A New Instance-Based Label Ranking Approach Using the Mallows Model

Weiwei Cheng & Eyke Hüllermeier
Knowledge Engineering & Bioinformatics Lab
Department of Mathematics and Computer Science
University of Marburg, Germany

Label Ranking (an example)

Learning customers' preferences on cars:

	label ranking
customer 1	MINI $\succ$ Toyota $\succ$ BMW
customer 2	BMW $\succ$ MINI $\succ$ Toyota
customer 3	BMW $\succ$ Toyota $\succ$ MINI
customer 4	Toyota $\succ$ MINI $\succ$ BMW
new customer	???

where the customers are typically described by feature vectors, e.g., (gender, age, place of birth, has child, …)

Label Ranking (an example)

Learning customers' preferences on cars:

	MINI	Toyota	BMW
customer 1	1	2	3
customer 2	2	3	1
customer 3	3	2	1
customer 4	2	1	3
new customer	?	?	?

$\pi(i)$ = position of the i-th label in the ranking

1: MINI   2: Toyota   3: BMW

Label Ranking (more formally)

Given:

- a set of training instances $\{\mathbf{x}_k \mid k = 1 \ldots m\} \subseteq \mathbf{X}$
- a set of labels $L = \{l_i \mid i = 1 \ldots n\}$
- for each training instance $\mathbf{x}_k$: a set of *pairwise preferences* of the form $l_i \succ_{\mathbf{x}_k} l_j$

Find:

- a ranking function ($\mathcal{X} \to \Omega$ mapping) that maps each $\mathbf{x} \in \mathbf{X}$ to a ranking $\succ_{\mathbf{x}}$ of *L* (permutation $\pi_{\mathbf{x}}$)

Model-Based Approaches

… essentially reduce label ranking to classification:

- Ranking by pairwise comparison (RPC)
  Fürnkranz and Hüllermeier, ECML-03

- Constraint classification (CC)
  Har-Peled , Roth and Zimak, NIPS-03

- Log linear models for label ranking (LL)
  Dekel, Manning and Singer, NIPS-03

Instance-Based Approach (this work)

- Lazy learning: Instead of eagerly inducing a model from the data, simply store the observations.
- Target functions are estimated on demand in a local way, no need to define the $\mathcal{X} \rightarrow \Omega$ mapping explicitly.
- Core part is the aggregation of preference (order) information from neighbored examples.

Related Work

Case-based Label Ranking (Brinker and Hüllermeier, ECML-06)

Aggregation of *complete* rankings is done by

- median ranking
- Borda count

Our aggregation method is based on a probabilistic model and can handle both *complete* and *incomplete* rankings.

Instance-Based Prediction

Basic assumption: Distribution of output is (approximately) constant in the neighborhood of the query; consider outputs of neighbors as an i.i.d. sample.

Conventional classification:

→ discrete distribution on class labels

→ estimate probabilities by relative class frequencies

→ class prediction by majority vote

Probabilistic Model for Ranking

Mallows model (Mallows, Biometrika, 1957)

$$\mathcal{P}(\sigma|\theta,\pi) = \frac{\exp(-\theta d(\pi,\sigma))}{\phi(\theta,\pi)}$$

with

center ranking $\pi \in \Omega$

spread parameter $\theta > 0$

and $d(\cdot)$ is a right invariant metric on permutations

$$\boxed{\forall \pi,\sigma,\nu \in \Omega,\ d(\pi,\sigma) = d(\pi\nu,\sigma\nu).}$$

Inference (full rankings)

We have observed $\boldsymbol{\sigma} = \{\sigma_1, \dots, \sigma_k\}$ from the neighbors.

$$\mathcal{P}(\boldsymbol{\sigma}|\theta,\pi) = \prod_{i=1}^{k} \mathcal{P}(\sigma_i|\theta,\pi)$$

$$= \prod_{i=1}^{k} \frac{\exp\left(-\theta d(\sigma_i,\pi)\right)}{\phi(\theta)}$$

$$= \frac{\exp\big(-\theta(d(\sigma_1,\pi)+\ldots+d(\sigma_k,\pi))\big)}{\phi^k(\theta)}$$

$$= \frac{\exp\left(-\theta \sum_{i=1}^{k} d(\sigma_i,\pi)\right)}{\left(\prod_{j=1}^{n} \frac{1-\exp(-j\theta)}{1-\exp(-\theta)}\right)^k}.$$

ML

$$\hat{\pi} = \arg\min_{\pi} \sum_{i=1}^{k} d(\sigma_i,\pi)$$

$$\frac{1}{k}\sum_{i=1}^{k} d(\sigma_i,\hat{\pi}) = \frac{n\exp(-\theta)}{1-\exp(-\theta)} - \sum_{j=1}^{n} \frac{j\exp(-j\theta)}{1-\exp(-j\theta)}$$

monotone in θ

Inference (incomplete ranking)

“marginal” distibution $\mathcal{P}(E(\sigma_i)) = \sum_{\sigma \in E(\sigma_i)} \mathcal{P}(\sigma|\theta, \pi)$

where $E(\sigma_i)$ denotes all consistent extensions of σ_i.

Example for label set $\{a,b,c\}$:

Observation σ	Extensions $E(\sigma)$
$a \succ b$	$a \succ b \succ c$ $a \succ c \succ b$ $c \succ a \succ b$

Inference (incomplete ranking) cont.

The corresponding likelihood:

$$\begin{aligned}\mathcal{P}(\boldsymbol{\sigma}|\theta,\pi) &= \prod_{i=1}^{k} \mathcal{P}(E(\sigma_i)|\theta,\pi) \\ &= \prod_{i=1}^{k} \sum_{\sigma \in E(\sigma_i)} \mathcal{P}(\sigma|\theta,\pi) \\ &= \frac{\prod_{i=1}^{k} \sum_{\sigma \in E(\sigma_i)} \exp\left(-\theta d(\sigma,\pi)\right)}{\left(\prod_{j=1}^{n} \frac{1-\exp(-j\theta)}{1-\exp(-\theta)}\right)^{k}}.\end{aligned}$$

ML estimation $(\hat{\pi}, \hat{\theta}) = \arg\max_{\pi,\theta} \mathcal{P}(\boldsymbol{\sigma}|\theta,\pi)$ becomes more difficult.

Inference

Not only the estimated ranking $\hat{\pi}$ is of interest …

… but also the spread parameter $\hat{\theta}$, which is a measure of precision and, therefore, reflects the confidence/reliability of the prediction (just like the variance of an estimated mean).

The bigger $\hat{\theta}$, the more peaked the distribution around the center ranking.

Experimental Setting

Data sets

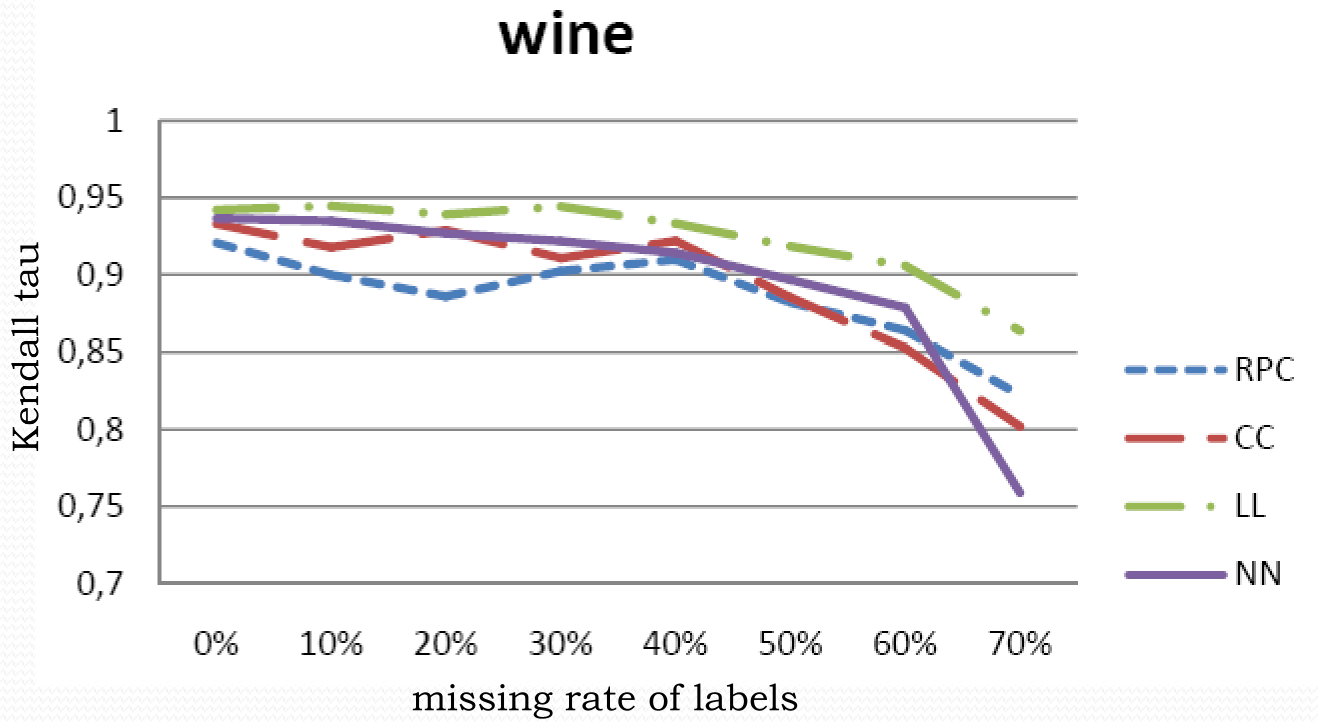
Name	#instances	#features	#labels
Iris[1]	150	4	3
Wine[1]	178	13	3
Glass[1]	214	9	6
Vehicle[1]	846	18	4
Dtt[2]	2465	24	4
Cold[2]	2465	24	4

[1] UCI data sets.

[2] Phylogenetic profiles and DNA microarray expression data.

Accuracy (Kendall tau)

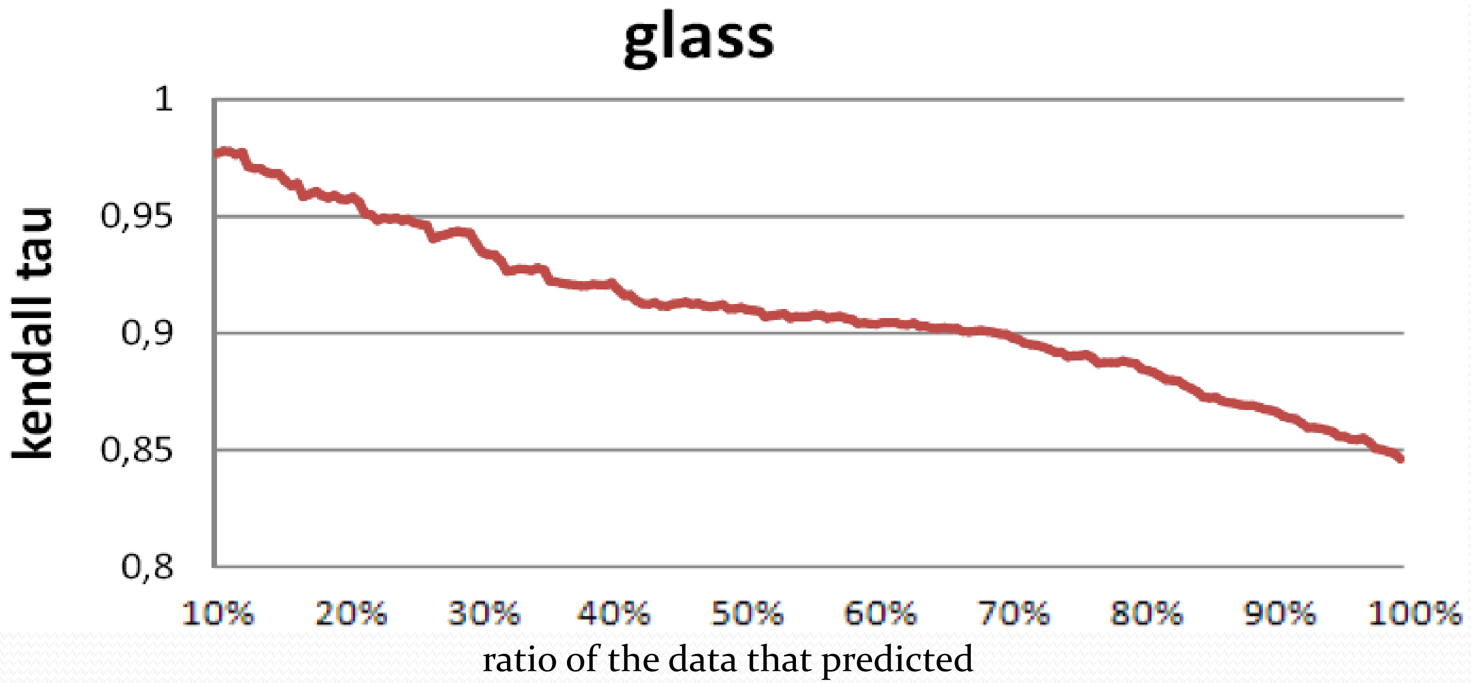
A typical run:



Main observation: Our approach is quite competitive with the state-of-the-art model based approaches.

Accuracy-Rejection Curve

θ as a measure of the reliability of a prediction



Main observation: Decreasing curve confirms that θ is a reasonable measure of confidence.

Take-away Messages

- An instance-based label ranking approach using a probabilistic model.
- Suitable for complete and incomplete rankings.
- Comes with a natural measure of the reliability of a prediction.

- More efficient inference for the incomplete case.
- Generalization: distance-weighted prediction.
- Dealing with variants of the label ranking problem, such as calibrated label ranking and multi-label classification.

Thanks!

Median rank

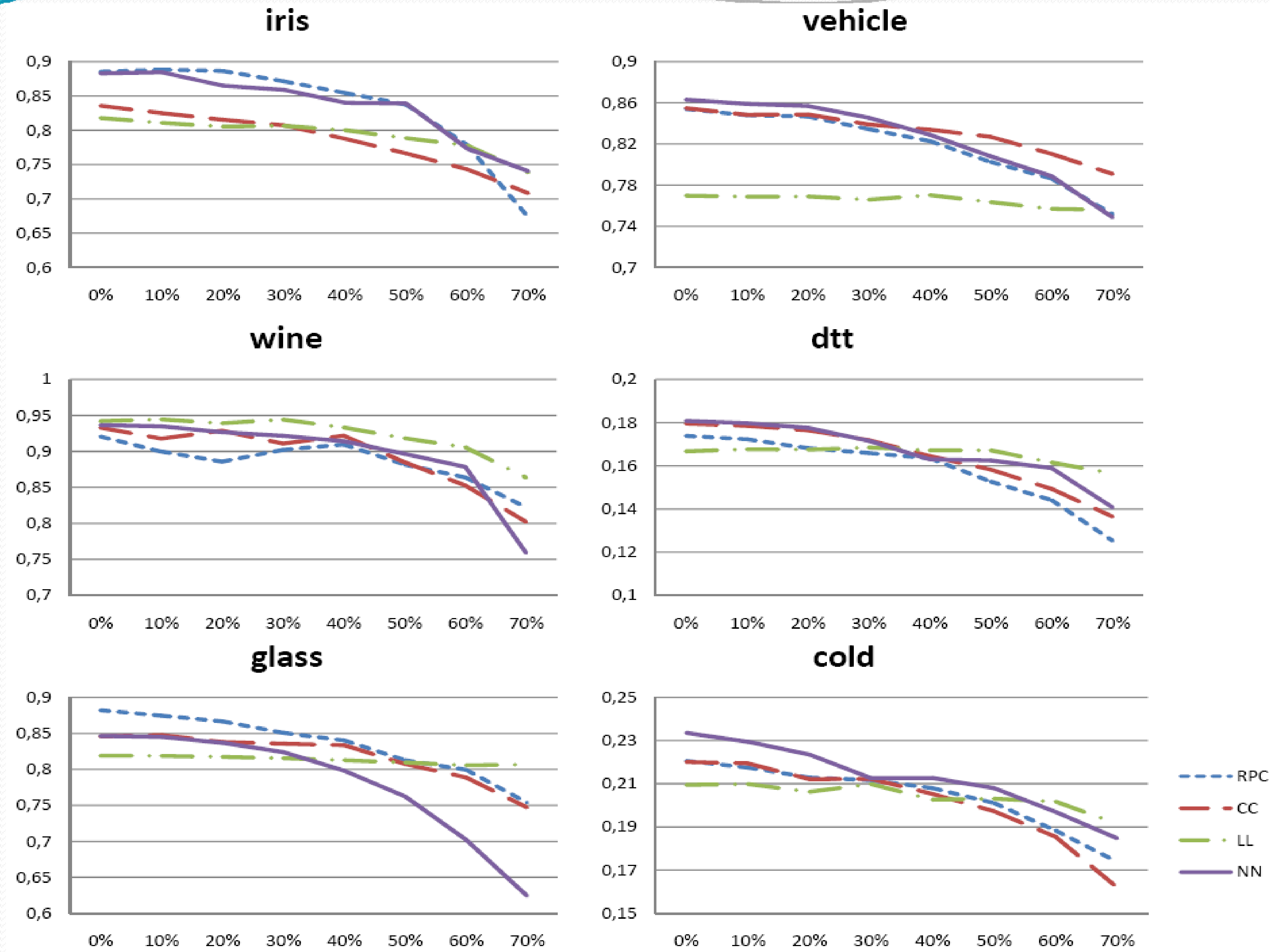
	MINI	Toyota	BMW
ranking 1	1	3	2
ranking 2	2	3	1
ranking 3	3	2	1
median rank	2	3	1

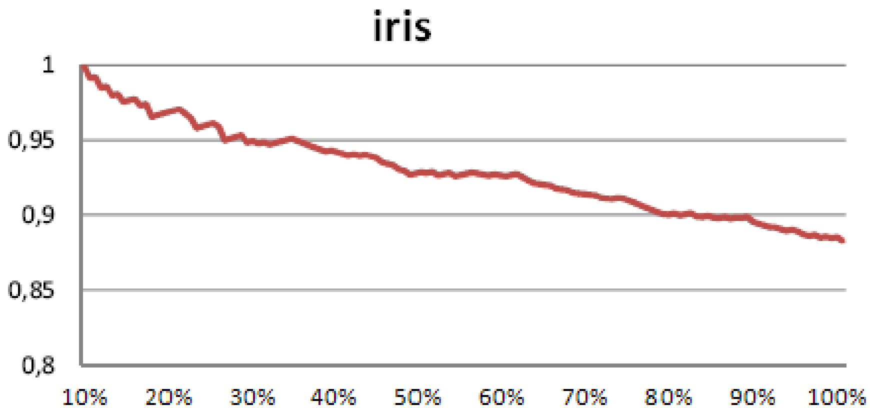
- tends to optimize Spearman footrule $\sum_{i=1}^{n} |\pi(i) - \hat{\pi}(i)|$

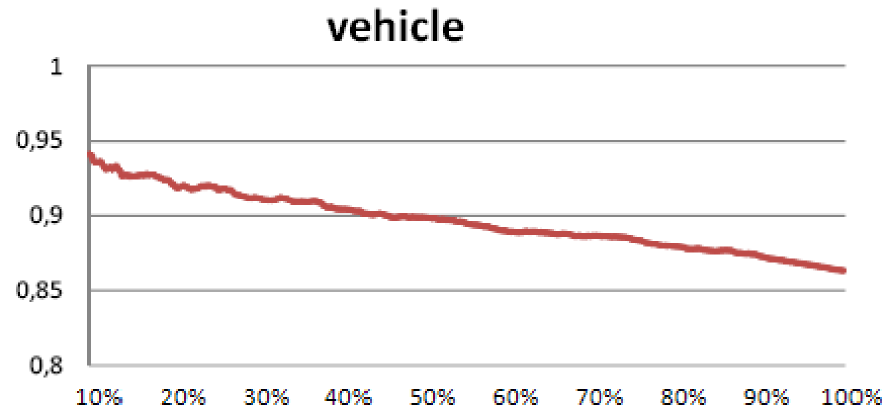
Borda count

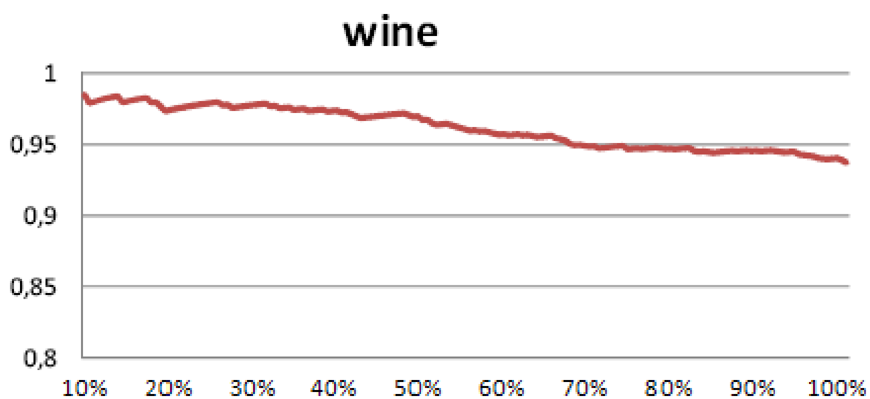
ranking 1	MINI $\succ$ Toyota $\succ$ BMW
ranking 2	BMW $\succ$ MINI $\succ$ Toyota
ranking 3	BMW $\succ$ Toyota $\succ$ MINI
Borda count	BMW: 4 MINI: 3 Toyota: 2

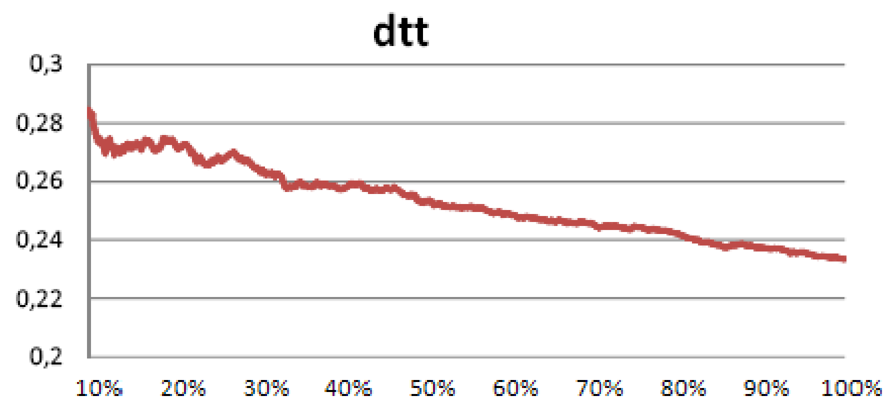
- tends to optimize Spearman rank correlation $\sum_{i=1}^{n} (\pi(i) - \hat{\pi}(i))^2$

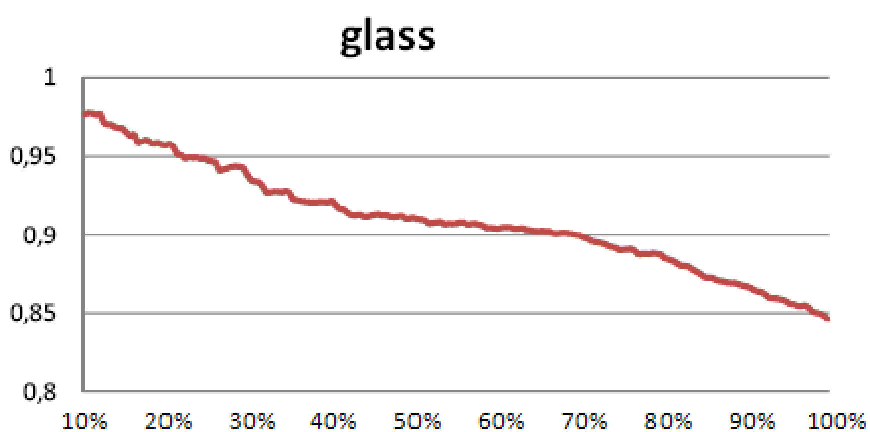

iris
vehicle
wine
dtt
glass
cold
0,9
0,85
0,8
0,75
0,7
0,65
0,6
0,86
0,82
0,78
0,74
1
0,95
0,2
0,18
0,16
0,14
0,12
0,1
0,25
0,23
0,21
0,19
0,17
0,15
0%
10%
20%
30%
40%
50%
60%
70%
RPC
CC
LL
NN

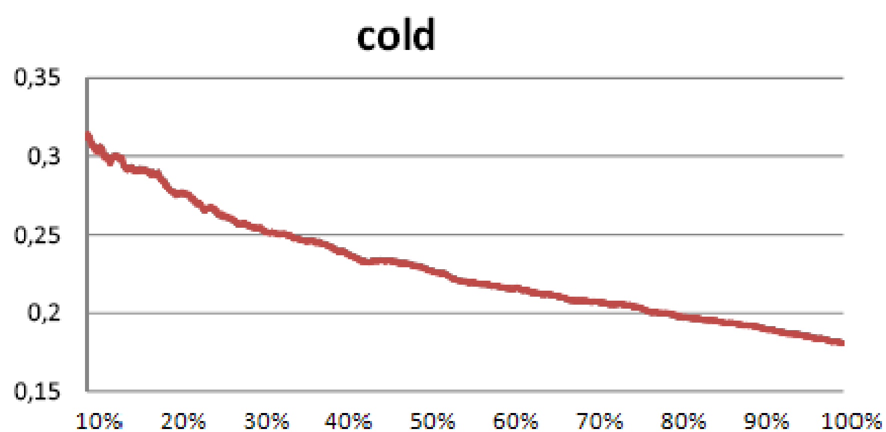

iris
1
0,95
0,9
0,85
0,8
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%


vehicle
1
0,95
0,9
0,85
0,8
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%


wine
1
0,95
0,9
0,85
0,8
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%


dtt
0,3
0,28
0,26
0,24
0,22
0,2
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%


glass
1
0,95
0,9
0,85
0,8
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%


cold
0,35
0,3
0,25
0,2
0,15
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%