Using Argumentation Frameworks to promote Fairness and Rationality in Multi-Experts Multi-Criteria Decision Making

Stefano Bistarelli[1], Martine Ceberio[2], Joel A. Henderson[2], and Francesco Santini[3]

[1] Department of Mathematics and Computer Science
University of Perugia
Via Vanvitelli, 1 – 06123 Perugia, Italy
bista@dmi.unipg.it

[2] Computer Science Department
The University of Texas at El Paso
500 West University – El Paso, TX 79968, USA
mceberio@utep.edu, jahenderson@miners.utep.edu

[3] Institute of Informatics and Telematics, CNR-Pisa
Via Moruzzi, 1 – 56124, Pisa, Italy
francesco.santini@iit.cnr.it

Abstract. In this work, we focus on multi-criteria decision making and in particular, in the case of multiple experts (ME-MCDM). The problem of making decisions when multiple (possibly conflicting) criteria are involved often boils down to identifying an aggregation function that will combine all appreciations of the multiple dimensions of the problem. In the case of Multiple Experts, decisions even already exist and the goals are to (1) make a decision based on the potentially multiple views of multiple experts and/or (2) use these decisions are informations about what the aggregation function of a "super" expert should be in the aim to make future decisions. Unfortunately, this line of approaches tends to overlook the irrationality and/or lack of fairness of experts, aggregating all available prior information regardless of quality.

In this work, we propose to model Multi-Experts Multi-Criteria Decision-Making (MEMCDM) problems using argumentation frameworks. We specifically design our proposed model so as to emulate fairness and rationality in decisions. For instance, when, of two expert's decisions, one is unfair, we impose an attack between these two decisions, forcing one of the two decisions out of the argumentation network's resulting extensions. Similarly, we specifically put irrational decisions in opposition to force one out. In doing so, we aim to enable the prediction of decisions that are themselves fair and rational. Our model is illustrated on two toy examples.

Keywords: Multi-Experts Multi-Criteria Decision Making, Disagreement, Fairness, Rationality, Argumentation Framework, Model.

1 Introduction

Expert analysis and decisions arguably provide high-quality and highly-valued support for action and policy making in a wide variety of fields, from social services, to medicine, to engineering, to grant funding committees, and so on. However, the use of experts can be prohibitive due to either lack of availability, high cost, or limited time frame for action – this is the case particularly more so in impoverished areas. As such, it is desirable to be able to replicate / predict such decisions when beneficial even in the absence of experts. Unfortunately there are many obstacles that still hinder an accurate simulation of expert decisions. First, it is hard to understand, and therefore replicate, the way each expert "aggregates" information/assessment along several criteria. In addition, even if we had a reasonable insight about it, any expert may make inconsistent decisions across similar scenarios. Finally, in the case of multiple experts, despite looking at the same information, two (or more) experts may disagree on the decisions to be made.

In spite of such challenges, traditional approaches seek to combine prior known decisions of experts into a classification of scenarios (machine learning approaches) or into some aggregation function that allows to best replicate the experts' decisions. Unfortunately, this line of approaches tends to overlook the irrationality and/or lack of fairness of experts, aggregating all available prior information regardless of quality.

In this work, we propose to model Multi-Experts Multi-Criteria Decision-Making (MEMCDM) problems using argumentation frameworks. We specifically design our proposed model so as to emulate fairness and rationality in decisions. For instance, when, of two expert's decisions, one is unfair, we impose an attack between these two decisions, forcing one of the two decisions out of the argumentation network's resulting extensions. Similarly, we specifically put irrational decisions in opposition to force one out. In doing so, we aim to enable the prediction of decisions that are themselves fair and rational. Our model is illustrated on two toy examples.

In what follows, we start by recalling preliminary notions, then we proceed with describing our model in details and illustrate our model in the case of Software Quality Assessment by multiple experts along multiple criteria.

2 Preliminary Notions

2.1 Multi-Criteria Decision Making (MCDM)

Multi-criteria decision-making (MCDM) involves selecting one of several different alternatives, based on a set of criteria that describe the alternatives. However, there are numerous problems that make comparing these alternatives difficult. For instance, very often, decisions are based on several conflicting criteria; e.g., which car to buy that is cheap and energy efficient. In addition, what happens when we have a group of decision makers that must come to some sort of consensus? This is known as multi-expert multi-criteria decision making (MEMCDM).

In MEMCDM, there are several new problems to be addressed. One such problem is how to handle expert disagreement and come to a consensus/decision in the first place. Another problem, as stated earlier, is that of predicting future decisions based on decision data from multiple experts along multiple criteria. Again, the question of "which expert/decision-making process to follow?" is a major challenge in solving such problems.

Approaches to MCDM In general, on a daily basis, when the decision is not critical, in order to reach a decision, we mentally "average / sort" these criteria along with their satisfaction levels. This corresponds to aggregating values of satisfaction with weights on each criterion, reflecting its importance in the overall score (a.k.a. additive aggregation), that is, calculating the overall score of an alternative with the weighted sum of the criterion scores. In other words, weights assigned to different sets of criteria in the weighted average approach form an "additive measure". Additive aggregation, however, assumes that criteria are independent, which is seldom the case [**?**]. Non-linear approaches also prove to lead to solutions that are not completely relevant [**?**].

This should change when considering possible dependence between criteria. For example, if two criteria are strongly dependent, it means that both criteria express, in effect, the same attribute. As a result, when we consider the set consisting of these two criteria, we should assign to this set the same weight as to each of these criteria – and not double the weight as in the weighted sum approach. In general, the weight associated to different sets should be different from the sum of the weights associated to individual criteria. In mathematics, such non-additive functions assigning numbers to sets are known as non-additive (fuzzy) measures. It is therefore reasonable to describe the dependence between different criteria by using an appropriate non-additive (fuzzy) measure. Combining the fuzzy measure values with the criteria satisfaction can be done using the Choquet integral, which integrals are actively used in Multi-Criteria Decision Making [**?**].

However, to make this happen, fuzzy measures need to be determined: they can either be identified by a decision maker/expert or by an automated system that extracts them from sample data. Since human expertise might not always be available and getting accurate fuzzy values (even from an expert) might be tedious [**?**], fuzzy measures are usually automatically extracted from prior decision decision data. To the original problem, this approach adds an optimization problem that can be tedious to solve. Although it was solved with success for some data sets [**?**], the overall prediction quality is not satisfactory and the approach limits the number of criteria that can be taken into account (the number of variables to determine is exponential in the number of criteria) [**?**].

2.2 Argumentation Frameworks

In this section we briefly summarise the background information related to classical AAFs [**?**]. We focus on the basic definition of an AAF (see Def. 1), on the notion of defense (Def. 2), and on extension-based semantics (Def. 3).

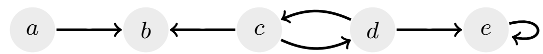

Fig. 1. An example of AAF.

Definition 1. *An Abstract Argumentation Framework (AAF) is a pair* $F = \langle A, R \rangle$ *of a set* A *of arguments and a binary relation* $R \subseteq A \times A$*, called the attack relation.* $\forall a, b \in A$*,* $aR\,b$ *(or,* $a \rightarrowtail b$*) means that* a *attacks* b*. An AAF may be represented by a directed graph (an interaction graph) whose nodes are arguments and edges represent the attack relation. A set of arguments* $S \subseteq A$ *attacks an argument* a*, i.e.,* $S \rightarrowtail a$*, if* a *is attacked by an argument of* S*, i.e.,* $\exists b \in S . b \rightarrowtail a$*.*

Definition 2. *Given an AAF,* $F = \langle A, R \rangle$*, an argument* $a \in A$ *is defended (in* F*) by a set* $S \subseteq A$ *if for each* $b \in A$*, such that* $b \rightarrowtail a$*, also* $S \rightarrowtail b$ *holds. Moreover, for* $S \subseteq A$*, we denote by* S^+_R *the set* $S \cup \{b \mid S \rightarrowtail b\}$*.*

The "acceptability" of an argument [**?**] depends on its membership to some sets, called *extensions*: these extensions, or semantics, characterise a collective "acceptability". Respectively, *stb*, *adm*, *prf*, *gde*, *com*, and *sem*, stand for stable, admissible, preferred, grounded, complete, and semi-stable semantics.

Definition 3. *Let* $F = \langle A, R \rangle$ *be an AAF. A set* $S \subseteq A$ *is conflict-free (in* F*), denoted* $S \in cf(F)$*, iff there are no* $a, b \in S$*, such that* $(a, b), (b, a) \in R$*. For* $S \in cf(F)$*,it holds that:*

- $S \in stb(F)$*, if foreach* $a \in A \setminus S$*,* $S \rightarrowtail a$*, i.e.,* $S^+_R = A$*;*
- $S \in adm(F)$*, if each* $a \in S$ *is defended by* S*;*
- $S \in prf(F)$*, if* $S \in adm(F)$ *and there is no* $T \in adm(F)$ *with* $S \subset T$*;*
- $S = gde(F)$ *if* $S \in com(F)$ *and there is no* $T \in com(F)$ *with* $T \subset S$*;*
- $S \in com(F)$*, if* $S \in adm(F)$ *and for each* $a \in A$ *defended by* S*,* $a \in S$ *holds;*
- $S \in sem(F)$*, if* $S \in adm(F)$ *and there is no* $T \in adm(F)$ *with* $S^+_R \subset T^+_R$*.*

We recall that for each AF, $stb(F) \subseteq sem(F) \subseteq prf(F) \subseteq com(F) \subseteq adm(F)$ holds, and that for each of the considered semantics σ (except stable) $\sigma(F) \neq \emptyset$ holds. Finally, $gde(F)$ and $ide(F)$ are always unique, and $ide(F) \in com(F)$.

Consider the $F = \langle A, R \rangle$ in Fig. 1, with $A = \{a, b, c, d, e\}$ and $R = \{(a, b), (c, b), (c, d), (d, c), (d, e), (e, e)\}$. We have that $stb(F) = sem(F) = \{\{a, d\}\}$, and $gde(F) = ide(F) = \{a\}$. The admissible sets of F are $\emptyset, \{a\}, \{c\}, \{d\}, \{a, c\}, \{a, d\}$, and $prf(F) = \{\{a, c\}, \{a, d\}\}$. The complete extensions are $\{a\}, \{a, c\}, \{a, d\}$.

3 Proposed Model for MEMCDM using Argumentation Frameworks

Here, we describe our model: given an MEMCDM problem with n criteria and p experts, how do we "translate"/model it as an AAF? In other words, which arguments and attacks should compose it?

3.1 Arguments

• What does the data we use (i.e., experts' evaluation of software in this case) tell us about the arguments to add to the network? We differentiate arguments that come from the data (i.e., Expert i said that Software j is good) from arguments that are implicit (i.e., Software k is Poor).

1. **Expert i gives Item j a total quality** D_{ij} (which, in the case of Software Quality Assessment – SQA, can be Bad, Poor, Fair, Good, or Excellent):

   $$\textbf{Argument} \quad (E_i, S_j, D_{ij})$$

   Let us call such arguments, arguments of type ESD.

2. **Expert i judges that Item j satisfies criterion m up to quality** D_{ijm}

   $$\textbf{Argument} \quad (E_i, S_j, c_m, D_{ijm})$$

   Let us call such arguments, arguments of type EScD.

• Which implicit arguments should be part of the argumentation network for this specific type of problem?

1. For each item, independently from what experts say, there will be a decision made. This decision will be in the form of a final ranking, ranging over all possibly ranking values (in the case of SQA: Bad, Poor, Fair, Good, Excellent). So regardless of ESD arguments, we add to the argumentation network the following arguments:

   $$\forall \text{ item } S_i, \forall \text{ ranking } D_j : \ \textbf{Argument} \quad (S_i, D_j)$$

   Let us call such arguments, **arguments of type SD**.

2. For each criterion of evaluation, regardless of which item is being evaluated and of what experts will decide, a ranking will be associated. So regardless of EScD arguments, we add to the argumentation network the following arguments:

   $$\forall \text{ item } c_k, \forall \text{ ranking } D_m : \ \textbf{Argument} \quad (c_k, D_m)$$

   Let us call such arguments, **arguments of type cD**. *Such arguments are expected to be useful for prediction of the decision of experts on items not part of the original data, but for which we do have an indication of their quality per criterion.*

• **Coalitions of Arguments** Here we aim to model the fact the n decisions of any expert on the n criteria of the problem at hand belong together: they together form the support for the expert's final decision on the given item. As a result, for any expert E_i and any item S_j, we define a coalition of "supporting" decisions as:

$$\forall E_i, \forall S_j, \ \textbf{Coalition:}\ \{(E_i, S_j, c_k, D_{i,j,k}),\ k \in \{1, \ldots, n\}\}$$

Let us call such coalitions of EScDs, extended arguments of type **CoEScD**. The result of modeling such coalitions is that all arguments in the coalition will be forced to be altogether either in or out of extensions. *Per se, we are enforcing an equality constraint on the belonging of these arguments to any extension.*

3.2 Attacks

In this subsection, we answer the following question: What are the **attacks** (*edges of the network*) between these arguments (*nodes*)? *Note:* All attacks we define are reciprocal, hence the edges are always set bidirectionally.

For attacks too, we differentiate between attacks that come from inconsistencies in the decision data (disagreement between experts, inconsistency in decisions of a single expert,lack of fairness, irrationality). An assumption that we make in designing the network model is that experts should be rational: in this, we mean that even if they are not (which we know), they should be and we aim to elicit decisions that are as rational as can be.

• **Attacks derived from lack of fairness** Here, we assume that if an expert is fair, then s/he should derive the same final ranking from the same criteria rankings. For instance, if there are 3 criteria (c_1, c_2, and c_3) to assess items and an expert E has the following decision history:

$$\begin{cases} E, S_i, c_1, D_1 \\ E, S_i, c_2, D_1 \\ E, S_i, c_3, D_1 \end{cases} \longrightarrow\ E, S_i, D$$

and: (with $S_i \neq S_j$)

$$\begin{cases} E, S_j, c_1, D_1 \\ E, S_j, c_2, D_1 \\ E, S_j, c_3, D_1 \end{cases} \longrightarrow\ E, S_j, D'$$

where $D \neq D'$, then we should see arguments (E, S_i, D) and (E, S_j, D') are a lack of fairness in judgment and therefore add the following attack in the argumentation network: $(E, S_i, D) \longleftrightarrow (E, S_j, D')$.

More generally, assuming that the criteria that are considered by the experts are c_k, with $k \in K$, and that the possible rankings are denoted by D_r, with $r \in R$, then we add the following rule to our model:

$$\forall S_i,\ S_j,\ E \text{ s.t. } i \neq j \text{ and } \forall k \in K, \exists r \in R,\ (S_i, E, c_k, D_r) \text{ and } (S_j, E, c_k, D_r):$$

$$\begin{array}{c}\text{if } (S_i, E, D_i) \text{ and } (S_j, E, D_j) \text{ and } D_i \neq D_j \\ \text{then } \textbf{Attack } (S_i, E, D_i) \longleftrightarrow (S_j, E, D_j)\end{array}$$

• **Attacks derived from lack of rationality** Let us recall that we assume that the rankings D_r, with $r \in R$, are totally ordered. However, with n criteria, the set of n-tuples of rankings is only partially ordered:

$$\begin{array}{l}(D_1, D_2, \ldots, D_n) \ \prec (D'_1, D'_2, \ldots, D'_n) \\ \text{iff :} \\ \forall i \in \{1, \ldots, n\}: \ (D_i \neq D'_i) \ \longrightarrow \ D_i < D'_i\end{array}$$

Now: $\forall E_i$ and $\forall S_j$, we denote by $(D_{1,i,j}, \ldots, D_{n,i,j})$ the set of n decisions made by Expert E_i on each of the criteria $c_1, \ldots, c_n$ for Item S_j, and by $D_{i,j}$ the final decision of Expert E_i on Item S_j.

Being rational for any given expert E_i means that if for Item S_j, s/he ranks criteria lower (w.r.t. above partial order) than s/he ranks the criteria of Item S_k, then his/her final ranking of S_j should not be higher than his/her ranking of S_k. Formally, it is expressed as follows:

$$\begin{array}{l}\forall E_i,\ \forall S_j,\ \forall S_k (j \neq k): \\ \text{if: } (D_{1,i,j}, \ldots, D_{n,i,j}) \prec (D_{1,i,k}, \ldots, D_{n,i,k}) \text{ and: } D_{i,j} > D_{i,k} \\ \text{then: } \textbf{Attack } (S_j, E_i, D_{i,j}) \ \longleftrightarrow \ (S_k, E_i, D_{i,k})\end{array}$$

• **Attack related to implicit arguments: SD and cD** In this subsection, we describe the following attacks:

- attacks between implicit arguments SD (resp. cD); and
- attacks across SD and ESD (resp. cD and EScD).

1. Attacks among SDs: SD Arguments associate an item with a ranking. For each item S_i, there is p SD arguments if there are p possible ranking levels. Each of these p arguments attack each other (they form a complete subgraph). In other words:

$$\forall S_i, \forall r_1, r_2 \in R, \text{ with } r_1 \neq r_2, \textbf{Attack: } (S_i, D_{r_1}) \longleftrightarrow (S_i, D_{r_2})$$

2. Attacks among cDs: In a fashion similar to attacks among SDs, we have:

$$\forall c_j, \forall r_1, r_2 \in R, \text{ with } r_1 \neq r_2, \textbf{Attack: } (c_j, D_{r_1}) \longleftrightarrow (c_j, D_{r_2})$$

3. Attacks between SDs and ESDs: For any given item S_i, an argument saying that S_i is evaluated D_j is in contradiction (and therefore attacks – and vice-versa) any argument (E, S_i, D_k) as soon as $D_j \neq D_k$. As a result:

$$\forall E,\ \forall S_i,\ (D_j \neq D_k) \ \rightarrow \ \textbf{Attack: } (S_i, D_j) \longleftrightarrow (E, S_i, D_k)$$

4. Attacks between cDs and EScDs: Similarly as above, for any given criterion c_m, an argument saying that c_m is evaluated D_j is in contradiction (and therefore attacks – and vice-versa) any argument (E, S_i, c_m, D_k) as soon as $D_j \neq D_k$. As a result:

$$\forall E,\ \forall S,\ \forall c_m, (D_j \neq D_k)\ \rightarrow\ \textbf{Attack:}\ (c_m, D_j) \longleftrightarrow (E, S, c_m, D_k)$$

• **Attacks between Coalitions and ESDs** Here we aim to model the fact that coalitions of decisions on criteria support experts' decisions. In order word:

$$\forall E_i, \forall S_j,\ \{(E_i, S_j, c_k, D_{i,j,k}),\ k \in \{1, \ldots, n\}\}\ \textbf{supports}\ (E_i, S_j, D_{i,j})$$

In terms of attacks, this is expressed as follows:

$$\begin{array}{l} \forall E_i, E_j \forall S_k :\ D_{i,k} \neq D_{j,k}\ \rightarrow \\ \textbf{Attack:}\ \{(E_i, S_k, c_l, D_{i,k,l}),\ k \in \{1, \ldots, n\}\} \longleftrightarrow (E_j, S_k, D_{j,k}) \end{array}$$

4 An Example

Here, let us look at a scenario in which experts independently assess given pieces of software, based on several given evaluation criteria. We describe the resulting argumentation networks (arguments/nodes and attack/edges).

JOEL: We need an illustration here.

5 Conclusion and Future Work

In this work, we proposed a model for MEMCDM problems, based on Bistarelli et Al.'s AAFs, that allows to emulate fairness and rationality. This allows discrimination among input decision data (from experts' prior decisions) between data of value and data that should just not be taken into account. Next steps include operationalizing the whole process (from input processing to results filtering) and then adding weights to the attacks to simulate the extent of disagreements and allow lineance towards small errors (e.g., unfairness / irrationality that are really minimal, minor disagreements). Also part of future work, we plan to explicitely acknowledge in the AAF that disagreement can be at two different levels: epistemic and pragmatic, and to make use of argumentation frameworks to identify disagreement configurations (epistemic and pragmatic, epistemic only, pragmatic only).

Acknowledgments. M. Ceberio's work was partially supported by the National Science Foundation, NSF CCF grant 0953339 and the American Association for the Advancement of Science, AAAS MIRC (agreement date 112612).