New perspectives on the problem of learning how to order high dimensional data

Nicolas Vayatis (ENS Cachan)

-

joint work with Stéphan Clémençon (Institut Telecom),
and Marine Depecker (CEA), Sylvain Robbiano (Institut Telecom)

EURO 2013 - Sapienza University, Roma

-

July 1st, 2013

Motivations

- **Rank!**

  Learn an order relation on a high dimensional space, *e.g.* $\mathbb{R}^d$

  $$x \preceq x' \ , \qquad \text{for} \ \ x, x' \in \mathbb{R}^d$$

- **Drop logistic regression!**

  Alternative approach to parametric modeling of the posterior probability

- **Less is more!**

  The *statistical scoring* problem...

  ... somewhere between classification and regression function estimation

Predictive Ranking/Scoring

- **Training data:** past data $\{X_1, \dots, X_n\}$ in $\mathbb{R}^d$ and *some* feedback on the ordering
- **Input:** new data $\{X'_1, \dots, X'_m\}$ with no feedback
- **Goal:** predict a ranking $(X'_{i_1}, \dots, X'_{i_m})$ from *best* to *worst*
- **Our approach:** build a scoring rule: $s \;:\; \mathbb{R}^d \to \mathbb{R}$
- **Key question:** when shall we be happy?
- **Answer:** study optimal elements and performance metrics

Nature of feedback information

- **Preference model:** label Z on pair (X, X')
- **Plain regression:** individual label Y over $\mathbb{R}$
- **Bipartite ranking:** binary classification data (X, Y), $Y \in \{-1, +1\}$
- **K-partite ranking:** ordinal labels Y over $\{1, \ldots, K\}$, $K > 2$

Optimal elements
for statistical scoring

- Bipartite case
  [Clémençon and V., IEEE IT, 2009]
- K-partite case
  [Clémençon, Robbiano and V., MLJ, 2013]
- Local AUC
  [Clémençon and V., JMLR, 2007]

Optimal elements for scoring ($K = 2$)

- Probabilistic modeling: $(X, Y) \sim P$ over $\mathbb{R}^d \times \{-1, +1\}$
- Key theoretical quantity (posterior probability)

$$\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\} \ , \quad \forall x \in \mathbb{R}^d$$

- Optimal scoring rules:

  $\Rightarrow$ increasing transforms of η (by Neyman-Pearson's Lemma)

Representation of optimal scoring rules ($K = 2$)

- Note that if $U \sim \mathcal{U}([0,1])$

$$\forall x \in \mathbb{R}^d \ , \quad \eta(x) = \mathbb{E}\left(\mathbb{I}\{\eta(x) > U\}\right)$$

- If $s^* = \psi \circ \eta$ with ψ strictly increasing, then:

$$\forall x \in \mathbb{R}^d \ , \quad s^*(x) = c + \mathbb{E}\left(w(V) \cdot \mathbb{I}\{\eta(x) > V\}\right)$$

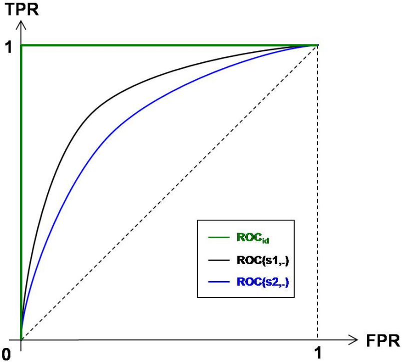
  for some:
  - $c \in \mathbb{R}$,
  - V continuous random variable in $[0,1]$
  - $w : [0,1] \to \mathbb{R}_+$ integrable.

- Optimal scoring amounts to recovering the level sets of η:

$$\{x : \eta(x) > q\}_{q \in (0,1)}$$

Classical performance measures for scoring ($K = 2$)

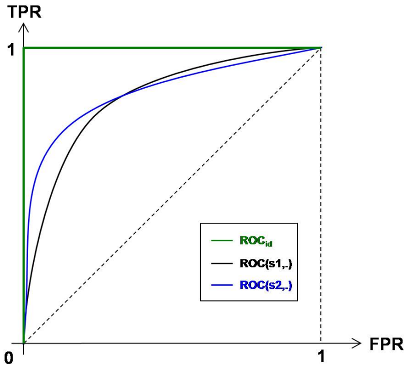
- Curve:
  - **ROC curve**
- Summaries (global vs. best scores):
  - **AUC** (global measure)
  - Partial AUC (Dodd and Pepe '03)
  - **Local AUC** (Clémençon and Vayatis '07)



ROC curves.

Classical performance measures for scoring ($K = 2$)

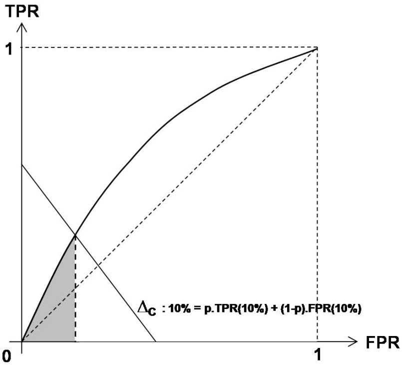
- Curve:
  - **ROC curve**
- Summaries (global vs. best scores):
  - **AUC** (global measure)
  - Partial AUC (Dodd and Pepe '03)
  - **Local AUC** (Clémençon and Vayatis '07)



ROC curves.

Classical performance measures for scoring ($K = 2$)

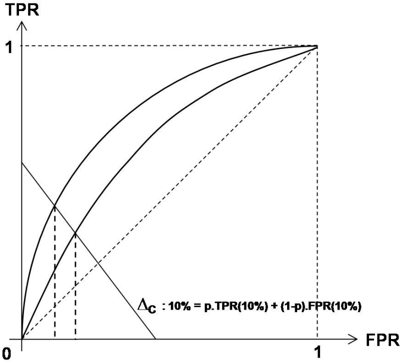
- Curve:
  - **ROC curve**
- Summaries (global vs. best scores):
  - **AUC** (global measure)
  - Partial AUC (Dodd and Pepe '03)
  - **Local AUC** (Clémençon and Vayatis '07)



Partial AUC.

Classical performance measures for scoring ($K = 2$)

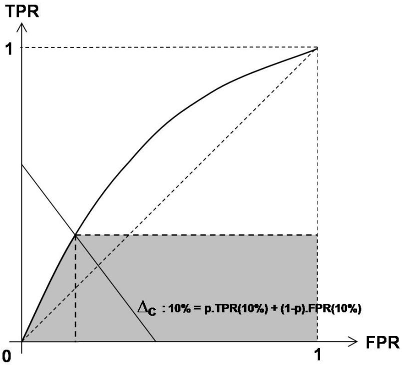
- Curve:
  - **ROC curve**
- Summaries (global vs. best scores):
  - **AUC** (global measure)
  - Partial AUC (Dodd and Pepe '03)
  - **Local AUC** (Clémençon and Vayatis '07)



Inconsistency of Partial AUC.

Classical performance measures for scoring ($K = 2$)

- Curve:
  - **ROC curve**
- Summaries (global vs. best scores):
  - **AUC** (global measure)
  - Partial AUC (Dodd and Pepe '03)
  - **Local AUC** (Clémençon and Vayatis '07)



Local AUC.

Optimal elements ($K > 2$)

- Recall for $K = 2$:

$$s^* = T \circ \eta = \tilde{T} \circ \left(\frac{\eta}{1 - \eta} \right)$$

is optimal for any strictly increasing transform T (or $\tilde{T}$).

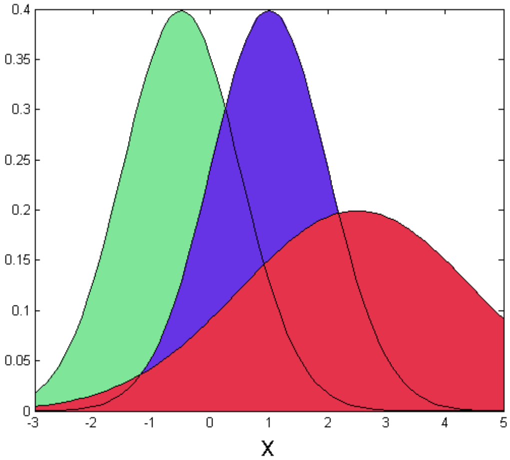
- For $K > 2$, define an optimal element as s^* by:

$\forall l < k$, $\exists T_{l,k}$ strictly increasing such that:

$$s^* = T_{l,k} \circ \left(\frac{\eta_k}{\eta_l} \right)$$

where $\eta_k(x) = \mathbb{P}(Y = k \mid X = x)$.

Counterexample for optimality with $K = 3$



Assumption **(H1)** - Monotonicity condition

- For any $1 \leq l < k \leq K-1$, we have: for x, x',

$$\frac{\eta_{k+1}}{\eta_k}(x) < \frac{\eta_{k+1}}{\eta_k}(x') \Rightarrow \frac{\eta_{l+1}}{\eta_l}(x) < \frac{\eta_{l+1}}{\eta_l}(x') \qquad \textbf{(H1)}$$

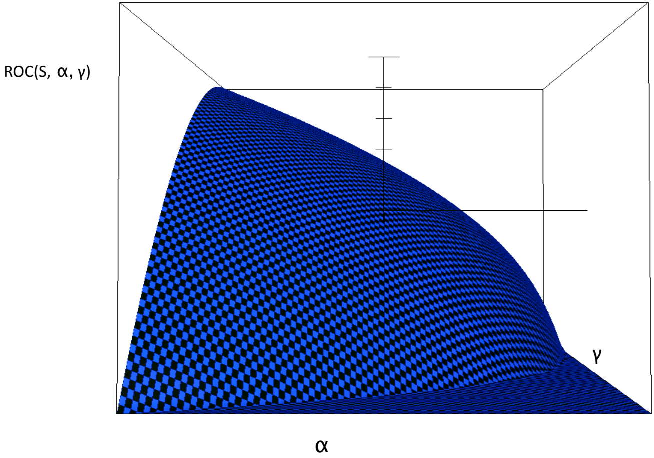
- Sufficient and necessary condition for the existence of an optimal scoring rule.

- Then, the regression function

$$\eta(x) = \mathbb{E}(Y \mid X = x) = \sum_{k=1}^{K} k \cdot \eta_k(x)$$

is optimal.

Assess performance for $K = 3$ - ROC surface and VUS



Aggregation principle for scoring

- Bipartite case
  [Clémençon, Depecker and V., JMLR, 2013]
- K-partite case
  [Clémençon, Robbiano and V., MLJ, 2013]

Motivations

- $K > 2$

  Mimic multiclass classification strategies based on binary decision rules (one vs. one, one against all, ...)

- $K = 2$

  Mimic bagging-like strategies for boosting performance and increase robustness

Agreement with Kendall τ

- Let X, X' i.i.d.and s_1 and s_2 real-valued scoring rules :

$$\begin{aligned}\tau(s_1, s_2) = \mathbb{P}\left\{\left(s_1(X) - s_1(X')\right) \cdot \left(s_2(X) - s_2(X')\right) > 0\right\} \\ + \frac{1}{2}\mathbb{P}\left\{s_1(X) \neq s_1(X'),\ s_2(X) = s_2(X')\right\} \\ + \frac{1}{2}\mathbb{P}\left\{s_1(X) = s_1(X'),\ s_2(X) \neq s_2(X')\right\} \ .\end{aligned}$$

- Define pseudo-distance between scoring rules:

$$d_\tau(s_1, s_2) = \frac{1}{2}(1 - \tau(s_1, s_2))$$

Median scoring rule

- Weak scoring rules $\Sigma_N = \{s_1, \ldots, s_N\}$
- Candidate class $\mathcal{S}$ for median scoring rule (aggregate)
- Median scoring rule $\bar{s}$ with respect to $(\mathcal{S}, \Sigma_N)$:

$$\sum_{j=1}^{N} d_\tau\left(\bar{s}, s_j\right) = \inf_{s \in \mathcal{S}} \sum_{j=1}^{N} d_\tau\left(s, s_j\right)$$

  (if the inf is reached).

- Link with the AUC ($K = 2$):

$$\mid \mathrm{AUC}(\mathrm{s}_1) - \mathrm{AUC}(\mathrm{s}_2) \mid \leq \frac{1}{2\mathrm{p}_+\mathrm{p}_-} \mathrm{d}_\tau(\mathrm{s}_1, \mathrm{s}_2)$$

Inverse control under low noise assumption **(H2)**

- The posterior probability $\eta(X)$ is a continuous random variable and there exist $c < \infty$ and $a \in (0,1)$ such that

$$\forall x \in \mathbb{R}^d \ , \ \ \mathbb{E}\left[|\eta(X) - \eta(x)|^{-a}\right] \leq c \ . \qquad \textbf{(H2)}$$

- Sufficient condition: $\eta(X)$ has bounded density function

- Inverse control under **(H2)**:

$$d_\tau(s, s^*) \leq C(\mathrm{AUC}^* - \mathrm{AUC(s)})^{\mathrm{a}/(1+\mathrm{a})}$$

for some $C = C(a, c, p_+)$.

Main results - Aggregation does not hurt!

- $K > 2$

  Aggregation permits to derive a consistent scoring rule for the K-partite problem from consistent rules on the pairwise bipartite subproblems.

- $K = 2$

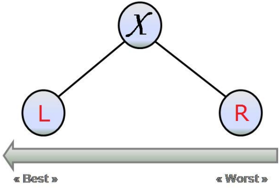
  Aggregation of consistent randomized scoring rules preserves AUC consistency .

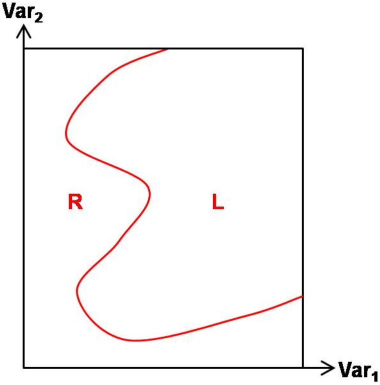
The TreeRank algorithm

- Plain TreeRank
  [Clémençon and V., IEEE IT, 2009]
- Optimized TreeRank
  [Clémençon, Depecker and V., MLJ, 2011]
- Aggregate TreeRank = Ranking Forests
  [Clémençon, Depecker and V., JMLR, 2013]

TreeRank - building ranking (binary) trees

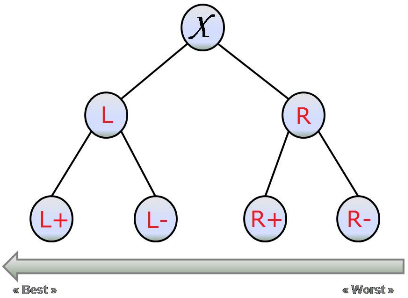
- Input domain $[0,1]^d$

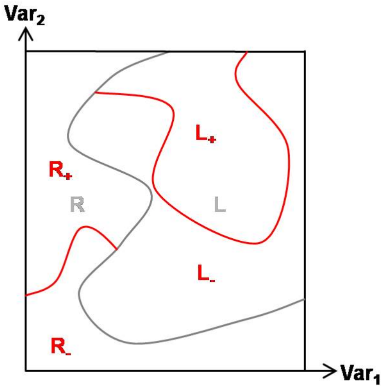




TreeRank - building ranking (binary) trees

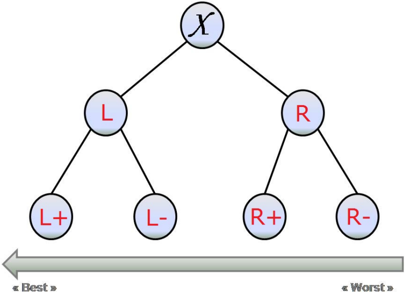
- Input domain $[0, 1]^d$

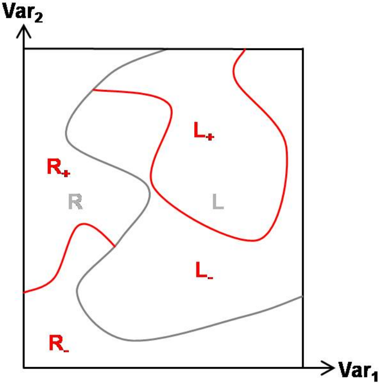




TreeRank - building ranking (binary) trees

- Input domain $[0,1]^d$



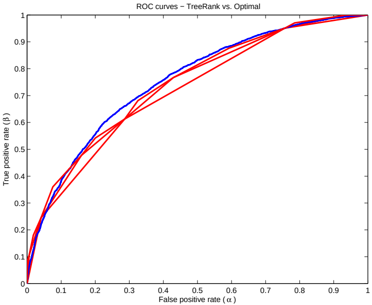


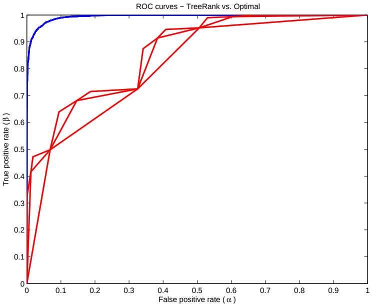
- A wiser option: use orthogonal splits!

Empirical performance of TreeRank

Gaussian mixture with orthogonal split

easy with overlap **vs.** difficult and no overlap



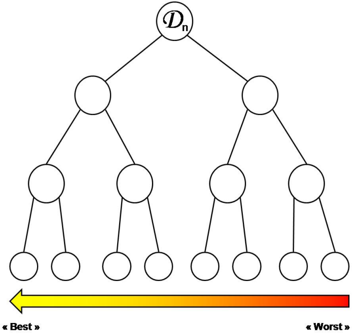


TreeRank and the problem with recursive partitioning

- The TreeRank algorithm:
  - implements an empirical version of local AUC maximization procedure
  - yields AUC- **and** ROC- **consistent** scoring rules (Clémençon-Vayatis '09)
  - boils down to solving a collection of **nested** optimization problems

- **Main goal:**
  - Global performance in terms of the ROC curve

- **Main issue:**
  - Recursive partitioning not so good when the nature of the problem is not local



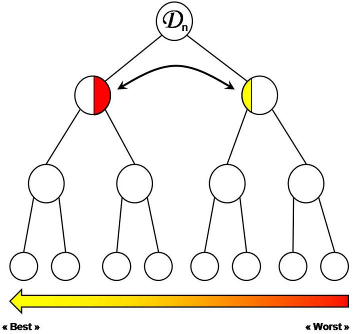
- **Key point:** choice of a splitting rule for the AUC optimization step

TreeRank and the problem with recursive partitioning

- The TreeRank algorithm:
  - implements an empirical version of local AUC maximization procedure
  - yields AUC- **and** ROC- **consistent** scoring rules (Clémençon-Vayatis '09)
  - boils down to solving a collection of **nested** optimization problems

- **Main goal:**
  - Global performance in terms of the ROC curve

- **Main issue:**
  - Recursive partitioning not so good when the nature of the problem is not local



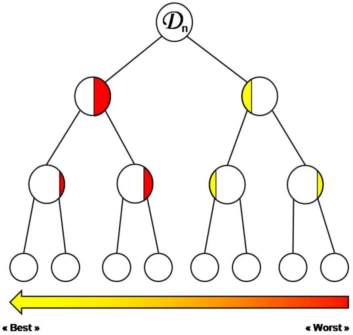
- **Key point:** choice of a splitting rule for the AUC optimization step

TreeRank and the problem with recursive partitioning

- The TreeRank algorithm:
  - implements an empirical version of local AUC maximization procedure
  - yields AUC- **and** ROC- **consistent** scoring rules (Clémençon-Vayatis '09)
  - boils down to solving a collection of **nested** optimization problems

- **Main goal:**
  - Global performance in terms of the ROC curve

- **Main issue:**
  - Recursive partitioning not so good when the nature of the problem is not local



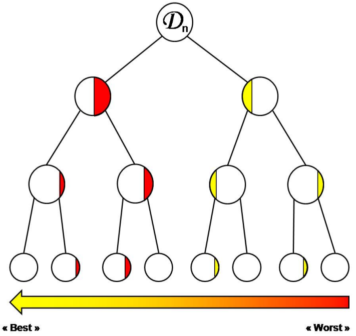
- **Key point:** choice of a splitting rule for the AUC optimization step

TreeRank and the problem with recursive partitioning

- The TreeRank algorithm:
  - implements an empirical version of local AUC maximization procedure
  - yields AUC- **and** ROC- **consistent** scoring rules (Clémençon-Vayatis '09)
  - boils down to solving a collection of **nested** optimization problems

- **Main goal:**
  - Global performance in terms of the ROC curve

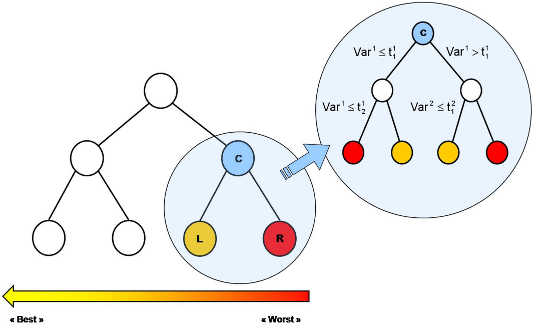
- **Main issue:**
  - Recursive partitioning not so good when the nature of the problem is not local

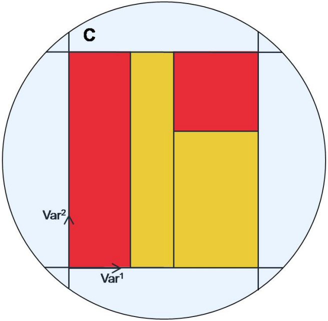


- **Key point:** choice of a splitting rule for the AUC optimization step

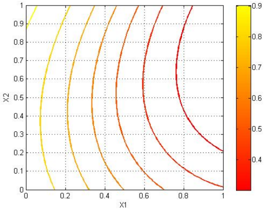
Nonlocal splitting rule - The LeafRank Procedure

- Any classification method can be used as a splitting rule
- Our choice: the LeafRank procedure
  - Use classification tree with orthogonal splits (CART)
  - Find optimal cell permutation for a fixed partition
  - Improves representation capacity and still permits interpretability

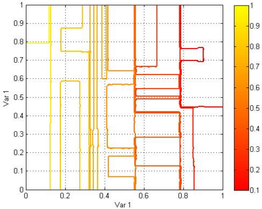




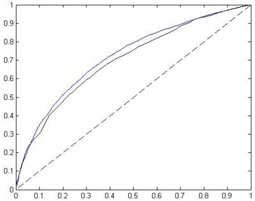
Iterative TreeRank in action- synthetic data set



a. Level sets of the true regression function η.



b. Level sets of the estimated regression function η.



c. True (blue) and Estimated (black) Roc Curve.

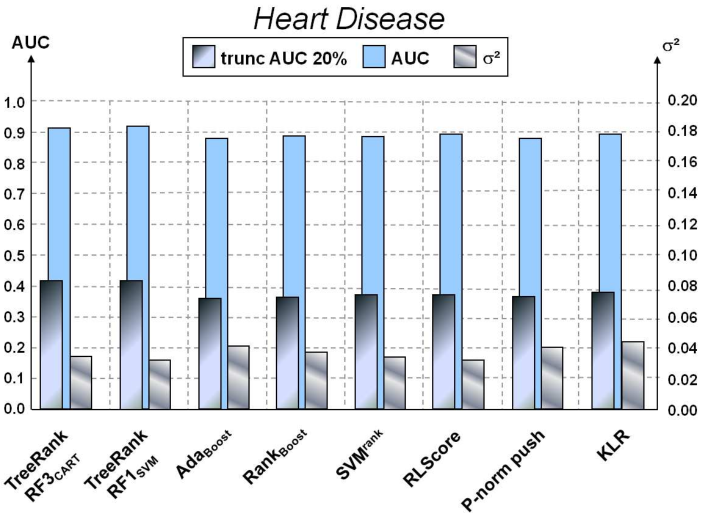
TreeRank in action!

- Extended comparison
  [Clémençon, Depecker and V., PAA, 2012]

RankForest and competitors on UCI data sets (1)

- Data sets from the UCI Machine Learning repository
  - Australian Credit
  - Ionosphere
  - Breast Cancer
  - Heart Disease
  - Hepatitis
- **Competitors:**
  - AdaBoost (Freund and Schapire '95)
  - RankBoost (Freund *et al.* '03)
  - RankSvm (Joachims '02, Rakotomamonjy '04)
  - RankRLS (Pahikkala *et al.* '07)
  - KLR (Zhu and Hastie '01)
  - P-normPush (Rudin '06)

RankForest and competitors (2)



Local AUC $u = 0.5$ $u = 0.2$ $u = 0.1$	TreeRank	RankBoost	RankSVM
Australian Credit	0.425 (±0.012) 0.248 (±0.039) 0.111 (±0.002)	0.412 (±0.014) 0.206 (±0.013) 0.103 (±0.011)	0.404 (±0.024) 0.204 (±0.013) 0.103 (±0.010)
Ionosphere	0.494 (±0.062) 0.156 (±0.002) 0.078 (±0.001)	0.288 (±0.005) 0.144 (±0.003) 0.072 (±0.003)	0.263 (±0.044) 0.131 (±0.024) 0.065 (±0.014)
Breast Cancer	0.559 (±0.010) 0.442 (±0.076) 0.146 (±0.010)	0.534 (±0.018) 0.265 (±0.012) 0.132 (±0.014)	0.537 (±0.017) 0.271 (±0.009) 0.137 (±0.012)
Heart Disease	0.416 (±0.027) 0.273 (±0.070) 0.118 (±0.017)	0.361 (±0.041) 0.176 (±0.027) 0.089 (±0.017)	0.371 (±0.035) 0.188 (±0.022) 0.094 (±0.011)
Hepatitis	0.572 (±0.240) 0.413 (±0.138) 0.269 (±0.190)	0.504 (±0.225) 0.263 (±0.115) 0.133 (±0.057)	0.526 (±0.248) 0.272 (±0.125) 0.137 (±0.062)

Perspectives

- Nonparametric multivariate homogeneity tests
- Application to experimental design
- Statistical theory (rates of convergence? analysis of R-processes?)