Statistical Model for Rough Set Approach to Multicriteria Classification

Krzysztof Dembczyński[1], Salvatore Greco[2], Wojciech Kotłowski[1] and Roman Słowiński[1,3]

[1] Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland
{kdembczynski, wkotlowski, rslowinski}@cs.put.poznan.pl
[2] Faculty of Economics, University of Catania, 95129 Catania, Italy
salgreco@unict.it
[3] Institute for Systems Research, Polish Academy of Sciences, 01-447 Warsaw, Poland

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 multicriteria classification. However, in real-life problems, in the presence of noise, the notions of rough approximations were found to be excessively restrictive, which led to the proposal of the Variable Consistency variant of DRSA. In this paper, we introduce a new approach to variable consistency that is based on maximum likelihood estimation. For two-class (binary) problems, it leads to the isotonic regression problem. The approach is easily generalized for the multi-class case. Finally, we show the equivalence of the variable consistency rough sets to the specific risk-minimizing decision rule in statistical decision theory.

1 Introduction

In decision analysis, a multicriteria classification problem is considered that consists in assignment of objects to m *decision classes* Cl_t, $t \in T = \{1, \ldots, m\}$. The classes are preference ordered according to an increasing order of class indices, i.e. for all $r, s \in T$, such that $r > s$, the objects from Cl_r are strictly preferred to objects from Cl_s. Objects are evaluated on a set of *condition criteria*, i.e. attributes with preference ordered value sets. It is assumed that a better evaluation of an object on a criterion, with other evaluations being fixed, should not worsen its assignment to a decision class. In order to construct a preference model, one can induce it from a *reference* (*training*) set of objects U already assigned to decision classes. Thus, multicriteria classification problem resembles typical classification problem considered in machine learning [6, 11] under monotonicity constraints: the expected decision value increases with increasing values on condition attributes. However, it still may happen that in U, there exists an object x_i not worse than another object x_k on all condition attributes, however, x_i is assigned to a worse class than x_k; such a situation violates the

monotone nature of data, so we shall call objects x_i and x_k *inconsistent with respect to dominance principle.*

Rough set theory [13] has been adapted to deal with this kind of inconsistency and the resulting methodology has been called *Dominance-based Rough Set Approach* (DRSA) [7, 8]. 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 decision 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) [9].

In this paper, we look at DRSA from a different point of view, identifying its connections with statistics and statistical decision theory. Using the maximum likelihood estimation we introduce a new variable consistency variant of DRSA. It leads to the statistical problem of isotonic regression [14], which is then solved by the optimal object reassignment problem [5]. Finally, we explain the approach as being a solution to the problem of finding a decision minimizing the empirical risk [1].

Notation. We assume that we are given a set $U = \{(x_1, y_1), \ldots, (x_\ell, y_\ell)\}$, consisting of ℓ training objects, with their decision values (class assignments), where each $y_i \in T$. Each object is described by a set of n condition criteria $Q = \{q_1, \ldots, q_n\}$ and by $\mathrm{dom}q_i$ we mean the set of values of attribute q_i. For each i, $\mathrm{dom}q_i$ is ordered by some weak preference relation, here we assume for simplicity $\mathrm{dom}q_i \subseteq \mathbb{R}$ and the order relation is a linear order $\geq$. We denote the evaluation of object x_i on attribute q_j by $q_j(x_i)$. Later on we will abuse a bit the notation, identifying each object x with its evaluations on all the condition criteria, $x \equiv (q_1(x), \ldots q_n(x))$ and denote $X = \{x_1, \ldots, x_\ell\}$. By *class* $Cl_t \subset X$, we mean a set of objects, such that $y_i = t$, i.e. $Cl_t = \{x_i \in X : y_i = t, 1 \leq i \leq \ell\}$.

2 Classical variable precision rough set approach

The classical rough set approach [13] (which does not take into account any monotonicity constraints) is based on the assumption that objects having the same description are indiscernible (similar) with respect to the available information [13, 8]. The indiscernibility relation I is defined as:

$$I = \{(x_i, x_j) \in X \times X: \ \ q_k(x_i) = q_k(x_j) \ \ \forall q_k \in Q\} \tag{1}$$

The equivalence classes of I (denoted $I(x)$ for some object $x \in X$) are called *granules.* The lower and upper approximations of class Cl_t are defined, respectively, by:

$$\underline{Cl}_t = \{x_i \in X : I(x_i) \subseteq Cl_t\} \qquad \overline{Cl}_t = \bigcup_{x_i \in Cl_t} I(x_i) \tag{2}$$

For application to the real-life data, a less restrictive definition was introduced under the name of *variable precision rough set model* (VPRS) [16] and is expressed in the probabilistic terms. Let $\Pr(Cl_t|I(x))$ be a probability that an object x_i from granule $I(x)$ belongs to the class Cl_t. The probabilities are unknown, but are estimated by frequencies $\Pr(Cl_t|I(x)) = \frac{|Cl_t \cap I(x)|}{|I(x)|}$. Then, the lower approximation of class Cl_t is defined as:

$$\underline{Cl_t} = \bigcup_{I(x):x \in X} \{I(x): \Pr(Cl_t|I(x)) \geq u\} \tag{3}$$

so it is the sum of all granules, for which the probability of class Cl_t is at least equal to some threshold u.

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^t be the number of objects from this class in G. Then the decision value y has a multinomial distribution when conditioned on granule G. Let us denote those probabilities $\Pr(y = t|G)$ by p_G^t. Then, the conditional probability of observing $n_G^1, \ldots n_G^t$ objects in G (conditional likelihood) is given by $L(p; n_G|G) = \prod_{t=1}^{m} (p_G^t)^{n_G^t}$, so that the log-likelihood is given by $\mathcal{L}(p; n_G|G) = \ln L(n; p, G) = \sum_{t=1}^{m} n_G^t \ln p_G^t$. The maximization of $\mathcal{L}(p; n_G|G)$ with additional constraint $\sum_{t=1}^{m} p_G^t = 1$ leads to the well-known fomula for ML estimators $\hat{p}_G^t$ in multinomial distribution:

$$\hat{p}_G^t = \frac{n_G^t}{n_G} \tag{4}$$

which are exactly the frequencies used in VPRS. This observation will lead us in section 4 to the definition of the variable consistency for dominance-based rough set approach.

3 Dominance-based Rough Set Approach (DRSA)

Within DRSA [7, 8], we define the *dominance* relation D as a binary relation on X in the following way: for any $x_i, x_k \in X$ we say that x_i *dominates* x_k, $x_i D x_k$, if on every condition criterion from Q, x_i has evaluation not worse than x_k, $q_j(x_i) \geq q_j(x_k)$, for $j = 1, \ldots, n$. The dominance relation D is a partial pre-order on X, i.e. it is reflexive and transitive. The *dominance principle* can be expressed as follows:

$$x_i D x_j \Longrightarrow y_i \geq y_j \tag{5}$$

for any $x_i, x_j \in X$. We say that two objects $x_i, x_j \in X$ are consistent if they satisfy the dominance principle. We say that object x_i is consistent, if it is consistent with every other object from X.

The rough approximations concern granules resulting from information carried out by the decisions. The decision granules can be expressed by upward and downward unions of decision classes, respectively:

$$Cl_t^{\geq} = \{x_i \in X : y_i \geq t\} \qquad Cl_t^{\leq} = \{x_i \in X : y_i \leq t\} \tag{6}$$

The condition granules are dominating and dominated sets defined, respectively, for each $x \in X$, as:

$$D^+(x) = \{x_i \in X : x_i D x\} \qquad D^-(x) = \{x_i \in X : x D x_i\} \tag{7}$$

Lower approximations of $Cl_t^{\geq}$ and $Cl_t^{\leq}$ are defined as:

$$\underline{Cl}_t^{\geq} = \{x_i \in X : D^+(x_i) \subseteq Cl_t^{\geq}\} \qquad \underline{Cl}_t^{\leq} = \{x_i \in X : D^-(x_i) \subseteq Cl_t^{\leq}\} \tag{8}$$

Upper approximations of $Cl_t^{\geq}$ and $Cl_t^{\leq}$ are defined as:

$$\overline{Cl}_t^{\geq} = \{x_i \in X : D^-(x_i) \cap Cl_t^{\geq} \neq \emptyset\} \quad \overline{Cl}_t^{\leq} = \{x_i \in X : D^+(x_i) \cap Cl_t^{\leq} \neq \emptyset\} \tag{9}$$

4 Statistical model of variable consistency in DRSA

In this section, we introduce a new model of variable consistency DRSA (VC-DRSA), by miming the ML estimation shown in section 2. The name *variable consistency* instead of *variable precision* is used in this chapter only to be consistent with the already existing theory [9].

In section 2, although it was not mentioned straightforward, while estimating the probabilities, we have made the assumption that in a single granule $I(x)$, each object $x \in G$ has the same conditional probability distribution, $\Pr(y = t|I(x)) \equiv p_G^t$. 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:

$$x_i D x_j \Longrightarrow p_i^t \geq p_j^t \qquad \forall t \in T, \;\; \forall x_i, x_j \in X \tag{10}$$

where p_i^t is a probability (conditioned on x_i) of decision value at least t, $\Pr(y \geq t|x_i)$. 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 (10) will be called *stochastic dominance principle*.

In this section, we will restrict the analysis to two-class (binary) problem, so we assume $T = \{0, 1\}$ (indices start with 0 for simplicity). Notice, that $\underline{Cl}_0^{\geq}$ and $\underline{Cl}_1^{\leq}$ are trivial, 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. We relax the definition of lower approximations for $T = \{0, 1\}$ in the following way (in analogy to the classical variable precision model):

$$\underline{Cl}_t = \{x_i \in X : p_i^t \geq \alpha\}, \tag{11}$$

where $\alpha \in (0.5, 1]$ is a chosen *consistency level*. Since we do not know probabilities p_i^t, we will use instead their ML estimators $\hat{p}_i^t$. The conditional likelihood function (probability of decision values with X being fixed) is a product of binomial distributions and is given by $\prod_{i=1}^{\ell}(p_i^1)^{y_i}(p_i^0)^{1-y_i}$, or using $p_i \equiv p_i^1$ (since $p_i^0 = 1 - p_i$), is given by $\prod_{i=1}^{\ell}(p_i)^{y_i}(1-p_i)^{1-y_i}$. The log-likelihood is then

$$\mathcal{L}(p; y|X) = \sum_{i=1}^{\ell} (y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)) \tag{12}$$

The stochastic dominance principle (10) simplifies to:

$$x_i D x_j \Longrightarrow p_i \geq p_j \qquad \forall x_i, x_j \in X \tag{13}$$

To obtain probability estimators $\hat{p}_i$, we need to maximize (12) subject to constraints (13). This is exactly the problem of statistical inference under the order restriction [14]. Before investigating properties of the problem, we state the following theorem:

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

Using Theorem 1 we can set $\hat{p}_i = y_i$ for each consistent object $x_i \in X$ and optimize (12) 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 (12) boils down to the isotonic regression problem.

5 Isotonic regression

For the purpose of this paper we consider the simplified version of the *isotonic regression problem* (IRP) [14]. Let $X = \{x_1, \ldots, x_\ell\}$ be a finite set with some pre-order relation $D \subseteq X \times X$. Suppose also that $y\colon X \to \mathbb{R}$ is some function on X, where $y(x_i)$ is shortly denoted y_i. A function $y^*\colon X \to \mathbb{R}$ is an *isotonic regression* of y if it is the optimal solution to the problem:

$$\begin{aligned} &\text{minimize} \ \sum_{i=1}^{\ell}(y_i - p_i)^2 \\ &\text{subject to} \ \ x_i D x_j \Longrightarrow p_i \geq p_j \qquad \forall 1 \leq i, j \leq \ell \end{aligned} \tag{14}$$

so that it minimizes the squared error in the class of all *isotonic* functions p (where we denoted $p(x_i)$ as p_i in (14)). In our case, the ordering relation D is the dominance relation, the set X and values of function y on X, i.e. $\{y_1, \ldots, y_\ell\}$ will have the same meaning as before. Although squared error in (14) seems to be arbitrarily chosen, it can be shown that minimizing many other error functions leads to the same function y^* as in the case of (14). 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. For each $u, v \in I$ define the function $\Delta_\Phi(u,v) = \Phi(u) - \Phi(v) - (u-v)\phi(v)$. Then the following theorem holds:

Theorem 2. *[14] Let y^* be an isotonic regression of y on X, i.e. y^* solves (14). Then it holds:*

$$\sum_{x_i \in X} \Delta_\Phi(y_i, f(x_i)) \geq \sum_{x_i \in X} \Delta_\Phi(y_i, y^*(x_i)) + \sum_{x_i \in X} \Delta_\Phi(y^*(x_i), f(x_i)) \tag{15}$$

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

$$\sum_{x_i \in X} \Delta_\Phi(y_i, f(x_i)) \tag{16}$$

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

It was shown in [14] that by using the function:

$$\Phi(u) = \begin{cases} u \ln u + (1-u)\ln(1-u) & \text{for } u \in (0,1) \\ 0 & \text{for } u \in \{0,1\} \end{cases} \tag{17}$$

in Theorem 2, we end up with the problem of maximizing (12) subject to constraints (13). Thus, we can find solution to the problem (12) subject to (13) by solving the IRP (14).

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. *[14] 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, $y^*(x)$ equal to the average of y over all the objects having the same value $y^*(x)$. 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.

6 Minimal reassignment problem

In this section we briefly describe the *minimal reassignment problem* (MRP), introduced in [5]. We define the reassignment of an object $x_i \in X$ as changing its decision value 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. To find minimal reassignment, one can formulate a linear program. Such problems were already considered in [3] (under the name *isotonic separation*, in the context of binary and multi-class classification) and also in [2] (in the context of boolean regression).

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

$$x_i D x_j \Longrightarrow d_i \geq d_j \qquad \forall 1 \leq i,j \leq \ell \tag{18}$$

Notice, that (18) has the form of the stochastic dominance principle (13). 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 MRP) is given by $\sum_{i=1}^{\ell} |y_i - d_i| = \sum_{i=1}^{\ell} (y_i(1-d_i) + (1-y_i)d_i)$, 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 (18) is totally unimodular, so we can relax the integer condition for d_i reformulating it as $0 \leq d_i \leq 1$, and get a linear program [3, 12]. Moreover, constraint $0 \leq d_i \leq 1$ can be dropped, since if there were any $d_i > 1$ (or $d_i < 0$) 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. Finally, for the purpose of the paper, we rewrite the problem in the following form:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{\ell} |y_i - d_i| \\ \text{subject to} \quad & x_i D x_j \Longrightarrow d_i \geq d_j \qquad \forall 1 \leq i,j \leq \ell \end{aligned} \tag{19}$$

Comparing (19) with (14), we notice that, although both problems emerged in different context, they look very similar and the only difference is in the objective function (L_1-norm in MRP instead of L_2-norm in IRP). In fact, both problems are closely connected, which will be shown in the next section.

7 Connection between IRP and MRP

To show the connection between IRP and MRP 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 (WMRP) is given by

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{\ell} w_{y_i} |y_i - d_i| \\ \text{subject to} \quad & x_i D x_j \Longrightarrow d_i \geq d_j \qquad \forall 1 \leq i,j \leq \ell \end{aligned} \tag{20}$$

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

Theorem 4. *Suppose* $\hat{p} = \{\hat{p}_1, \ldots, \hat{p}_\ell\}$ *is an optimal solution to IRP (14). Choose some value* $\alpha \in [0, 1]$ *and define two functions:*

$$l(p) = \begin{cases} 0 & \text{if } p \leq \alpha \\ 1 & \text{if } p > \alpha \end{cases} \tag{21}$$

and

$$u(p) = \begin{cases} 0 & \text{if } p < \alpha \\ 1 & \text{if } p \geq \alpha \end{cases} \tag{22}$$

Then the solution $\hat{d}^l = \{\hat{d}^l_1, \ldots, \hat{d}^l_\ell\}$ *such that* $\hat{d}^l_i = l(\hat{p}_i)$ *for each* $i \in \{1, \ldots, \ell\}$, *and the solution* $\hat{d}^u = \{\hat{d}^u_1, \ldots, \hat{d}^u_\ell\}$ *such that* $\hat{d}^u_i = u(\hat{p}_i)$ *for each* $i \in \{1, \ldots, \ell\}$, *are the optimal solutions to WMRP (20) with weights:*

$$\begin{aligned} w_0 &= p \\ w_1 &= 1 - p \end{aligned} \tag{23}$$

Moreover, if $\hat{d} = \{\hat{d}_1, \ldots, \hat{d}_\ell\}$ *is an optimal integer solution to WMRP with weights (23), it must hold* $\hat{d}^l_i \leq \hat{d}_i \leq \hat{d}^u_i$, *for all* $i \in \{1, \ldots, \ell\}$. *In particular, if* $\hat{d}^l \equiv \hat{d}^u$, *the solution to the WMRP is unique.*

Theorem 4 clearly states, that if the optimal value for a variable $\hat{p}_i$ in IRP (14) is greater (or smaller) than α, then the optimal value for the corresponding variable $\hat{d}_i$ in the WMRP (20) with weights (23) is 1 (or 0). In particular, for $\alpha = \frac{1}{2}$ we have $w_0 = w_1 = 1$, so we obtain MRP (19). 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 IRP (14), the optimal solution to the WMRP (20) is unique. It follows from Theorem 3 (and discussion after it), that $\hat{p}$ can take only finite number of values, which must be of the form $\frac{r}{r+s}$, where $r < \ell_1$ and $s < \ell_1$ are integer (ℓ_0 and ℓ_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 WMRP is easier to solve than IRP. It is linear, so that one can use linear programming, it can also be transformed to the network flow problem [3] and solved in $O(n^3)$. In the next section, we show, that to obtain lower and upper approximations for the VC-DRSA, it is enough to solve IRP and solves two reassignment problems instead.

8 Summary of the statistical model for DRSA

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

$$\underline{Cl}_t = \{x_i \in X : p^t_i \geq \alpha\} \tag{24}$$

for $t \in \{0,1\}$. The probabilities p^t 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 IRP.

As it was stated in the previous section we choose α to be proper, so that the definition (24) can be equivalently stated as:

$$\begin{aligned} \underline{Cl}_1 &= \{x_i \in X : \hat{p}_i > \alpha\} \\ \underline{Cl}_0 &= \{x_i \in X : 1 - \hat{p}_i > \alpha\} = \{x_i \in X : \hat{p}_i < 1 - \alpha\} \end{aligned} \tag{25}$$

where we replaced the probabilities by their ML estimators. It follows from Theorem 4, that to obtain $\underline{Cl}_0$ and $\underline{Cl}_1$ we do not need to solve IRP. Instead we solve two weighted minimal reassignment problems (20), 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 decision value (optimal assignment) $\hat{d}_i = 1$ in the first problem form $\underline{Cl}_1$, while objects with new decision value $\hat{d}_i = 0$ in the second problem form $\underline{Cl}_0$. It is easy to show that the boundary between classes (defined as $X - (\underline{Cl}_1 \cup \underline{Cl}_0)$) is composed of objects, for which new decision values are different in those two problems.

9 Extension to the multi-class case

Till now, we focused on binary classification problems considered within DRSA. Here we show, how to solve the general problem with m decision classes.

We proceed as follows. We divide the problem into $m-1$ binary problems. In tth binary problem, we estimate the lower approximations of upward union for class $t+1$, $\underline{Cl}^{\geq}_{t+1}$, and the lower approximation of downward union for class t, $\underline{Cl}^{\leq}_{t}$ using the theory stated in the section 8 for two-class problem with $Cl_0 = Cl^{\leq}_t$ and $Cl_1 = Cl^{\geq}_{t+1}$. Notice, that for the procedure to be consistent, it must hold if $t' > t$ than $\underline{Cl}^{\geq}_{t'} \subseteq \underline{Cl}^{\geq}_{t}$ and $\underline{Cl}^{\leq}_{t} \subseteq \underline{Cl}^{\leq}_{t'}$. 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, we have:

Theorem 5. *For each $t = 1, \ldots, m-1$, let $\underline{Cl}^{\leq}_{t}$ and $\underline{Cl}^{\geq}_{t+1}$ be the sets obtained from solving two-class isotonic regression problem with consistency level α for binary classes $Cl_0 = Cl^{\leq}_t$ and $Cl_1 = Cl^{\geq}_{t+1}$. Then, we have:*

$$t' \geq t \Longrightarrow \underline{Cl}^{\leq}_{t} \subseteq \underline{Cl}^{\leq}_{t'} \tag{26}$$

$$t' \geq t \Longrightarrow \underline{Cl}^{\geq}_{t'+1} \subseteq \underline{Cl}^{\geq}_{t+1} \tag{27}$$

10 Decision-theoretical view

In this section we look at the problem of VPRS and VC-DRSA from the point of view of statistical decision theory [1, 11]. A decision-theoretic approach has

already been proposed in [15] (for VPRS) and in [10] (for DRSA). The theory presented here for VPRS is slightly different than in [15], 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* $L(f(x), y)$ is introduced for penalizing the prediction error. Since x and y are random variables, the overall measure of the classifier $f(x)$ is the *expected loss* or *risk*, which is defined as a functional:

$$R(f) = E[L(y, f(x))] = \int L(y, f(x))dP(y, x) \tag{28}$$

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:

$$R_e(f) = \sum_{i=1}^{\ell} L(y_i, f(x_i)). \tag{29}$$

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 VPRS. Assume, that we allow the classifier to give no answer, which is denoted as $f(x) =?$. The loss function suitable for the problem is the following:

$$L_c(f(x), y) = \begin{cases} 0 & \text{if } f(x) = y \\ 1 & \text{if } f(x) \neq y \\ a & \text{if } f(x) =? \end{cases} \tag{30}$$

There is a penalty a 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:

$$x_i, x_j \in G \Longrightarrow f(x_i) = f(x_j) \qquad \forall x_i, x_j \in X \tag{31}$$

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

Theorem 6. *The function f^* minimizing the empirical risk (29) with loss function (30) between all functions satisfying (31) 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 $u = 1 - a$, otherwise $f^*(G) =?$.*

Concluding, the VPRS can be derived by considering the class of functions constant in each granule and choosing the function f^*, which minimizes the empirical risk (29) for loss function (56) with parameter $a = 1 - u$. As we see, classical rough set theory suits well for considering the problems when the classification procedure is allowed not to give predictions for some x.

We now turn back to DRSA. Assume, that to each point x, the classifier f assigns the interval of classes, denoted $[l(x), u(x)]$. The lower and upper ends of each interval are supposed to be consistent with the dominance principle:

$$\begin{aligned} x_i D x_j \Longrightarrow l(x_i) \geq l(x_j) \qquad & \forall x_i, x_j \in X \\ x_i D x_j \Longrightarrow u(x_i) \geq u(x_j) \qquad & \forall x_i, x_j \in X \end{aligned} \tag{32}$$

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

$$L(f(x), y) = a(u(x) - l(x)) + I(y \notin [l(x), u(x)]) \min\{|y - l(x)|, |y - u(x)|\} \tag{33}$$

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

Theorem 7. *The function f^* minimizing the empirical risk (29) with loss function (33) between all interval functions satisfying (32) is equivalent to the statistical VC-DRSA with consistency level $\alpha = 1 - a$ in the sense, that for each $x \in X$, $x \in \underline{Cl}_t^{\geq}$ or $x \in \underline{Cl}_t^{\leq}$ if and only if $t \in f^*(x)$.*

Concluding, the statistical VC-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 (29) with loss function (33) with parameter $a = 1 - \alpha$.

11 Conclusions

The paper introduced a new variable consistency theory for Dominance-based Rough Set Approach. 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 new variable consistency model, it is enough to solve minimal reassignment problem (which is linear), instead of the isotonic regression problem (quadratic). The approach has also been extended to the multi-class case by solving $m - 1$ 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 decision values (in DRSA case). However, rough set theory itself has a rather small generalization capacity, due to its nonparametric character, which was shown in section 10. The plans for further research are to investigate some restricted classes of functions which would allow to apply rough set theory directly for classification.

References

1. Berger, J., Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York (1993)
2. Boros, E., Hammer, P. L., Hooker, J. N., Boolean regression. Annals of Operations Research, **58**, 3 (1995).
3. Chandrasekaran ,R., Ryu, Y. U., Jacob, V., Hong, S.: Isotonic separation. INFORMS J. Comput. **17** 2005 462–474
4. Dembczyński, K., Greco, S., Kotłowski, W., Słowiński, R.: Quality of Rough Approximation in Multi-Criteria Classification Problems. Lecture Notes in Computer Science **4259**, Springer 2006 318–327.
5. Dembczyński, K., Greco, S., Kotłowski, W., Słowiński, R.: Optimized Generalized Decision in Dominance-based Rough Set Approach. Lecture Notes in Computer Science, Springer 2007
6. Duda, R., Hart, P., Pattern Classification. Wiley-Interscience, New York (2000).
7. Greco S., Matarazzo, B., Słowiński, R.: Rough approximation of a preference relation by dominance relations. European Journal of Operational Research, **117** (1999) 63–83
8. Greco S., Matarazzo, B., Słowiński, R.: Rough sets theory for multicriteria decision analysis. European Journal of Operational Research, **129** 1 (2001) 1–47
9. 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
10. Greco, S., Słowiński, R., Yao, Y.: Bayesian Decision Theory for Dominance-based Rough Set Approach. Lecture Notes in Computer Science **4481** (2007), 134–141
11. Hastie, T., Tibshirani, R., Friedman J., The Elements of Statistical Learning. Springer 2003.
12. Papadimitriou, C. H., Steiglitz, K., Combinatorial Optimization. Dover Publications, New York 1998.
13. Pawlak, Z.: Rough sets. International Journal of Information & Computer Sciences, **11** (1982) 341–356
14. Robertson, T., Wright, F. T., Dykstra, R. L., Order Restricted Statistical Inference. John Wiley & Sons (1998).
15. Yao, Y., Wong, S., A decision theoretic Framework for approximating concepts. International Journal of Man-machine Studies, **37**, 6 (1992), 793–809.
16. Ziarko, W., Probabilistic Rough Sets. Lecture Notes in Artificial Intelligence **3641**, Springer-Verlag 2005, 283–293.

Appendix: Proofs of the Theorems

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 $x_j D x_i$ and $y_j = 0$ (otherwise, it would violate dominance principle and consistency of x_i as well). Thus, for every x_j, such that $x_j D x_i$, $y_j = 1$ 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 (12) for those objects while satisfying the constraints (13)

Now, suppose $\hat{p}_i = 1$ and assume the contrary that x_i is not consistent, i.e. there exists x_j, $x_j D x_i$, but $y_j = 0$. Then, due to the monotonicity constraints (13), $\hat{p}_j \geq \hat{p}_i = 1$, so $\hat{p}_j = 1$, and the log-likelihood (12) 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).□

Theorem 4. *Suppose $\hat{p} = \{\hat{p}_1, \ldots, \hat{p}_\ell\}$ is an optimal solution to the problem of isotonic regression (14). Choose some value $\alpha \in [0, 1]$ and define two functions:*

$$l(p) = \begin{cases} 0 & \text{if } p \leq \alpha \\ 1 & \text{if } p > \alpha \end{cases} \tag{34}$$

and

$$u(p) = \begin{cases} 0 & \text{if } p < \alpha \\ 1 & \text{if } p \geq \alpha \end{cases} \tag{35}$$

Then the solution $\hat{d}^l = \{\hat{d}^l_1, \ldots, \hat{d}^l_\ell\}$ such that $\hat{d}^l_i = l(\hat{p}_i)$ for each $i \in \{1, \ldots, \ell\}$, and the solution $\hat{d}^u = \{\hat{d}^u_1, \ldots, \hat{d}^u_\ell\}$ such that $\hat{d}^u_i = u(\hat{p}_i)$ for each $i \in \{1, \ldots, \ell\}$, are the optimal solutions to the problem of weighted minimal reassignment (20) with weights:

$$\begin{aligned} w_0 &= \alpha \\ w_1 &= 1 - \alpha \end{aligned} \tag{36}$$

Moreover, if $\hat{d} = \{\hat{d}_1, \ldots, \hat{d}_\ell\}$ is an optimal integer solution to the problem of weighted minimal reassignment with weights (36), it must hold $\hat{d}^l_i \leq \hat{d}_i \leq \hat{d}^u_i$, for all $i \in \{1, \ldots, \ell\}$. In particular, if $\hat{d}^l \equiv \hat{d}^u$, then the solution is unique.

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

$$\Phi(u) = \begin{cases} \alpha(u - \alpha) & \text{for } x \geq \alpha \\ (1 - \alpha)(\alpha - u) & \text{for } x < \alpha \end{cases} \tag{37}$$

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:

$$\Delta_\Phi(u, v) = \Phi(u) - \Phi(v) - (u - v)\phi(v) \tag{38}$$

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 (34) of function l, we can comprehensively write those results as:

$$\Delta_\Phi(u, v) = w_u |l(v) - u| \tag{39}$$

for $u \in \{0, 1\}$, where w_u are given by (36). Thus, according to Theorem 2, $\hat{p}$ is the optimal solution to the problem:

$$\text{minimize} \sum_{i=1}^{\ell} w_{y_i} |l(p_i) - y_i| \tag{40}$$

$$\text{subject to } x_i D x_j \Longrightarrow p_i \geq p_j \qquad \forall 1 \leq i, j \leq \ell \tag{41}$$

Notice, that $\hat{d}^l = l(\hat{p})$ is also the optimal solution to the problem (40)-(41), because l is a nondecreasing function, so if $\hat{p}$ satisfies constraints (41), then so does $\hat{d}^l$. Moreover, $l(l(x)) = l(x)$, so the value of the objective function (40) is the same for both $\hat{p}$ and $\hat{d}^l$. But $\hat{d}^l$ is integer and, for integer solutions, problems (40)-(41) and (20) are the same, so $\hat{d}^l$ is a solution to the problem (20) with the lowest objective value among all the integer solutions to this problem. But, from the analysis of the unimodularity of constraints matrix of (20) we know that if $\hat{d}^l$ is the solution to (20) with the lowest objective value among the integer solutions, it is also the optimal solution, since there exists an optimal solution to (20), 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 (20).

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:

$$\sum_{i=1}^{\ell} \Delta_\Phi(y_i, f(x_i)) \geq \sum_{i=1}^{\ell} \Delta_\Phi(y_i, \hat{p}_i) + \sum_{i=1}^{\ell} \Delta_\Phi(\hat{p}_i, f(x_i)) \tag{42}$$

for any isotonic function f in the range $[0, 1]$. Notice that if the last term in (42) is nonzero, then f cannot be optimal to the problem (40)-(41) (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 (20). But then it is also the solution to the problem (40)-(41) 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 (40)-(41) 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 (42) must be zero, so for each $i \in \{1, \ldots, \ell\}$ 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:

$$\hat{p}_i > \alpha \Longrightarrow \hat{d}_i = 1 \tag{43}$$

$$\hat{p}_i < \alpha \Longrightarrow \hat{d}_i = 0 \tag{44}$$

for all $i \in \{1, \ldots, \ell\}$. 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 (43)-(44) we conclude that:

$$\hat{d}^l_i \leq \hat{d}_i \leq \hat{d}^u_i \tag{45}$$

for all $i \in \{1, \ldots, \ell\}$, for any optimal integer solution $\hat{d}$ to problem (20). □

Lemma 1. *Let $\hat{p}$ be the optimal solution to the isotonic regression problem (14) for decision values y. Suppose, we introduce a new vector of decision values y', such that $y'_i \geq y_i$ for all $i \in \{1, \ldots, \ell\}$. 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 $i \in \{1, \ldots, \ell\}$.*

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:

$$\hat{p}_i^+ = \max\{\hat{p}_i, \hat{p}'_i\}, \tag{46}$$

$$\hat{p}_i^- = \min\{\hat{p}_i, \hat{p}'_i\}. \tag{47}$$

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 (14). Suppose $x_i D x_j$. 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 (14) as $F(y, p) = \sum_{i=1}^{\ell}(y_i - p_i)^2$. Then, we have:

$$\begin{aligned} F(y', \hat{p}^+) - F(y', \hat{p}') &= \sum_{i=1}^{\ell} \left(\hat{p}_i^{+2} - \hat{p}_i'^2 - 2y'_i\hat{p}_i^+ - 2y'_i\hat{p}'_i\right) = \\ &= \sum_{i=1}^{\ell} \left((\hat{p}_i^+ - \hat{p}'_i)(\hat{p}_i^+ + \hat{p}'_i) - 2y'_i(\hat{p}_i^+ - \hat{p}'_i)\right) \end{aligned} \tag{48}$$

Since it holds that $\hat{p}_i^+ - \hat{p}'_i \geq 0$ and $y'_i \geq y_i$, we have:

$$\sum_{i=1}^{\ell} 2y'_i(\hat{p}_i^+ - \hat{p}'_i) \geq \sum_{i=1}^{\ell} 2y_i(\hat{p}_i^+ - \hat{p}'_i) \tag{49}$$

Finally, it holds that $\hat{p}_i^+ + \hat{p}_i^- = \hat{p}'_i + \hat{p}_i$, so that:

$$\hat{p}_i^+ - \hat{p}'_i = \hat{p}_i - \hat{p}_i^- \tag{50}$$

and

$$\hat{p}_i^+ + \hat{p}'_i = 2(\hat{p}'_i - \hat{p}_i^-) + (\hat{p}_i + \hat{p}_i^-). \tag{51}$$

Putting (49)-(51) into (48), we finally obtain:

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

$$= \sum_{i=1}^{\ell} 2(\hat{p}_i - \hat{p}_i^-)(\hat{p}'_i + \hat{p}_i^-) + F(y, \hat{p}) - F(y, \hat{p}^-)$$
$$< \sum_{i=1}^{\ell} 2(\hat{p}_i - \hat{p}_i^-)(\hat{p}'_i + \hat{p}_i^-) \tag{52}$$

since by the assumption $\hat{p}$ is the isotonic regression of y, so it is the unique optimal solution for decision values 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:

$$F(y', \hat{p}^+) - F(y', \hat{p}') < 0 \tag{53}$$

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

Theorem 5. *For each $t = 1, \ldots, m-1$, let $\underline{Cl}_t^{\leq}$ and $\underline{Cl}_{t+1}^{\geq}$ be the sets obtained from solving two-class isotonic regression problem with consistency level α for binary classes $Cl_0 = Cl_t^{\leq}$ and $Cl_1 = Cl_{t+1}^{\geq}$. Then, we have:*

$$t' \geq t \Longrightarrow \underline{Cl}_t^{\leq} \subseteq \underline{Cl}_{t'}^{\leq} \tag{54}$$
$$t' \geq t \Longrightarrow \underline{Cl}_{t'+1}^{\geq} \subseteq \underline{Cl}_{t+1}^{\geq} \tag{55}$$

Proof. Suppose we have solved the problem for some t. Denote $y_i = 1$ if $x_i \in Cl_{t+1}^{\geq}$ and $y_i = 0$ if $x_i \in Cl_t^{\leq}$. Suppose we have also solved the problem for some $t' \geq t$. Denote $y'_i = 1$ if $x_i \in Cl_{t'+1}^{\geq}$ and $y'_i = 0$ if $x_i \in Cl_{t'}^{\leq}$. Clearly, from the definition of $Cl_t^{\leq}, Cl_t^{\geq}$ it follows that $y_i \geq y'_i$ for each $i \in \{1, \ldots, \ell\}$. Then, according to Lemma 1, if $x_i \in \underline{Cl}_t^{\leq}$ (so that $\hat{p}_i < \alpha$), then also $x_i \in \underline{Cl}_{t'}^{\leq}$ (since then $\hat{p}'_i \leq \hat{p}_i < \alpha$). Analogously, if $x_i \in \underline{Cl}_{t'+1}^{\geq}$, then also $x_i \in \underline{Cl}_{t+1}^{\geq}$. This proves the theorem. $\square$

Theorem 6. *The function f^* minimizing the empirical risk (29) with loss function (30) between all functions satisfying (31) is equivalent to the VPRS for classical rough set theory 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 $u = 1 - a$, otherwise $f^*(G) =?$.*

Proof. Since apart from (31) 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 label $t \in T$, let us denote n_G^t as the number of objects from class t in G. It is clear that the total loss of some function f in granule G is the following:

$$Lf(G) = \begin{cases} n_G - n_G^t & \text{if } f(G) = t \\ a \cdot n_G & \text{if } f(G) =? \end{cases} \tag{56}$$

This follows from the fact that if $f(G) = t$, then for each $x_i \in G$ such that $y_i \neq t$, f suffers loss 1. On the other hand, if $f(G) =?$, for each $x_i \in G$, function f suffers loss a. 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 $n_G - n_G^t \leq a n_G$ or:

$$a \geq 1 - \frac{n_G^t}{n_G} \tag{57}$$

Otherwise, if no t satisfies this relation, $f^*(G) =?$ is chosen. Comparing this results with section 2, one can see that the decision $f^*(G) = t$ is chosen if granule G belongs to the lower approximation of class t with the precision threshold $u = 1 - a$. Clearly, from (3) with probabilities estimated by (4) 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^* gives no answer.□

Theorem 7. *The function f^* minimizing the empirical risk (29) with loss function (33) between all interval functions satisfying (32) is equivalent to the variable consistency model for DRSA with consistency level $\alpha = 1 - a$ in the sense, that for each $x \in X$, $x \in \underline{Cl}_t^{\geq}$ or $x \in \underline{Cl}_t^{\leq}$ if and only if $t \in f^*(x)$.*

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 $i \in \{1, \ldots, \ell\}, k \in \{2, \ldots, m\}$. We code the ranges of interval $f(x_i)$ as $l(x_i) = 1 + \sum_{k=2}^{m} l_{ik}$ and $u(x_i) = 1 + \sum_{k=2}^{m} u_{ik}$. 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:

$$u_{ik} \geq l_{ik} \qquad \forall i \in \{1, \ldots, \ell\}, k \in \{2, \ldots, m\} \tag{58}$$

$$l_{ik} \geq l_{ik'} \qquad \forall i \in \{1, \ldots, \ell\}, k < k' \tag{59}$$

$$u_{ik} \geq u_{ik'} \qquad \forall i \in \{1, \ldots, \ell\}, k < k' \tag{60}$$

Moreover, for dominance principle (32) to hold, we must also have:

$$x_i D x_j \Longrightarrow l_{ik} \geq l_{ik} \qquad \forall i \in \{1, \ldots, \ell\}, k \in \{2, \ldots, m\} \tag{61}$$

$$x_i D x_j \Longrightarrow u_{ik} \geq u_{ik} \qquad \forall i \in \{1, \ldots, \ell\}, k \in \{2, \ldots, m\} \tag{62}$$

It is not hard to verify, that the loss function (33) for object x_i can be written as:

$$L_i = L(f(x_i), y_i) = a \sum_{k=2}^{m} (u_{ik} - l_{ik}) + \sum_{k=y_i+1}^{m} l_{ik} + \sum_{k=2}^{y_i} (1 - u_{ik}) \tag{63}$$

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

$$
\begin{aligned}
L_i = (1-a)\sum_{k=2}^{m} l_{ik}(1-y_{ik}) - a\sum_{k=2}^{m} l_{ik}y_{ik} + \\
a\sum_{k=2}^{m} u_{ik}(1-y_{ik}) - (1-a)\sum_{k=2}^{m} u_{ik}y_{ik} + \sum_{k=2}^{m} y_{ik} \\
= \sum_{k=2}^{m} w^{II}_{y_{ik}}|l_{ik}-y_{ik}| + \sum_{k=2}^{m} w^{I}_{y_{ik}}|u_{ik}-y_{ik}| + C \qquad (64)
\end{aligned}
$$

where C is constant term (which does not depend on l_{ik} and u_{ik}) and weights has the form $w^I_0 = a, w^I_1 = 1-a, w^{II}_0 = 1-a, w^{II}_1 = a$. But it follows from (64), that minimizing empirical risk $R_e = \sum_{i=1}^{\ell}$ is equivalent to solving the sequence of $m-1$ pairs of weighted minimal reassignment, as described in section 9 (solving multi-class case as $m-1$ binary problems) and in section 8 (obtaining lower approximations by solving pair of weighted minimal reassignment problems) with the penalty a equal to $1-\alpha$, but with additional constraints (58)-(60). We now show that those constraints are in fact not needed.

Suppose now, we remove constraints (58)-(60). Then we obtain $2(k-1)$ separate problems, since variables $\{l_{i2}\}_{i=1}^{\ell}, \{u_{i2}\}_{i=1}^{\ell}, \ldots, \{l_{im}\}_{i=1}^{\ell}, \{u_{im}\}_{i=1}^{\ell}$ are now independent sets and their optimal values can be obtained separately. This is exactly the constructing of statistical VC-DRSA in multi-class case as described before. But it follows from Theorem 5, that constraints (59) and (60) are satisfied at optimality. Moreover, from Theorem 4 and analysis in section 9 it follows that also the constraints (58) are satisfied at optimality. Thus, the optimal solution to problem without constraints (58)-(60) is also the solution to the problem with constraints (58)-(60).□