Statistical Framework for Dominance-based Rough Set Approach

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

[1]Institute of Computing Science, Poznań University of Technology

[2]Faculty of Economics, University of Catania

[3]Institute for System Research, Polish Academy of Sciences

CSM 2007, IIASA, Laxenburg

Outline

Mutlicriteria classification problem

- The problem of classifying objects to m *decision classes* $\{1, \ldots, m\}$, which are *preference ordered*.
- Objects described by n *condition criteria* (attributes with preference-ordered domains).
- Monotone relationship between evaluations of object on condition criteria and its decision value (class): a better evaluation of object on a criterion with other evaluations being fixed should not worsen its decision value.
- The model is obtained from the *training* set of objects.

Notation

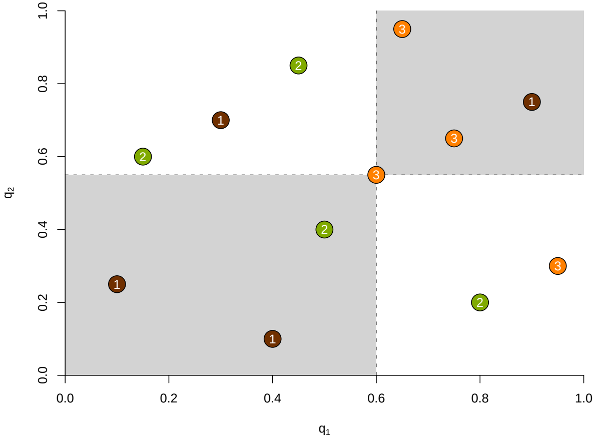
object	$\boldsymbol{q_1}$	$\boldsymbol{q_2}$	$\boldsymbol{\ldots}$	$\boldsymbol{q_n}$	$\boldsymbol{y}$
x_1	$q_1(x_1)$	$q_2(x_1)$	$\ldots$	$q_n(x_1)$	y_1
x_2	$q_1(x_2)$	$q_2(x_2)$	$\ldots$	$q_n(x_2)$	y_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_ℓ	$q_1(x_\ell)$	$q_2(x_\ell)$	$\ldots$	$q_n(x_\ell)$	y_ℓ

- A training set $\{(x_1, y_1), \ldots, (x_\ell, y_\ell)\}$ of ℓ objects x_i with their decision values (class indices) y_i.
- Each object x_i described by n condition criteria $Q = \{q_1, \ldots, q_n\}$, where $q_i(x_j)$ is the value of object x_j on criterion q_i.
- Each decision value $y_i \in T = \{1, \ldots, m\}$ (m classes).

Dominance relation

Definition

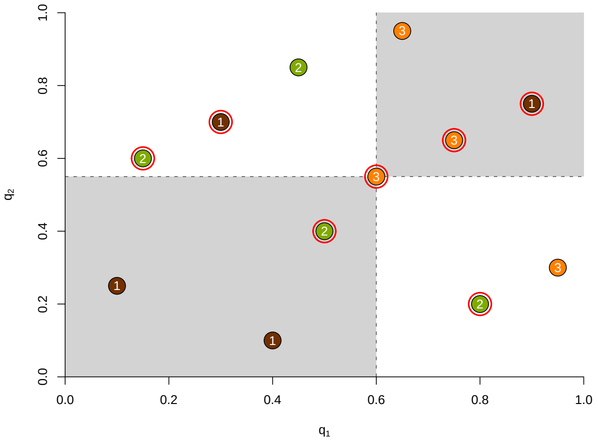
For each $x_i, x_j \in X$, x_i *dominates* x_j if x_i has better or equal evaluations on all the condition criteria: $\forall_{q_k \in Q}\ q_k(x_i) \geq q_k(x_j)$



Dominance principle

Definition

If x_i dominates x_j then x_i is preferred to x_j: $x_i D x_j \rightarrow y_i \geq y_j$.
Objects violating dominance principle are *inconsistent*.



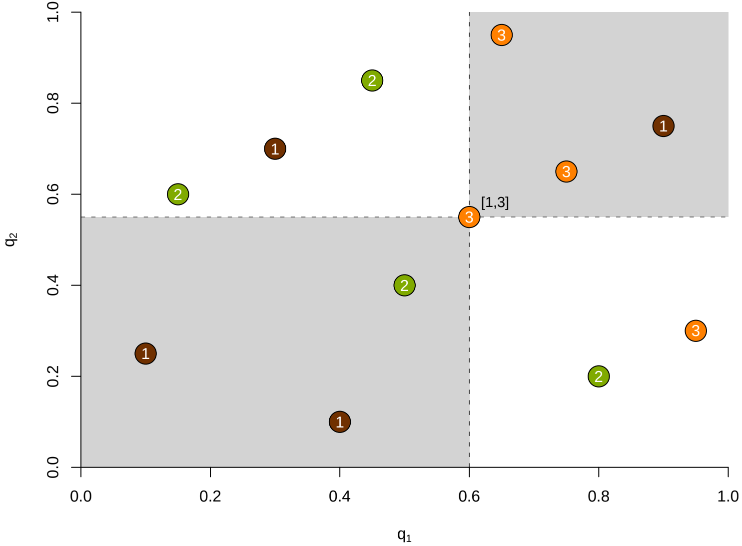
Dominance-based Rough Set Approach (DRSA)

- Handling inconsistencies by the notion of *lower approximations* and *boundary regions* of classes.
- A *generalized decision* $\delta(x_i) = [l_i, u_i]$ is assigned to each object such that:

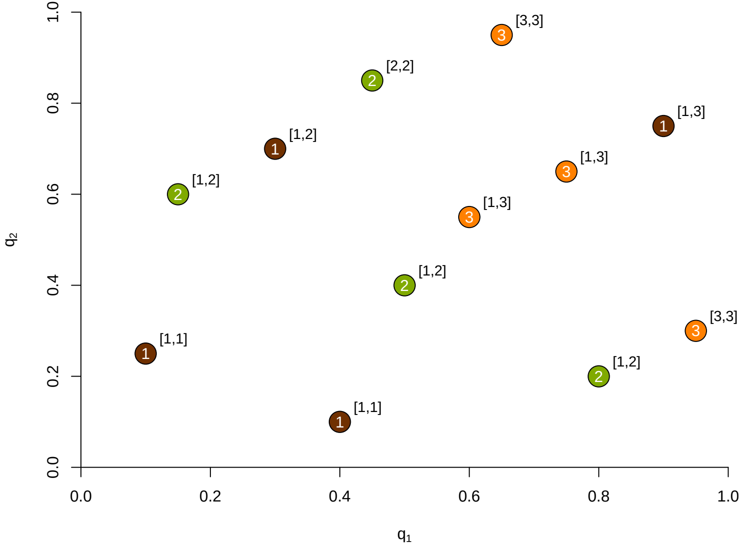
$$l_i = \min_{x_j \in X:\, x_j D x_i} \{y_j\}$$
$$u_i = \max_{x_j \in X:\, x_i D x_j} \{y_j\}$$

- It is the interval of decision values to which an object may belong due to the inconsistencies with respect to the dominance principle – the measure of the degree of imprecision.

Generalized decisions – example



Generalized decisions – example



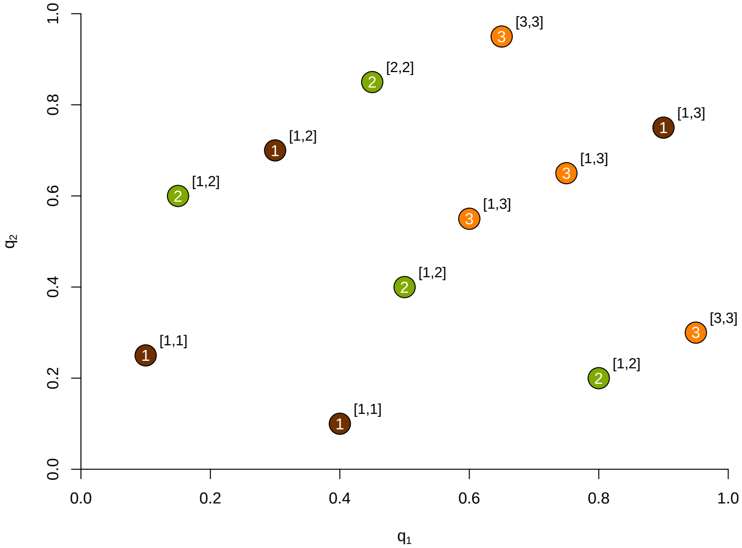
Properties of generalized decision

- Define the order relation $\succeq$ on the intervals:

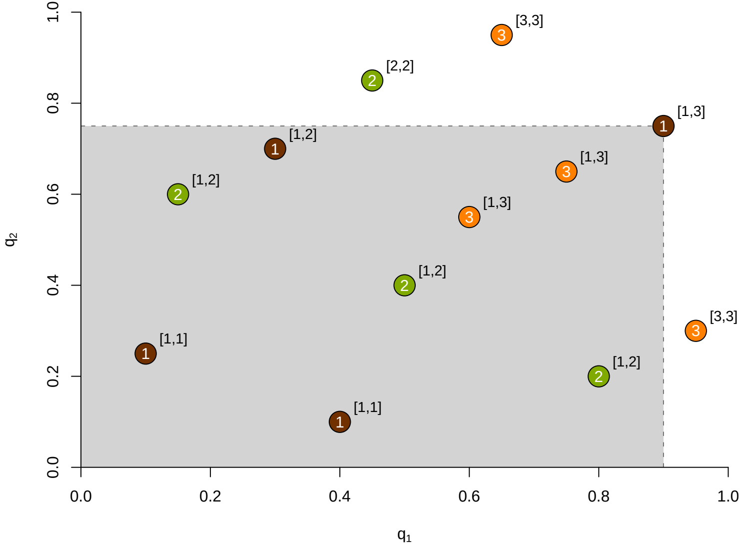
$$[l_i, u_i] \succeq [l_j, u_j] \iff l_i \geq l_j \wedge u_i \geq u_j$$

- Generalized decisions are consistent with dominance principle in the sense that if object x_i dominates x_j then $\delta(x_i) \succeq \delta(x_j)$.
- For each $x_i \in X$, we have $l_i \leq y_i \leq u_i$ and $l_i = y_i = u_i$ iff x_i is consistent (surely belong to one class).
- Generalized decision is a sort of confidence interval obtained using only monotonicity constraints.

Generalized decisions – example



Problems with generalized decision



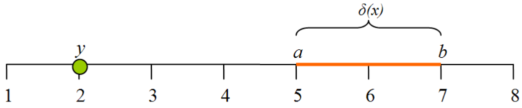
The new methodology

Motivation

- Objects which make many other objects inconsistent probably have wrong class assignments.
- Idea: assign to each object decision intervals with possible error correction.
- The procedure will be stated as a problem of minimization a specific loss function on the dataset (Empirical Risk Minimization).

Loss function

- Suppose we assign to object (x_i, y_i) a decision interval $\delta(x_i) = [l_i, u_i]$
- The loss function $L(y_i, \delta(x_i))$ consist of two parts:
  - penalty for imprecision, proportional to the interval size: $\alpha(u_i - l_i)$,
  - penalty for misclassification: 0 if $y_i \in \delta(x_i)$ or distance to the interval: $\min\{|y_i - l_i|, |y_i - u_i|\}$.
- Example (8 classes)



$$L(y, \delta(x)) = 3 + 2\alpha$$

Stochastic DRSA

Problem statement

Minimize:

$$\sum_{i=1}^{\ell} L(y_i, \delta(x_i))$$

such that $x_i D x_j \rightarrow \delta(x_i) \geq \delta(x_j)$

Properties

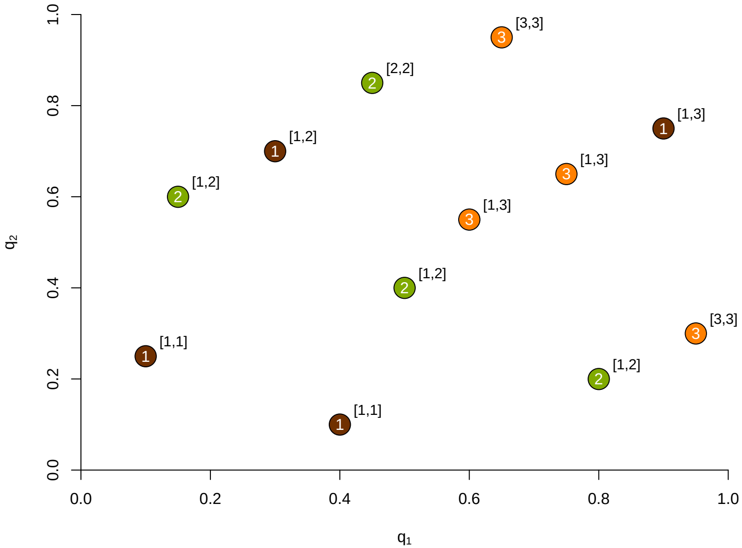
- solved by linear programming
- due to unimodularity of constraints matrix, integer constraints can be dropped
- possible strong reduction of problem size

Analysis of the solution

- as $\alpha \to 0$ we obtain classical DRSA generalized decision,
- when $\alpha \geq 0.5$ all decision intervals shrink to one point – only error corrections,
- when $\alpha \in (0, 0.5)$ a trade-off between interval size and correcting the class assignments,
- for almost each $\alpha \in (0, 0.5)$ unique solution.

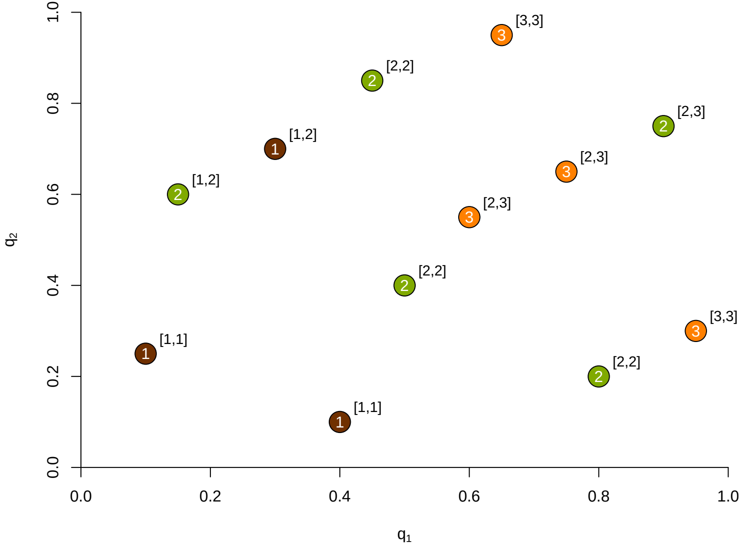
Stochastic DRSA – example

$\alpha = 0$



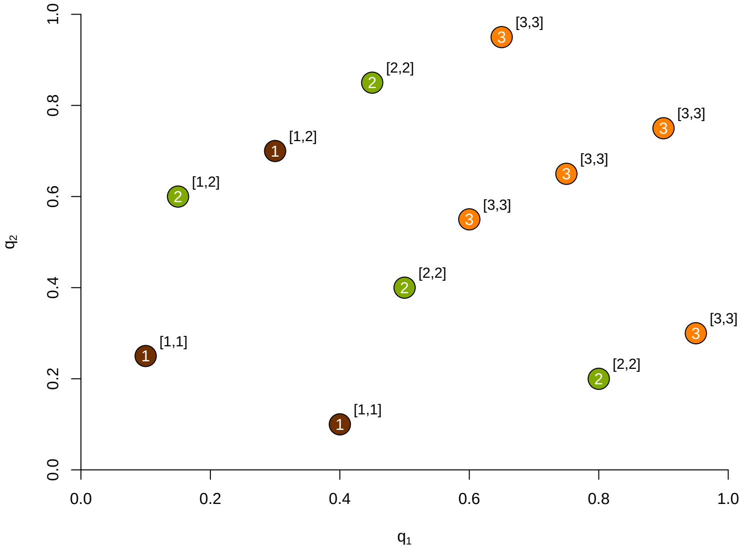
Stochastic DRSA – example

$\alpha = 1/5$



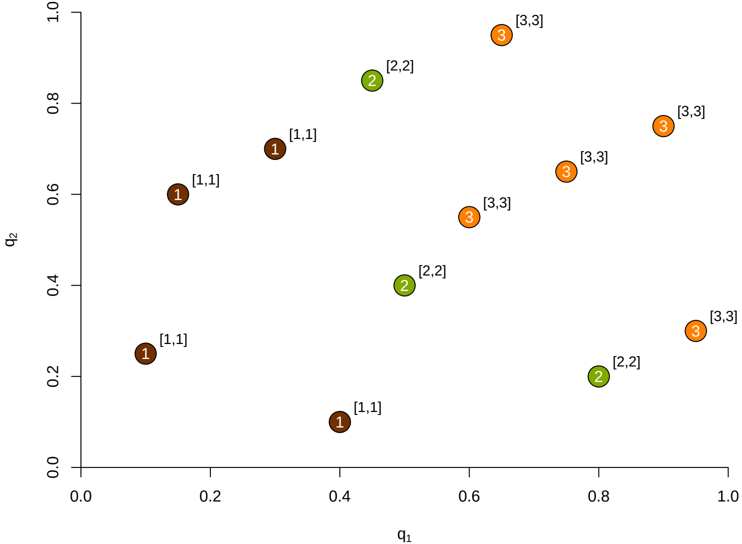
Stochastic DRSA – example

$\alpha = 1/3$



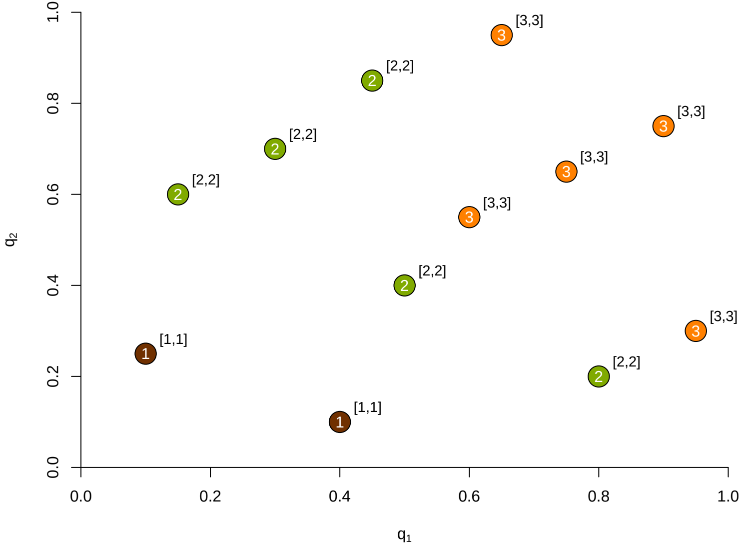
Stochastic DRSA – example

$\alpha = 1/2$



Stochastic DRSA – example

$\alpha = 1/2$



Stochastic dominance

Assume objects (x, y) are generated independently according to some unknown probability distribution $P(x, y)$

Stochastic dominance principle

If object x_i dominates object x_j, class distribution conditioned at x_i *stochastically dominates* class distribution conditioned at x_j:

$$x_i D x_j \rightarrow P(y \geq k | x_i) \geq P(y \geq k | x_j) \quad \forall k = 1 \ldots m$$

Statistical properties of stochastic DRSA

Under stochastic dominance principle (and some mild distribution assumptions):

- For a given α, as training set size grows (in the limit $\ell \to \infty$), we obtain confidence intervals around the median such that l is α quantile, and u is $1 - \alpha$ quantile of $P(y|x)$,
- For two-class problem the procedure gives results identical to maximum likelihood estimate (isotonic regression).
- For $\alpha = 0.5$, as $\ell \to \infty$, the procedure converges to the best possible classifier, i.e. minimizing the risk (so called Bayesian classifier).

Experimental Results

- Seven datasets for which the monotone relationships between decision and condition attributes are observed: `Wisconsin breast cancer`, `Ljubljana breast cancer`, `Windsor housing`, `Boston housing`, `Den Bosch housing`, `cpu`, `bankruptcy risk`.
- Three typical classifiers (SVM, j48, AdaBoost) and one monotone classifier (ensemble decision rules, ENDER) learned on data with and without using stochastic DRSA approach ($\alpha = 0.5$).
- The measure of accuracy for ordinal classification: mean absolute error (MAE).
- Random split into training (66%) and testing (33%) sets, repeated 1000 times.

Results

dataset	SVM	j48	Adaboost	ENDER
housing Den Bosch	0.2029	0.1754	0.1539	0.1418
	0.2065	0.1658	0.1382	0.1241
CPU	0.6517	0.1593	0.5874	0.0981
	—	—	—	0.0923
breast cancer Wisconsin	0.0313	0.0497	0.0486	0.0337
	0.0305	0.0494	0.0449	0.0357
bankruptcy risk	0.2196	0.2254	0.3838	0.1921
	—	—	—	0.1805
breast cancer Ljubljana	0.2944	0.2718	0.2669	0.2859
	0.2487	0.2712	0.2453	0.2530
housing Boston	0.5290	0.3951	0.6599	0.3043
	0.5168	0.3698	0.6574	0.2993
housing Windsor	0.6752	0.6517	0.7691	0.5741
	0.6511	0.5957	0.7854	0.5174

Summary

In rough set theory:

- An extension of DRSA based on stochastic model, robust to noise in the data.
- Explanation of DRSA in terms of statistics.

In machine learning:

- Nonparametric method of confidence interval estimation in the presence of monotonicity constraints.
- Method of error correction based on domain knowledge (monotonicity).