Some Insights from Decision Making under Strict Uncertainty for Machine Learning

Th. Augustin, C. Jansen, M. Nalenz, G. Schollmeyer

Foundations of Statistics and Their Applications
Department of Statistics, LMU Munich
augustin@stat.uni-muenchen.de

DA2PL 2022

Central Question

What can the

foundations of statistics

and decision theory

contribute to

machine learning?

Table of contents

Table of contents

Decision Theory

Find

optimal actions

under uncertainty about the "nature"

by optimally utilizing information from data available!

Here Two Aspects // Application Areas

- Some speculative ideas on optimality of algorithms and their construction (loss function externally given)
- Evaluation of benchmark study of given algorithms: formulate preferences under multiplicity of algorithms and evaluation criteria (multi-criteria decision making; preference system)

Decision Theory

Find

optimal actions

under uncertainty about the "nature"

by optimally utilizing information from data available!

Decision Theory

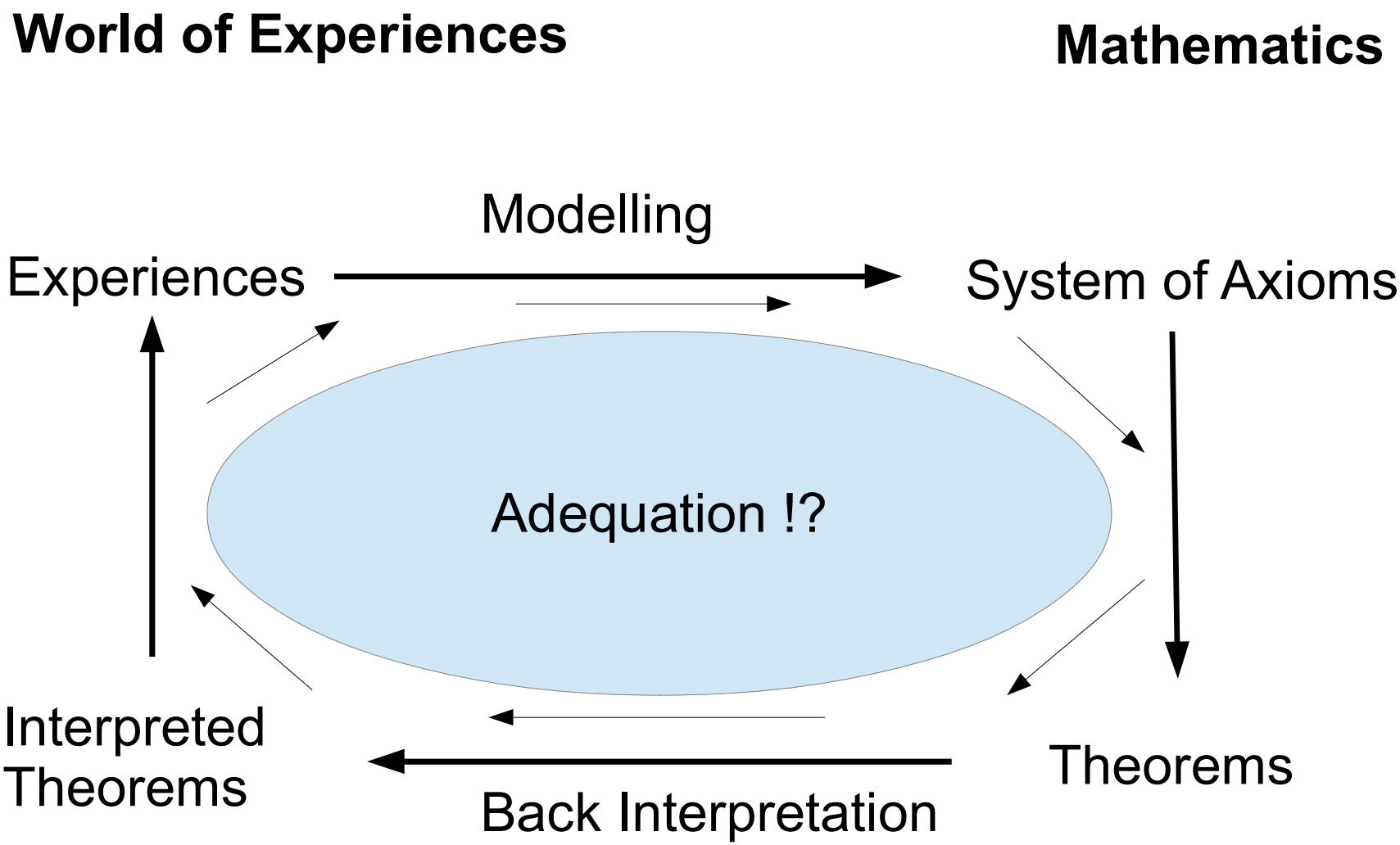
Find

optimal actions $\longrightarrow$: **utility/loss, preferences on consequences**

under uncertainty . . . $\longrightarrow$: **probability**

by optimally utilizing information from data available! $\longrightarrow$: **probability, data quality**

Modelling (e.g,, Behnen & Neuhaus (1984, Teubner))



Basic Model Classical Decision Theory

idealized model of a decision situation, fundamental assumptions

U The utility/losses involved are expressible on a cardinal (metrical) scale.

P All uncertainties involved can be described by classical (σ–additive) probabilities.

Modelling

Our conception determines our actions.

1

[1] Hauskatze_langhaar.jpg, Von Chatennoir–Eigenes Werk, CC0, https://commons.wikimedia.org/w/index.php?curid=14893649
Taken from https://de.wikipedia.org/wiki/Hauskatze, November 16th, 2022

[2] Creative Commons Attribution-Share Alike 4.0 International license; attribution: Charles James Sharp, from https://de.wikipedia.org/wiki/Tiger, November 16th, 2022

Modelling

Our conception determines our actions.

1

2

[1] Hauskatze_langhaar.jpg, Von Chatennoir–Eigenes Werk, CC0, https://commons.wikimedia.org/w/index.php?curid=14893649
Taken from https://de.wikipedia.org/wiki/Hauskatze, November 16th, 2022

[2] Creative Commons Attribution-Share Alike 4.0 International license; attribution: Charles James Sharp, from https://de.wikipedia.org/wiki/Tiger, November 16th, 2022

Modelling and Manski's Law of Decreasing Credibility

Credibility ?

"The credibility of inference decreases with the strength of the assumptions maintained." (Manski (2003, p. 1))

Charles Manski[3]

[3] http://faculty.wcas.northwestern.edu/~cfm754/; [Nov 17th, 2022]

Consequences

- Be careful when assumptions had just been made for mathematical convenience or "to get at least any solution"'!

Consequences

- Be careful when assumptions had just been made for mathematical convenience or "to get at least any solution"'!
- Develop frameworks that allow to reflect/extract the full information provided by weakly structured settings!

Consequences

- Be careful when assumptions had just been made for mathematical convenience or "to get at least any solution"'!
- Develop frameworks that allow to reflect/extract the full information provided by weakly structured settings!
- One important step in this direction: Set-based perspective: Consider the set of all classical models compatible with the available information (data and tenable assumptions properly derived from domain knowledge).

In Decision Theory

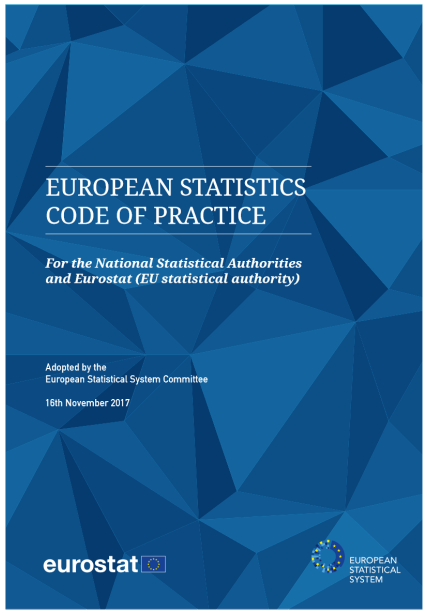
- More general utility representations, sets of potential utility functions
- imprecise probabilities $\sim$ sets of prior probabilities, sets of sampling models (for instance, neighborhood models)

In Decision Theory

- More general utility representations, sets of potential utility functions
- imprecise probabilities ~ sets of prior probabilities, sets of sampling models (for instance, neighborhood models)
- Criteria
  - Worst-case scenarios (Gamma-Maximin approach)
  - **choice-sets:** E-Admissibility (all potential Bayes actions), maximality (survivors of set-valued pairwise comparisons)

Table of contents

One Source: Special Quality Requirements of Official Statistics



4

- QF4SA: A quality framework for statistical algorithms: Yung et al. (2022, Statistical Journal of the IAOS)
- Workshop: Quality Aspects of Machine Learning – Official Statistics between Specific Quality Requirements and Methodological Innovation, September 6th to 8th, 2022, in Munich[5]

[4] https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142 [last access: November 117th, 2022]

[5] https://www.statistiknetzwerk.bayern.de/mam/themen/workshops/2022_cfp_en.pdf [last access: November 17th, 2022]

Example: The Epistemic Artificial Intelligence Project

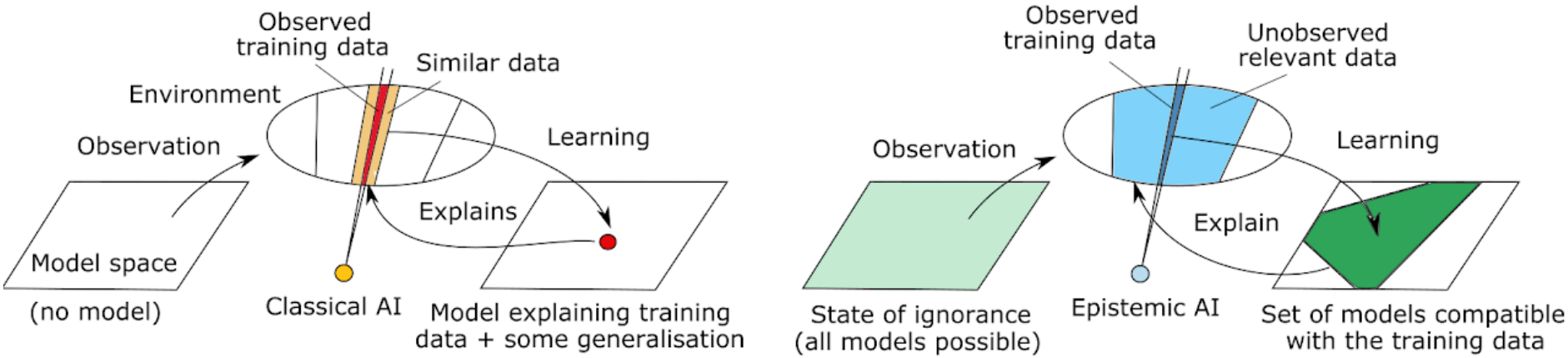
- Horizon 2020 project
- Here: AI taken as identical to Machine Learning
- Oxford Brookes, TU Delft, KU Leuven
- Fabio Cuzzolin: computer vision in autonomous driving
- From the project homepage:

  "Epistemic AI's overall objective is to create a new paradigm for a next-generation artificial intelligence providing worst-case guarantees on its predictions thanks to a proper modelling of real-world uncertainties.[...]

  Mathematically, Epistemic AI's principle translates into **seeking to learn sets of hypotheses compatible with the (scarce) data** *[bold: TA] available, rather than individual models. A set of models can provide, given new data, a robust set of predictions among which the most cautious one can be adopted, thus avoiding catastrophic results."*[6]

[6] `https://www.epistemic-ai.eu` [last access: November 17th, 2022]

The Epistemic Artificial Intelligence Project



"Illustration of the concept of epistemic artificial intelligence. Epistemic AI's notion of learning (right), as opposed to that of traditional machine learning/artificial intelligence (left)."[7]

[7] `https://www.epistemic-ai.eu` [last access: November 17th, 2022]

The Credal Point of View

- Levi (1980, MIT Press) epistemology, reasoning with sets of belief structures
- **Sets** of basic models

For instance: predict a **set** of classes, instead of a single one

The Credal Point of View

- Levi (1980, MIT Press) epistemology, reasoning with sets of belief structures

- **Sets** of basic models

For instance: predict a **set** of classes, instead of a single one

- still carries substantial information
- seize of set data-dependent: separate the clear cases (still single class prediction) from the hard cases (several classes highly plausible)
- efficient allocation of human expertise!?
- intuitively "clear": higher robustness of results
- may also enable results in situations where otherwise no automatic classification would have been performed at all

Machine Learning (Classification, for the Moment)

- **knowledge structure** $\mathfrak{k}$, fixed
  - features $X_1, \ldots, X_n$, $X_i \in \mathcal{X}$
  - targets $Y_1, \ldots, Y_n$, $Y_i \in \mathcal{Y}$, categorical for the moment

  - structural assumptions on the data (e.g. i.i.d.)
  - prior judgments
  - loss function

Machine Learning (Classification, for the Moment)

- **knowledge structure** $\mathfrak{k}$, fixed
  - features $X_1, \ldots, X_n$, $X_i \in \mathcal{X}$
  - targets $Y_1, \ldots, Y_n$, $Y_i \in \mathcal{Y}$, categorical for the moment
  - structural assumptions on the data (e.g. i.i.d.)
  - prior judgments
  - loss function
- Find optimal $\hat{f}$:

$$\hat{f} : \mathcal{X} \longrightarrow \mathcal{Y}$$

$$\Longleftrightarrow$$

Find for every feature value x the optimal class $y = \hat{f}(x)$

Machine Learning (Classification, for the Moment)

- **knowledge structure** $\mathfrak{k}$, fixed
  - features $X_1, \ldots, X_n$, $X_i \in \mathcal{X}$
  - targets $Y_1, \ldots, Y_n$, $Y_i \in \mathcal{Y}$, categorical for the moment
  - structural assumptions on the data (e.g. i.i.d.)
  - prior judgments
  - loss function

- Find optimal $\hat{F}$:

$$\hat{F} : \mathcal{X} \longrightarrow \not{\mathcal{Y}} \; \mathcal{P}(\mathcal{Y})$$

$$\Longleftrightarrow$$

Find for every feature value x the optimal ~~class~~ set of classes $\mathcal{S} = \hat{F}(x)$

Optimal Sets of Classes

- sets of classes with guaranteed coverage: **conformal prediction**, Vovk & Shafer (2008, JMLR); valid inference models, Stella & Martin (2022, IJAR)

Optimal Sets of Classes

- sets of classes with guaranteed coverage: **conformal prediction**, Vovk & Shafer (2008, JMLR); valid inference models, Stella & Martin (2022, IJAR)

- optimal accuracy under **penalization** of seize of the set: e.g., Yang, Destercke & Masson (2017, IEEE Trans Cybernetics),

Optimal Sets of Classes

- sets of classes with guaranteed coverage: **conformal prediction**, Vovk & Shafer (2008, JMLR); valid inference models, Stella & Martin (2022, IJAR)
- optimal accuracy under **penalization** of seize of the set: e.g., Yang, Destercke & Masson (2017, IEEE Trans Cybernetics),
- **sets** of knowledge structure corresponding to a concrete problem: **credal knowledge structures**

Credal Knowledge Structures

Often, a concrete problem corresponds to a **set** $\mathfrak{K}$ of knowledge structures $\mathfrak{k}$

- **knowledge structure** $\mathfrak{k}$, fixed
  - features $X_1, \ldots, X_n$, $X_i \in \mathcal{X}$
  - targets $Y_1, \ldots, Y_n$, $Y_i \in \mathcal{Y}$,

  - structural assumptions on the data (e.g. i.i.d.)
  - prior judgments
  - loss function

  - Find optimal $\hat{f}_{\mathfrak{k}}$:

$$\hat{f}_{\mathfrak{k}} : \mathcal{X} \longrightarrow \mathcal{Y}$$

$$\Longleftrightarrow$$

Find for every feature value x the optimal prediction $y = \hat{f}_{\mathfrak{k}}(x)$

A Natural Construction of Credal Procedures

- Collect all optimal solutions $\hat{f}_{\mathfrak{k}}$ over $\mathfrak{K}$

$$\hat{F}_{\mathfrak{K}} := \left\{ \hat{f}_{\mathfrak{k}} \,\middle|\, \hat{f}_{\mathfrak{k}} \text{ is optimal for at least one } \mathfrak{k} \in \mathfrak{K} \right\}$$

- set-valued prediction for concrete feature value x

$$\hat{F}_{\mathfrak{K}}(x) := \left\{ \hat{f}_{\mathfrak{k}}(x) \,\middle|\, \hat{f}_{\mathfrak{k}} \in \hat{F}_{\mathfrak{K}} \right\}$$

- classification: $|\hat{F}_{\mathfrak{K}}(x)|$ measure of stability of learning?
- if, and only if, $\hat{F}_{\mathfrak{K}}(x)$ singleton: trustworthy precision
- Note that any prediction $\mathcal{S} \subsetneq \hat{F}_{\mathfrak{K}}(x)$, i.e. that is more concise than $\hat{F}_{\mathfrak{K}}(x)$, is not supported by the credal knowledge structure. It would need additional justification.

Credal Knowledge Structures: Sets of Loss Fcts / Priors

- consider several loss functions
- sets of prior distributions
  - credal classification from generalized Bayesian networks: IPG@IDSIA, e.g., for software, Cabañas, Antonucci, A., Huber & Zaffalon, M. (2020, PGM); review: Mauá & Cozman (2020, IJAR)
  - credal discriminant analysis: Carranza Alarcón & Destercke (2021, Pattern Recognition)

Credal Knowledge Structures: Neighborhood Models

- Utilize neighborhood models from robust statistics?

- evaluation under neighborhood of the empirical distribution
- first steps in that direction: work following Abellan & Moral (2003, Int.J.Gen.Systems)
  - imprecise classification relying on the so-called IDM: contamination neighborhood
  - minimax entropy approach
- distributionally robust stochastic optimization!? e.g. Gao & Kleywegt (2022, Mathematics of Operation Research

Credal Information Structure Induced by 'Partial Data'

- A particular important type of credal information structures arises from **'partial data'**, i.e. situations where the data at hand defacto describe a whole **set of potential data**. Typical examples include

Credal Information Structure Induced by 'Partial Data'

- A particular important type of credal information structures arises from **'partial data'**, i.e. situations where the data at hand defacto describe a whole **set of potential data**. Typical examples include
- coarsened data
  - coarse data by indecisiveness between categories (e.g. Kreiß & Nalenz & Augustin (2020, SUM 2020); Rodemann, Kreiß, Hüllermeier & Augustin (2022, SUM))
  - coarsened / rounded by questionnaire design (e.g. PASS data: Trappmann et al (2010, Schmollers JB); Plass et al (2019, IntStatRev))
  - ex-post coarsened / rounded for reasons of data protection
  - nonrandomly missing data !?

Credal Information Structure Induced by 'Partial Data'

- A particular important type of credal information structures arises from **'partial data'**, i.e. situations where the data at hand defacto describe a whole **set of potential data**. Typical examples include
- coarsened data
  - coarse data by indecisiveness between categories (e.g. Kreiß & Nalenz & Augustin (2020, SUM 2020); Rodemann, Kreiß, Hüllermeier & Augustin (2022, SUM))
  - coarsened / rounded by questionnaire design (e.g. PASS data: Trappmann et al (2010, Schmollers JB); Plass et al (2019, IntStatRev))
  - ex-post coarsened / rounded for reasons of data protection
  - nonrandomly missing data !?
- micro-aggregated data

Credal Information Structure Induced by 'Partial Data'

- A particular important type of credal information structures arises from **'partial data'**, i.e. situations where the data at hand defacto describe a whole **set of potential data**. Typical examples include
- coarsened data
  - coarse data by indecisiveness between categories (e.g. Kreiß & Nalenz & Augustin (2020, SUM 2020); Rodemann, Kreiß, Hüllermeier & Augustin (2022, SUM))
  - coarsened / rounded by questionnaire design (e.g. PASS data: Trappmann et al (2010, Schmollers JB); Plass et al (2019, IntStatRev))
  - ex-post coarsened / rounded for reasons of data protection
  - nonrandomly missing data !?
- micro-aggregated data
- yet to be combined data sets: statistical matching and record linkage

Statistical Matching and Record Linkage as Partial Data

- statistical matching (D'Orazio, Di Zio & Scanu, Mauro (2006, Wiley)), combine data sets
  - with partially overlapping variables
  - disjoint sets of units

  - micro (data set) and macro approach (joint distribution) under-identified
  - except under exact specification of the (unoberservable!) correlation structure
  - CIA: conditional independence assumptions
  - partial identification in the macro approach: Di Zio & Vantaggi (2017, IJAR)
  - imprecise imputation for the micro approach: Endres, Fink & Augustin (2019, JOffStat)

Statistical Matching and Record Linkage as Partial Data

- statistical matching
- record linkage, combine data sets
  - with disjoint sets of [major] variables
  - overlapping/identical sets of units
  - under-identified under weak/missing identifier
  - for many units several potential matches

Some Major Aspects of the Technical Handling of Credal Knowledge Structures I

- values x_i and y_i induce sets $\mathfrak{x}_i$ and $\mathfrak{y}_i$ of values, potentially under further constraints
- resulting sets $\dot{\mathfrak{X}}$ and $\dot{\mathfrak{Y}}$
- every pair in $(\dot{x}, \dot{y}) \in \dot{\mathfrak{X}} \times \dot{\mathfrak{Y}}$ is a "potential data set"

Some Major Aspects of the Technical Handling of Credal Knowledge Structures I

- values x_i and y_i induce sets $\mathfrak{x}_i$ and $\mathfrak{y}_i$ of values, potentially under further constraints
- resulting sets $\dot{\mathfrak{X}}$ and $\dot{\mathfrak{Y}}$
- every pair in $(\dot{x}, \dot{y}) \in \dot{\mathfrak{X}} \times \dot{\mathfrak{Y}}$ is a "potential data set"
- Given a certain model structure (e.g., GLM), find for every potential data set the optimal procedure.

Some Major Aspects of the Technical Handling of Credal Knowledge Structures I

- values x_i and y_i induce sets $\mathfrak{x}_i$ and $\mathfrak{y}_i$ of values, potentially under further constraints
- resulting sets $\dot{\mathfrak{X}}$ and $\dot{\mathfrak{Y}}$
- every pair in $(\dot{x}, \dot{y}) \in \dot{\mathfrak{X}} \times \dot{\mathfrak{Y}}$ is a "potential data set"
- Given a certain model structure (e.g., GLM), find for every potential data set the optimal procedure.
- For instance, the set $\hat{\Theta}$ of all maximum likelihood estimators can be characterized as the set of all zeros of the corresponding score function $\dot{\psi}()$:

$$\hat{\Theta} = \left\{ \hat{\vartheta}(\dot{x}, \dot{y}) \,\middle|\, \dot{\psi}(\hat{\vartheta}(\dot{x}, \dot{y}), \dot{x}, \dot{y}) = 0 \right\}$$

Some Major Aspects of the Technical Handling of Credal Knowledge Structures II

Characterize $\hat{\Theta}$ by maxima/minima (of linear combinations) of its components $\vartheta[\ell]$, e.g.

$$\vartheta[\ell] \rightarrow \min_{(\dot{x},\dot{y})}$$

subject to

$$\psi(\hat{\vartheta}(\dot{x},\dot{y}),\dot{x},\dot{y}) = 0$$

$$(\dot{x},\dot{y}) \in \dot{\mathfrak{X}} \times \dot{\mathfrak{Y}}$$

Some Major Aspects of the Technical Handling of Credal Knowledge Structures II

Characterize $\hat{\Theta}$ by maxima/minima (of linear combinations) of its components $\vartheta[\ell]$, e.g.

$$\vartheta[\ell] \to \min_{(\dot{x},\dot{y})}$$

subject to

$$\psi(\hat{\vartheta}(\dot{x},\dot{y}),\dot{x},\dot{y}) = 0$$

$$(\dot{x},\dot{y}) \in \dot{\mathfrak{X}} \times \dot{\mathfrak{Y}}$$

- For rounding and many microaggregation procedures, the set $\dot{\mathfrak{X}} \times \dot{\mathfrak{Y}}$ is a convex polyhedron.
- Substantial simplifications for special situations (GLM, linear models; in particular when only the dependent variable is affected) is possible.
- Transfer of situations with not connected values, like matching, record linkage: use Boolean variables as weights

Concluding Remarks of the General Part

- credal view opens new possibilities
- comprehensive reflection of what is (justifiably assumed to be) known
- technical treatment by applying, and extending, techniques for handling credal sets in statistics and decision theory

Table of contents

This Section is Mainly Based on

Christoph Jansen, Malte Nalenz, Georg Schollmeyer and Thomas Augustin
Statistical Comparisons of Classifiers by Generalized Stochastic Dominance, Arxiv, `https://arxiv.org/abs/2209.01857`

SCAN ME

[8]

[8] Photo Christoph Jansen (left, private) and Malte Nalenz (`https://avatars.githubusercontent.com/u/22354882?v=4`)

Table of contents

Comparing Classifiers

accuracy, *area under the curve* and *Brier score*

- $\mathcal{C}$ set of classifiers
- $\mathcal{D}$ set of data sets
- $\phi_1, \ldots, \phi_n : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$ criteria to measure the goodness of classification, ordinal / metrical, like accuracy, area under the curve, Brier score

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

Challenges Comparing Classifiers

(At least) three fundamental fundamental challenges

- multiplicity of quality criteria
- multiplicity of data sets
- and the randomness / arbitrariness of the selected data sets.

Idea: Choice Theory, Preference Aggregation

- Arrow's impossibility theorem Arrow, 1958: Any aggregation satisfying a Pareto and an independence condition can be shown to be dictatorial: implicitly, only one of the quality criteria is considered.
- This includes the Borda's rule (Borda (1781)), in which classifiers are compared on the basis of their average ranks with respect to the various quality criteria.

- Outside Arrow's framework: Condorcet rule (Condorcet (1785)): ranking between any pair of classifiers is derived by counting which one performs better with respect to more quality criteria may produce potentially intransitive and – even worse – cyclic rankings.

Table of contents

Preference Systems: $\mathcal{A} = [A, R_1, R_2]$

- Prior work: Jansen, Schollmeyer, Augustin (2018, IJAR) decision theoretic, Jansen, Blocher, Augustin, Schollmeyer (2022): design efficient elicitation strategies

- Expressing locally varying scale of measurement
- Two relations
  - R_1 ordinal part: pre-order
  - R_2 potentially in addition a cardinal part: pre-order on R_1, locally measuring strength of order

Originally behavioristic interpretation:

- $(a, b) \in R_1$: a is at least as desirable as b
- $((a, b), (c, d)) \in R_2$: exchanging b by a is at least as desirable as exchanging d by c

Later in addition, regularization parameter δ ('granularity'): least notable difference.

Definition: Consistency

Regulating the interaction between R_1 and R_2
The preference system $\mathcal{A} = [A, R_1, R_2]$ is **consistent** if there exists a function $u : A \to [0, 1]$ such that for all $a, b, c, d \in A$ we have:

i) If $(a, b) \in R_1$, then $u(a) \geq u(b)$ with = iff $(a, b) \in I_{R_1}$ (indifference part).

ii) If $((a, b), (c, d)) \in R_2$, then $u(a) - u(b) \geq u(c) - u(d)$ with = iff $((a, b), (c, d)) \in I_{R_2}$.

The set of all such **representations** u satisfying i) and ii) is denoted by $\mathcal{U}_{\mathcal{A}}$.
Analogous definitions for $\delta > 0$

Generalizing the Choice Function, $(\mathcal{A}, \mathcal{M}, \delta)$-dominance

Theory for optimal decision making based on the sets $\mathcal{U}_{\mathcal{A}}$ and a credal set $\mathcal{M}$ and corresponding as well as efficient computation from Jansen, Schollmeyer, Augustin (2018, IJAR). Here focus on $(\mathcal{A}, \mathcal{M}, \delta)$-dominance
Let $\mathcal{A} = [A, R_1, R_2]$ be δ-consistent and $\mathcal{M}$ a credal set on $(S, \sigma(S))$.
For measurable X, Y, we say that Y is $(\mathcal{A}, \mathcal{M}, \delta)$-**dominated** by X if

$$\mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$$

for all $u \in \mathcal{N}_{\mathcal{A}}^{\delta}$ and $\pi \in \mathcal{M}$.

Denote the induced relation by $\geq_{(\mathcal{A},\mathcal{M},\delta)}$.

Prominent Special cases of $\geq_{(\mathcal{A},\mathcal{M},\delta)}$

- $\mathcal{M} = \{\pi\}$ and $R_2 = \varnothing$

  $\rightarrow$ Reduction to (generalized first-order) **stochastic dominance**
  Mosler & Scarsini (1991)

- $\mathcal{M} = \{\pi\}$ and R_1 and R_2 guaranteeing utility unique up to plts

  $\rightarrow$ Reduction to comparing expected utilities

- R_1 and R_2 guaranteeing utility unique up to plts

  $\rightarrow$ Reduction to **Bewley dominance**

Checking for $\geq_{(\mathcal{A},\mathcal{M},\delta)}$

Can be done by linear optimization as long as the credal set $\mathcal{M}$ is finitely generated (convex polyhedron, set of extreme points)

Assumptions:

- All Q_i are of at least ordinal scale with preference order $\geq_i$.
- All Q_i possess minimal and maximal elements w.r.t. $\geq_i$.
- $(Q_j)_{j\leq k}$, where $k \leq n$, are of metric scale with metric $d_i : Q_i \times Q_i \to \mathbb{R}$.

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

Defining the Preference System

We define a preference system on the set of all quality vectors:

Ordinal part:

$$R_1 := \Big\{ (q,p) \in \mathcal{Q} \times \mathcal{Q} : q_i \geq_i p_i \text{ for all } i = 1, \ldots, n \Big\}$$

Cardinal (metric) part:

$$R_2 := \Big\{ ((q,p),(r,s)) \in R_1 \times R_1 : d_i(q_i,p_i) \geq d_i(r_i,s_i) \text{ for all } i = 1, \ldots, k \Big\}$$

Induced preference system:

$$\mathbb{C} = [\mathcal{Q}, R_1, R_2]$$

The Criterion of δ-Dominance

We can now transfer the decision criterion from before to our specific setting.

For that, assume the law π generating the data sets from $\mathcal{D}$ to be known.
δ-Dominance (theoretical version) Let $\mathbb{C}$ be δ-consistent and $\mathcal{C}$ be such that $\{\phi(C,\cdot) : C \in \mathcal{C}\} \subseteq \mathcal{F}_{(\mathbb{C},\mathcal{D})}$.

Call C_j δ-**dominated** by C_i, if $\phi(C_j,\cdot)$ is $(\mathbb{C},\{\pi\},\delta)$-dominated by $\phi(C_i,\cdot)$.

Denote the induced binary relation by $\gtrsim_\delta$.
Challenge: The true law π on the and the set $\mathcal{D}$ will often be inaccessible and we will only have an i.i.d. sample $D_1,\ldots,D_s \sim \pi$ of data sets from $\mathcal{D}$.
δ-Dominance (empirical version) Replace $\mathcal{D}$ by $\hat{\mathcal{D}}_s := \{D_1,\ldots,D_s\}$ and π by the empirical law $\hat{\pi}$.

We call C_j δ-**dominated (in sample)** by C_i, if $\phi(C_j,\cdot)$ is $(\mathbb{C},\{\hat{\pi}\},\delta)$-dominated by $\phi(C_i,\cdot)$. Denote the induced binary relation by $\gtrsim_\delta$ (sloppy!).

Checking for (in-sample) δ-Dominance

We can adapt our algorithm for checking (in-sample) δ-dominance.

Wlog: $\phi(\mathcal{C} \times \hat{\mathcal{D}}_s) = \{q_1, \ldots, q_d\}$ s.t. q_1 and q_2 min and max w.r.t. R_1.
Corollary For $C_i, C_j \in \mathcal{C}$, we consider the linear programming problem

$$\sum_{\ell=1}^{d} v_\ell \cdot [\hat{\pi}(\phi(C_i, \cdot)^{-1}(\{q_\ell\})) - \hat{\pi}(\phi(C_j, \cdot)^{-1}(\{q_\ell\}))] \longrightarrow \min_{(v_1,\ldots,v_d)\in\mathbb{R}^d}$$

with constraints $(v_1, \ldots, v_d) \in \nabla^{\delta}_{\mathbb{C}}$.

Denote by opt_{ij} the optimal value of this programming problem.

It then holds:

$$C_i \gtrsim_\delta C_j \quad \Leftrightarrow \quad opt_{ij} \geq 0.$$

Table of contents

Application Example: Setup

- We use 16 binary classification benchmark data sets all taken from the UCI machine learning repository.

Application Example: Setup

- We use 16 binary classification benchmark data sets all taken from the UCI machine learning repository.
- For classifier comparison, we consider **accuracy**, **AUC** and **Brier score**.

Application Example: Setup

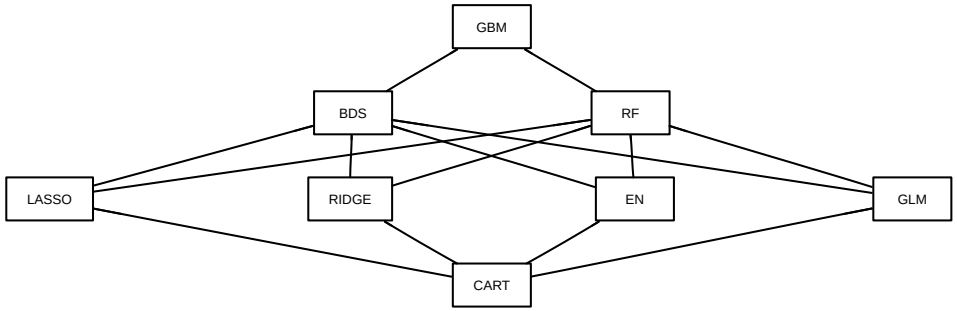
- We use 16 binary classification benchmark data sets all taken from the UCI machine learning repository.
- For classifier comparison, we consider **accuracy**, **AUC** and **Brier score**.
- We compare the algorithms
  - Classification and regression trees (**CART**)
  - Random forests (**RF**)
  - Gradient boosted trees (**GBM**)
  - Boosted decision stumps (**BDS**)
  - Generalized linear models (**GLM**)
  - Lasso regression (**LASSO**)
  - Elastic net (**EN**)
  - Ridge regression (**RIDGE**)

Application Example: Setup

- We use 16 binary classification benchmark data sets all taken from the UCI machine learning repository.
- For classifier comparison, we consider **accuracy**, **AUC** and **Brier score**.
- We compare the algorithms
  - Classification and regression trees (**CART**)
  - Random forests (**RF**)
  - Gradient boosted trees (**GBM**)
  - Boosted decision stumps (**BDS**)
  - Generalized linear models (**GLM**)
  - Lasso regression (**LASSO**)
  - Elastic net (**EN**)
  - Ridge regression (**RIDGE**)
- All three criteria are assumed to be metric.

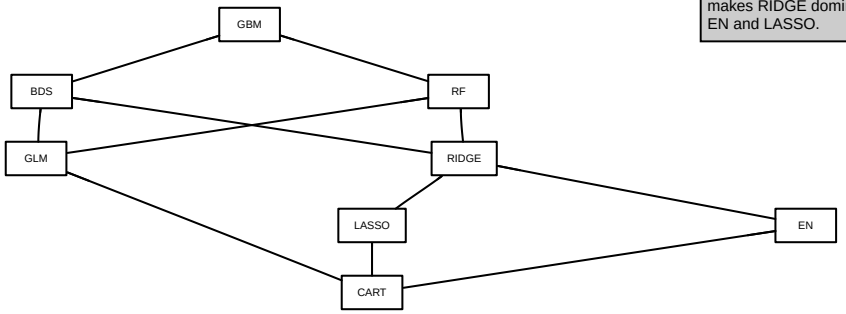
Application Example: Results

$\delta = 0$



$\delta = 0.004$

Raising the threshold to 0.004 makes RIDGE dominate both EN and LASSO.



$\delta = 0.0077$

Raising the threshold to 0.0077 makes GLM dominate RIDGE, EN and LASSO.

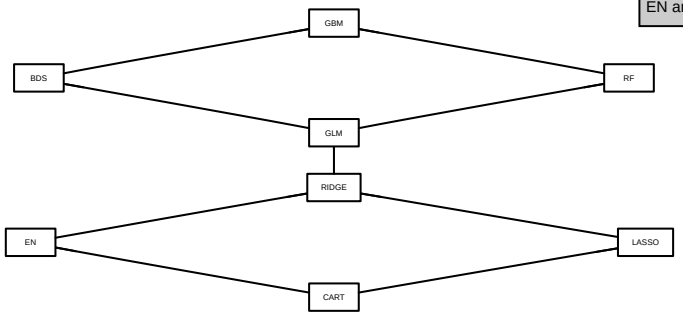


Table of contents

Discussion: How to Address Additionally the Statistical Uncertainty

- **Good news:** In-sample δ-Dominance resolves the problems appearing at the Levels 1 and 2 at the same time.
- **Bad news:** Level 3 is still a problem, i.e., changing the sample of data sets will, in general, change the order among the classifiers!

- **Idea:** Construct a statistical test for checking whether in-sample orderings are statistically significant. Use opt_{ij} as a test statistic for a test with the null hypothesis

$$H_0 : C_j \gtrsim_\delta C_i$$

  Reject H_0 if this value is larger than a critical value c.
- **Challenge:** The distribution of opt_{ij} cannot be analyzed straightforwardly.
- **Solution:** Use a two-sample observation-randomization test (permutation-based, non-parametric) instead.

Resampling Scheme Underlying the Permutation Test

The procedure for evaluating opt_{ij} has the following five steps:

Step 1: Produce two separate samples $(x_1, \ldots, x_s)$ and $(y_1, \ldots, y_s)$, where $x_l := \phi(C_i, D_l)$ and $y_l := \phi(C_j, D_l)$.

Step 2: Take the pooled sample $z = (x_1, \ldots, x_s, y_1, \ldots, y_s)$.

Step 3: Take all $I \subseteq \{1, \ldots, 2s\}$ of size s and compute opt_{ij}^I for the permuted data $(z_i)_{i \in I}$ and $(z_i)_{i \in \{1,\ldots,2s\} \setminus I}$.

Step 4: Sort all opt_{ij}^I in increasing order.

Step 5: Reject H_0 if opt_{ij} is greater than the $\lceil (1-\alpha) \cdot \binom{2s}{s} \rceil$-th value of the increasingly ordered values opt_{ij}^I, where α is the confidence level.

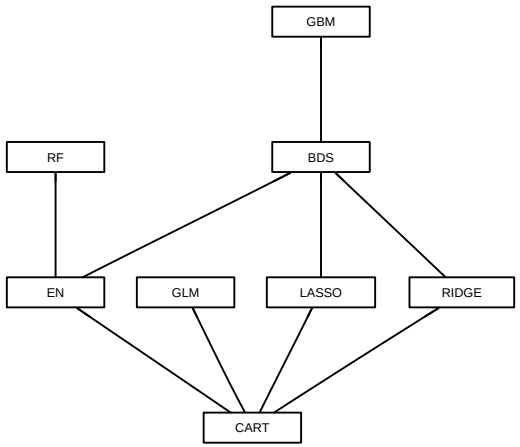
If $\binom{2s}{s}$ is too large, one can alternatively compute opt_{ij}^I only for a large enough number N of randomly drawn index sets I.

Application Example: Results for Tests

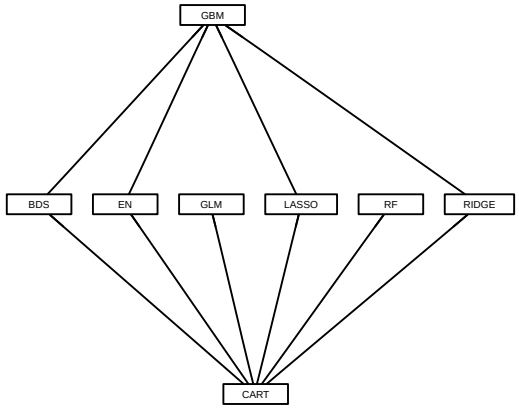
Results of the resample tests with $\delta = 10^{-5}$ and $N = 1000$ for all binary comparisons. A line symbolizes a value strictly below 0.95.

	BDS	CART	EN	GBM	GLM	LASSO	RF	RIDGE
BDS	–	1.000	0.976	–	–	0.967	–	0.951
CART	–	–	–	–	–	–	–	–
EN	–	0.998	–	–	–	–	–	–
GBM	0.998	1.000	0.998	–	–	0.999	–	0.997
GLM	–	1.000	–	–	–	–	–	–
LASSO	–	0.997	–	–	–	–	–	–
RF	–	1.000	0.953	–	–	–	–	–
RIDGE	–	0.999	–	–	–	–	–	–

Significant orders:



without correction for multiple testing



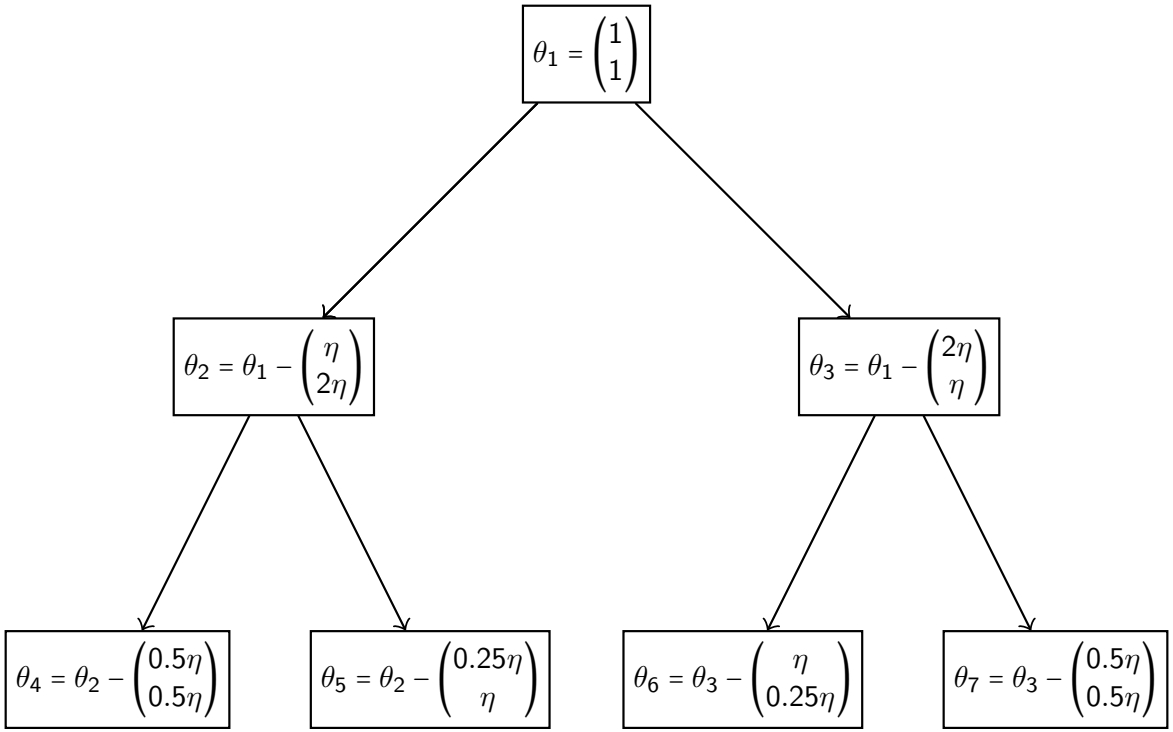
with correction for multiple testing

Table of contents

Simulation:Setup

Seven simulated classifiers $C_1, \ldots, C_7$ with expected performance $\theta_i \in [0,1]^2$ on two two cardinal quality criteria are compared.

Groundtruth:



Performances x_{ij} of C_i on data set D_j are i.i.d. drawn from a normal distribution, i.e., $x_{ij} \sim \mathcal{N}_2(\theta_i, \Sigma_\epsilon)$, where $\Sigma_\epsilon = \sigma_\epsilon I$ and σ_ϵ rules the extent of a noise term.

Simulation: Competitors

Demsar (2006, JMLR) proposed a test for systematical differences between classifiers w.r.t. *one single* quality criterion.

We add two multi-dimensional adaptations of this test to our study:

all-test: Classifier C_i is considered better than C_j if it performs significantly better on each quality criterion w.r.t. the above test.

one-test: C_i is better than C_j if C_i performs significantly better in at least one dimension and if the converse is not true for any other dimension.

Moreover, we add our proposed test for $\delta = 0$ and $\delta = 10^{-5}$.

Question: Which of the tests performs best in significantly revealing the true ordering structure?

Simulation: Results (Bonferroni corrected)

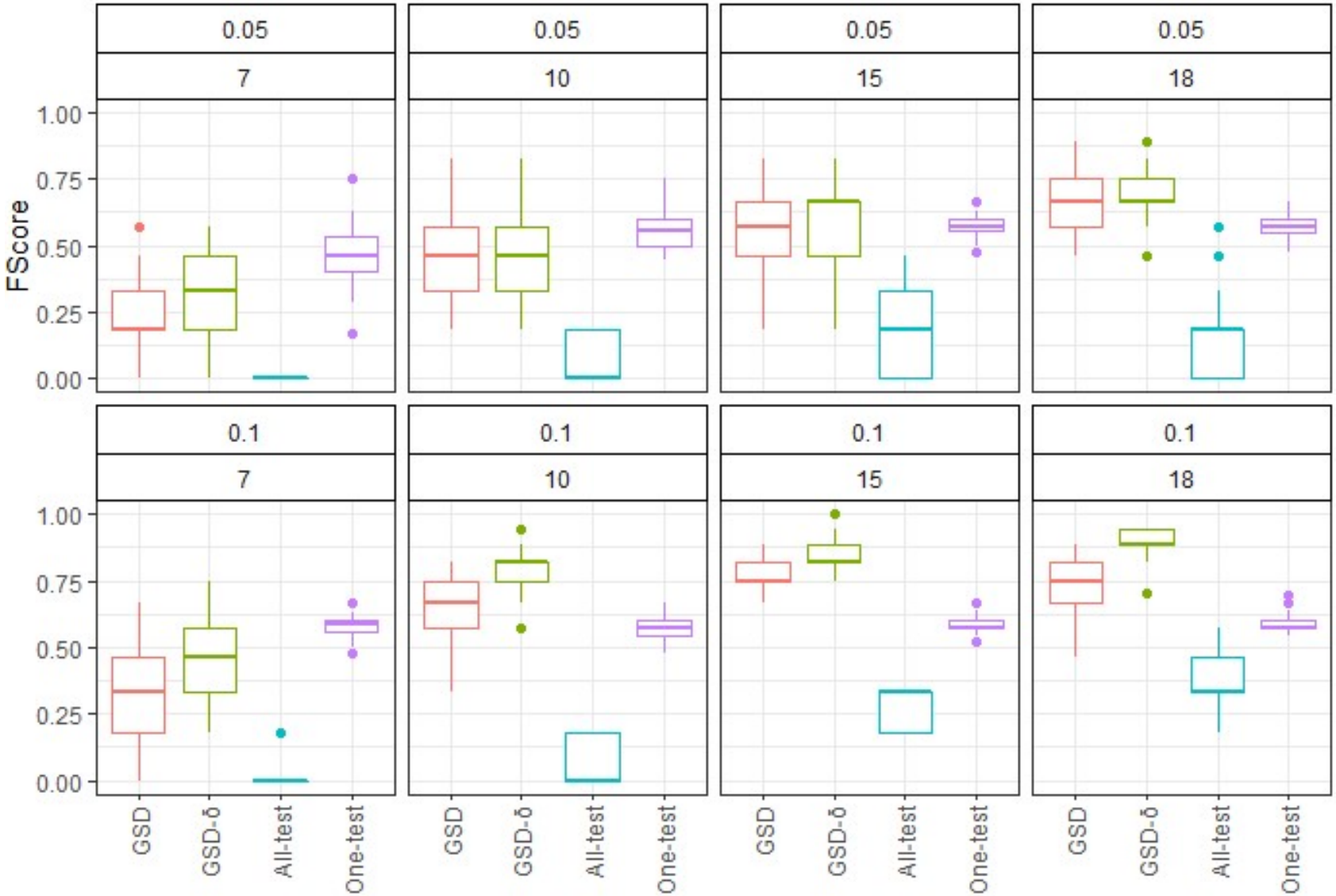


Table of contents

Concluding Remarks

- Some speculative ideas on optimality of algorithms and their construction (loss function externally given)
- Evaluation of benchmark study of given algorithms: formulate preferences under multiplicity of algorithms and evaluation criteria (multi-criteria decision making; preference system)

augustin@stat.uni-muenchen.de