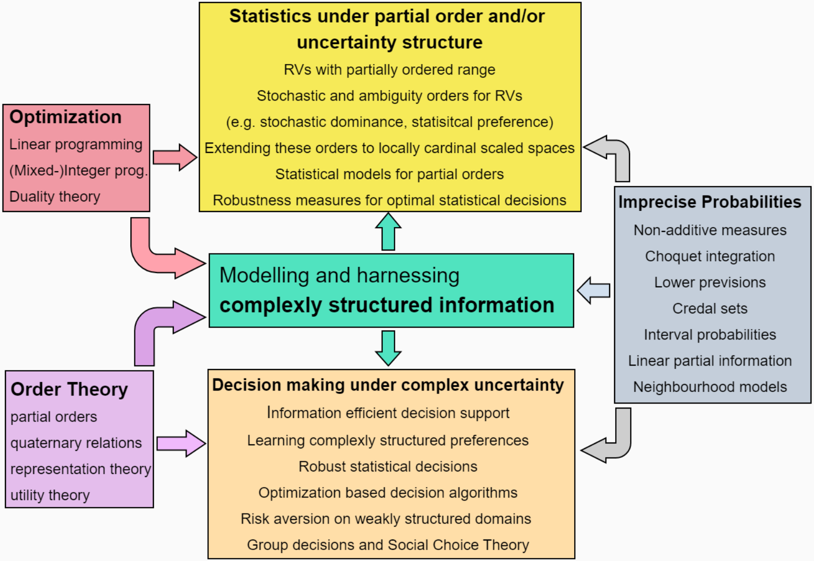
Decision Making under Complex Information

with Applications to Statistics and Machine Learning

Christoph Jansen
Department of Statistics, LMU Munich

Institutskolloquium, 2022/10/19

What to expect



Outline

Classical Decision Theory

- probability π on the states
- cardinal utility u on the consequences
- evaluation by expected utility w.r.t. u and π

PART I

Weakly Structured Information

$$\pi \longrightarrow \mathcal{M}$$

$$u \longrightarrow \mathcal{A} = [A, R_1, R_2] \longrightarrow \mathcal{U}$$

Project I:

Efficient elicitation of

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

and consequences for decision making.

PART II

Project II:

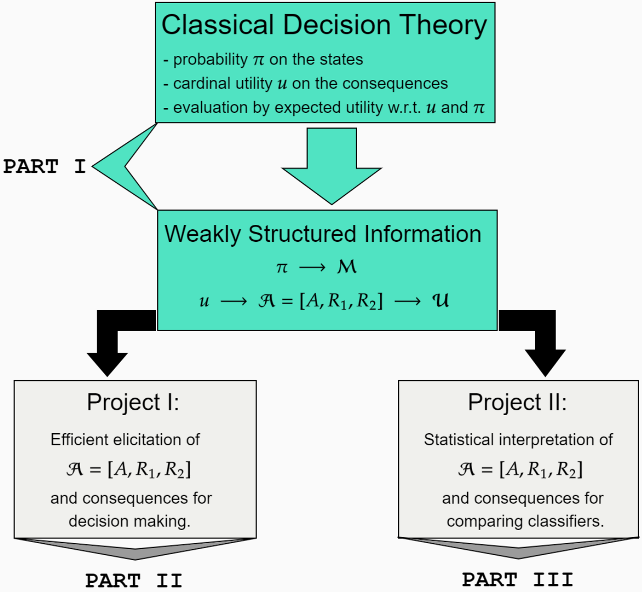
Statistical interpretation of

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

and consequences for comparing classifiers.

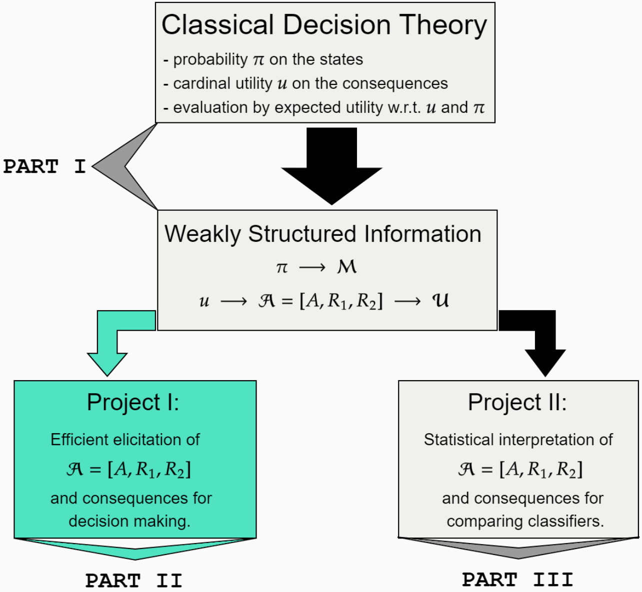
PART III

Outline

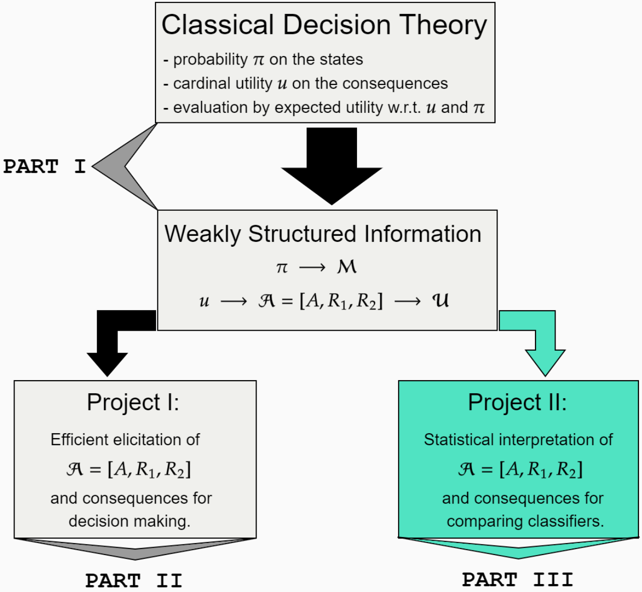


Classical Decision Theory
- probability π on the states
- cardinal utility u on the consequences
- evaluation by expected utility w.r.t. u and π
PART I
Weakly Structured Information
$\pi \longrightarrow \mathcal{M}$
$u \longrightarrow \mathcal{A} = [A, R_1, R_2] \longrightarrow \mathcal{U}$
Project I:
Efficient elicitation of
$\mathcal{A} = [A, R_1, R_2]$
and consequences for decision making.
PART II
Project II:
Statistical interpretation of
$\mathcal{A} = [A, R_1, R_2]$
and consequences for comparing classifiers.
PART III

Outline


Classical Decision Theory
- probability π on the states
- cardinal utility u on the consequences
- evaluation by expected utility w.r.t. u and π
PART I
Weakly Structured Information
$\pi \longrightarrow \mathcal{M}$
$u \longrightarrow \mathcal{A} = [A, R_1, R_2] \longrightarrow \mathcal{U}$
Project I:
Efficient elicitation of
$\mathcal{A} = [A, R_1, R_2]$
and consequences for
decision making.
Project II:
Statistical interpretation of
$\mathcal{A} = [A, R_1, R_2]$
and consequences for
comparing classifiers.
PART II
PART III

Outline


Classical Decision Theory
- probability π on the states
- cardinal utility u on the consequences
- evaluation by expected utility w.r.t. u and π
PART I
Weakly Structured Information
$\pi \longrightarrow \mathcal{M}$
$u \longrightarrow \mathcal{A} = [A, R_1, R_2] \longrightarrow \mathcal{U}$
Project I:
Efficient elicitation of
$\mathcal{A} = [A, R_1, R_2]$
and consequences for decision making.
PART II
Project II:
Statistical interpretation of
$\mathcal{A} = [A, R_1, R_2]$
and consequences for comparing classifiers.
PART III

Decision Theory in a Nutshell

Classical Decision Theory

Informal description of the model:

- An agent has to choose among different acts X from a set $\mathcal{G}$.
- The consequence that choosing X yields depends on which state of nature s from a set S is the true one.

Classical Decision Theory

Informal description of the model:

- An agent has to choose among different acts X from a set $\mathcal{G}$.
- The consequence that choosing X yields depends on which state of nature s from a set S is the true one.

Formal description of the model:

- Let A denote some non-empty set of consequences.
- Each act X corresponds to a mapping $X : S \to A$.
- The set $\mathcal{G}$ is a subset of $A^S = \{X : S \to A\}$.

Classical Decision Theory

Informal description of the model:

- An agent has to choose among different acts X from a set $\mathcal{G}$.
- The consequence that choosing X yields depends on which state of nature s from a set S is the true one.

Formal description of the model:

- Let A denote some non-empty set of consequences.
- Each act X corresponds to a mapping $X : S \to A$.
- The set $\mathcal{G}$ is a subset of $A^S = \{X : S \to A\}$.

Goal: Determining a choice function

$$ch : 2^{\mathcal{G}} \to 2^{\mathcal{G}} \text{ with } ch(\mathcal{D}) \subseteq \mathcal{D} \text{ for all } \mathcal{D} \in 2^{\mathcal{G}}$$

that best possibly utilizes the available information.

Statistical Decision Theory as a Special Case

Additional information: Data $\mathbf{Z} : \Omega \to \mathcal{Z}$ with $\mathbf{Z} \sim P_s$ given that $s \in S$ is the true state, i.e. S parametrizes our model.

Statistical Decision Theory as a Special Case

Additional information: Data $\mathbf{Z} : \Omega \to \mathcal{Z}$ with $\mathbf{Z} \sim P_s$ given that $s \in S$ is the true state, i.e. S parametrizes our model.

Induced Statistical Decision Problem:

- Instead of directly choosing acts X from $\mathcal{G}$, we now consider decision functions $d : \mathcal{Z} \to \mathcal{G}$ from a suitable $\mathbb{D} \subset \mathcal{G}^{\mathcal{Z}}$.

- The choice of $d \in \mathbb{D}$ under $s \in S$ (i.e. $\mathbf{Z} \sim P_s$) is then evaluated by an element $C(d, P_s) \in A^*$ using the distribution information.

- Every $d \in \mathbb{D}$ can then be identified with a mapping

$$X_d : S \to A^* \quad , \quad s \mapsto C(d, P_s)$$

  yielding again a data-free decision problem $\mathcal{G}^* = \{X_d : d \in \mathbb{D}\}$.

Statistical Decision Theory as a Special Case

Additional information: Data $\mathbf{Z} : \Omega \to \mathcal{Z}$ with $\mathbf{Z} \sim P_s$ given that $s \in S$ is the true state, i.e. S parametrizes our model.

Induced Statistical Decision Problem:

- Instead of directly choosing acts X from $\mathcal{G}$, we now consider decision functions $d : \mathcal{Z} \to \mathcal{G}$ from a suitable $\mathbb{D} \subset \mathcal{G}^{\mathcal{Z}}$.
- The choice of $d \in \mathbb{D}$ under $s \in S$ (i.e. $\mathbf{Z} \sim P_s$) is then evaluated by an element $C(d, P_s) \in A^*$ using the distribution information.
- Every $d \in \mathbb{D}$ can then be identified with a mapping
$$X_d : S \to A^* \quad , \quad s \mapsto C(d, P_s)$$
yielding again a data-free decision problem $\mathcal{G}^* = \{X_d : d \in \mathbb{D}\}$.

Choice function: Use elements $[C(d, s)]_{d,s}$ to construct a choice function that selects optimal decision functions (tests, estimators, classifiers,...).

Constructing Choice Functions for Decision Making

Classical assumptions: (e.g., [von Neumann et al., 1944, Savage, 1954]))

(I) The agent's preferences among the elements of A are characterized by a cardinal utility function $u : A \to \mathbb{R}$.

(II) The uncertainty among the states from S is described by some classical probability measure π.

Under (I) and (II), there is strong consensus for comparing acts X and Y by comparing their Expected Utilities $\mathbb{E}_\pi(u \circ X)$ and $\mathbb{E}_\pi(u \circ Y)$.

Constructing Choice Functions for Decision Making

Classical assumptions:

(I) The agent's preferences among the elements of A are characterized by a cardinal utility function $u : A \to \mathbb{R}$.

(II) The uncertainty among the states from S is described by some classical probability measure π.

Under (I) and (II), there is strong consensus for comparing acts X and Y by comparing their **E**xpected **U**tilities $\mathbb{E}_\pi(u \circ X)$ and $\mathbb{E}_\pi(u \circ Y)$.

Standard Choice Function:

This induces a choice function by setting, for all $\mathcal{D} \in 2^{\mathcal{G}}$,

$$ch_{u,\pi}(\mathcal{D}) = \Big\{ Y \in \mathcal{D} : \mathbb{E}_\pi(u \circ Y) \geq \mathbb{E}_\pi(u \circ X) \text{ for all } X \in \mathcal{D} \Big\},$$

i.e., by choosing that acts from $\mathcal{G}$ that maximize expected utility.

Weakly structured Information

Maximizing Expected Utility?

Problem: Both (I) and (II) require **strong axiomatic assumptions**.

These assumptions explicitly dismiss the following settings:

- Purely ordinal or partial preferences
  (e.g. random variables with locally varying scale of measurement).
  (e.g., [Seidenfeld et al., 1995, Nau, 2006])
- Agents with partial probabilistic beliefs
  (e.g. Robust Bayesian analysis, uncertainty quantification).
  (e.g., [Kikuti et al., 2011, Shaker and Hüllermeier, 2021])
- Problems of group decision making
  (e.g. ensemble methods).
  (e.g., [Bradley, 2019])

These are **highly relevant situations** to investigate!

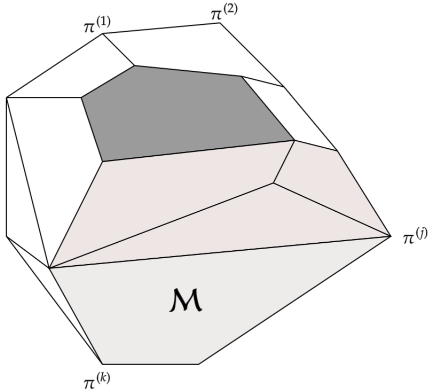
Relaxing (I) and (II): Weakly structured Information

Two different sources of complexity:

Relaxing (I) and (II): Weakly structured Information

Two different sources of complexity:

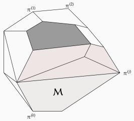
(I)′ Imprecise probabilistic models: If it isn't possible to specify **one** probability on S, we still can work with the *set* $\mathcal{M}$ of all probabilities **compatible with the information**.


$\pi^{(1)}$
$\pi^{(2)}$
$\pi^{(j)}$
$\pi^{(k)}$
$\mathcal{M}$

Relaxing (I) and (II): Weakly structured Information

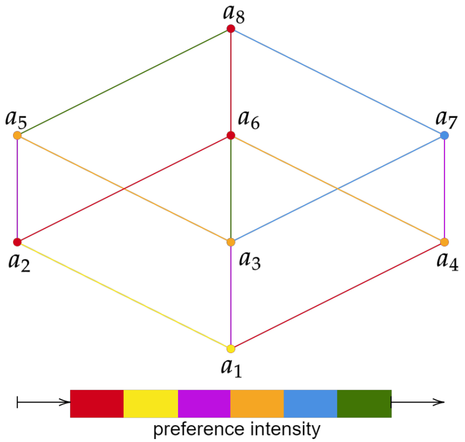
Two different sources of complexity:

(I)′ Imprecise probabilistic models: If it isn't possible to specify **one** probability on S, we still can work with the *set* $\mathcal{M}$ of all probabilities **compatible with the information**.



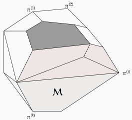
(II)′ Complexly ordered consequences: A cardinal utility demands the agent to satisfy **very restrictive axioms**. If these are too restrictive, we still can work with the **set** $\mathcal{U}$ of all utilities **compatible with the information**.

Relaxing (I) and (II): Weakly structured Information

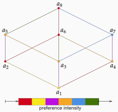


Two different sources of complexity:

(I)′ Imprecise probabilistic models: If it isn't possible to specify **one** probability on S, we still can work with the *set* $\mathcal{M}$ of all probabilities **compatible with the information**.



(II)′ Complexly ordered consequences: A cardinal utility demands the agent to satisfy **very restrictive axioms**. If these are too restrictive, we still can work with the **set** $\mathcal{U}$ of all utilities **compatible with the information**.



Notation: Binary relation R has **strict part** P_R and **indifference part** I_R.

Preference system & Consistency

Let A denote a set of consequences. Let further

- $R_1 \subseteq A \times A$ be a binary relation on A
- $R_2 \subseteq R_1 \times R_1$ be a binary relation on R_1

The triplet $\mathcal{A} = [A, R_1, R_2]$ is called a **preference system** on A.

We call $\mathcal{A}$ **consistent** if there is $u : A \to [0,1]$ with for all $a, b, c, d \in A$:

$$(a, b) \in R_1 \Rightarrow u(a) \geq u(b) \qquad (\text{with} = \textit{iff} \in I_{R_1}).$$

$$((a, b), (c, d)) \in R_2 \Rightarrow u(a) - u(b) \geq u(c) - u(d) \qquad (\text{with} = \textit{iff} \in I_{R_2}).$$

The set of all representations u of $\mathcal{A}$ is denoted by $\mathcal{U}_{\mathcal{A}}$.

Interpretation of the components of $\mathcal{A}$:

- $(a, b) \in R_1$: *"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 d by c"*

Modelling $\mathcal{U}$: Preference Systems I

Notation: Binary relation R has **strict part** P_R and **indifference part** I_R.

Preference system & Consistency

Let A denote a set of consequences. Let further

$R_1 \subseteq A \times A$ be a binary relation on A

$R_2 \subseteq R_1 \times R_1$ be a binary relation on R_1

The triplet $\mathcal{A} = [A, R_1, R_2]$ is called a **preference system** on A.

We call $\mathcal{A}$ **consistent** if there is $u : A \to [0,1]$ with for all $a, b, c, d \in A$:

- $(a, b) \in R_1 \Rightarrow u(a) \geq u(b)$ (with $=$ *iff* $\in I_{R_1}$).
- $((a, b), (c, d)) \in R_2 \Rightarrow u(a) - u(b) \geq u(c) - u(d)$ (with $=$ *iff* $\in I_{R_2}$).

The set of all representations u of $\mathcal{A}$ is denoted by $\mathcal{U}_{\mathcal{A}}$.

Normalization & Regularization

Let $\mathcal{A} = [A, R_1, R_2]$ be consistent and assume there exist $a_*, a^* \in A$ such that $(a^*, a) \in R_1$ and $(a, a_*) \in R_1$ for all $a \in A$. Then

$$\mathcal{N}_{\mathcal{A}} := \Big\{ u \in \mathcal{U}_{\mathcal{A}} : u(a_*) = 0 \ \wedge \ u(a^*) = 1 \Big\}$$

is called the **normalized representation set** of $\mathcal{A}$.

Further, for $\delta \in [0, 1)$, we denote by $\mathcal{N}_{\mathcal{A}}^{\delta}$ the set of all $u \in \mathcal{N}_{\mathcal{A}}$ satisfying

$$u(a) - u(b) \geq \delta \quad \wedge \quad u(c) - u(d) - u(e) + u(f) \geq \delta$$

for all $(a, b) \in P_{R_1}$ and for all $((c, d), (e, f)) \in P_{R_2}$.

We call $\mathcal{A}$ δ-**consistent** if $\mathcal{N}_{\mathcal{A}}^{\delta} \neq \emptyset$.

Modelling $\mathcal{M}$: Credal sets

Credal set

The uncertainty among the elements of S is described by a polyhedral *credal set* of probability measures of the form

$$\mathcal{M} = \left\{ \pi \in \mathcal{P} : \underline{b}_\ell \leq \mathbb{E}_\pi(f_\ell) \leq \overline{b}_\ell \text{ for } \ell = 1, \ldots, r \right\}$$

where $\mathcal{P}$ is the set of all probability measures on $(S, \sigma(S))$ and

- $f_1, \ldots, f_r : S \to \mathbb{R}$ are real-valued mappings and
- $\underline{b}_\ell \leq \overline{b}_\ell$, $\ell = 1, \ldots, r$, are lower and upper expectation bounds.

Credal set

The uncertainty among the elements of S is described by a polyhedral *credal set* of probability measures of the form

$$\mathcal{M} = \left\{ \pi \in \mathcal{P} : \underline{b}_\ell \leq \mathbb{E}_\pi(f_\ell) \leq \overline{b}_\ell \text{ for } \ell = 1, \ldots, r \right\}$$

where $\mathcal{P}$ is the set of all probability measures on $(S, \sigma(S))$ and

- $f_1, \ldots, f_r : S \to \mathbb{R}$ are real-valued mappings and
- $\underline{b}_\ell \leq \overline{b}_\ell$, $\ell = 1, \ldots, r$, are lower and upper expectation bounds.

Description: Such $\mathcal{M}$ is a convex polyhedron with extreme points

$$\mathcal{E}(\mathcal{M}) = \{\pi^{(1)}, \ldots \pi^{(K)}\}$$

Modelling $\mathcal{M}$: Credal sets

Credal set

The uncertainty among the elements of S is described by a polyhedral *credal set* of probability measures of the form

$$\mathcal{M} = \left\{\pi \in \mathcal{P} : \underline{b}_\ell \leq \mathbb{E}_\pi(f_\ell) \leq \overline{b}_\ell \text{ for } \ell = 1, \ldots, r\right\}$$

where $\mathcal{P}$ is the set of all probability measures on $(S, \sigma(S))$ and

- $f_1, \ldots, f_r : S \to \mathbb{R}$ are real-valued mappings and
- $\underline{b}_\ell \leq \overline{b}_\ell$, $\ell = 1, \ldots, r$, are lower and upper expectation bounds.

Description: Such $\mathcal{M}$ is a convex polyhedron with extreme points

$$\mathcal{E}(\mathcal{M}) = \{\pi^{(1)}, \ldots \pi^{(K)}\}$$

Special cases: *Classical probability – Interval probability – Lower previsions – Linear partial information – Neighbourhood models*

(e.g., [Levi, 1974, Walley, 1991, Weichselberger, 2001, Augustin et al., 2014]))

Generalizing the Choice Function

Theory for optimal decision making based on the sets $\mathcal{U}_{\mathcal{A}}$ and $\mathcal{M}$ as well as efficient computation algorithms have been developed in:


Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences ☆,☆☆

C. Jansen *, G. Schollmeyer, T. Augustin

Generalizing the Choice Function

Theory for optimal decision making based on the sets $\mathcal{U}_{\mathcal{A}}$ and $\mathcal{M}$ as well as efficient computation algorithms have been developed in:

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences ☆,☆☆

C. Jansen *, G. Schollmeyer, T. Augustin

We focus on only one decision criterion from the paper:

$(\mathcal{A}, \mathcal{M}, \delta)$-dominance

Let $\mathcal{A} = [A, R_1, R_2]$ be δ-consistent and $\mathcal{M}$ a credal set on $(S, \sigma(S))$. Define

$$\mathcal{F}_{(\mathcal{A},S)} := \left\{ X \in A^S : u \circ X \text{ is } \sigma(S)\text{-}\mathcal{B}_{\mathbb{R}}([0,1])\text{-measurable for all } u \in \mathcal{U}_{\mathcal{A}} \right\}.$$

For $X, Y \in \mathcal{F}_{(\mathcal{A},S)}$, 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 $\geq_{(\mathcal{A},\mathcal{M},\delta)}$.

Some Special Cases

The relation $\geq_{(\mathcal{A},\mathcal{M},\delta)}$ has some prominent special cases.

For $\delta = 0$ and ...

- ... and $\mathcal{M} = \{\pi\}$ and $R_2 = \emptyset$

  $\rightarrow$ Reduction to (first-order) **stochastic dominance**

  (see, e.g., [Mosler and Scarsini, 1991]))

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

  $\rightarrow$ Reduction to comparing **expected utilities**.

  (see, e.g., [Krantz et al., 1971]))

- ... and R_1 and R_2 guaranteeing utility unique up to plts

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

  (see, e.g., [Troffaes, 2007]))

Checking for $(\mathcal{D}, \mathcal{M}, \delta)$-dominance: Preparation

Now, let

- $\mathcal{A} = [A, R_1, R_2]$ be a δ-consistent decision system,
- $A = \{a_1, \ldots, a_n\}$, $S = \{s_1, \ldots, s_m\}$, and
- $a_{k_1}, a_{k_2} \in A$ such that $(a_{k_1}, a) \in R_1$ and $(a, a_{k_2}) \in R_1$ for all $a \in A$.

A vector $(v_1, \ldots, v_n)$ containing exactly the images of a utility function $u \in \mathcal{N}_{\mathcal{A}}^{\delta}$ is then describable by the system of linear (in-)equalities given through

- $v_{k_1} = 1$ and $v_{k_2} = 0$,
- $v_i = v_j$ for every pair $(a_i, a_j) \in I_{R_1}$,
- $v_i - v_j \geq \delta$ for every pair $(a_i, a_j) \in P_{R_1}$,
- $v_k - v_l = v_p - v_q$ for every pair of pairs $((a_k, a_l), (a_p, a_q)) \in I_{R_2}$ and
- $v_k - v_l - v_p + v_q \geq \delta$ for every pair of pairs $((a_k, a_l), (a_p, a_q)) \in P_{R_2}$.

Denote by $\nabla_{\mathcal{A}}^{\delta}$ the set of all $(v_1, \ldots, v_n) \in [0, 1]^n$ satisfying these (in)equalities.

Checking for $(\mathcal{D}, \mathcal{M}, \delta)$-dominance: Preparation

Now, let

Under finitely many consequences and states...

A vector $(v_1, \ldots, v_n)$ containing exactly the images of a utility function $u \in \mathcal{N}_{\mathcal{A}}$ is then describable by the system of linear (in-)equalities given through

...the set of admissible utilities is describable
by finitely many linear constraints.

Denote by $\nabla_{\mathcal{A}}$ the set of all $(v_1, \ldots, v_n) \in [0, 1]^n$ satisfying these (in)equalities.

Checking for $(\mathcal{A}, \mathcal{M}, \delta)$-Dominance: Algorithm

Theorem

Consider the same situation as described above.

For $X_i, X_j \in \mathcal{G}$ and $t \in \{1, \ldots, K\}$, we consider the linear program

$$\sum_{\ell=1}^{n} v_\ell \cdot [\pi^{(t)}(X_i^{-1}(\{a_\ell\})) - \pi^{(t)}(X_j^{-1}(\{a_\ell\}))] \longrightarrow \min_{(v_1,\ldots,v_n)\in\mathbb{R}^n}$$

with constraints $(v_1, \ldots, v_n) \in \nabla_{\mathcal{A}}^{\delta}$.

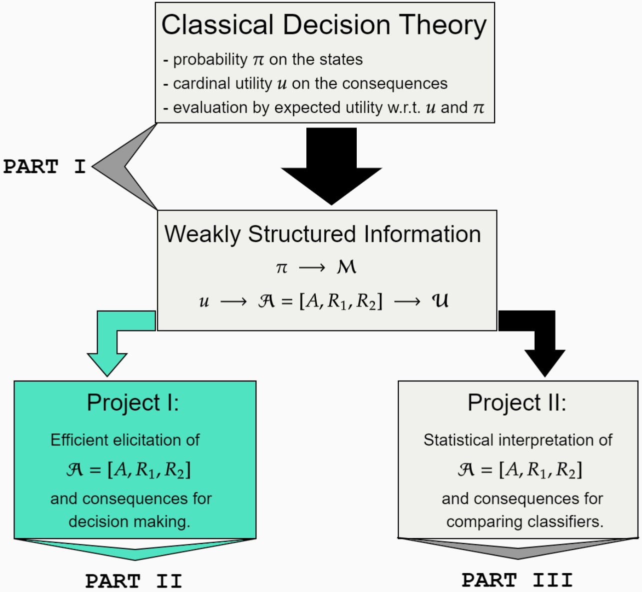
Denote by $opt_{ij}(t)$ the optimal value of this programming problem.

It then holds:

$$X_i \geq_{(\mathcal{A},\mathcal{M},\delta)} X_j \quad \Leftrightarrow \quad \min\{opt_{ij}(t) : t = 1, \ldots, K\} \geq 0.$$

Project I: Elicitation

Efficient Elicitation of Preference Systems



Efficient Elicitation of Preference Systems

Important question: Similar as in classical utility theory, the question of how to receive an agent's preference system in practice is of vast importance!

Efficient Elicitation of Preference Systems

Important question: Similar as in classical utility theory, the question of how to receive an agent's preference system in practice is of vast importance!

Idea: Design efficient elicitation strategies for preference systems.

Efficient Elicitation of Preference Systems

Important question: Similar as in classical utility theory, the question of how to receive an agent's preference system in practice is of vast importance!

Idea: Design efficient elicitation strategies for preference systems.

Challenges:

- How exactly?
- What does efficiency mean in this context?

Efficient Elicitation of Preference Systems

Important question: Similar as in classical utility theory, the question of how to receive an agent's preference system *in practice* is of vast importance!

Idea: Design efficient *elicitation strategies* for preference systems.

Challenges:

- How exactly?
- What does efficiency mean in this context?

These questions are addressed in the paper:


Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

ELSEVIER

Information efficient learning of complexly structured preferences: Elicitation procedures and their application to decision making under uncertainty

C. Jansen*, H. Blocher, T. Augustin, G. Schollmeyer

Department of Statistics, LMU Munich, Ludwigsstr. 33, 80539 Munich, Germany

Outline of the Paper

Goal: Elicit (the relevant parts of) an agent's preference system

$$\mathcal{A}^* = [A, R_1^*, R_2^*]$$

by asking as few as possible ranking questions about R_1^*.

Outline of the Paper

Goal: Elicit (the relevant parts of) an agent's preference system

$$\mathcal{A}^* = [A, R_1^*, R_2^*]$$

by asking as few as possible ranking questions about R_1^*.

Two different approaches are considered:

Procedure 1 utilizes the agent's consideration times.

Procedure 2 collects labels of preference strength.

Outline of the Paper

Goal: Elicit (the relevant parts of) an agent's preference system

$$\mathcal{A}^* = [A, R_1^*, R_2^*]$$

by asking as few as possible ranking questions about R_1^*.

Two different approaches are considered:

Procedure 1 utilizes the agent's consideration times.

Procedure 2 collects labels of preference strength.

Main contributions of the paper:

(I) Methods for eliciting $\mathcal{A}$ by only asking ranking questions about R_1.

(II) Data-driven guidance of elicitation with previous user experience.

(III) Utilizing elicitation methods for information efficient decision making between acts $X : S \to A$ taking values in A.

Focus today:

Procedure 2: Collecting labels of preference strength.

$\rightarrow$ **Label elicitation**

Procedure 2: Label elicitation

Setup: Agent assigns a label $\ell_r^{ij} \in \mathcal{L}_r := \{\mathbf{n}, \mathbf{c}, 0, 1, \dots, r\}$ to every (a_i, a_j) by some labelling function $\ell_r : A \times A \to \mathcal{L}_r$:

$\mathbf{n}$:	non-comparable
$\mathbf{c}$:	strict preference of unknown strength
0 :	indifferent
$1, \dots, r$:	strict preference of increasing strength

Label elicitation

Input: $A = \{a_1, \dots, a_n\}$; $R_1 = \emptyset$; number of labels r;

Output: $\mathcal{A} = [A, R_1, R_2]$;

Procedure: Present all pairs $(a_i, a_j) \in A \times A$.

i) If $\ell_r^{ij} \in \mathcal{L}_r \setminus \{\mathbf{n}, 0\}$, set $R_1 = R_1 \cup \{(a_i, a_j)\}$.

ii) If $\ell_r^{ij} = 0$, set $R_1 = R_1 \cup \{(a_i, a_j), (a_j, a_i)\}$.

iii) If $\ell_r^{ij} = \mathbf{n}$, set $R_1 = R_1$.

Define R_2 by setting $((a_i, a_j), (a_k, a_l)) \in R_2 \quad :\Leftrightarrow \quad \ell_r^{ij} > \ell_r^{kl} \;\vee\; \ell_r^{ij} = \ell_r^{kl} = 0$

Procedure 2: Assumptions

Assumption 1

i) $(a_i, a_j) \in I_{R_1^*} \quad \Leftrightarrow \quad \ell_r^{ij} = 0$

ii) $(a_i, a_j) \in P_{R_1^*} \quad \Leftrightarrow \quad \ell_r^{ij} \in \mathcal{L}_r \setminus \{\mathbf{n}, 0\} \ \wedge \ \ell_r^{ji} = \mathbf{n}$

iii) $(a_i, a_j) \in C_{R_1^*} \quad \Leftrightarrow \quad \ell_r^{ij} = \ell_r^{ji} = \mathbf{n}$

Assumption 2

For all $(a_i, a_j), (a_k, a_l) \in R_1^*$ the following holds:

i) $\ell_r^{ij} > \ell_r^{kl} \quad \Rightarrow \quad ((a_i, a_j), (a_k, a_l)) \in P_{R_2^*}$

ii) $\ell_r^{ij} = \ell_r^{kl} = 0 \quad \Rightarrow \quad ((a_i, a_j), (a_k, a_l)) \in I_{R_2^*}$

iii) $\ell_r^{ij} = \mathbf{c} \ \vee \ \ell_r^{kl} = \mathbf{c} \quad \Leftrightarrow \quad ((a_i, a_j), (a_k, a_l)) \in C_{R_2^*}$

Assumption 3

For all $((a_i, a_j), (a_k, a_l)) \in P_{R_2^*}$ the statement $\ell_r^{ij} = \ell_r^{kl} = x \notin \{0, \mathbf{n}, \mathbf{c}\}$ implies that $\{1, \ldots, r\} \subset \ell_r(A \times A)$.

Procedure 2: Assumptions

Assumption 1

ordinal part is reported truthfully

Assumption 2

cardinal part is reported best possibly

Assumption 3

labels are interpreted purely ordinal

Procedure 2: Findings

Theorem

The following two statements hold true:

i) If, for some $r \in \mathbb{N}$, $\ell_r : A \times A \to \mathcal{L}_r$ satisfies Assumptions 1 and 2, then Procedure 2 produces a sub-system of $\mathcal{A}^*$.

ii) There exists $r_0 \in \mathbb{N}$ such that if $\ell_{r_0} : A \times A \to \mathcal{L}_{r_0}$ satisfies Assumptions 1, 2 and 3, then Procedure 2 produces the true $\mathcal{A}^*$.

Procedure 2: Findings

Theorem

The following two statements hold true:

i) If, for some $r \in \mathbb{N}$, $\ell_r : A \times A \to \mathcal{L}_r$ satisfies Assumptions 1 and 2, then Procedure 2 produces a sub-system of $\mathcal{A}^*$.

ii) There exists $r_0 \in \mathbb{N}$ such that if $\ell_{r_0} : A \times A \to \mathcal{L}_{r_0}$ satisfies Assumptions 1, 2 and 3, then Procedure 2 produces the true $\mathcal{A}^*$.

Challenge: Although the Theorem guarantees that Procedure 2 reproduces $\mathcal{A}^*$ for some r^*, labelling may be too demanding if r^* is large.

Procedure 2: Findings

Theorem

The following two statements hold true:

i) If, for some $r \in \mathbb{N}$, $\ell_r : A \times A \to \mathcal{L}_r$ satisfies Assumptions 1 and 2, then Procedure 2 produces a sub-system of $\mathcal{A}^*$.

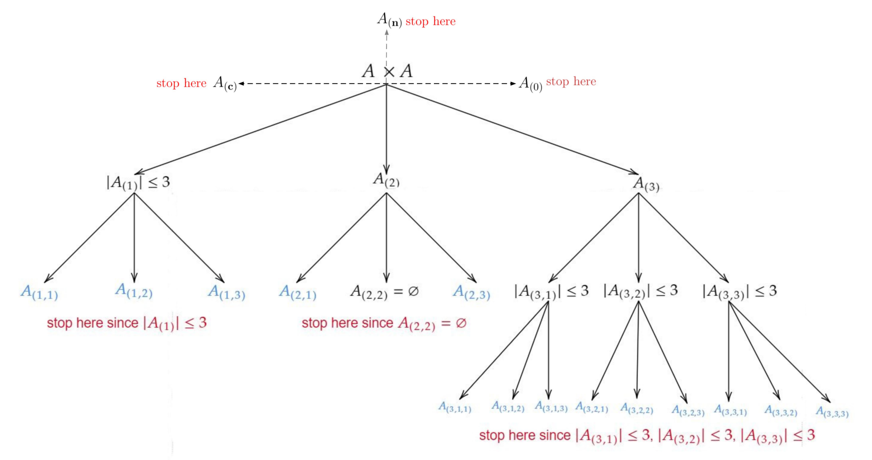
ii) There exists $r_0 \in \mathbb{N}$ such that if $\ell_{r_0} : A \times A \to \mathcal{L}_{r_0}$ satisfies Assumptions 1, 2 and 3, then Procedure 2 produces the true $\mathcal{A}^*$.

Challenge: Although the Theorem guarantees that Procedure 2 reproduces $\mathcal{A}^*$ for some r^*, labelling may be too demanding if r^* is large.

Solution: Use a relatively small r and restart elicitation on pairs with equal label. Stop as soon as you know that equal labels originate from indifference.

Procedure 2: Hierarchical version

Graphical intuition:



Hierarchical version: Findings

For the hierarchical version of label elicitation to work, we need to assume that the agent is able to adapt the labelling function to arbitrary subsets.

Hierarchical version: Findings

For the hierarchical version of label elicitation to work, we need to assume that the agent is able to adapt the labelling function to arbitrary subsets.

Formally, we arrive at:

Assumption 4

For every $N \subseteq A \times A$ the labels on the restricted set of pairs N are given w.r.t. a labelling function $\ell_{(N,r)} : N \to \mathcal{L}_r$ satisfying Assumptions 1, 2 and 3.

Hierarchical version: Findings

For the hierarchical version of label elicitation to work, we need to assume that the agent is able to adapt the labelling function to arbitrary subsets.

Formally, we arrive at:

Assumption 4

For every $N \subseteq A \times A$ the labels on the restricted set of pairs N are given w.r.t. a labelling function $\ell_{(N,r)} : N \to \mathcal{L}_r$ satisfying Assumptions 1, 2 and 3.

This indeed allows the following Proposition:

Theorem

Let Assumption 4 hold true. For $n = |A|$ consequences and $r \geq 2$ labels, the hierarchical version of Procedure 2 terminates in $\mathcal{A}^*$ after at most

$$\max\{1, \lceil \tfrac{n^2 - r}{r - 1} \rceil + 1\}$$

elicitation rounds.

Application to decision making under uncertainty

We now return to decision under uncertainty:

- Consider the decision problem $\mathcal{G}$ under uncertainty model $\mathcal{M}$.
- Suppose $\mathcal{A}^*$ is elicited by either Procedure 1 or 2 (or some variant).
- Let $\mathcal{A}_1, \mathcal{A}_2, \dots$ be the preference system after elicitation step $1, 2, \dots$.

Application to decision making under uncertainty

We now return to decision under uncertainty:

- Consider the decision problem $\mathcal{G}$ under uncertainty model $\mathcal{M}$.
- Suppose $\mathcal{A}^*$ is elicited by either Procedure 1 or 2 (or some variant).
- Let $\mathcal{A}_1, \mathcal{A}_2, \dots$ be the preference system after elicitation step $1, 2, \dots$.

Theorem

Let the assumptions of the used procedure be satisfied. Then, for any k :

$$X \in ch_{\mathcal{A}_k,\mathcal{M}}(\mathcal{G}) \quad \Rightarrow \quad X \in ch_{\mathcal{A}^*,\mathcal{M}}(\mathcal{G})$$

Here:

$$ch_{\mathcal{A}_k,\mathcal{M}}(\mathcal{G}) := \Big\{ Y \in \mathcal{G} : \forall X \in \mathcal{G}, u \in \mathcal{U}_{\mathcal{A}}, \pi \in \mathcal{M} \text{ it holds } \mathbb{E}_\pi(u \circ Y) \geq \mathbb{E}_\pi(u \circ X) \Big\}.$$

Application to decision making under uncertainty

We now return to decision under uncertainty:

- Consider the decision problem $\mathcal{G}$ under uncertainty model $\mathcal{M}$.
- Suppose $\mathcal{A}^*$ is elicited by either Procedure 1 or 2 (or some variant).
- Let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be the preference system after elicitation step $1, 2, \ldots$.

Theorem

Let the assumptions of the used procedure be satisfied. Then, for any k :

$$X \in ch_{\mathcal{A}_k,\mathcal{M}}(\mathcal{G}) \quad \Rightarrow \quad X \in ch_{\mathcal{A}^*,\mathcal{M}}(\mathcal{G})$$

Here:

$$ch_{\mathcal{A}_k,\mathcal{M}}(\mathcal{G}) := \Big\{ Y \in \mathcal{G} : \forall X \in \mathcal{G}, u \in \mathcal{U}_{\mathcal{A}}, \pi \in \mathcal{M} \text{ it holds } \mathbb{E}_\pi(u \circ Y) \geq \mathbb{E}_\pi(u \circ X) \Big\}.$$

Why is this good?

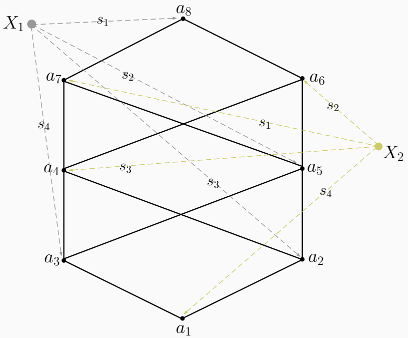
If an act is optimal w.r.t. the preference system $\mathcal{A}_k$ elicited so far, we can conclude it is optimal w.r.t. the true preference system $\mathcal{A}^*$.

A small example

Consider the following decision problem:

	s_1	s_2	s_3	s_4
X_1	a_8	a_5	a_2	a_3
X_2	a_7	a_6	a_4	a_1

Decision problem



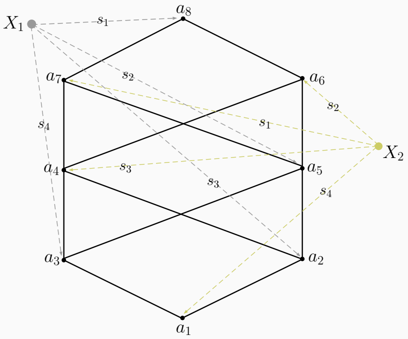
Hasse diagram of R_1^*

A small example

Consider the following decision problem:

	s_1	s_2	s_3	s_4
X_1	a_8	a_5	a_2	a_3
X_2	a_7	a_6	a_4	a_1

Decision problem



Hasse diagram of R_1^*

R_2^* is the transitive hull of (where $e_{ij} := (a_i, a_j)$):

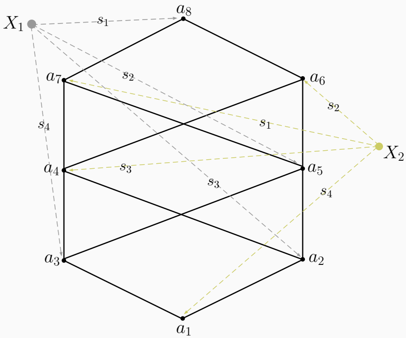
$$e_{31} P_{R_2^*} e_{52} P_{R_2^*} e_{74} P_{R_2^*} e_{21} I_{R_2^*} e_{64} I_{R_2^*} e_{42} I_{R_2^*} e_{86} P_{R_2^*} e_{87} P_{R_2^*} e_{53} P_{R_2^*} e_{75} P_{R_2^*} e_{65} P_{R_2^*} e_{43}$$

A small example

Consider the following decision problem:

	s_1	s_2	s_3	s_4
X_1	a_8	a_5	a_2	a_3
X_2	a_7	a_6	a_4	a_1

Decision problem



Hasse diagram of R_1^*

R_2^* is the transitive hull of (where $e_{ij} := (a_i, a_j)$):

$$e_{31} P_{R_2^*} e_{52} P_{R_2^*} e_{74} P_{R_2^*} e_{21} I_{R_2^*} e_{64} I_{R_2^*} e_{42} I_{R_2^*} e_{86} P_{R_2^*} e_{87} P_{R_2^*} e_{53} P_{R_2^*} e_{75} P_{R_2^*} e_{65} P_{R_2^*} e_{43}$$

$\mathcal{M} = \{\pi\}$ with π the uniform distribution.

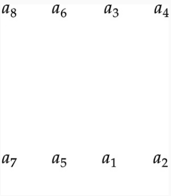
A small example, continued

Procedure 2 with $r = 5$ is applied and the first four steps look as follows:

A small example, continued

Procedure 2 with $r = 5$ is applied and the first four steps look as follows:

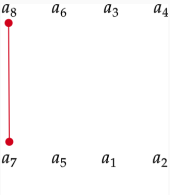
Step	Pair	Label



A small example, continued

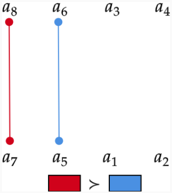
Procedure 2 with $r = 5$ is applied and the first four steps look as follows:

Step	Pair	Label
1	(a_8, a_7)	$\ell_5^{87} = 2$



Procedure 2 with $r = 5$ is applied and the first four steps look as follows:

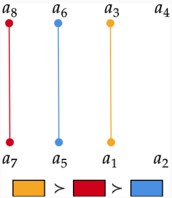
Step	Pair	Label
1	(a_8, a_7)	$\ell_5^{87} = 2$
2	(a_6, a_5)	$\ell_5^{65} = 1$



A small example, continued

Procedure 2 with $r = 5$ is applied and the first four steps look as follows:

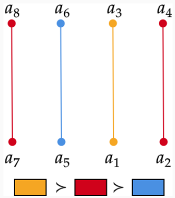
Step	Pair	Label
1	(a_8, a_7)	$\ell_5^{87} = 2$
2	(a_6, a_5)	$\ell_5^{65} = 1$
3	(a_3, a_1)	$\ell_5^{31} = 3$



A small example, continued

Procedure 2 with $r = 5$ is applied and the first four steps look as follows:

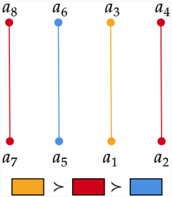
Step	Pair	Label
1	(a_8, a_7)	$\ell_5^{87} = 2$
2	(a_6, a_5)	$\ell_5^{65} = 1$
3	(a_3, a_1)	$\ell_5^{31} = 3$
4	(a_4, a_2)	$\ell_5^{42} = 2$



A small example, continued

Procedure 2 with $r = 5$ is applied and the first four steps look as follows:

Step	Pair	Label
1	(a_8, a_7)	$\ell_5^{87} = 2$
2	(a_6, a_5)	$\ell_5^{65} = 1$
3	(a_3, a_1)	$\ell_5^{31} = 3$
4	(a_4, a_2)	$\ell_5^{42} = 2$



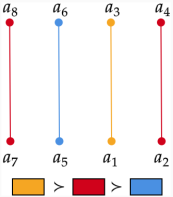
Then, for every $u \in \mathcal{U}_{\mathcal{A}_4}$ (where $u_i := u(a_i)$):

$$4 \cdot (\mathbb{E}_\pi(u \circ X_1) - \mathbb{E}_\pi(u \circ X_2)) = \underbrace{(u_8 - u_7) - (u_6 - u_5)}_{>0,\ since\ (e_{87}, e_{65}) \in P_{R_2}} + \underbrace{(u_3 - u_1) + (u_4 - u_2)}_{>0,\ since\ (e_{31}, e_{42}) \in P_{R_2}} > 0$$

A small example, continued

Procedure 2 with $r = 5$ is applied and the first four steps look as follows:

Step	Pair	Label
1	(a_8, a_7)	$\ell_5^{87} = 2$
2	(a_6, a_5)	$\ell_5^{65} = 1$
3	(a_3, a_1)	$\ell_5^{31} = 3$
4	(a_4, a_2)	$\ell_5^{42} = 2$



Then, for every $u \in \mathcal{U}_{\mathcal{A}_4}$ (where $u_i := u(a_i)$):

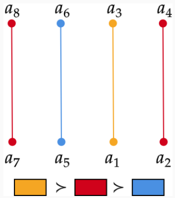
$$4 \cdot (\mathbb{E}_\pi(u \circ X_1) - \mathbb{E}_\pi(u \circ X_2)) = \underbrace{(u_8 - u_7) - (u_6 - u_5)}_{>0,\ since\ (e_{87}, e_{65}) \in P_{R_2}} + \underbrace{(u_3 - u_1) + (u_4 - u_2)}_{>0,\ since\ (e_{31}, e_{42}) \in P_{R_2}} > 0$$

Thus $X_1 \in ch_{\mathcal{A}_4, \mathcal{M}}(\mathcal{G})$.

A small example, continued

Procedure 2 with $r = 5$ is applied and the first four steps look as follows:

Step	Pair	Label
1	(a_8, a_7)	$\ell_5^{87} = 2$
2	(a_6, a_5)	$\ell_5^{65} = 1$
3	(a_3, a_1)	$\ell_5^{31} = 3$
4	(a_4, a_2)	$\ell_5^{42} = 2$



Then, for every $u \in \mathcal{U}_{\mathcal{A}_4}$ (where $u_i := u(a_i)$):

$$4 \cdot (\mathbb{E}_\pi(u \circ X_1) - \mathbb{E}_\pi(u \circ X_2)) = \underbrace{(u_8 - u_7) - (u_6 - u_5)}_{>0,\ \textit{since}\ (e_{87}, e_{65}) \in P_{R_2}} + \underbrace{(u_3 - u_1) + (u_4 - u_2)}_{>0,\ \textit{since}\ (e_{31}, e_{42}) \in P_{R_2}} > 0$$

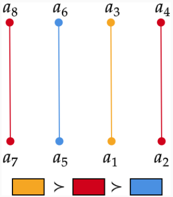
Thus $X_1 \in ch_{\mathcal{A}_4, \mathcal{M}}(\mathcal{G})$.

Thus $X_1 \in ch_{\mathcal{A}^*, \mathcal{M}}(\mathcal{G})$ by our Theorem.

A small example, continued

Procedure 2 with $r = 5$ is applied and the first four steps look as follows:

Step	Pair	Label
1	(a_8, a_7)	$\ell_5^{87} = 2$
2	(a_6, a_5)	$\ell_5^{65} = 1$
3	(a_3, a_1)	$\ell_5^{31} = 3$
4	(a_4, a_2)	$\ell_5^{42} = 2$



Then, for every $u \in \mathcal{U}_{\mathcal{A}_4}$ (where $u_i := u(a_i)$):

$$4 \cdot (\mathbb{E}_\pi(u \circ X_1) - \mathbb{E}_\pi(u \circ X_2)) = \underbrace{(u_8 - u_7) - (u_6 - u_5)}_{>0,\ since\ (e_{87}, e_{65}) \in P_{R_2}} + \underbrace{(u_3 - u_1) + (u_4 - u_2)}_{>0,\ since\ (e_{31}, e_{42}) \in P_{R_2}} > 0$$

Thus $X_1 \in ch_{\mathcal{A}_4, \mathcal{M}}(\mathcal{G})$.

Thus $X_1 \in ch_{\mathcal{A}^*, \mathcal{M}}(\mathcal{G})$ by our Theorem.

!! *We concluded that X_1 is optimal by asking four simple ranking questions.* !!

There's more in the Paper!

Beyond the concepts just shown, we ...

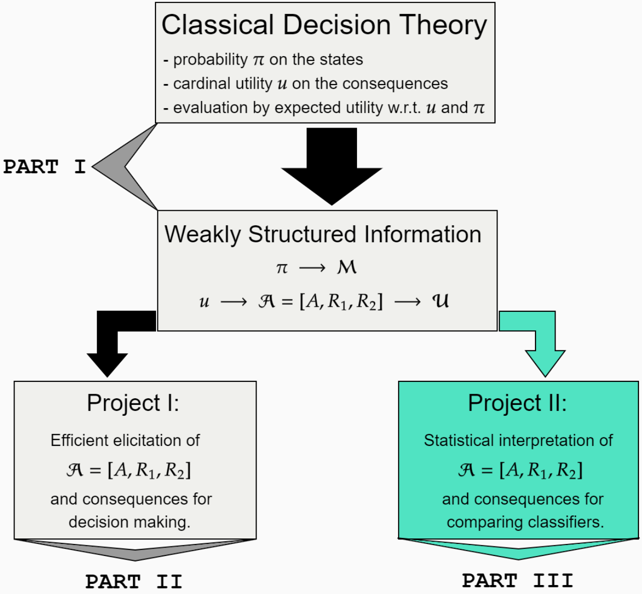
- ... introduced a second elicitation scheme based on consideration times.
- ... gave more efficient versions of our algorithms based on ...
    1. ... purely order-theoretic considerations, amd
    2. ... data-driven elicitation with previous user experience..

Promising lines of future research:

- Improving prediction of promising pairs.
- Explicitly incorporating the choice function into the prediction.
- Mixing hierarchical and non-hierarchical procedures.

Project II: Statistical Applications

Comparing Classifiers by Generalized Stochastic Dominance



Comparing Classifiers by Generalized Stochastic Dominance

Question of interest: How to utilize our decision-theoretical approach for comparing classifiers under multiplicity of quality criteria and data sets?

Setup: Let

- $\mathcal{D}$ denote the set of all relevant data sets,
- $\mathcal{C}$ denote the set of all relevant classifiers,
- $\left(\phi_i : \mathcal{C} \times \mathcal{D} \to Q_i\right)_{i \in \{1,\ldots,n\}}$ denote a family of quality criteria,
- $\phi := (\phi_1, \ldots, \phi_n) : \mathcal{D} \times \mathcal{C} \to \mathcal{Q}$, where $\mathcal{Q} := Q_1 \times \cdots \times Q_n$.

Comparing Classifiers by Generalized Stochastic Dominance

Question of interest: How to utilize our decision-theoretical approach for comparing classifiers under multiplicity of quality criteria and data sets?

Setup: Let

- $\mathcal{D}$ denote the set of all relevant data sets,
- $\mathcal{C}$ denote the set of all relevant classifiers,
- $\big(\phi_i : \mathcal{C} \times \mathcal{D} \to Q_i\big)_{i \in \{1,\dots,n\}}$ denote a family of quality criteria,
- $\phi := (\phi_1, \dots, \phi_n) : \mathcal{D} \times \mathcal{C} \to \mathcal{Q}$, where $\mathcal{Q} := Q_1 \times \cdots \times Q_n$.

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 $d_i : Q_i \times Q_i \to \mathbb{R}$.

Comparing Classifiers by Generalized Stochastic Dominance

Three levels of problems when comparing classifiers w.r.t. multiple quality criteria on multiple data sets simultaneously.

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

Comparing Classifiers by Generalized Stochastic Dominance

Three levels of problems when comparing classifiers w.r.t. multiple quality criteria on multiple data sets simultaneously.

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

Level 1: On a **fixed** data set D it may hold

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

Comparing Classifiers by Generalized Stochastic Dominance

Three levels of problems when comparing classifiers w.r.t. multiple quality criteria on multiple data sets simultaneously.

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	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.9 \\ \vdots \\ \phi_n(C_q, D_s) \end{pmatrix}$

Level 2: 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

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

Comparing Classifiers by Generalized Stochastic Dominance

Three levels of problems when comparing classifiers w.r.t. multiple quality criteria on multiple data sets simultaneously.

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	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$

Level 3: Even if a decision can be made for a sample $(D_1, \dots, D_s)$ of data sets,

Comparing Classifiers by Generalized Stochastic Dominance

Three levels of problems when comparing classifiers w.r.t. multiple quality criteria on multiple data sets simultaneously.

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	$\begin{pmatrix} 0.85 \\ \vdots \\ 0.67 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.33 \\ \vdots \\ 0.98 \end{pmatrix}$

Level 3: 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 $(D_1^*, \dots, D_s^*)$.

Comparing Classifiers by Generalized Stochastic Dominance

All three levels of problems are at the same time addressed by a generalized notion of stochastic dominance in our recent paper

arXiv > stat > arXiv:2209.01857

Statistics > Machine Learning

[Submitted on 5 Sep 2022]

Statistical Comparisons of Classifiers by Generalized Stochastic Dominance

Christoph Jansen (1), Malte Nalenz (1), Georg Schollmeyer (1), Thomas Augustin (1) ((1) Ludwig-Maximilians-Universität Munich)

Short cut:

SCAN ME

Outline of the Paper

Goal of the project: Framework comparing classifiers w.r.t. multiple quality criteria on multiple data sets simultaneously and suitable statistical tests.

Outline of the Paper

Goal of the project: Framework comparing classifiers w.r.t. multiple quality criteria on multiple data sets simultaneously and suitable statistical tests.

Motivation:

- Existing approaches mostly not account for multiplicity of criteria.
- Decision-theoretic framework addresses multiplicity naturally.

Outline of the Paper

Goal of the project: Framework comparing classifiers w.r.t. multiple quality criteria on multiple data sets simultaneously and suitable statistical tests.

Motivation:

- Existing approaches mostly not account for multiplicity of criteria.
- Decision-theoretic framework addresses multiplicity naturally.

Main contributions of the paper:

(I) Criterion for comparing *classifiers* w.r.t. multiple quality criteria on multiple data sets simultaneously.

(II) An optimization approach for evaluating this criterion.

(III) A statistical test to check in-sample differences for significance.

Defining the Preference System

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

Ordinal part:

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

Cardinal (metric) part:

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

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 $\succsim_\delta$.

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 $\succsim_\delta$.

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 $D_1,\dots,D_s \sim \pi$ of data sets from $\mathcal{D}$.

δ-Dominance (empirical version)

Replace $\mathcal{D}$ by $\hat{\mathcal{D}}_s := \{D_1,\dots,D_s\}$ 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 $\succsim_\delta$ (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 $(v_1, \ldots, v_d) \in \nabla_{\mathbb{C}}^{\delta}$.

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

It then holds:

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

Application Example: Setup

The setup of the application example is as follows:

Application Example: Setup

The setup of the application example is as follows:

- We use 16 binary classification benchmark data sets all taken from the UCI machine learning repository. (see [Dua and Graff, 2017])

Application Example: Setup

The setup of the application example is as follows:

- We use 16 binary classification benchmark data sets all taken from the UCI machine learning repository. (see [Dua and Graff, 2017])
- For classifier comparison, we consider **accuracy**, **AUC** and **Brier score**.

Application Example: Setup

The setup of the application example is as follows:

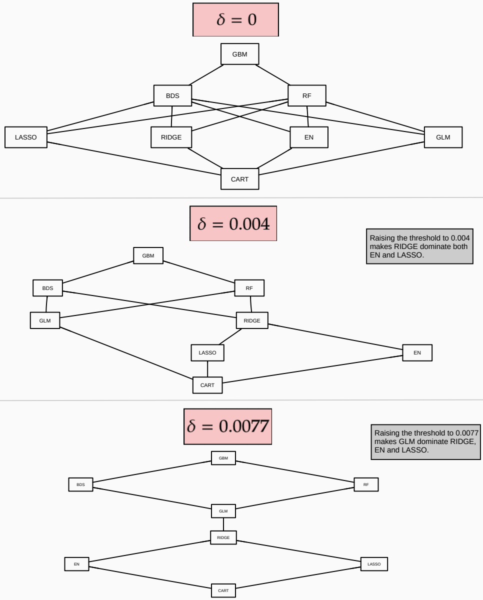
- We use 16 binary classification benchmark data sets all taken from the UCI machine learning repository. (see [Dua and Graff, 2017]))
- For classifier comparison, we consider **accuracy**, **AUC** and **Brier score**.
- We compare the algorithms
  - Classification and regression trees (**CART**)
  - 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

The setup of the application example is as follows:

- We use 16 binary classification benchmark data sets all taken from the UCI machine learning repository. (see [Dua and Graff, 2017]))
- 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



Discussion: How to address Level 3

Good news: In-sample δ-Dominance resolves the problems appearing at the Levels 1 and 2 at the same time.

Discussion: How to address Level 3

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!

Discussion: How to address Level 3

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

$$H_0 : C_j \succsim_\delta C_i$$

Reject H_0 if this value is larger than a critical value c.

Discussion: How to address Level 3

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

$$H_0 : C_j \succsim_\delta C_i$$

Reject H_0 if this value is larger than a critical value c.

Challenge: The distribution of opt_{ij} cannot be analyzed straightforwardly.

Discussion: How to address Level 3

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

$$H_0 : C_j \succsim_\delta C_i$$

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. (see, e.g., [Pratt and Gibbons, 2012]))

Resampling Scheme

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^I_{ij} for the permuted data $(z_i)_{i \in I}$ and $(z_i)_{i \in \{1,\ldots,2s\} \setminus I}$.

Step 4: Sort all opt^I_{ij} 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^I_{ij}, where α is the confidence level.

If $\binom{2s}{s}$ is too large, one can alternatively compute opt^I_{ij} 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.

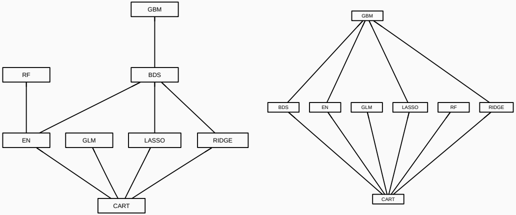
	BDS	CART	EN	GBM	GLM	LASSO	RF	RIDGE
BDS	—	1.000	0.976	—	—	0.967	—	0.951
CART	—	—	—	—	—	—	—	—
EN	—	0.998	—	—	—	—	—	—
GBM	0.998	1.000	0.998	—	—	0.999	—	0.997
GLM	—	1.000	—	—	—	—	—	—
LASSO	—	0.997	—	—	—	—	—	—
RF	—	1.000	0.953	—	—	—	—	—
RIDGE	—	0.999	—	—	—	—	—	—

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:

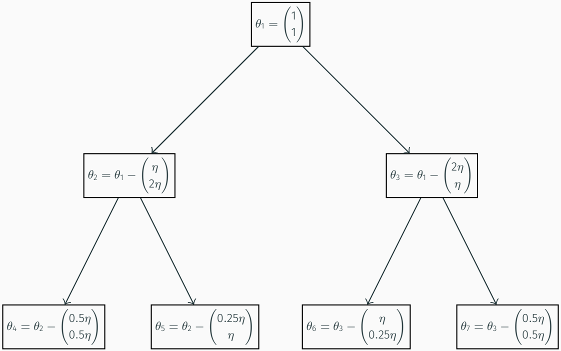


without correction for multiple testing

with correction for multiple testing

Seven simulated classifiers $C_1, \dots, C_7$ 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 is a noise term.

Simulation: Competitors

[Demšar, 2006] proposes a test for systematical differences between classifiers w.r.t. one single quality criterion.

We add two multidimensional adaptations of this test to our study:

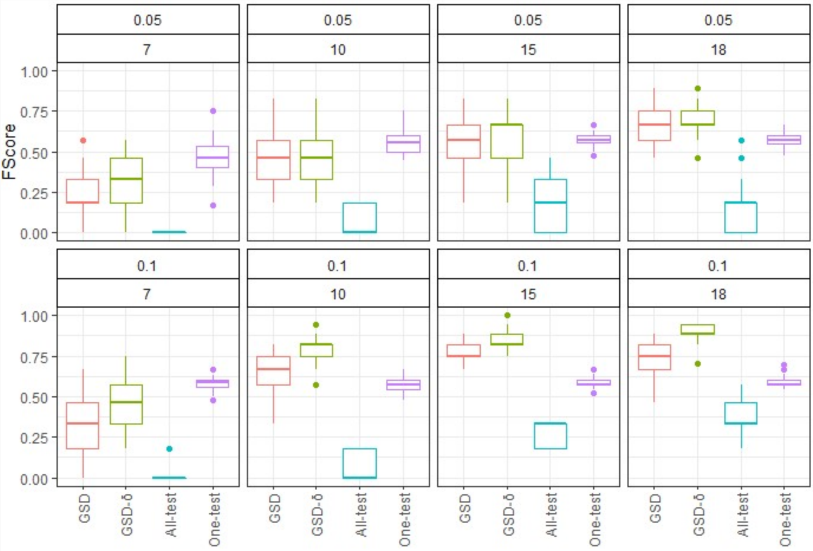
all-test: Classifier C_i is considered better than C_j if it performs significantly better on each quality criterion w.r.t. the above test.

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 enravelling the true ordering structure?

Simulation: Results (Bonferroni corrected)



Future Research

There are several promising directions for future research:

- Incorporating classification difficulty: Specifying data set specific loss functions in advance could account for classification difficulty.
- Reducing computational complexity for special cases: See if costs can be reduced if more constraints on the preference system are imposed.
- Extension to multi-criteria decision making: Our framework straightforwardly generalizes to multi-criteria decision problems under uncertainty.
- Robustifying comparisons: Framework can straightforwardly be extended to generalized uncertainty models, making comparisons more robust.

More Recent Work on Weakly Structured Information

State-dependent preference systems:

C. Jansen and T. Augustin (2022): Decision making with state-dependent preference systems. *Communications in Computer and Information Science*, vol 1601, Springer.

More Recent Work on Weakly Structured Information

State-dependent preference systems:

C. Jansen and T. Augustin (2022): Decision making with state-dependent preference systems. *Communications in Computer and Information Science*, vol 1601, Springer.

Risk analysis on weakly structured domains:

J. Baccelli, G. Schollmeyer and C. Jansen (2022): Risk aversion over finite domains. *Theory and Decision*, 93(3): 371 - 397.

More Recent Work on Weakly Structured Information

State-dependent preference systems:

C. Jansen and T. Augustin (2022): Decision making with state-dependent preference systems. *Communications in Computer and Information Science*, vol 1601, Springer.

Risk analysis on weakly structured domains:

J. Baccelli, G. Schollmeyer and C. Jansen (2022): Risk aversion over finite domains. *Theory and Decision*, 93(3): 371 - 397.

Statistical models for partial orders:

H. Blocher, G. Schollmeyer and C. Jansen (2022): Statistical models for partial orders based on data depth and formal concept analysis. *Communications in Computer and Information Science*, vol 1602, Springer.

More Recent Work on Weakly Structured Information

State-dependent preference systems:

C. Jansen and T. Augustin (2022): Decision making with state-dependent preference systems. *Communications in Computer and Information Science*, vol 1601, Springer.

Risk analysis on weakly structured domains:

J. Baccelli, G. Schollmeyer and C. Jansen (2022): Risk aversion over finite domains. *Theory and Decision*, 93(3): 371 - 397.

Statistical models for partial orders:

H. Blocher, G. Schollmeyer and C. Jansen (2022): Statistical models for partial orders based on data depth and formal concept analysis. *Communications in Computer and Information Science*, vol 1602, Springer.

Uncertainty quantification in decision making:

C. Jansen, G. Schollmeyer and T. Augustin (2022): Quantifying Degrees of E-Admissibility in Decision Making with Imprecise Probabilities. *Theory and Decision Library A*, vol 54. Springer.

Augustin, T., Coolen, F., de Cooman, G., and Troffaes, M., editors (2014).
Introduction to Imprecise Probabilities.
Wiley, Chichester.

Bradley, S. (2019).
Aggregating belief models.
In *Proceedings of ISIPTA 2019*, Proceedings of Machine Learning Research.

Demšar, J. (2006).
Statistical comparisons of classifiers over multiple data sets.
The Journal of Machine Learning Research, 7:1–30.

Dua, D. and Graff, C. (2017).
UCI machine learning repository.

Kikuti, D., Cozman, F., and Filho, R. (2011).
Sequential decision making with partially ordered preferences.
Artificial Intelligence, 175:1346 – 1365.

Krantz, D., Luce, R., Suppes, P., and Tversky, A. (1971).
Foundations of Measurement. Volume I: Additive and Polynomial Representations.
Academic Press, San Diego and London.

Levi, I. (1974).
On indeterminate probabilities.
The Journal of Philosophy, 71:391–418.

Mosler, K. and Scarsini, M. (1991).
Some theory of stochastic dominance.
In Mosler, K. and Scarsini, M., editors, *Stochastic Orders and Decision under Risk*, pages 203–212. Institute of Mathematical Statistics, Hayward, CA.

Nau, R. (2006).
The shape of incomplete preferences.
Annals of Statistics, 34:2430––2448.

Pratt, J. and Gibbons, J. (2012).
Concepts of Nonparametric Theory.
Springer.

Savage, L. (1954).
The Foundations of Statistics.
Wiley.

Seidenfeld, T., Kadane, J., and Schervish, M. (1995).
A representation of partially ordered preferences.
Annals of Statistics, 23:2168–2217.

Shaker, M. H. and Hüllermeier, E. (2021).
Ensemble-based uncertainty quantification: Bayesian versus credal inference.
CoRR, abs/2107.10384.

Troffaes, M. (2007).
Decision making under uncertainty using imprecise probabilities.
International Journal of Approximate Reasoning, 45:17–29.

von Neumann, J., Morgenstern, O., Kuhn, H., and Rubinstein, A. (1944).
Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition).
Princeton University Press.

Walley, P. (1991).
Statistical Reasoning with Imprecise Probabilities.
Chapman and Hall, London.

Weichselberger, K. (2001).
Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept.
Physica.