Lancaster University Leipzig

Statistical Multicriteria Benchmarking via the GSD-Front

Presentation given at the **AI4Science Elevator Pitch Event**
Held on 10.07.2026 at Lancaster University Leipzig

Christoph Jansen
School of Computing & Communications
Lancaster University Leipzig

The Role of Benchmarking in Science

Claim 1: **Benchmarking is crucial for scientific progress**

Benchmarking machine learning models is essential for ensuring reliable, comparable, and reproducible performance across disciplines.

The Role of Benchmarking in Science

Claim 1: **Benchmarking is crucial for scientific progress**

Benchmarking machine learning models is essential for ensuring reliable, comparable, and reproducible performance across disciplines.

Claim 2: **Benchmarking is genuinely multidimensional**

Frameworks that consider multiple performance metrics better reflect the trade-offs inherent in machine learning.

The Role of Benchmarking in Science

Claim 1: **Benchmarking is crucial for scientific progress**

Benchmarking machine learning models is essential for ensuring reliable, comparable, and reproducible performance across disciplines.

Claim 2: **Benchmarking is genuinely multidimensional**

Frameworks that consider multiple performance metrics better reflect the trade-offs inherent in machine learning.

Claim 3: **Benchmarking needs inferential guarantees**

Without inferential guarantees, benchmarking results may reflect noise rather than true performance differences.

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

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	$\cdots$	D_s
C_1	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.7 \end{pmatrix}$	$\cdots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
C_q	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.8 \end{pmatrix}$	$\cdots$	$\begin{pmatrix} \phi_1(C_q, D_s) \\ \vdots \\ \phi_n(C_q, D_s) \end{pmatrix}$

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	$\cdots$	D_s
C_1	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.8 \end{pmatrix}$	$\cdots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$		
C_q	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$		

Challenge 2: Conflicting datasets

Even if, for all $i \in \{1, \ldots, n\}$, 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

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

Five Challenges in (Multicriteria) Benchmarking

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

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	$\begin{pmatrix} \phi_1(C_1, D_1) \\ \vdots \\ \phi_n(C_1, D_1) \end{pmatrix}$	$\cdots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_n(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} \phi_1(C_q, D_1) \\ \vdots \\ \phi_n(C_q, D_1) \end{pmatrix}$	$\cdots$	$\begin{pmatrix} \phi_1(C_q, D_s) \\ \vdots \\ \phi_n(C_q, D_s) \end{pmatrix}$

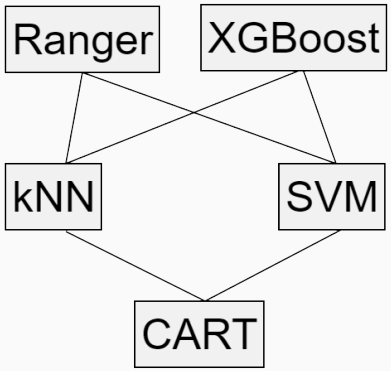
Challenge 5: Non-robustness under deviations from i.i.d.

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

What I have to offer

Start with the GSD relation $\gtrsim$ among classifiers

(borrowing ideas from decision theory ideas)



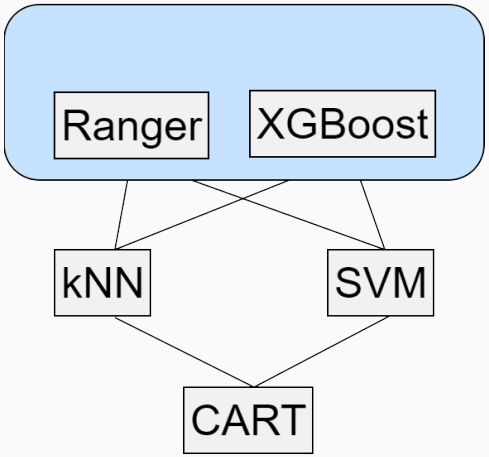
What I have to offer

Start with the GSD relation $\gtrsim$ among classifiers

(borrowing ideas from decision theory ideas)

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

(more informative than Pareto)



What I have to offer

Start with the GSD relation $\gtrsim$ among classifiers

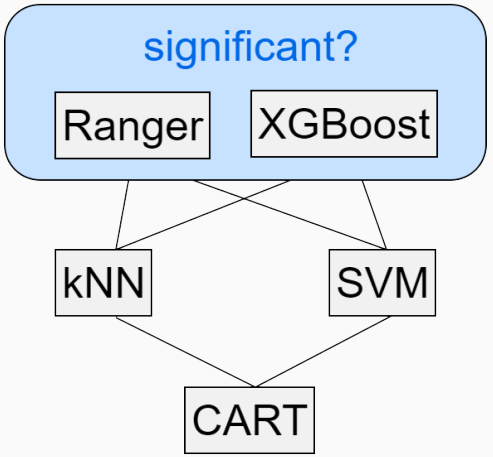
(borrowing ideas from decision theory ideas)

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

(more informative than Pareto)

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

(providing inferential guarantees under i.i.d)



What I have to offer

Start with the GSD relation $\gtrsim$ among classifiers

(borrowing ideas from decision theory ideas)

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

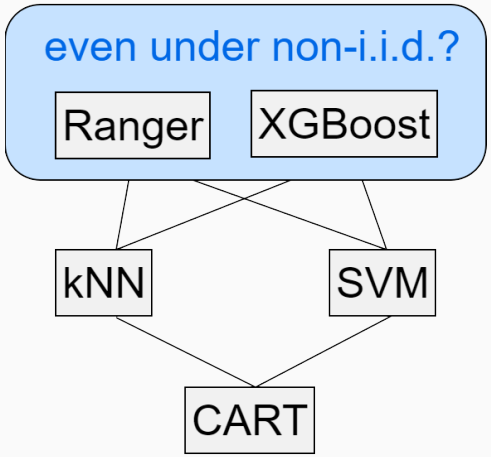
(more informative than Pareto)

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

(providing inferential guarantees under i.i.d)

Robustify the test for (H_0, H_1) to deviations from i.i.d.

(using ideas from imprecise probability theory)



What I have to offer

Start with the GSD relation $\gtrsim$ 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)

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

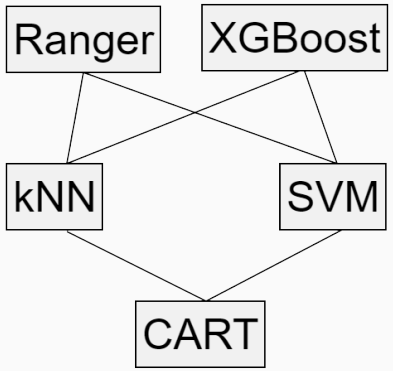
(providing inferential guarantees under i.i.d)

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!

Key papers:

- Garces-Arias, Blocher, Rodemann, Aßenmacher, Jansen: Statistical Multicriteria Evaluation of LLM-Generated Text, **INLG 2025**
- Jansen, Schollmeyer, Rodemann, Blocher, Augustin: Statistical Multicriteria Benchmarking via the GSD-Front, **NeurIPS 2024**.
- Jansen, Schollmeyer, Blocher, Rodemann, Augustin: Robust Statistical Comparison of Random Variables with Locally Varying Scale of Measurement, **UAI 2023**
- Jansen, Nalenz, Schollmeyer, Augustin: Statistical Comparisons of Classifiers by Generalized Stochastic Dominance, **JMLR 2022**