}| + C \qquad (87) \end{aligned}$$

where C is constant term (which does not depend on l_{ik} and u_{ik}) and weights has the form $w_0^I = \beta, w_1^I = 1-\beta, w_0^{II} = 1-\beta, w_1^{II} = \beta$. But it follows from (87), that minimizing empirical risk $R_e = \sum_{i=1}^{n} L_i$ is equivalent to solving the sequence of $K-1$ pairs of weighted minimal reassignment, as described in section 10 (solving multi-class case as $m-1$ binary problems) and in section 9 (obtaining lower approximations by solving pair of weighted minimal reassignment problems) with the penalty β equal to $1-\alpha$, but with additional constraints (81)-(83). We now show that those constraints are in fact not needed.

Suppose now, we remove constraints (81)-(83). Then we obtain $2(K-1)$ separate problems, since variables $\{l_{i2}\}_{i=1}^n$, $\{u_{i2}\}_{i=1}^n$, ..., $\{l_{iK}\}_{i=1}^n$, $\{u_{iK}\}_{i=1}^n$ are now independent sets and their optimal values can be obtained separately. This is exactly the constructing of stochastic lower approximations in the multi-class case as described before. But it follows from Theorem 6, that constraints (82) and (83) are satisfied at optimality. Moreover, from Theorem 4 and analysis in section 10 it follows that also the constraints (81) are satisfied at optimality. Thus, the optimal solution to problem without constraints (81)-(83) is also the solution to the problem with constraints (81)-(83).

Since constraints are not needed (they are satisfied at optimality), the empirical risk minimization of loss function (80) corresponds to the obtaining stochastic lower approximations (25)-(26). One can check, that the function minimizing the risk, $f^*(x) = [l^*(x), u^*(x)]$ is a stochastic decision defined by (18)-(19).

■

Concluding, the stochastic DRSA can be derived by considering the class of interval functions, for which the lower and upper ends of interval are isotonic (consistent with the dominance principle) and choosing the function f^*, which minimizes the empirical risk (74) with loss function (80) with parameter $a = 1 - \alpha$. For each $x \in X$, $f^*(x)$ is stochastic decision.

12 Conclusions

The paper introduced a new stochastic theory for Dominance-based Rough Set Approach. Application of the theory to the dataset results in estimating a class interval for each object (so called stochastic decision). For a given object x_i such class interval $[l(x_i), u(x_i)]$ have the property that $k \leq l(x_i) \iff \Pr(y \geq k|x_i) \geq \alpha$ and $k \geq u(x_i) \iff \Pr(y \leq k|x_i) \geq \alpha$. In other words, it reflects an interval of classes, to which class index y_i probably belong. On the other hand, such class intervals has the form of confidence intervals and follows from empirical risk minimization of the specific loss function.

In order to obtain stochastic lower approximation we had to consider a problem of probability estimation. Starting from the general remarks about the estimation of probabilities in the classical rough set approach (which appears to be maximum likelihood estimation), we used the same statistical procedure for DRSA, which led us to the isotonic regression problem. The connection between isotonic regression and minimal reassignment solutions was considered and it was shown that in the case of the stochastic lower approximations, it is enough to solve minimal reassignment problem (which is linear), instead of the isotonic regression problem (quadratic) and obtain stochastic lower approximations directly, without estimating probabilities. The approach has also been extended to the multi-class case by solving $K - 1$ binary subproblems for the class unions. The proposed theory has an advantage of basing on well investigated maximum likelihood estimation method – its formulation is clear and simple, it unites seemingly different approaches for classical and dominance-based case.

Finally notice that a connection was established between statistical decision theory and rough set approach. It follows from the analysis that rough set theory can serve as a tools for constructing classifiers, which can abstain from assigning a new object to a class in case of doubt (in classical case) or give imprecise prediction in the form of interval of classes (in DRSA case). However, rough set theory itself has a rather small generalization capacity, due to its nonparametric character, which was shown in section 11. Therefore, the methodology can only be regarded as a analysis of inconsistencies (following from monotonicity constraints) on the training error. For classification of unseen objects, some generalizing classification function must be constructed.

References

[1] Berger, J.: Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York (1993)

[2] Błaszczyński, J., Słowiński, R., Szeląg, M.: Monotonic Variable Consistency Rough Set Approaches. Lecture Notes in Computer Science, **4481** (2007), 126–133

[3] Boros, E., Hammer, P. L., Hooker, J. N.: Boolean regression. Annals of Operations Research, **58**, 3 (1995).

[4] Chandrasekaran ,R., Ryu, Y. U., Jacob, V., Hong, S.: Isotonic separation. INFORMS J. Comput. **17** 2005 462–474

[5] Dembczyński, K., Greco, S., Słowinski, R.: Second-order Rough Approximations in Multi-criteria Classification with Imprecise Evaluations and Assignments, LNAI, **3641** (2005) 54–63

[6] Dembczyński, K., Greco, S., Kotłowski, W., Słowiński, R.: Optimized Generalized Decision in Dominance-based Rough Set Approach. Lecture Notes in Computer Science, **4481** (2007), 118-125

[7] Doumpos, M., Pasiouras, F.: Developing and testing models for replicating credit ratings: A multicriteria approach. Computational Economics **25**(4) (2005) 327–341

[8] Duda, R., Hart, P.: Pattern Classification. Wiley-Interscience, New York (2000).

[9] Greco, S., Matarazzo, B., Słowiński, R.: A new rough set approach to evaluation of bankruptcy risk. In Zopounidis, C., ed.:

Operational Tools in the Management of Financial Risks. Kluwer Academic Publishers, Dordrecht (1998) 121–136

[10] Greco S., Matarazzo, B., Słowiński, R.: Rough approximation of a preference relation by dominance relations. European Journal of Operational Research, **117** (1999) 63–83

[11] Greco S., Matarazzo, B., Słowiński, R.: Rough sets theory for multicriteria decision analysis. European Journal of Operational Research, **129** 1 (2001) 1–47

[12] Greco, S., Matarazzo, B., Słowiński, R., Stefanowski, J.: Variable consistency model of dominance-based rough set approach. [In]: W.Ziarko, Y.Yao (eds.): Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence, vol. 2005, Springer-Verlag, Berlin, 2001, pp. 170-181

[13] Greco, S., Słowiński, R., Yao, Y.: Bayesian Decision Theory for Dominance-based Rough Set Approach. Lecture Notes in Computer Science **4481** (2007), 134–141.

[14] Hastie, T., Tibshirani, R., Friedman J.: The Elements of Statistical Learning. Springer 2003.

[15] Papadimitriou, C. H., Steiglitz, K.: Combinatorial Optimization. Dover Publications, New York 1998.

[16] Pawlak, Z.: Rough sets. International Journal of Information & Computer Sciences, **11** (1982) 341–356

[17] Pawlak, Z.: Rough sets and intelligent data analysis. Information Sciences **147**, 1-4 (2002) 1–12.

[18] Pawlak, Z., Skowron, A.: Rough sets. Some extensions. Information Sciences **177**, 1 (2007) 341–356.

[19] Potharst, R., Feelders, A.J.: Classification trees for problems with monotonicity constraints. SIGKDD Explorations **4**(1) (2002) 1–10

[20] Robertson, T., Wright, F. T., Dykstra, R. L.: Order Restricted Statistical Inference. John Wiley & Sons (1998).

[21] Ryu, Y.U., Chandrasekaran, R., Jacob, V.: Data classification using the isotonic seeparation technique: Application to breast cancer prediction. European Journal of Operational Research **181**(2) (2007) 842–854

[22] Sill, J.: Monotonicity and connectedness in learning systems. Ph.D. dissertation, California Institute of Technology (1997)

[23] Yao, Y., Wong, S.: A decision theoretic Framework for approximating concepts. International Journal of Man-machine Studies, **37**, 6 (1992), 793–809.

[24] Ziarko, W.: Set approximation quality measures in the variable precision rough set model. Soft Computing Systems, Management and Applications, IOS Press, 2001, 442–452.

[25] Ziarko, W.: Probabilistic Rough Sets. Lecture Notes in Artificial Intelligence **3641**, Springer-Verlag 2005, 283–293.