Statistical Multicriteria Benchmarking via the GSD-Front

Christoph Jansen[1*], Georg Schollmeyer[2*], Julian Rodemann[2*], Hannah Blocher[2*], Thomas Augustin[2]

[1]School of Computing & Communications, Lancaster University Leipzig

[2]Department of Statistics, Ludwig-Maximilians-Universität München

* marks equal contribution

Why use multiple criteria in benchmark studies?

Reason 1: **Performance is a latent construct**

The application at hand suggests a very clear evaluation concept, which is too complex to be expressed in terms of a single metric.

Example: *Robustness* as stability under perturbations of both X and Y.

Why use multiple criteria in benchmark studies?

Reason 1: **Performance is a latent construct**

The application at hand suggests a very clear evaluation concept, which is too complex to be expressed in terms of a single metric.

Example: *Robustness* as stability under perturbations of both X and Y.

Reason 2: **Quality is a multidimensional concept**

It may be desirable to trade-off various competing quality dimensions.

Example: Trade-off between *accuracy* and *computation time*.

Why use multiple criteria in benchmark studies?

Reason 1: **Performance is a latent construct**

The application at hand suggests a very clear evaluation concept, which is too complex to be expressed in terms of a single metric.

Example: *Robustness* as stability under perturbations of both X and Y.

Reason 2: **Quality is a multidimensional concept**

It may be desirable to trade-off various competing quality dimensions.

Example: Trade-off between *accuracy* and *computation time*.

Take-away:

*Using **multiple criteria should be standard** rather than the exception.*

Five Challenges in (Multicriteria) Benchmarking

Setup: Let

- $\mathcal{D}$ denote the universe of data sets,
- $\mathcal{C}$ denote the finite set of all relevant classifiers,
- $\left(\phi_i : \mathcal{C} \times \mathcal{D} \to [0,1]\right)_{i \in \{1,\ldots,n\}}$ denote a family of quality criteria,
- $\Phi := (\phi_1, \ldots, \phi_n) : \mathcal{D} \times \mathcal{C} \to [0,1]^n$ be the mulidimensional criterion.

Five Challenges in (Multicriteria) Benchmarking

Setup: Let

- $\mathcal{D}$ denote the set of all relevant data sets,
- $\mathcal{C}$ denote the finite set of all relevant classifiers,
- $(\phi_i : \mathcal{C} \times \mathcal{D} \to [0,1])_{i \in \{1,\ldots,n\}}$ denote a family of quality criteria,
- $\Phi := (\phi_1, \ldots, \phi_n) : \mathcal{D} \times \mathcal{C} \to [0,1]^n$ be the mulidimensional criterion.

Assumptions:

- For $0 \leq z \leq n$, the criteria $\phi_1, \ldots, \phi_z$ are of cardinal scale.
- The remaining criteria are of purely ordinal scale.

Five Challenges in (Multicriteria) Benchmarking

classifier \ data sets	D_1	$\dots$	D_s
C_1	$\begin{pmatrix} \phi_1(C_1, D_1) \\ \vdots \\ \phi_n(C_1, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} \phi_1(C_q, D_1) \\ \vdots \\ \phi_n(C_q, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_q, D_s) \\ \vdots \\ \phi_n(C_q, D_s) \end{pmatrix}$

Five Challenges in (Multicriteria) Benchmarking

classifier \ data sets	D_1	$\dots$	D_s
C_1	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.7 \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.8 \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_q, D_s) \\ \vdots \\ \phi_n(C_q, D_s) \end{pmatrix}$

Challenge 1: Intra-dataset incomparability

On a **fixed** data set D it may hold

$$\phi_1(C_1, D) > \phi_1(C_2, D) \ \wedge \ \phi_2(C_1, D) < \phi_2(C_2, D).$$

Five Challenges in (Multicriteria) Benchmarking

classifier \ data sets	D_1	$\dots$	D_s
C_1	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.8 \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$		
C_q	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$		

Challenge 2: Conflicting datasets

Even if, for all $i \in \{1, \dots, n\}$, we have

$$\phi_i(C_1, D_1) > \phi_i(C_2, D_1)$$

Five Challenges in (Multicriteria) Benchmarking

classifier \ data sets	D_1	$\dots$	D_s
C_1	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.8 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.6 \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.9 \\ \vdots \\ \phi_n(C_q, D_s) \end{pmatrix}$

Challenge 2: Conflicting datasets

Even if, for all $i \in \{1, \dots, n\}$, we have

$$\phi_i(C_1, D_1) > \phi_i(C_2, D_1)$$

there may exists some $i_0 \in \{1, \dots, n\}$ such that

$$\phi_{i_0}(C_1, D_2) < \phi_{i_0}(C_2, D_2).$$

Observation: Under challenges 1 and 2, commonly the Pareto-front will consist of all classifiers in $\mathcal{C}$ and not allow for a meaningful analysis.

Five Challenges in (Multicriteria) Benchmarking

classifier \ data sets	D_1	$\dots$	D_s
C_1	$\begin{pmatrix} \texttt{medium} \\ \vdots \\ 0.8 \end{pmatrix}$	$\dots$	$\begin{pmatrix} \texttt{bad} \\ \vdots \\ 0.7 \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} \texttt{good} \\ \vdots \\ 0.93 \end{pmatrix}$	$\dots$	$\begin{pmatrix} \texttt{excellent} \\ \vdots \\ 0.64 \end{pmatrix}$

Challenge 3: Mixed-scaled quality metrics

Even if some of the quality metrics are only of ordinal scale, we still want to capture the entire information encoded in the metrics with cardinal scale.

Five Challenges in (Multicriteria) Benchmarking

classifier \ data sets	D_1	$\dots$	D_s
C_1	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.8 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.8 \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$

Challenge 4: Lack of inferential guarantees

Even if a decision can be made for a sample $(D_1, \dots, D_s)$ of data sets,

Five Challenges in (Multicriteria) Benchmarking

classifier \ data sets	D_1^*	$\dots$	D_s^*
C_1	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.9 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.75 \\ \vdots \\ 0.4 \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} 0.85 \\ \vdots \\ 0.67 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.33 \\ \vdots \\ 0.98 \end{pmatrix}$

Challenge 4: Lack of inferential guarantees

Even if a decision can be made for a sample $(D_1, \dots, D_s)$ of data sets, no clear decision might be possible for a different sample $(D_1^*, \dots, D_s^*)$.

Five Challenges in (Multicriteria) Benchmarking

classifier \ data sets	D_1	i.i.d.!!	D_s
C_1	$\begin{pmatrix} \phi_1(C_1, D_1) \\ \vdots \\ \phi_n(C_1, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} \phi_1(C_q, D_1) \\ \vdots \\ \phi_n(C_q, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_q, D_s) \\ \vdots \\ \phi_n(C_q, D_s) \end{pmatrix}$

Challenge 5: Non-robustness under deviations from i.i.d.

Even if our classifier ranking comes with inferential guarantees under i.i.d. sampling of data sets,

Five Challenges in (Multicriteria) Benchmarking

classifier \ data sets	D_1^*	contamination!!	D_s^*
C_1	$\begin{pmatrix} \phi_1(C_1, D_1) \\ \vdots \\ \phi_n(C_1, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} \phi_1(C_q, D_1) \\ \vdots \\ \phi_n(C_q, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_q, D_s) \\ \vdots \\ \phi_n(C_q, D_s) \end{pmatrix}$

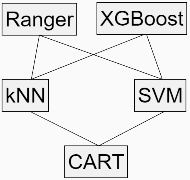
Challenge 5: Non-robustness under deviations from i.i.d.

Even if our classifier ranking comes with inferential guarantees under i.i.d. sampling of data sets, these are invalid under contaminated sampling.

Our Contribution

Start with the GSD relation $\succsim$ among classifiers

(borrowing ideas from decision theory ideas)



Our Contribution

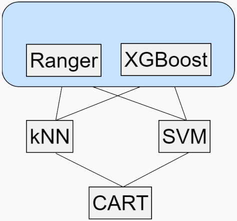
Start with the GSD relation $\succsim$ among classifiers

(borrowing ideas from decision theory ideas)

↓

$$\text{gsd}(\mathbb{C}) = \left\{ C : \nexists C' \succ C \right\}$$

(more informative than Pareto)



Our Contribution

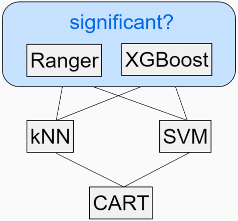
Start with the GSD relation $\gtrsim$ among classifiers
(borrowing ideas from decision theory ideas)

$$\text{gsd}(\mathcal{C}) = \left\{ C : \nexists C' \succ C \right\}$$

(more informative than Pareto)

$$H_0 : C \notin gsd(\mathcal{C}) \quad vs. \quad H_1 : C \in gsd(\mathcal{C})$$

(providing inferential guarantees under i.i.d)



Our Contribution

Start with the GSD relation $\succsim$ among classifiers
(borrowing ideas from decision theory ideas)

↓

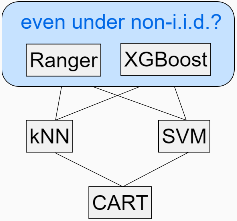
$$\text{gsd}(\mathcal{C}) = \left\{ C : \nexists C' \succ C \right\}$$
(more informative than Pareto)

↓

$$H_0 : C \notin gsd(\mathcal{C}) \quad vs. \quad H_1 : C \in gsd(\mathcal{C})$$
(providing inferential guarantees under i.i.d)

↓

Robustify the test for (H_0, H_1) to deviations from i.i.d.
(using ideas from imprecise probability theory)



Our Contribution

Start with the GSD relation $\succsim$ among classifiers
(borrowing ideas from decision theory ideas)

Challenges 1, 2, 3

$$\text{gsd}(\mathbb{C}) = \left\{C : \nexists C' \succ C\right\}$$

(more informative than Pareto)

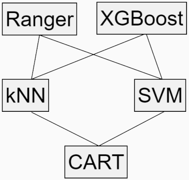
$$H_0 : C \notin gsd(\mathbb{C}) \quad vs. \quad H_1 : C \in gsd(\mathbb{C})$$

(providing inferential guarantees under i.i.d)

Challenge 4

Robustify the test for (H_0, H_1) to deviations from i.i.d.
(using ideas from imprecise probability theory)

Challenge 5



Thank you for your attention!

We hope to see many of you at our poster.

Statistical Multicriteria Benchmarking via the GSD-Front

Christoph Jansen[1,*]
c.jansen@lancaster.ac.uk

Georg Schollmeyer[2,*]
georg.schollmeyer@stat.uni-muenchen.de

Julian Rodemann[2,*]
julian@stat.uni-muenchen.de

Hannah Blocher[2,*]
hannah.blocher@stat.uni-muenchen.de

Thomas Augustin[2]
thomas.augustin@stat.uni-muenchen.de