A Model for Accountable Ordinal Sorting

Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, Wassila Ouerdane

▶ **To cite this version:**

Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, Wassila Ouerdane. A Model for Accountable Ordinal Sorting. 26th International Joint Conference on Artificial Intelligence (IJCAI-2017), Aug 2017, Melbourne, Australia. hal-01523779

HAL Id: hal-01523779
https://hal.science/hal-01523779

Submitted on 15 Oct 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Model for Accountable Ordinal Sorting

K. Belahcene[1], **C. Labreuche**[2], **N. Maudet**[3], **V. Mousseau**[1], **W. Ouerdane**[1]
[1]LGI, CentraleSupélec, Université Paris-Saclay, Chatenay Malabry, France.
[2]Thales Research & Technology, 91767 Palaiseau Cedex, France.
[3]Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 75005 Paris.
khaled.belahcene, vincent.mousseau, wassila.ouerdane@centralesupelec.fr
christophe.labreuche@thalesgroup.com
nicolas.maudet@lip6.fr

Abstract

We address the problem of multicriteria ordinal sorting through the lens of *accountability*, *i.e.* the ability of a human decision-maker to own a recommendation made by the system. We put forward a number of model features that would favor the capability to support the recommendation with a convincing explanation. To account for that, we design a recommender system implementing and formalizing such features. This system outputs explanations under the form of specific *argument schemes* tailored to represent the specific rules of the model. At the end, we discuss possible and promising argumentative perspectives.

1 Introduction

While algorithmic automated decisions or recommendations are nowadays pervasive, there is a growing demand of institutions and citizens to make these recommendations *transparent* and *trustworthy*, while system designers seek *persuasive* recommendations [Tintarev, 2007]. The recent regulation adopted by the European Parliament (known as the General Data Protection Regulation, GDPR) goes further by adding a "right to explanation". According to [Goodman and Flaxman, 2016] *"the GDPR's requirements could require a complete overhaul of standard and widely used algorithmic techniques"*. We interpret this requirement in the strong sense of *accountability*, its litmus test being the ability of the recipient of the recommendation to defend it before other, skeptical, stakeholders of the decision (whereas *trust* requires the recommendation to be consistently accurate, but eventually asks for delegation of the decision to the system; *transparency* simply provides access to the underlying algorithm without concern for technical literacy [Burell, 2016]; and *persuasiveness* is hardly transferable: someone persuaded by a recommendation may not be a good persuader).

Our aim in this paper is thus to build an accountable, ordinal, multicriteria classifier, mapping a *candidate* object to a *recommendation* consisting in one or more categories among a predefined, ordered collection of these. In a multicriteria decision aiding (MCDA) context, the only indisputable relation between objects is the Pareto dominance, occurring when an object outperforms another on all criteria. As the situation is seldom so clear, the rules permitting the comparison of objects need to be enriched, taking into account the knowledge and values of the decision-maker, collected under the label *preference information*, which is also considered as an input of the classifier. We also consider an additional output, an *explanation* aimed at the decision-maker, supporting the recommendation and enabling the accountability sought for. In order to reach this goal of accountability, we make two important assumptions about the recommender system. These *design principles* are as follows:

- *No jargon*. A first step in a MCDA process is to collect decision-maker's preferences information. In order to accurately represent the specific decision process, we opt for an indirect elicitation [Dias *et al.*, 2002]: the decision-maker is never asked any questions about artifacts of the model (e.g. weights). Instead she should express preferences directly in the language of the actual decision situation, i.e. providing direct assignments of typical examples, *reference objects*, to categories.
- *No arbitrariness*. MCDA usually proceeds by representing the reasoning of the decision-maker with a formal parametric model, describing a specific stance. The values of the *preference parameters* are often fitted during an elicitation process, up to a certain point. While many methods proceed by picking a specific, so-called *representative* value of the parameters, we opt for a *robust* approach (to the lack of preference information) [Vincke, 1999; Greco *et al.*, 2008], formulating a –possibly partial – recommendation that cannot be refuted by any judgment function consistent with the preference information.

On top of these principles, we make three further assumptions about the MCDA model, proceeding from the willingness to keep the model accessible to human reasoning.

- *No compensation*. This assumption deals with the interpretation of collected data –the evaluation of objects on various criteria. We assume they are always used comparatively, in a purely ordinal manner: on a given criterion, an alternative is either as good as another one, or strictly worse. Hence, only the *set* of criteria for which an alternative is better is important, regardless of the specific values, and being very good on some criterion cannot compensate for low performance on others. This feature enables the algorithm to proceed without performing any algebraic

computation, which makes it particularly suited for explanation. It is shared with established non-compensatory ordinal sorting models used in the field of MCDA (eg. NCS) [Bouyssou and Marchant, 2007]. Moreover, the use of a 2-valued comparison ($\geq, <$) is similar to [Bouyssou, 1986] rather than [Fishburn, 1976] who proposes a 3-valued one ($<, =, >$).

- *No values.* At the heart of the recommender system is a *preference structure* encoding the comparison of alternatives. There are two main families of structures: those based on *value* [Keeney and Raiffa, 1976], and those based on *outranking relations* [Roy, 1991]. We opt for the latter, as they eschew the construction of a scoring function. An outranking relation naturally provides four outcomes when comparing two alternatives: preference for the former, for the latter, indifference, or incomparability; also, it does not enforce transitivity of preference.

- *No frontiers*. In MCDA, most classifiers link the preference structure and the recommendation of a class by introducing an explicit frontier between classes, defining the limit of each class (a single value for value-based models, a limiting profile for outranking-based ones, e.g. [Leroy *et al.*, 2011]). We do without this construct, as for instance models based on Logical Analysis of Data (LAD) techniques [Crama *et al.*, 1988] which output classification rules. We shall use simple rules permitting to classify a new object by comparing it to a set of already classified *reference objects* (see Sect.2.3).

The general philosophy of these principles must be clear to the reader: accountability should exclude in principle the use of any model artifact that the decision-maker may not properly handle, but at the same time provide enough understanding of the model so as to allow the decision-maker to defend the recommendation *as if it was her own*. Following this, our approach is to enforce these principles by design, and to investigate how far we can get with the resulting sorting model. This approach differs from the recent work of [Ribeiro *et al.*, 2016] which adopts a model-agnostic approach, and builds explanations adapted to virtually any classifier. They obtain extremely promising results in terms of trust. As expected, the explanation cannot be fully faithful to the model (they are "locally" faithful though). It also differs from [Datta *et al.*, 2016] which seeks to extract how influential are input parameters, but keeping a black-box access to the model. While for the trust requirement these approaches are sufficient, our notion of accountability requires to get to grips with the model.

The rest of this paper is as follows. We propose a model implementing and formalizing the different principles, decomposing it in a learning phase (Section 2) and a recommendation phase (Section 3). We provide formal explanations of the recommendation in most cases, in the form of *argument schemes* tailored to represent the specific rules of the model. Section 4 introduces some insights on the description of the sorting problem through an argumentation system. Section 5 concludes the paper, by putting its findings into perspective.

2 Formal description

In this section, we define a recommender system following the design principles and assumptions, and describe some of its properties.

2.1 The recommender system

We consider a multicriteria ordinal sorting problem : a collection of objects are evaluated on a set of criteria N. We note $\mathbb{B} := \{0, 1\}$, so that elements of $\mathbb{B}^N$ are at the same time vectors with binary coordinates, and subsets of N, partially ordered by inclusion. The maximal element of $\mathbb{B}^N$ is the unanimous coalition N, also denoted $(1, \ldots, 1)$. The minimal element of $\mathbb{B}^N$ is the empty coalition $\emptyset$, also denoted $(0, \ldots, 0)$. Each criterion $i \in N$ maps an object to a performance value in a totally ordered set $\mathbb{X}_i$, the higher the better. Consequently, each object is described by a performance vector in the partially ordered set $\mathbb{X} = \prod_{i \in N} \mathbb{X}_i$. The objects are to be assigned to some class chosen among an ordered set $\mathbb{K} = \{k_1 \prec \cdots \prec k_p\}$, so that assignment to a class with a high index is desirable.

Formally, let us describe the recommender system as a function mapping a pair $\langle z, \mathcal{P} \rangle$ to a pair $\langle K, \mathcal{E} \rangle$, where:

- The object $z \in \mathbb{X}$ is a *candidate* for sorting;
- $\mathcal{P}$ denotes *preference information* collected from the decision-maker consisting of typical classification examples, a collection of *reference objects* $\mathbb{X}^\star \subset \mathbb{X}$, and their assigned categories $Class : \mathbb{X}^\star \to \mathbb{K}$. For syntactic reasons, we represent it by a set of object-assignment pairs $\mathcal{P} \subset \mathbb{X} \times \mathbb{K}$.

$$\mathcal{P} := \bigcup_{x^\star \in \mathbb{X}^\star} (x^\star, Class(x^\star))$$

- $K \subset \mathbb{K}$ is the *recommendation*, concerning the classes that could be assigned to the candidate (see Sect. 3);
- $\mathcal{E}$ is an *explanation* yet unspecified, supporting the recommendation K (see for instance [Labreuche *et al.*, 2012; Belahcene *et al.*, 2017]), and addressed by Sect. 3.

Example 1. *Objects are evaluated according to four criteria* a, b, c, d *(higher is better). Six reference objects:* $\mathbb{X}^\star := \{A_1, A_2, B_1, B_2, C_1, C_2\}$*, described by the performance table below, are assigned to three classes:* $\mathbb{K} := \{\star \prec \star\star \prec \star\star\star\}$ *and make up the preference information* $\mathcal{P}$*. We consider two candidates:* X, Y *and try to assign them to some possible classes.*

Object	a	b	c	d	*Assignment*
A_1	3	3	2.5	0	$\star\star\star$
A_2	3	2	2.1	1	$\star\star\star$
B_1	2	2	1.3	1	$\star\star$
B_2	3	1	3.7	0	$\star\star$
C_1	2	1	1.6	1	$\star$
C_2	1	1	4.1	0	$\star$
X	2	2	1.1	0	?
Y	2	3	1.8	0	?

2.2 The reasoning of the decision-maker

A non-compensatory outranking relation can be represented by a Boolean composite function:

$$\forall x, y \in \mathbb{X},\ x S_\phi y \iff \phi \circ O_N(x, y) = 1$$

where the *observation function* O_N maps a pair of objects to its *concordance set*, and the consistent judgment of the decision-maker, based on these concordance sets, is represented by the *judgment function* ϕ mapping a concordance set to a truth value.

$$\begin{array}{rccl} O_N: & \mathbb{X}\times\mathbb{X} & \to & \mathbb{B}^N \\ & (x,y) & \mapsto & \{i\in N : x_i \geq y_i\} \end{array}$$

Antecedents of 1 by ϕ, called *true points* in the language of the LAD [Crama *et al.*, 1988], represent *sufficient coalitions of criteria*, while antecedents of 0 by ϕ are *false points* or *insufficient coalitions of criteria*. ϕ is supposed *non-decreasing*, meaning that a superset of a sufficient coalition of criteria is also sufficient, and a subset of an insufficient coalition is also insufficient. Compatibility of the outranking relation S to the Pareto dominance imposes that a unanimous support of criteria is always sufficient, so $\phi(N) = 1$. Conversely, $\phi(\emptyset) = 0$ must hold, so the relation S is not reduced to generalized indifference. Finally, we define the set of any possible judgment function :

$$\phi \in \widehat{\Phi} := \{\phi : \mathbb{B}^N \to \mathbb{B} \ : \phi \nearrow \text{ and } \phi(N) = 1 \text{ and } \phi(\emptyset) = 0\}$$

2.3 Learning from the assignment examples

To assign a new object to a category, we shall use the following classification rules:

- *(R1)* an object cannot outrank any object assigned to a strictly better class;
- *(R2)* an object outranks objects assigned to a strictly worse class;
- *(R3)* objects in the same class can be in any position with respect to outranking.

To account for that, we first denote $\succsim_{\mathcal{P}}$ the complete preorder between reference objects induced by $\mathcal{P}$:

$$\left\{\begin{array}{rcl} x^\star \succsim_{\mathcal{P}} y^\star & \iff & Class(x^\star) \succsim Class(y^\star) \\ x^\star \succ_{\mathcal{P}} y^\star & \iff & Class(x^\star) \succ Class(y^\star) \\ x^\star \sim_{\mathcal{P}} y^\star & \iff & Class(x^\star) = Class(y^\star) \end{array}\right.$$

We consider the strict enforcement of the model rules for reference objects:

- *(R1)* : $\forall x^\star, y^\star \in \mathbb{X}^\star, x^\star \succ_{\mathcal{P}} y^\star \Rightarrow Not\ (y^\star S_\phi x^\star)$;
- *(R2)* : $\forall x^\star, y^\star \in \mathbb{X}^\star, x^\star \succ_{\mathcal{P}} y^\star \Rightarrow x^\star S_\phi y^\star$.

Hence, the assignment of reference objects expressed by $\mathcal{P}$ places upper (by *(R1)*) and lower (by *(R2)*) bounds upon the outranking relation between reference objects. so that:

$$\succ_{\mathcal{P}} \subseteq S_\phi \cap (\mathbb{X}^\star)^2 \subseteq \succsim_{\mathcal{P}}$$

These constraints transfer to the judgment functions. Each pair $(x^\star, y^\star)$ is mapped by the observation function O_N to a coalition of criteria. The observed coalitions $O_N(\mathbb{X}^\star \times \mathbb{X}^\star)$ serve as a learning set for the judgment function ϕ. They are sorted between three sets, yielding necessary conditions on ϕ:

- insufficient coalitions $O_N(\prec_{\mathcal{P}})$ should be mapped to 0;
- sufficient coalitions $O_N(\succ_{\mathcal{P}})$ should be mapped to 1;
- $O_N(\sim_{\mathcal{P}})$, which images by ϕ are not constrained.

Consequently, we define the set $\Phi(\mathcal{P})$ of judgment functions compatible to the preference information $\mathcal{P}$:

$$\Phi(\mathcal{P}) := \{\phi \in \widehat{\Phi} \ : \ \phi \circ O_N(\succ_{\mathcal{P}}) = 1 \text{ and } \phi \circ O_N(\prec_{\mathcal{P}}) = 0\}$$

Example 2. *(ex. 1 continued) In the following table, we detail all the relevant observed coalitions. Sufficient coalitions appear in the upper right side, boldfaced, while insufficient coalitions are in the lower left side.* ***N*** *stands for unanimity, which is self-explanatory.*

	$\star\star\star$		$\star\star$		$\star$	
	A_1	A_2	B_1	B_2	C_1	C_2
A_1	–	–	**abc**	**abd**	**abc**	**abd**
A_2	–	–	**N**	**abd**	**N**	**abd**
B_1	d	bd	–	–	**abd**	**abd**
B_2	acd	ac	–	–	**abc**	**abd**
C_1	d	d	acd	bd	–	–
C_2	cd	c	c	bcd	–	–

2.4 Consistency of judgment

The set $\Phi(\mathcal{P})$ is empty if, and only if, Pareto dominance is contradicted ($\exists\ x^\star, y^\star \in \mathbb{X}^\star,\ \forall i \in N,\ x_i^\star \geq y_i^\star$ and $Class(x^\star) < Class(y^\star)$), or some coalition of criteria $M \in \mathbb{B}^N$ observed as being sufficient is weaker (for inclusion) than some coalition $M' \in \mathbb{B}^N$ observed as being insufficient. In such a case, we call the preference information $\mathcal{P}$ *inconsistent*; otherwise, it is consistent and $\Phi(\mathcal{P})$ is a *partially defined Boolean function* [Crama *et al.*, 1988]. Combining the constraints on the judgment functions expressed by $\widehat{\Phi}$ and by $\mathcal{P}$, we can compute the true points of $\Phi(\mathcal{P})$. They are the antecedents of 1 common to every judgment function $\phi \in \Phi(\mathcal{P})$, and represent the coalitions *established as sufficient*, by the virtue of being at least as strong as an observed sufficient coalition.

$$\mathcal{T}_{\mathcal{P}} := \{t \in \mathbb{B}^N : \exists\ t_{obs} \in O_N(\succ_{\mathcal{P}}),\ t_{obs} \subseteq t\}$$

Conversely, the false points are the antecedents of zero common to every $\phi \in \Phi(\mathcal{P})$ and represent the coalitions established as insufficient.

$$\mathcal{F}_{\mathcal{P}} := \{f \in \mathbb{B}^N : \exists\ f_{obs} \in O_N(\prec_{\mathcal{P}}),\ f_{obs} \supseteq f\}$$

Proposition 1 details three manners to express inconsistency:

Proposition 1. *For any $\mathcal{P} \subset \mathbb{X}\times\mathbb{K}$, the three following conditions are equivalent and characterize inconsistency:*

1. *Absence of compatible judgment function:* $\Phi(\mathcal{P}) = \emptyset$
2. *Conflicting constraints:* $\mathcal{T}_{\mathcal{P}} \cap \mathcal{F}_{\mathcal{P}} \neq \emptyset$
3. *Explicit contradiction:* $\exists t \in O_N(\succ_{\mathcal{P}}), \exists f \in O_N(\prec_{\mathcal{P}}) : t \subseteq f$

Example 3. *(ex. 2 continued) Coalitions are sorted according to the observations, and monotonicity:*
$O_N(\succ_{\mathcal{P}}) = \{N, abc, abd\} = \mathcal{T}_{\mathcal{P}}$
$O_N(\prec_{\mathcal{P}}) = \{c, d, ac, bd, cd, acd, bcd\}$
$\mathcal{F}_{\mathcal{P}} = \{\emptyset, a, b, c, d, ac, ad, bc, bd, acd, bcd\}$
There is no dispute, as $\mathcal{T}_{\mathcal{P}} \cap \mathcal{F}_{\mathcal{P}} = \emptyset$, but the coalition ab is left undecided.

3 Recommendations and explanations

In the previous section, we saw how the decision-maker interprets pairwise comparisons between reference objects belonging to different classes as sufficient or insufficient coalitions of criteria. Here comes a new candidate, $z \in \mathbb{X}$. It

gauges every reference object in $\mathbb{X}^\star$, yielding $|\mathcal{P}|$ observations $\overrightarrow{o}(z,\mathcal{P}) := \bigcup_{x^\star \in \mathbb{X}^\star} O_N(z, x^\star)$, and is also evaluated by every reference object, yielding $|\mathcal{P}|$ other observations $\overleftarrow{o}(z,\mathcal{P}) := \bigcup_{x^\star \in \mathbb{X}^\star} O_N(x^\star, z)$. Each of these $2|\mathcal{P}|$ observations is interpreted as a *sufficient, insufficient* or *undecided* coalition of criteria.

Example 4. *(ex. 3 continued) The following table augments the one presented in example 2 with the coalitions resulting from comparisons between the reference objects $A_1, A_2, B_1, B_2, C_1, C_2$ and the candidates X, Y.*

	$\star\star\star$		$\star\star$		$\star$	
	A_1	A_2	B_1	B_2	C_1	C_2
A_1	$-$	$-$	**abc**	**abd**	**abc**	**abd**
A_2	$-$	$-$	**N**	**abd**	**N**	**abd**
B_1	d	bd	$-$	$-$	**abd**	**abd**
B_2	acd	ac	$-$	$-$	**abc**	**abd**
C_1	d	d	acd	bd	$-$	$-$
C_2	cd	c	c	bcd	$-$	$-$

	?	?
	X	Y
A_1	**N**	**N**
A_2	**N**	acd
B_1	**N**	ad
B_2	acd	acd
C_1	acd	ad
C_2	cd	cd

X	d	b	(ab)	bd	(ab)	**abd**
Y	bd	b	**abc**	bd	**abc**	**abd**

Non-bracketed coalitions have already been sorted according to the preference information: boldfaced coalitions are those previously established as sufficient, the others are insufficient. Bracketed coalitions are yet undecided. $\forall z \in \{X, Y\}$, $\overrightarrow{o}(z,\mathcal{P})$ appears in the corresponding line, and $\overleftarrow{o}(z,\mathcal{P})$ in the appropriate column.

In this section, we specify the mapping between these observations and the output of the classifier system, the *recommendation* $K(z,\mathcal{P}) \subset \mathbb{K}$ and an *explanation* $\mathcal{E}(k,\mathcal{P})$ supporting it.

3.1 Possible assignments

As defined by the works of [Greco *et al.*, 2010] about *necessary* and *possible* preference relations, the definition of *possible assignments* is closely related to the notion of *consistency* of an assignment with respect to the corpus of preference information. Defining, as we did in Section 2, $\Phi(\mathcal{P})$ as the set of preference parameters compatible to $\mathcal{P}$, and assuming it is not empty:

- *necessary* assignments are yielded by *every* possible completion of these preference parameters;
- *possible* assignments are yielded by *some* possible completion of these preference parameters;
- *impossible* assignments are yielded by *no* possible completion of these preference parameters;

These sets of assignments are concisely described referring to the set:

$$\widehat{K}(z,\mathcal{P}) := \{k \in \mathbb{K} :\ \Phi(\mathcal{P} \cup \{(z,k)\}) \neq \emptyset\}$$

A possible assignment is in $\widehat{K}(z,\mathcal{P})$, an impossible one is not. When $\widehat{K}(z,\mathcal{P})$ boils down to a singleton, then it is a necessary assignment for z.

This definition of *possible assignment* is straightforward to implement, simply iterating through the set of possible assignments classes $k \in \mathbb{K}$, updating the preference information $\mathcal{P}' \leftarrow \mathcal{P} \cup \{(z,k)\}$, and checking the consistency of $\mathcal{P}'$. Unfortunately, it is a tricky notion when it comes to explaining. The actual unveiling of a Boolean judgment function compatible to the assignment is not very appealing, as it introduces at the same time elements of *jargon* –describing the judgment of the decision-maker as the partition of coalitions of criteria between sufficient and insufficient– and *arbitrariness*, as some coalitions may very well be undecided and should remain so. Consequently, we adopt the following principle: *"Everything is possible, unless proven otherwise"*.

Doing so shifts the burden of proof towards impossibility, focusing on the exhibition of constraints restricting the set $\widehat{K}(z,\mathcal{P})$. We aim at *explaining* these constraints thanks to *statements* of the form $[premises : conclusions]_{scheme}$. We define several *argument schemes*, as formalized by [Walton, 1996] in order to capture stereotypical patterns of human reasoning. These schemes specify the nature and conditions imposed to both premises and conclusions, yielding to valid arguments. We are looking for *complete* explanations, so we must ensure the validity of the implication $premises \Rightarrow conclusions$, and provide *grounded* sets of statements, such that any premise is either the conclusion of another argument, or directly referencing the assumed available information (pairwise comparisons between the reference objects or the candidate, based on criteria or assignment).

In order to make apparent the cause of impossibility, we consider the potential consequences of assigning a candidate to a class through the *additional (in)sufficient coalitions conditional to the assignment of the candidate z to the class k*:

$$\mathbf{\Delta}\mathcal{T}_\mathcal{P}(z,k) := \mathcal{T}_{\mathcal{P}\cup\{(z,k)\}} \backslash \mathcal{T}_\mathcal{P};\ \mathbf{\Delta}\mathcal{F}_\mathcal{P}(z,k) := \mathcal{F}_{\mathcal{P}\cup\{(z,k)\}} \backslash \mathcal{F}_\mathcal{P}$$

We rewrite the impossibility of assigning the candidate z to the class k using the *conflicting constraints* characterization of inconsistency (see Prop. 1). We consider three potential sources of impossibility, sorted by evidence: $\widehat{K}(z,\mathcal{P}) = \bigcap_{i\in\{1,2,3\}} K_i(z,\mathcal{P})$ where:

- $K_1(z,\mathcal{P}) := \{k \in \mathbb{K} : \mathcal{T}_\mathcal{P} \cap \Delta\mathcal{F}_\mathcal{P}(z,k) = \emptyset\}$ highlights conflicts between established sufficient coalitions, and the assignment of z;
- $K_2(z,\mathcal{P}) := \{k \in \mathbb{K} : \Delta\mathcal{T}_\mathcal{P}(z,k) \cap \mathcal{F}_\mathcal{P} = \emptyset\}$ highlights conflicts between established insufficient coalitions, and the assignment of z;
- $K_3(z,\mathcal{P}) := \{k \in \mathbb{K} : \Delta\mathcal{T}_\mathcal{P}(z,k) \cap \Delta\mathcal{F}_\mathcal{P}(z,k) = \emptyset\}$ takes into account the least obvious situation where some assignment of z may be self-contradictory, without conflicting with any previously acknowledged information.

The next section details the impossibilities captured by the set $K_1(z,\mathcal{P})$, and proposes a supporting explanation $\mathcal{E}_1(z,\mathcal{P})$, while the other cases are briefly presented in section 3.3.

3.2 Assignments contradicting previously established sufficient coalitions

In this section, we focus on the set $K_1(z,\mathcal{P}) := \{k \in \mathbb{K} : \mathcal{T}_\mathcal{P} \cap \Delta\mathcal{F}_\mathcal{P}(z,k) = \emptyset\}$. As seen in the previous section this set provides a range of possible assignments for the candidate z, and partially implements the model described by the manifesto exposed in the introduction. We first describe $K_1(z,\mathcal{P})$

as an intersection of constraints, for which we provide a description based on arguments. We prove $K_1(z, \mathcal{P})$ is an interval of $\mathbb{K}$, and provide a short, yet complete, explanation accounting for this recommendation.

For increased readability, we introduce notations for particular sets of classes. For $k \in \mathbb{K}$, let $\mathbb{K}_{\precsim k}$ (resp. $\mathbb{K}_{\succsim k}$) the interval of classes not greater (resp. not lower) than k.

By construction, the recommended set $K_1(z, \mathcal{P})$ is built in order to reject some impossible assignments. To illustrate and understand its behavior, we make up a situation that specifically triggers this rejection flag. Suppose we know that:

(1) the coalition of criteria $T \in \mathbb{B}^N$ is already known to be sufficient, and

(2) the candidate $z \in \mathbb{X}$ is at least as good as the reference object $\underline{x}^\star \in \mathbb{X}^\star$, assigned to class $\underline{k} \in \mathbb{K}$, for all criteria in T.

Then, z outranks $\underline{x}^\star$ and cannot be assigned to a class strictly worse than $\underline{k}$ by application of (R1). This constraint is captured by the set $K_1(z, \mathcal{P})$, as the assignment of z to any class $k \prec \underline{k}$ would lead to conclude that the coalition of criteria $O_N(z, \underline{x}^\star)$ is insufficient, so that the coalition of criteria T would belong to both sets $\Delta\mathcal{F}_\mathcal{P}(z, k)$ and $\mathcal{T}_\mathcal{P}$. Consequently, $k \notin K_1(z, \mathcal{P})$.
If we replace the assumption (2) by:

(2') the reference object $\overline{x}^\star \in \mathbb{X}^\star$, assigned to class $\overline{k} \in \mathbb{K}$, is at least as good as the candidate $z \in \mathbb{X}$ for all criteria in T.

then $\overline{x}^\star \in \mathbb{X}^\star$ outranks z and z cannot be assigned to a class strictly better than $\overline{k}$, as

$$k \succ \overline{k} \Rightarrow \mathcal{T}_\mathcal{P} \ni T \subseteq O_N(\overline{x}^\star, z) \in \Delta\mathcal{F}_\mathcal{P}(z, k) \Rightarrow k \notin K_1(z, \mathcal{P})$$

Reciprocally, any assignment $k_0 \notin K_1(z, \mathcal{P})$ results in a non-empty intersection $\mathcal{T}_\mathcal{P} \cap \Delta\mathcal{F}_\mathcal{P}(z, k_0)$, which involves at least one sufficient coalition $T \in \mathcal{T}_\mathcal{P}$, as in assumption (1), and one stronger, insufficient coalition resulting either from the observations $\overrightarrow{o}(z, \mathcal{P})$, as in assumption (2), or from $\overleftarrow{o}(z, \mathcal{P})$, as in (2').

A statement of type (1) needs to be backed by evidence, so we introduce two argument schemes:

Definition 1. *For any reference objects $a^\star, b^\star \in \mathbb{X}^\star$ and any coalition of criteria $T \in \mathbb{B}^N$, we say the tuple $[a^\star, b^\star : T]_\mathcal{T}$ instantiates the argument scheme* SUFFICIENT COALITION$(\mathcal{P})$ *if, and only if, $T \supseteq O_N(a^\star, b^\star)$ and $a^\star \succ_\mathcal{P} b^\star$. We also say the tuple $[\emptyset : N]_1$ instantiates the argument scheme* WEAK DOMINANCE.

Proposition 2 (Argumentative structure of the sufficient coalitions)**.**

$$\mathcal{T}_\mathcal{P} = \{N\} \cup \bigcup_{[a^\star, b^\star : T]_\mathcal{T}} \{T\}$$

The sufficient coalitions are exactly the conclusions of the arguments instantiating the SUFFICIENT COALITION$(\mathcal{P})$ *scheme.*

In order to account for the atoms of reasoning (2) and (2') and present them to the recipient of the recommendation, we define the corresponding argument schemes.

Definition 2. *For any coalition of criteria $T \in \mathbb{B}^N$, any reference object $x^\star \in \mathbb{X}^\star$ and any class $c \in \mathbb{K}$, we say that:*

- *the tuple $[T, x^\star : \mathbb{K}_{\succsim c}]_{\mathcal{T}/\overrightarrow{o}}$ instantiates the argument scheme* OUTRANKING$(z, \mathcal{P})$ *if, and only if, $T \in \mathcal{T}_\mathcal{P}$ and $\forall i \in T, z_i \geq x^\star_i$ and $class(x^\star) = c$.*
- *the tuple $[T, x^\star : \mathbb{K}_{\precsim c}]_{\mathcal{T}/\overleftarrow{o}}$ instantiates the argument scheme* OUTRANKED$(z, \mathcal{P})$ *if, and only if, $T \in \mathcal{T}_\mathcal{P}$ and $\forall i \in T, x^\star_i \geq z_i$ and $class(x^\star) = c$*

Proposition 3 (Argumentative structure of the recommendation)**.**

$$K_1(z, p) = \mathbb{K} \cap \bigcap_{[T, \underline{x}^\star : \underline{k}]_{\mathcal{T}/\overrightarrow{o}}} \mathbb{K}_{\succsim \underline{k}} \cap \bigcap_{[T, \overline{x}^\star : \overline{k}]_{\mathcal{T}/\overleftarrow{o}}} \mathbb{K}_{\precsim \overline{k}}$$

Proposition 3 is a concise rewording of the necessary and sufficient conditions for a given class *not* to belong to the set $K_1(z, \mathcal{P})$ detailed previously. As a corollary, it shows that $K_1(z, \mathcal{P})$ is an interval of $\mathbb{K}$. Consequently, $K_1(z, \mathcal{P})$ can be completely described by a pair $(\underline{k_1}, \overline{k_1})$ such that:

- $K_1(z, \mathcal{P}) = \mathbb{K}_{\succsim \underline{k_1}} \cap \mathbb{K}_{\precsim \overline{k_1}}$
- the lower bound $\underline{k_1}$ is *maximal*, as there is no class strictly better than $\underline{k_1}$ which is supported by an argument instantiating the OUTRANKING$(z, \mathcal{P})$ scheme. It is *trivial* if $\underline{k_1} = \min \mathbb{K}$ (either when the set OUTRANKING$(z, \mathcal{P})$ is empty, or when it does not support a stronger outcome), in which case it does not need any explanation. If $\underline{k_1} \succ \min \mathbb{K}$, then it admits at least one *explanation* $\underline{E_1}$ composed of an argument $[T, \underline{x}^\star : \mathbb{K}_{\succsim \underline{k_1}}]_{\mathcal{T}/\overrightarrow{o}} \in$ OUTRANKING backed by an argument $[a^\star, b^\star : T]_\mathcal{T} \in$ SUFFICIENT COALITION;
- the upper bound $\overline{k_1}$ is *minimal*, as there is no class strictly worse than $\overline{k_1}$ which is supported by an argument instantiating the OUTRANKED$(z, \mathcal{P})$ scheme. It is *trivial* if $\overline{k_1} = \max \mathbb{K}$, in which case it does not need any explanation. If $\underline{k_1} \prec \max \mathbb{K}$, then it admits at least one *explanation* $\overline{E_1}$ composed of an argument $[T', \overline{x}^\star : \mathbb{K}_{\precsim \overline{k_1}}]_{\mathcal{T}/\overleftarrow{o}} \in$ OUTRANKED backed by an argument $[a^\star, b^\star : T']_\mathcal{T} \in$ SUFFICIENT COALITION.

Finally, the recommended interval $K_1(z, \mathcal{P})$ is supported by an explanation $\mathcal{E}_1$ in the form of a pair $(\underline{E_1}, \overline{E_1})$, where $\underline{E_1}$ and $\overline{E_1}$ can be either the empty set or a pair of arguments. Taken together, all these 0, 2 or 4 arguments prove that any assignment $k \in \mathbb{K} \setminus K_1(z, \mathcal{P})$ should be rejected as "impossible". Such explanation is not necessarily unique, and we denote by $\widehat{\mathcal{E}_1}(z, \mathcal{P})$ the set of suitable explanations.

Example 5. *(ex. 4 continued)*

Using the table presented in Example 4, the set K_1 can be interpreted as "a candidate cannot be assigned a class laying strictly on the right of, nor a class strictly above, a case containing a boldfaced coalition": Consequently,

- $\begin{cases} K_1(X, \mathcal{P}) = \{\star, \star\star\} \\ \mathcal{E}_1(X, \mathcal{P}) \ni (\emptyset, \{[\emptyset : N]_1, [N, B_1 : \precsim \star\star]_{\mathcal{T}/\overleftarrow{o}}\}) \end{cases}$
  X cannot be ranked higher than $\star\star$, because B_1 is rated $\star\star$ and dominates X.

- $\begin{cases} K_1(Y,\mathcal{P}) = \{\star\star, \star\star\star\} \\ \widehat{\mathcal{E}_1}(Y,\mathcal{P}) \ni (\{[A_1, C_1 : \mathtt{abc}]_{\mathcal{T}}, [\mathtt{abc}, B_1 : \succsim \star\star]_{\mathcal{T}/\vec{\sigma}}\}, \emptyset) \end{cases}$
Y cannot be ranked lower than $\star\star$, because it outranks B_1. Indeed, Y compares to B_1 the same way as A_1 to C_1: it is at least as good on the sufficient coalition of criteria $\mathtt{abc}$.

3.3 Other impossible assignments

The set $K_2(z,\mathcal{P})$ is defined symmetrically from $K_1(z,\mathcal{P})$ w.r.t. sufficient and insufficient coalitions. Assignments *not* in $K_2(z,\mathcal{P})$ result from the collision of a coalition of criteria known to be insufficient, and the observation of a candidate object resulting in an even weaker coalition, so outranking is excluded, and all the classes strictly above or below (depending on the direction of observation) the one of the reference object are therefore forbidden. *Mutatis mutandis*, we can define the argument schemes INSUFFICIENT COALITION($\mathcal{P}$), WEAKLY DOMINATED, NOT OUTRANKING($z,\mathcal{P}$), NOT OUTRANKED($z,\mathcal{P}$) and obtain the same structural results, leading to define similar explanations for the lower and upper bounds of the interval $K_2(z,\mathcal{P})$.

Example 6. *(ex. 4 continued)*
Using the table presented in Ex. 4, the set K_2 interprets the insufficient coalitions of the table, those not boldfaced nor parenthesized. A candidate cannot be assigned a class strictly below, nor strictly on the left, of such cases. For instance, $O_N(B_2, X) = \mathtt{acd} \in \mathcal{F}_{\mathcal{P}}$ (e.g. because $O_N(C_1 \prec_{\mathcal{P}} B_1) = \mathtt{acd}$), so X is not outranked by B_2 and should be at least assigned the same class ($\star\star$), and $O_N(X, B_2) = \mathtt{bd} \in \mathcal{F}_{\mathcal{P}}$ (e.g. because it is weaker than $\mathtt{bcd} = O_N(C_2 \prec_{\mathcal{P}} B_2)$), so X does not outrank B_2 and should not be assigned a strictly better class ($\star\star$). In terms of preference, objects X and B_2 are incomparable, and thus should be assigned the same class. Finally, $K_2(X,\mathcal{P}) = \{\star\star\}$.

The set $K_3(z,\mathcal{P})$ excludes inconsistent judgments on yet undecided coalitions of criteria. There is no guarantee that $K_3(z,\mathcal{P})$ has an interval structure. We omit this case due to space limitations.

4 An argumentative perspective

Along this paper, we proposed the construction of explanations supporting results of a multi-criteria sorting problem, as combinations of arguments schemes. Each instantiation of one of the six previous main schemes (see Def. 1, 2 and their symmetrical forms) provides one type of argument. These arguments may be conflicting, and two different relations can be distinguished:

Conflicting coalitions: we have evidence indicating that a given coalition is potentially at the same time sufficient and insufficient (i.e. there are two coalitions $t \subseteq f$ such that $[a^\star, b^\star : t]_{\mathcal{T}}$ and $[c^\star, d^\star : f]_{\mathcal{F}}$). This situation represents an explicit contradiction corresponding to an inconsistency situation (see Sec. 2.4). Such conflicts are not illustrated through the previous examples, however inconsistencies are classical situations within decision problems, as it concerns a human decision-maker.

Conflicting classification: it may occur that, for some candidate, arguments based on the outranking relation point towards an *empty* interval of possible assignments. This situation corresponds to the fact that the sets $K_1(z,\mathcal{P})$ and $K_2(z,\mathcal{P})$ are disjoint, which may happen when either is empty, or when the lower bound of one exceed the upper bound of the other.

Example 7. *(ex. 4 cont.) Y and A_2 are incomparable, Y and B_2 are incomparable, yet A_2 is preferred to B_2. In particular, A_2 ($\star\star\star$) does not outrank Y and Y does not outrank B_2 ($\star\star$) so $K_2(Y,\mathcal{P}) = \emptyset$.*

The impossibility to provide any recommendation is clearly critical from the point of view of decision aiding. These unfortunate situations cannot be ruled out in the general case, as they may stem from Condorcet paradoxes (failures of transitivity) concerning the necessary outranking relation or the necessary not-outranking relation (see e.g. [Köksalan *et al.*, 2009] for a discussion).

The argumentative treatment for our multi-criteria ordinal sorting problem is thus to construct arguments pro and against each possible assignment (of the reference object and the candidate), and to determine among conflicting arguments the *acceptable* ones. This can be done by taking two different perspectives. One way is to rely on the work of [Dung, 1995] - the next question being to identify which semantics are appropriate in our situation. This is close in spirit to an approach presented in [Amgoud and Serrurier, 2007] for classification in *unordered* classes (however in our context the relation between arguments would be symmetric [Coste-Marquis *et al.*, 2005]). Another perspective is to consider the construction of the argumentation system as a dialogue game and to rely on critical questions [Walton, 1996] to evaluate the arguments. This perspective has the advantage to keep the decision-maker in the loop, which is often essential in a decision situation [Labreuche *et al.*, 2015]. Both approaches look promising and are made possible thanks to the modeling presented in this paper.

5 Conclusion

We have presented a fully accountable multi-criteria ordinal sorting model, based on several design principles and assumptions. The strength of the model is that it solely relies on a simple set of classification rules, which means that each recommendation can be justified by instantiating and combining these rules–nothing else. Several argument schemes have been proposed for that purpose. Interestingly, some of these schemes have a flavour of analogical reasoning, which was studied in the context of classification [Hug *et al.*, 2016]. Now the simplicity of our model comes at a price: there are different situations where inconsistency might occur, and the model is not equipped yet to handle such situations. Facing this issue we can take two stances. The first one is to relax some of our design assumptions. For instance, we may decide that it is actually acceptable for the model to use a *frontier* between classes (allowing to eschew the Condorcet paradox). This would require original explanation techniques to maintain the desired accountability. Another avenue is to handle the inconsistencies thanks to defeasible and non-monotonic reasoning techniques [Brewka *et al.*, 2008]. Our discussion in Sect. 4 points to formal argumentation as a natural and promising opportunity for future research.

References

[Amgoud and Serrurier, 2007] L. Amgoud and M. Serrurier. Arguing and explaining classifications. In *Proceeding of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, page 160, 2007.

[Belahcene *et al.*, 2017] K. Belahcene, C. Labreuche, N. Maudet, V. Mousseau, and W. Ouerdane. Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision*, 82(2):151–183, 2017.

[Bouyssou and Marchant, 2007] D. Bouyssou and T. Marchant. An axiomatic approach to noncompensatory sorting methods in mcdm, i: The case of two categories. *EJOR*, 178(1):217–245, 2007.

[Bouyssou, 1986] D. Bouyssou. Some remarks on the notion of compensation in mcdm. *EJOR*, 26(1):150–160, 1986.

[Brewka *et al.*, 2008] G. Brewka, I. Niemelä, and M. Truszczynski. Nonmonotonic reasoning. In *Handbook of Knowledge Representation*, volume 3 of *Foundations of Artificial Intelligence*, pages 239–284. Elsevier, 2008.

[Burell, 2016] J. Burell. How the machine "thinks": understanding opacity in machine learning algorithms. *Big Data and Society*, 1(3), 2016.

[Coste-Marquis *et al.*, 2005] S. Coste-Marquis, C. Devred, and P. Marquis. Symmetric argumentation frameworks. In *Proceedings of the 8th European Conference Symbolic and Quantitative Approaches to Reasoning with Uncertainty ECSQARU*, pages 317–328. Springer, 2005.

[Crama *et al.*, 1988] Y. Crama, P. L. Hammer, and T. Ibaraki. Cause-effect relationships and partially defined boolean functions. *Annals of Operations Research*, 16(1):299–325, 1988.

[Datta *et al.*, 2016] A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *The 37th IEEE Symposium on Security and Privacy (Oakland)*, 2016.

[Dias *et al.*, 2002] L. Dias, V. Mousseau, J. Figueira, and J. Clímaco. An aggregation/disaggregation approach to obtain robust conclusions with electre tri. *EJOR*, 138(2):332–348, 2002.

[Dung, 1995] P. M. Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995.

[Fishburn, 1976] P. C. Fishburn. Noncompensatory preferences. *Synthese*, 33(2/4):393–403, 1976.

[Goodman and Flaxman, 2016] B. Goodman and S. Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". ArXiv e-prints: 1606.08813, June 2016.

[Greco *et al.*, 2008] S. Greco, V. Mousseau, and R. Słowiński. Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. *EJOR*, 191(2):416–436, 2008.

[Greco *et al.*, 2010] S. Greco, R. Słowinski, J. Figueira, and V. Mousseau. Robust ordinal regression. In *Trends in Multiple Criteria Decision Analysis*, pages 241–284. Springer Verlag, 2010.

[Hug *et al.*, 2016] N. Hug, H. Prade, G. Richard, and M. Serrurier. Analogical classifiers: A theoretical perspective. In *ECAI 2016 - 22nd European Conference on Artificial Intelligence*, pages 689–697, 2016.

[Keeney and Raiffa, 1976] R.L. Keeney and H. Raiffa. *Decisions with multiple objectives: Preferences and value tradeoffs*. J. Wiley, New York, 1976.

[Köksalan *et al.*, 2009] M. Köksalan, V. Mousseau, O. Ozpeynirci, and S. Bilgin Ozpeynirci. A new outranking-based approach for assigning alternatives to ordered classes. *Naval Research Logistics*, 56(1):74–85, 2009.

[Labreuche *et al.*, 2012] C. Labreuche, N. Maudet, and W. Ouerdane. Justifying Dominating Options when Preferential Information is Incomplete. In *ECAI'12*. IOS Press, 2012.

[Labreuche *et al.*, 2015] C. Labreuche, N. Maudet, W. Ouerdane, and S. Parsons. A dialogue game for recommendation with adaptive preference models. In *Proceedings of the 14th International Conference on Autonomous Agent and MultiAgent systems (AAMAS)*, pages 959–967, 2015.

[Leroy *et al.*, 2011] A. Leroy, V. Mousseau, and M. Pirlot. Learning the parameters of a multiple criteria sorting method. In *ADT*, pages 219–233. Springer, 2011.

[Ribeiro *et al.*, 2016] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[Roy, 1991] B. Roy. The outranking approach and the foundations of electre methods. *Theory and decision*, 31(1):49–73, 1991.

[Tintarev, 2007] N. Tintarev. Explanations of recommendations. In *Proc. ACM conference on Recommender systems*, pages 203–206, 2007.

[Vincke, 1999] Ph. Vincke. Robust solutions and methods in decision-aid. *Journal of multicriteria decision analysis*, 8(3):181, 1999.

[Walton, 1996] D. Walton. *Argumentation schemes for Presumptive Reasoning*. Mahwah, N. J.,Erlbaum, 1996.