Information Sciences 178 (2008) 4019–4037

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Stochastic dominance-based rough set model for ordinal classification

Wojciech Kotłowski [a], Krzysztof Dembczyński [a], Salvatore Greco [b], Roman Słowiński [a,c,*]

[a] Institute of Computing Science, Poznań University of Technology, Piotrowo 2, Poznań 60-965, Poland
[b] Faculty of Economics, University of Catania, 95129 Catania, Italy
[c] Institute for Systems Research, Polish Academy of Sciences, Warsaw 01-447, Poland

ARTICLE INFO

Article history:
Received 12 November 2007
Received in revised form 18 April 2008
Accepted 7 June 2008

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

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 generalize 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 a two-class (binary) case. The approach is easily generalized to a 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.

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 the value set of each attribute is a subset of $\mathbb{R}$ (even if the scale is purely ordinal, evaluation on attributes can be numbercoded) and the order relation is a linear order $\geqslant$, 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. One can induce a data model 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 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 indices $y_i = k$, $Cl_k = \{x_i \in X : y_i = k\}$.

Thus, ordinal classification problem with monotonicity constraints resembles a typical classification problem considered in machine learning [10,17], but requires two additional constraints. The first one is the assumption of the ordinal scale on each attribute and on class indices. The second constraint is the monotonicity property: the expected class index increases with increasing evaluations 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 [13], which has exactly the form

* Corresponding author. Address: Institute of Computing Science, Poznań University of Technology, Piotrowo 2, Poznań 60-965, Poland.
E-mail addresses: wkotlowski@cs.put.poznan.pl (W. Kotłowski), kdembczynski@cs.put.poznan.pl (K. Dembczyński), salgreco@unict.it (S. Greco), rslowinski@cs.put.poznan.pl (R. Słowiński).

0020-0255/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2008.06.013

of ordinal classification problem with monotonicity constraints. Moreover, in many different domains monotone properties follow from the domain knowledge about the problem and should not be neglected. They have been recognized in applications such as bankruptcy risk prediction [11], breast cancer diagnosis [25], house pricing [23], credit rating [9], liver disorder diagnosis [26] and many others.

As an example, consider the customer satisfaction analysis [15], which aims at determining customer preferences in order to optimize decisions about strategies for launching new products, or about improving the image of existing products. The monotonicity constraints are of fundamental importance here. Indeed, consider two customers, A and B, and suppose that the evaluations of a product by customer A on a set of attributes are better than the evaluations by customer B. In this case, it is reasonable to expect that also the comprehensive evaluation of this product (i.e. class, to which the product is assigned) by customer A is better (or at least not worse) than the comprehensive evaluation made by customer B. As another example, consider the problem of credit rating. One of the attributes could be the degree of regularity in paying previous debts by a consumer (with ordered value set, e.g. "unstable", "acceptable", "very stable"); on the other hand, the class attribute could be the evaluation of potential risk of lending money to a consumer, also with ordered value set (e.g. "high-risk", "medium-risk", "low-risk"); moreover, there exists a natural monotone relationship between the two attributes: the more stable the payment of the debt, the less risky the new credit is.

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 class worse than x_j; such situation violates the monotone properties of the data, so we shall call objects x_i and x_j *inconsistent*. Rough set theory [19,20,22] has been adapted to deal with this kind of inconsistency and the resulting methodology has been called *dominance-based rough set approach* (DRSA) [12,13]. 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, a *variable consistency* model within DRSA has been proposed (VC-DRSA) [14].

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 set theory and show that the variable-precision model [31,32] 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 objects reassignment problem. Finally, we explain the approach as being a solution to the problem of finding a decision function minimizing the empirical risk [2].

We stress that the theory presented in this paper is related to the training set only. In order to properly classify objects outside the training set, a 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 [19], which neither takes into account monotonicity constraints nor are the classes and attribute values 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 induces a partition of the universe into blocks of indiscernible objects, called *granules* [19,13]. The indiscernibility relation I is defined as

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

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 a class $Cl_k, k \in Y$. The lower and upper approximations of class Cl_k are defined, respectively, by

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

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

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 [19]. 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 also certainly an element of Cl_k, while if x belongs to $\overline{Cl}_k$, it may belong to class Cl_k.

For application to the real-life data, some less restrictive definitions were introduced under the name *variable consistency rough sets* (VPRS) [31,32,27]. The new definitions of approximations (where lower approximation is usually replaced by the term *positive region*, which, however, will not be used here) 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

$$\underline{Cl}_k = \{x \in X : \Pr(y = k|I(x)) \geqslant u\}, \tag{4}$$

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

$$\overline{Cl}_k = \{x \in X : \Pr(y = k|I(x)) \geqslant l\}, \tag{5}$$

where l is usually set to $1 - u$ for the complementarity reasons. An example of VPRS lower approximations for a binary-class problem is shown in Fig. 1.

It can be shown that frequencies used for estimating probabilities are the maximum likelihood (ML) estimators under the assumption of common class probability distribution for every object within each granule. The sketch of the derivation is the following. Let us choose a granule $G = I(x)$. Let n_G be the number of objects in G, and for each class Cl_k, 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

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

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

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

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. An example of VPRS lower approximations for a binary-class problem is shown in Fig. 1.

3. Dominance-based rough set approach (DRSA)

Within DRSA [12–14,6,28], 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 x_i has evaluation not worse than x_j on every attribute, $x_{it} \geqslant x_{jt}$, 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

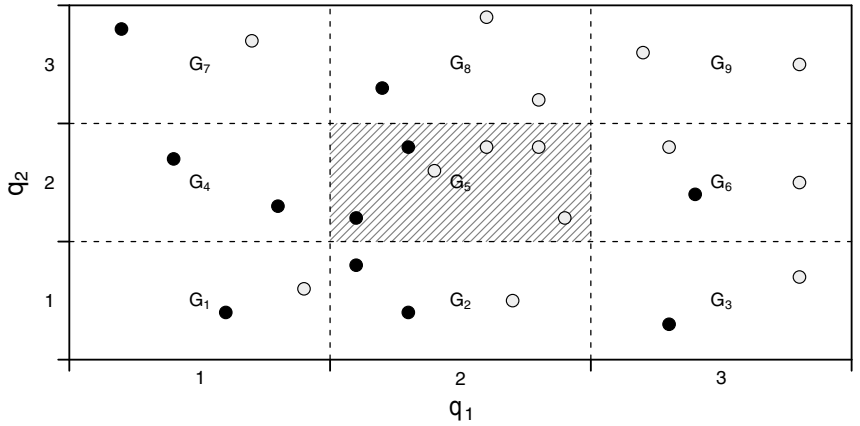

Fig. 1. Example of a two-class problem. Black points are objects from class 1, light points – from class 2. The value sets of two attributes q_1, q_2 are $\{1, 2, 3\}$. At granule G_5, $n_{G_5}^1 = 2$, $n_{G_5}^2 = 4$, so that $p_{G_5}^1 = 1/3, p_{G_5}^2 = 2/3$. We see, that for precision threshold $u = 2/3$ or lower, G_5 belongs to $\underline{Cl}_2$.

$$x_i \succeq x_j \Rightarrow y_i \geqslant y_j. \tag{9}$$

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 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 class indices and by attributes. These granules are called *decision* and *condition* granules, respectively.[1] The decision granules can be expressed by unions of classes

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

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

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 single classes, and dominating (dominated) sets instead of single objects, is a general property of most of the methodologies dealing with ordinal classification problem with monotonicity constraints and follows directly from the monotone nature of the data.

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

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

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

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

Notice, that for any $k \in Y$, we have $Cl_k^{\geqslant} \cup Cl_{k-1}^{\leqslant} = X$. It is not the case with the lower approximations. Therefore we define *the boundary (doubtful) region* [13] for class unions $Cl_k^{\geqslant}$ and $Cl_{k-1}^{\leqslant}$ as

$$B_k = X \setminus (\underline{Cl}_k^{\geqslant} \cup \underline{Cl}_{k-1}^{\leqslant}). \tag{16}$$

This region reflects the area which does not belong to lower approximations of class unions $Cl_k^{\geqslant}$ and $Cl_{k-1}^{\leqslant}$. Notice, that DRSA handles the analysis of inconsistencies by decomposition into $K-1$ separate binary problems: for each $k = 2, \ldots, K$ we have lower approximations $\underline{Cl}_k^{\geqslant}$, $\underline{Cl}_{k-1}^{\leqslant}$ and boundary B_k, which together form the whole set X. Such a decomposition will also be 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 [6]. Consider an object $x_i \in \underline{Cl}_k^{\geqslant}$; since the lower approximation of class union $Cl_k^{\geqslant}$ is a region in which objects certainly belong to $Cl_k^{\geqslant}$, we can state that the class index of x_i should be at least k. Choosing the greatest k for which $x_i \in \underline{Cl}_k^{\geqslant}$ holds (denoted by $l(x_i)$), we know that the class index of x_i must be at least $l(x_i)$; moreover, 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 $x_i \notin \underline{Cl}_k^{\geqslant}$). On the other hand, if $x_i \in \underline{Cl}_k^{\leqslant}$, we known that the class index of x_i must be at most k. By similarly choosing the lowest k for which $x_i \in \underline{Cl}_t^{\leqslant}$ (denoted by $u(x_i)$), we end up with the interval of classes $[l(x_i), u(x_i)]$, for which we know that object x_i must belong to. This interval is often denoted by $\delta(x_i)$, and is called a *generalized decision*[2]:

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

where

$$l(x_i) = \max\{k : x_i \in \underline{Cl}_k^{\geqslant}\}, \tag{18}$$

$$u(x_i) = \min\{k : x_i \in \underline{Cl}_k^{\leqslant}\}. \tag{19}$$

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) and (15) one can show, that generalized decision can be easily computed without reference to the lower approximation:

[1] Those names come from the fact, that in rough set theory the class index for a given object is called *a decision value* and the attributes are called *condition attributes*.

[2] We remind that the class assignments are called *decision values* in rough set theory.

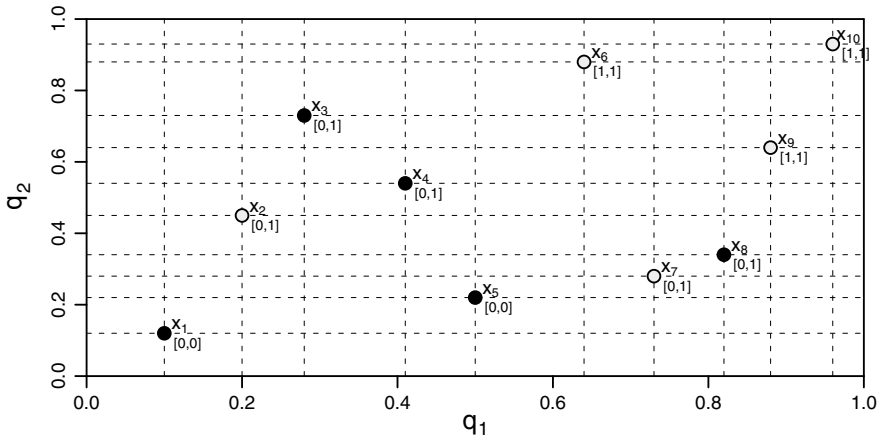

Fig. 2. Example of a two-class problem. Black points are objects from class 0, light points — from class 1 (indexing classes from 0 will be used throughout this paper in binary case). We have $\underline{Cl}_0^{\geqslant} = \underline{Cl}_1^{\leqslant} = X$, $\underline{Cl}_0^{\leqslant} = \{x_1, x_5\}$, $\underline{Cl}_1^{\geqslant} = \{x_6, x_9, x_6\}$. The generalized decisions for objects $x_1, \ldots, x_{10}$ were shown in brackets on the chart.

$$l(x_i) = \min\{y_j : x_j \succeq x_i, x_j \in X\}, \tag{20}$$

$$u(x_i) = \max\{y_j : x_i \succeq x_j, x_j \in X\}. \tag{21}$$

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, $l(x_i) \leqslant y_i \leqslant u(x_i)$ 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. The generalized decision, along with lower approximations for a binary-class problem, are shown in Fig. 2.

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

$$\underline{Cl}_k^{\geqslant} = \{x_i \in X : l(x_i) \geqslant k\}, \tag{22}$$

$$\underline{Cl}_k^{\leqslant} = \{x_i \in X : u(x_i) \leqslant k\}. \tag{23}$$

Finally, notice that the definitions of lower approximations and generalized decisions for DRSA are very restrictive. Suppose, there exists one object dominating many other objects from a dataset, but its class index is the lowest one (e.g. due to a mistake). Then, many of the objects will be included into boundary regions and their generalized decision will be broadened. Therefore, relaxed definitions of lower approximations have been introduced under the name of *variable consistency* DRSA (VC-DRSA) [14,3], 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 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

$$x_i \succeq x_j \Rightarrow \Pr(y \geqslant k|x_i) \geqslant \Pr(y \geqslant k|x_j) \quad \forall k \in Y,\ \forall x_i, x_j \in X, \tag{24}$$

where $\Pr(y \geqslant k|x_i)$ is a probability (conditioned on x_i) of class index at least k. In other words, if object x_i dominates object x_j, the probability distribution conditioned at point x_i *stochastically dominates* the probability distribution conditioned at x_j. Eq. (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

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

$$\begin{aligned}\underline{Cl}_k^{\leqslant} &= \{x_i \in X : \Pr(y \leqslant k|x_i) \geqslant \alpha\},\\ &= \{x=_i \in X : \Pr(y \geqslant k+1|x_i) \leqslant 1-\alpha\},\end{aligned} \tag{26}$$

where α is a fixed threshold. Thus, lower approximation of class union $Cl_k^{\geqslant}$ is a region in which objects are assigned to $Cl_k^{\geqslant}$ with high probability (at least α). The boundary region $B_k = X \setminus (\underline{Cl}_k^{\geqslant} \cup \underline{Cl}_{k-1}^{\leqslant})$ is the region in which objects belong to any of unions $Cl_k^{\geqslant}$ and $Cl_{k-1}^{\leqslant}$ with probability in the range $(1-\alpha, \alpha)$. Two special cases are important. When $\alpha = 1$, lower approximation reflects the certain region for a given class union (contains 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 α becomes close to $\frac{1}{2}$, only objects for which $\Pr(y \leqslant k-1|x_i) = \Pr(y \geqslant k|x_i) = \frac{1}{2}$ are in the boundary B_k, which corresponds to the Bayes boundary between classes [10].

Assume for a while that the probabilities are known so that we can obtain lower approximations for each class union. It may happen for an object x_i, that although it does not belong to the class union $Cl_k^{\geqslant}$, it belongs to $\underline{Cl}_k^{\geqslant}$ (because its class probability satisfies $\Pr(y \geqslant k|x_i) \geqslant \alpha$). The interpretation of this fact is the following: although the class index of x_i observed in the dataset is less than k, i.e. $y_i < k$, such event is less likely than the event $y_i \geqslant k$; hence we should change its class union to the more probable one. Therefore stochastic approximations lead to reassigning the objects.

To determine, what the range of classes to which an object x_i belongs with high probability should be, we must take the greatest class index k for which $x_i \in \underline{Cl}_k^{\geqslant}$ and the smallest class index k for which $x_i \in \underline{Cl}_k^{\leqslant}$. This is exactly the generalized decision defined in (18) and (19), but using the stochastic lower approximations (25) and (26) in the definition (so that Eqs. (20) and (21) do not hold any longer). To distinguish between the classical and stochastic definitions of the generalized decision we will refer to the latter one as a *stochastic decision*. Concluding, the stochastic decision reflects the classes, to which an object belongs 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 decisions and cover the whole probability distribution conditioned at a given object x_i – the real class of x_i is inside the interval with certainty. 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, next few sections will be devoted to the nonparametric estimation of probabilities under stochastic dominance assumption, which is a much harder task than in the VPRS case with indiscernibility relation. Since for each $k = 2, \ldots, K$ we need to obtain two lower approximations $\underline{Cl}_k^{\geqslant}$ and $\underline{Cl}_{k-1}^{\leqslant}$, we must solve $K-1$ binary problems, where in each problem the "positive" class corresponds to the class union $Cl_k^{\geqslant}$ and the "negative" class – to the class union $Cl_{k-1}^{\leqslant}$. Therefore one needs to estimate the probabilities only for the binary-class problems. This will be considered in Sections 5–7. 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 the splitting 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 the binary classification problem, so we assume $Y = \{0, 1\}$ (0 denotes "negative" class, while 1 – "positive"). Notice, that $\underline{Cl}_0^{\geqslant}$ and $\underline{Cl}_1^{\leqslant}$ are trivial (they are equal to X), so that only $\underline{Cl}_1^{\geqslant}$ and $\underline{Cl}_0^{\leqslant}$ 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 the boundary region.

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

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

for $k \in \{0, 1\}$, where α is a chosen threshold value. 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 their ML estimators $\hat{p}_i^k$ instead, and 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

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

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

$$x_i \succeq x_j \Rightarrow p_i \geqslant p_j \quad \forall x_i, x_j \in X. \tag{31}$$

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 [24].

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 would not be consistent). Hence, 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), $\hat{p}_j \geqslant \hat{p}_i = 1$, 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 [24] 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 [24]. 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 by y_i. Any function $p : X \to \mathbb{R}$ is called *isotonic*, if $p_i \geqslant p_j$ 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

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

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 [24].

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

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

Then the following theorem holds

Theorem 2 [24]. *Let y^* be an isotonic regression of y on X, i.e. y^* solves (32). Then it holds*

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

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 does not look like least squares at all [24]. 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 [24]

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

(see Fig. 3). One can show 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 terms in (38) depend only on y_i, so they can be removed from the objective function, thus leading to the problem of minimizing:

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

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). An example of isotonic regression can be found in Fig. 4.

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

Theorem 3 [24]. *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 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), then $y^*(x)$ 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 is a 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.

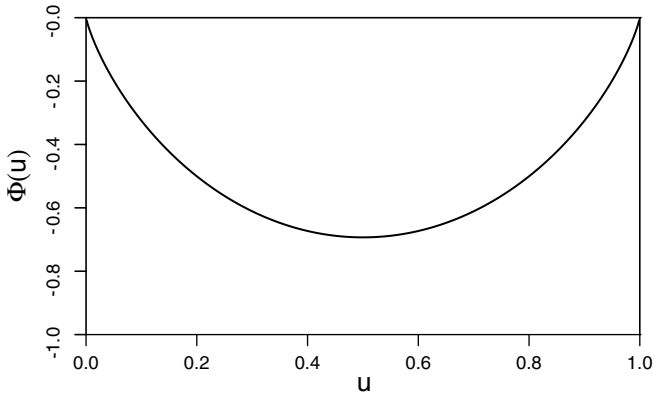

Fig. 3. Function $\Phi(u) = u\ln u + (1-u)\ln(1-u)$.

7. Minimal reassignment problem

In this section we briefly describe the problem of minimal reassignment, introduced in [8]. We focus only on the binary problem. Comparing to [8], 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 inconsistencies 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 [5] (under the name *isotonic separation*, in the context of binary and multi-class classification) and also in [4] (in the context of boolean regression). In [8] the similar problem was formulated, but with a different aim. An example of minimal reassignment for an illustrative binary problem is shown in Fig. 5.

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 demand that the class indices must be consistent with respect to the dominance principle implies:

$$x_i \succeq x_j \Rightarrow d_i \geqslant d_j \quad \forall 1 \leqslant i,j \leqslant n. \tag{40}$$

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

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

where 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 [5,18,8], so we can relax the integer condition for d_i reformulating it as $0 \leqslant d_i \leqslant 1$, and get a linear programming problem.

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n} (y_i(1-d_i) + (1-y_i)d_i) \\ &\text{subject to} && x_i \succeq x_j \Rightarrow d_i \geqslant d_j \qquad \forall 1 \leqslant i,j \leqslant n, \\ & && 0 \leqslant d_i \leqslant 1 \qquad \forall 1 \leqslant i \leqslant n \end{aligned} \tag{42}$$

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

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n} |y_i - d_i|, \\ &\text{subject to} && x_i \succeq x_j \Rightarrow d_i \geqslant d_j \qquad \forall 1 \leqslant i,j \leqslant n, \end{aligned} \tag{43}$$

where the last constraint $0 \leqslant d_i \leqslant 1$ has been dropped, because if there were any $d_i \geqslant 1$ (or $d_i \leqslant 0$) 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).

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 a more general form, allowing the cost of reassignment to be different for different classes. The *weighted* minimal reassignment problem is given by

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

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*

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

and

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

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* $\hat{d}^l_i \leqslant \hat{d}_i \leqslant \hat{d}^u_i$*, 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:

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

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 on what the values of u and v are

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. $u = 0$, $v \leqslant \alpha$; 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. $u = 1$, $v \leqslant \alpha$; 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 the 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

$$\text{minimize} \quad \sum_{i=1}^{n} w_{y_i} |l(p_i) - y_i|, \tag{51}$$

$$\text{subject to} \quad x_i \succeq x_j \Rightarrow p_i \geqslant p_j \qquad \forall 1 \leqslant i, j \leqslant n. \tag{52}$$

Notice, that $\hat{d}^l = l(\hat{p})$ is also the optimal solution to the problem (51) and (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 on what the values of u and v are

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

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

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) and (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) and (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 the optimal solution to the problem (51) and (52) and it is integer (so that there exists an integer solution which is optimal), $\hat{d}$ is also the optimal solution to this problem. Then, however, the last term in (53) must be zero, so for each $i \in \{1,\dots,n\}$ 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:

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

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

for all $i \in \{1,\dots,n\}$. 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) and (55) we conclude that

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

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

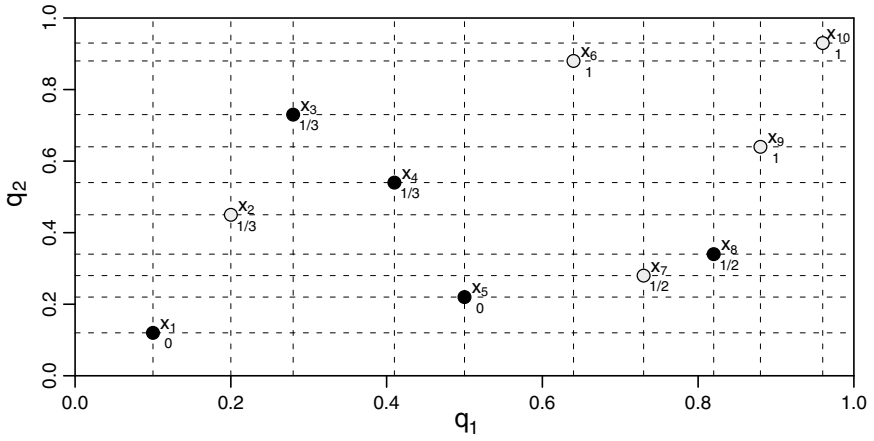

Fig. 4. Example of an isotonic regression problem. Black points are objects from class 0, light points — from class 1. For every x_i, the values of probabilities $\hat{p}_i$, where $\hat{p}$ is the optimal solution to the isotonic regression problem (32), are shown on the chart.

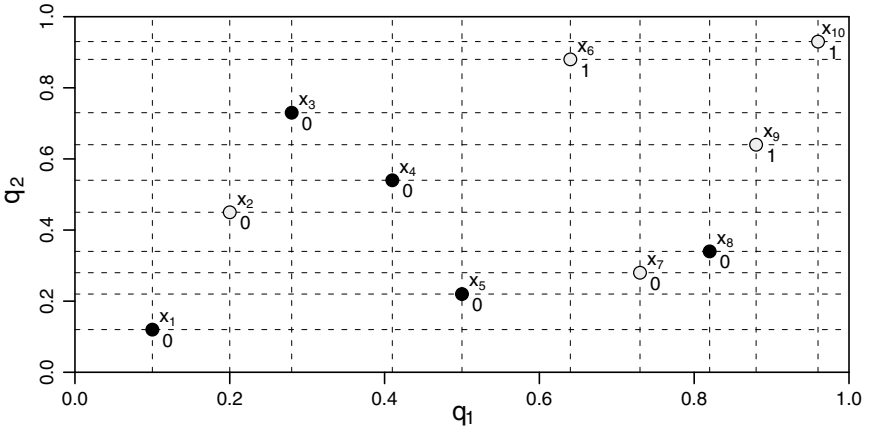

Fig. 5. Example of a minimal reassignment problem. Black points are objects from class 0, light points — from class 1. For every x_i, a new label $\hat{d}_i$ (where $\hat{d} = \{\hat{d}_1,\dots,\hat{d}_n\}$ is one of the optimal solutions to the minimal reassignment problem (43)) is shown on the chart. There is one more optimal solution $\hat{d}\prime$, which differs from $\hat{d}$ only for objects x_7, x_8, namely $\hat{d}'_7 = 1, \hat{d}'_8 = 1$.

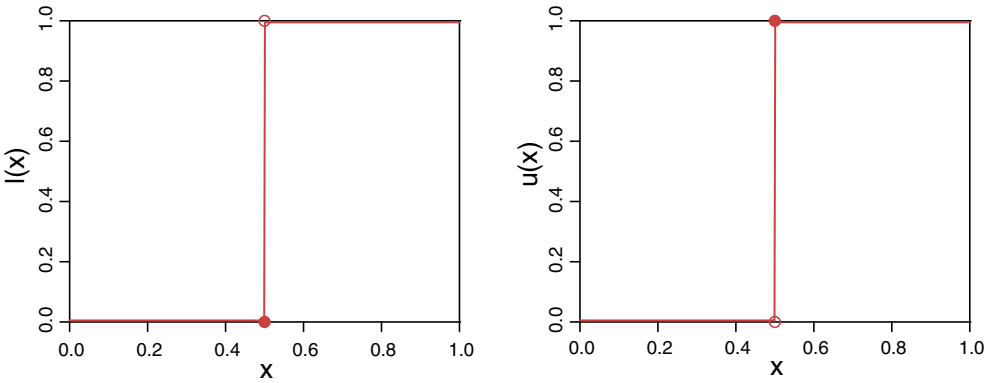

Fig. 6. Functions $l(x)$ and $u(x)$ defined in (45) and (46), for value $\alpha = \frac{1}{2}$.

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 the reassignment problem (43). Comparing Figs. 4 and 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 the Fig. 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 integers (n_0 and n_1 are numbers of objects from class 0 and 1, respectively). 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 a simple ratio, e.g. $2/3$ and adding some small number ϵ). We call such value of α to be *proper*.

It is worth noticing that the weighted minimal reassignment problem is easier to solve than the isotonic regression. It is linear, so that one can use linear programming, it can also be transformed to the network flow problem [5] and solved in $O(n^3)$. In the next section, we show, that to obtain stochastic lower approximations, one does not need to solve the isotonic regression problem, but only 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 a two-class problem) with threshold α

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

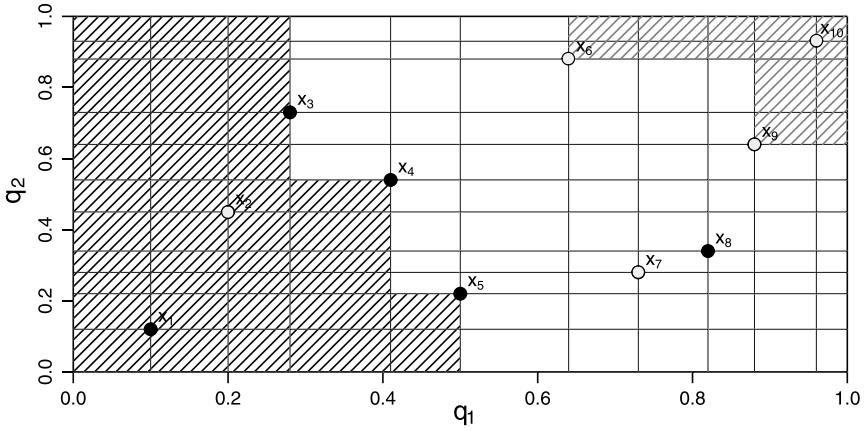

Fig. 7. Black points are objects from class 0, light points — from class 1. Lower approximations for threshold $\alpha = 0.6$ were shown on the chart (dashed regions): $\underline{Cl}_0 = \{x_1, x_2, x_3, x_4, x_5\}$, $\underline{Cl}_1 = \{x_6, x_9, x_{10}\}$. Notice, that x_7, x_8 do not belong to any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

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 use "$>$" instead of "$\geqslant$", 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 do not need to solve isotonic regression. Instead we solve two weighted minimal reassignment problems (44), the first one with weights $w_0 = \alpha$ and $w_1 = 1 - \alpha$, the 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 these two problems (see Fig. 7).

10. Extension to the multi-class case

Till now, we focused only on the 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:

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

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

$$x_i \succeq x_j \Rightarrow p_i^k \geqslant p_j^k \qquad \forall k \in Y, \forall x_i, x_j \in X. \tag{60}$$

Unfortunately, there is a serious problem with (59) – it has an 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 $i,j \in \{1,\ldots,n\}$. Then, usually we have only one value y_i to estimate the full probability distribution $\{p_i^1, \ldots, p_i^K\}$ at point x_i, from which the lack of strict convexity follows.

Here we propose a different approach, which always gives a 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 a K-class problem. Suppose, we want to calculate the lower approximations of upward union for class k, $\underline{Cl}_k^{\geqslant}$, and the lower approximation of downward union for class $k-1$, $\underline{Cl}_{k-1}^{\leqslant}$. Then we set the "negative" class to be $Cl_0 = Cl_{k-1}^{\leqslant}$, and the "positive" class to be $Cl_1 = Cl_k^{\geqslant}$. Having obtained the binary problem, we can solve it and get the lower approximations $\underline{Cl}_{k-1}^{\leqslant}$ and $\underline{Cl}_k^{\geqslant}$. Repeating the process $K-1$ times for $k = 2,\ldots,K$, we obtain the whole set of lower approximations for upward and downward unions (see Fig. 8).

Thus, we divide the problem into $K-1$ binary problems. This procedure gives a unique solution, since each binary subprocedure gives a unique solution. Notice, that for the procedure to be consistent, it must follow that for any $k' > k$, $\underline{Cl}_{k'}^{\geqslant} \subseteq \underline{Cl}_k^{\geqslant}$ and $\underline{Cl}_k^{\leqslant} \subseteq \underline{Cl}_{k'}^{\leqslant}$. 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 holds. 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 $y'_i \geqslant y_i$ for all $i \in \{1,\ldots,n\}$. Then, $\hat{p}'$, the isotonic regression of y' (the optimal solution to the isotonic regression problem for values y'), has the following property: $\hat{p}'_i \geqslant \hat{p}_i$, for all $i \in \{1,\ldots,n\}$.*

Proof. Assume the contrary: let $\hat{p}'$ be the isotonic regression of y', and let i be 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 $\hat{p}_i \geqslant \hat{p}_j$ and $\hat{p}'_i \geqslant \hat{p}'_j$. But from definition of $\hat{p}_i^+$ we have, that $\hat{p}_i^+ \geqslant \hat{p}_i$ and $\hat{p}_i^+ \geqslant \hat{p}'_i$, so it also holds that $\hat{p}_i^+ \geqslant \hat{p}_j$ and $\hat{p}_i^+ \geqslant \hat{p}'_j$. Then, $\hat{p}_i^+ \geqslant \max\{\hat{p}_j, \hat{p}'_j\} = \hat{p}_j^+$.

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

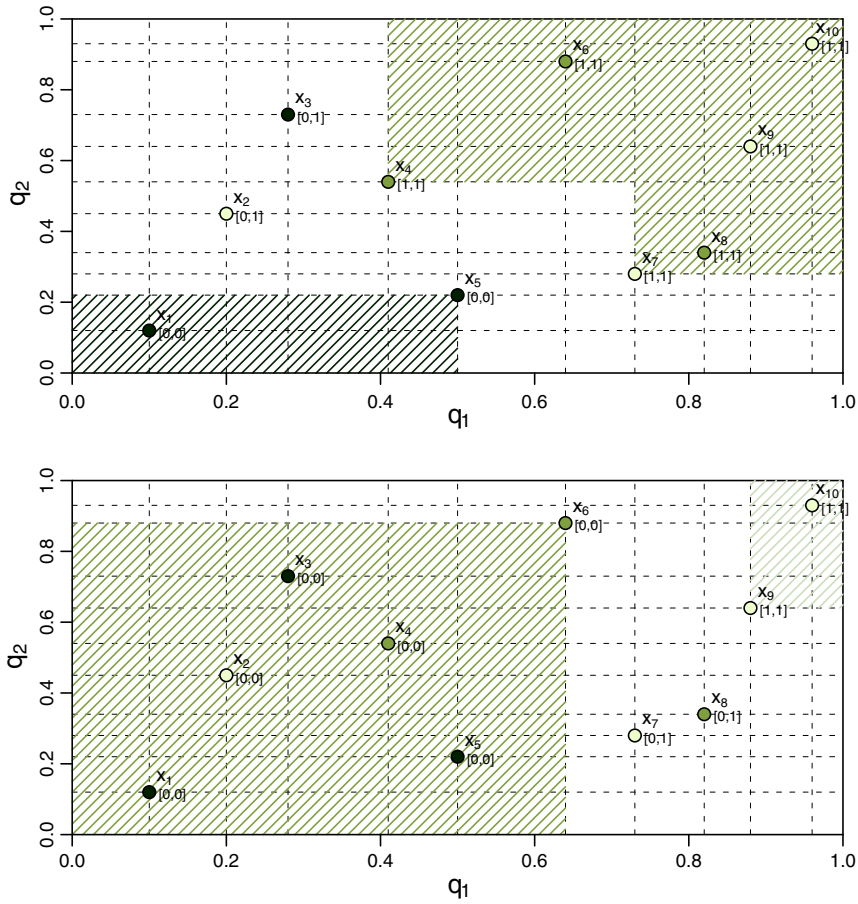

Fig. 8. Example of a three-class case. Black points are from class 1, green — from class 2, while light — from class 3. The threshold $\alpha = 0.6$. On the upper chart, the solution to the binary problem $Cl_1^{\leqslant}$ vs. $Cl_2^{\geqslant}$ is shown (in brackets there are shown the assignments of new labels in two weighted minimal reassignment problems, as described in Section 9). We see, that $\underline{Cl}_1^{\leqslant} = \{x_1, x_5\}, \underline{Cl}_2^{\geqslant} = \{x_4, x_6, x_7, x_8, x_9, x_{10}\}$. On the lower chart, the solution to the problem $Cl_2^{\leqslant}$ vs. $Cl_3^{\geqslant}$ is shown. Notice, that $\underline{Cl}_2^{\leqslant} = \{x_1, x_2, x_3, x_4, x_5, x_6\}, \underline{Cl}_3^{\geqslant} = \{x_9, x_{10}\}$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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

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

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

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

so that

$$F(y', \hat{p}^+) - F(y', \hat{p}') \leqslant \sum_{i=1}^{n} \left((\hat{p}_i^+ - \hat{p}_i')(\hat{p}_i^+ + \hat{p}_i') - 2y_i(\hat{p}_i^+ - \hat{p}_i') \right). \tag{65}$$

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

$$\hat{p}_i^+ - \hat{p}_i' = \hat{p}_i - \hat{p}_i^- \tag{66}$$

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

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

Putting (66) and (67) into (65), we finally obtain

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

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

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

which is a contradiction, since $\hat{p}'$ is the isotonic regression of y'. □

Now, we may state the following theorem.

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

$$k' \geqslant k \Rightarrow \underline{Cl}^{\leqslant}_{k-1} \subseteq \underline{Cl}^{\leqslant}_{k'-1}, \tag{70}$$

$$k' \geqslant k \Rightarrow \underline{Cl}^{\geqslant}_{k'} \subseteq \underline{Cl}^{\geqslant}_{k}. \tag{71}$$

Proof. Suppose we have solved the problem for some k. Denote $y_i = 1$ if $x_i \in Cl^{\geqslant}_{k}$ and $y_i = 0$ if $x_i \in Cl^{\leqslant}_{k-1}$. Suppose we have also solved the problem for some $k' \geqslant k$. Denote $y'_i = 1$ if $x_i \in Cl^{\geqslant}_{k'}$ and $y'_i = 0$ if $x_i \in Cl^{\leqslant}_{k'-1}$. Clearly, from the definition of $Cl^{\leqslant}_{k-1}, Cl^{\geqslant}_{k}$ it follows that $y_i \geqslant y'_i$ for each $i \in \{1,\ldots,n\}$. Then, according to Lemma 5, if $x_i \in \underline{Cl}^{\leqslant}_{k-1}$ (so that $\hat{p}_i < \alpha$), then also $x_i \in \underline{Cl}^{\leqslant}_{k'-1}$ (since then $\hat{p}'_i \leqslant \hat{p}_i < \alpha$). Analogously, if $x_i \in \underline{Cl}^{\geqslant}_{k}$, then also $x_i \in \underline{Cl}^{\geqslant}_{k}$. 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) and (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 approximations for a given threshold α (which is easier). Now, having obtained 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 the stochastic decision intervals have interesting decision-theoretic properties.

Notice that the probability estimation is done by minimizing the squared error in the class of all monotone functions. Such a class of functions can be too broad, especially when m (dimension of the attribute space) grows. Then, the dominance relation becomes sparse; this, in turn, makes the dataset more and more consistent because only few objects are comparable by dominance relation. This may deteriorate the estimation of probability and, in the extreme case, when the dataset is completely consistent, for each object x_i the probability distribution becomes concentrated on a single class y_i. This is one of the symptoms of the famous curse of dimensionality [1].

Generally, the quality of probability estimates depends on the number of objects which can be compared by dominance relation. If this number is high, the estimates are reliable, however, it is low in a high-dimensional space. In such cases, one should decrease the dimension of the space by removing some of the attributes. Since we are dealing with the problems in which the domain knowledge (in the form of monotonicity constraints) is present, the ideal process of attribute selection would be supervised by a domain expert. If such supervision is impossible, the attribute selection can be done by searching for reducts with respect to the *quality of approximation*. A measure of the quality of approximation, particularly useful for this purpose, was proposed in [7]. Here, however, we will not consider these issues in greater detail.

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 [2,17]. A decision-theoretic approach has already been proposed for VPRS [29,30,21] and for DRSA [16]. The theory presented here for VPRS is slightly different than in [29], 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 [17]. 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

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

for some probability measure $P(y,x)$. Since $P(y,x)$ is unknown in almost all cases, one usually minimizes 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 an 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

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

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 $f^*(G) = k$ if and only if granule G belongs to the lower approximation of class k 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 a 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 a function f in the 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 $y_i \neq k$, function 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 abstain from answer, depending on which loss is lower. The preferred strategy is to choose the majority class, if for a given k it holds $n_G - n_G^k \leqslant \beta n_G$ or

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

Otherwise, if no k satisfies this relation, the preferred strategy is to choose $f^*(G) = ?$. Comparing this result 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 inequality 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^* abstains from answer. □

Concluding, the variable precision rough sets can be derived by considering the class of functions 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 a given $k \in Y$, then $f^*(x) = k$ for each $x \in G$. Otherwise $f^*(x) = ?$ (abstaining 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 an 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:

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

The loss function $L(y, f(x))$ is composed of two terms. The first term is a penalty for the size of the interval (degree of imprecision) and equals to $\beta(u(x) - l(x))$. The 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 the 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) and (19).*

Proof. First we 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 lower and upper ends 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 $u(x_i) \geqslant l(x_i)$, the following properties are sufficient:

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

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

$$u_{ik} \geqslant u_{ik'} \quad \forall i \in \{1, \ldots, n\}, k < k'. \tag{83}$$

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

$$x_i \succeq x_j \Rightarrow l_{ik} \geqslant l_{ik} \quad \forall i \in \{1, \ldots, n\}, k \in \{2, \ldots, K\}, \tag{84}$$

$$x_i \succeq x_j \Rightarrow u_{ik} \geqslant u_{ik} \quad \forall i \in \{1, \ldots, n\}, k \in \{2, \ldots, K\}. \tag{85}$$

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

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

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

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

where C is a constant term (which does not depend on l_{ik} and u_{ik}), and $w^I_0 = \beta, w^I_1 = 1 - \beta, w^{II}_0 = 1 - \beta, w^{II}_1 = \beta$. 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 the multi-class case as $K - 1$ binary problems) and in Section 9 (obtaining lower approximations by solving a 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 $\{l_{i2}\}_{i=1}^{n}, \{u_{i2}\}_{i=1}^{n}, \ldots, \{l_{iK}\}_{i=1}^{n}, \{u_{iK}\}_{i=1}^{n}$ are now independent sets and their optimal values can be obtained separately. This is exactly the construction 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 the 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 obtaining stochastic lower approximations (25) and (26). One can check, that the function minimizing the risk, $f^*(x) = [l^*(x), u^*(x)]$ is a stochastic decision defined by (18) and (19). □

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

12. Conclusions

The paper introduced a new stochastic approach to dominance-based rough sets. Application of the approach results in estimating the class intervals for each object (so called stochastic decision). For a given object x_i, such class interval $[l(x_i), u(x_i)]$ has the property that $k \leqslant l(x_i) \iff \Pr(y \geqslant k|x_i) \geqslant \alpha$ and $k \geqslant u(x_i) \iff \Pr(y \leqslant k|x_i) \geqslant \alpha$. In other words, it reflects an interval of classes, to which class index y_i probably belongs. On the other hand, such a class interval has the form of a confidence interval and follows from the 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 general remarks about the estimation of probabilities in the classical rough set approach (which appears to be the 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 the minimal reassignment problem (which is linear), instead of the isotonic regression problem (quadratic), and obtain stochastic lower approximations directly, without estimating the 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 the advantage of basing on the well investigated maximum likelihood estimation method – its formulation is clear and simple, it unites seemingly different approaches for classical and dominance-based cases.

Finally, notice that a connection was established between the statistical decision theory and the rough set approach. It follows from the analysis that rough set theory can serve as a tool 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 an 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 the analysis of inconsistencies (following from the monotonicity constraints) on the training error. For classification of unseen objects, a generalizing classification function must be constructed.

References

[1] R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton University Press, 1961.
[2] J. Berger, Statistical Decision Theory and Bayesian Analysis, Springer-Verlag, New York, 1993.
[3] J. Błaszczyński, S. Greco, R. Słowiński, M. Szeląg, Monotonic variable consistency rough set approaches, Lecture Notes in Computer Science 4481 (2007) 126–133.
[4] E. Boros, P.L. Hammer, J.N. Hooker, Boolean regression, Annals of Operations Research 58 (1995) 3.
[5] R. Chandrasekaran, Y.U. Ryu, V. Jacob, S. Hong, Isotonic separation, INFORMS Journal of Computational 17 (2005) 462–474.
[6] K. Dembczyński, S. Greco, R. Słowinski, Second-order rough approximations in multi-criteria classification with imprecise evaluations and assignments, Lecture Notes in Artificial Intelligence 3641 (2005) 54–63.
[7] K. Dembczyński, S. Greco, W. Kotłowski, R. Słowiński, Quality of Rough Approximation in Multi-Criteria Classification Problems, Lecture Notes in Computer Science, 4259, Springer, 2006. pp. 318–327.
[8] K. Dembczyński, S. Greco, W. Kotłowski, R. Słowiński, Optimized generalized decision in dominance-based rough set approach, Lecture Notes in Computer Science 4481 (2007) 118–125.
[9] M. Doumpos, F. Pasiouras, Developing and testing models for replicating credit ratings: a multicriteria approach, Computational Economics 25 (4) (2005) 327–341.
[10] R. Duda, P. Hart, Pattern Classification, Wiley-Interscience, New York, 2000.
[11] S. Greco, B. Matarazzo, R. Słowiński, A new rough set approach to evaluation of bankruptcy risk, in: C. Zopounidis (Ed.), Operational Tools in the Management of Financial Risks, Kluwer Academic Publishers, Dordrecht, 1998, pp. 121–136.
[12] S. Greco, B. Matarazzo, R. Słowiński, Rough approximation of a preference relation by dominance relations, European Journal of Operational Research 117 (1999) 63–83.
[13] S. Greco, B. Matarazzo, R. Słowiński, Rough sets theory for multicriteria decision analysis, European Journal of Operational Research 129 (1) (2001) 1–47.
[14] S. Greco, B. Matarazzo, R. Słowiński, J. Stefanowski, 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.
[15] S. Greco, B. Matarazzo, R. Słowiński, Rough set approach to customer satisfaction analysis, Lecture Notes in Computer Science 4259 (2006) 284–295.
[16] S. Greco, R. Słowiński, Y. Yao, Bayesian decision theory for dominance-based rough set approach, Lecture Notes in Computer Science 4481 (2007) 134–141.
[17] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, 2003.
[18] C.H. Papadimitriou, K. Steiglitz, Combinatorial Optimization, Dover Publications, New York, 1998.
[19] Z. Pawlak, Rough sets, International Journal of Information and Computer Sciences 11 (1982) 341–356.
[20] Z. Pawlak, Rough sets and intelligent data analysis, Information Sciences 147 (1–4) (2002) 1–12.
[21] Z. Pawlak, Rough sets, decision algorithms and Bayes' theorem, European Journal of Operational Research 136 (1) (2002) 181–189.

[22] Z. Pawlak, A. Skowron, Rough sets. Some extensions, Information Sciences 177 (1) (2007) 341–356.
[23] R. Potharst, A.J. Feelders, Classification trees for problems with monotonicity constraints, SIGKDD Explorations 4 (1) (2002) 1–10.
[24] T. Robertson, F.T. Wright, R.L. Dykstra, Order Restricted Statistical Inference, John Wiley & Sons, 1998.
[25] Y.U. Ryu, R. Chandrasekaran, V. Jacob, Data classification using the isotonic seeparation technique: application to breast cancer prediction, European Journal of Operational Research 181 (2) (2007) 842–854.
[26] J. Sill, Monotonicity and connectedness in learning systems. Ph.D. dissertation, California Institute of Technology, 1997.
[27] D. Ślezak, W. Ziarko, The investigation of the Bayesian rough set model, International Journal of Approximate Reasoning 40 (1–2) (2005) 81–91.
[28] X. Yang, J. Yang, C. Wu, D. Yu, Dominance-based rough set approach and knowledge reductions in incomplete ordered information system, Information Sciences 178 (4) (2008) 1219–1234.
[29] Y. Yao, S. Wong, A decision theoretic framework for approximating concepts, International Journal of Man–Machine Studies 37 (6) (1992) 793–809.
[30] Y. Yao, Probabilistic approaches to rough sets, Expert Systems 20 (5) (2003) 287–297.
[31] W. Ziarko, Set approximation quality measures in the variable precision rough set model, Soft Computing Systems, Management and Applications, IOS Press, 2001. pp. 442–452.
[32] W. Ziarko, Probabilistic Rough Sets, Lecture Notes in Artificial Intelligence, 3641, Springer-Verlag, 2005. pp. 283–293.