Recourse under Model Multiplicity via Argumentative Ensembling (Extended Abstract)

Junqi Jiang, **Francesco Leofante**, **Antonio Rago** and **Francesca Toni**
Department of Computing, Imperial College London, United Kingdom
{junqi.jiang, f.leofante, a.rago, f.toni}@imperial.ac.uk

Introduction

Model Multiplicity (MM), also known as predictive multiplicity or the Rashomon Effect, refers to a scenario where multiple, equally performing machine learning (ML) models may be trained to solve a prediction task (Black, Raghavan, and Barocas 2022). The literature has identified that these models may differ greatly in their internals and might thus produce inconsistent predictions when deployed (Breiman 2001; Marx, Calmon, and Ustun 2020).

Ensembling techniques are commonly used to deal with MM scenarios (Black, Leino, and Fredrikson 2022; Black, Raghavan, and Barocas 2022). An example of such a technique is *naive ensembling* (Black, Leino, and Fredrikson 2022), where the predictions of several models are aggregated to produce a single outcome that reflects the opinion of a majority of the models. While ensembling methods have been shown to be effective in practice, their application to consequential decision-making tasks raises some important challenges. Specifically, these methods tend to ignore the need to provide avenues for recourse to users negatively impacted by the models' outputs, which the literature typically achieves via the provision of *counterfactual explanations* (CEs) for the predictions (see (Guidotti 2022) for a recent overview).

Dealing with MM while also taking CEs into account is non-trivial. Standard algorithms designed to generate CEs for single models typically fail to produce recourse recommendations that are valid across equally performing models (Jiang et al. 2024b). This phenomenon may have troubling implications as a lack of robustness may lead users to question whether a CE is actually explaining the underlying decision-making task and is not just an artefact of a (subset of) model(s).

In this extended abstract, we summarise our recent contributions in (Jiang et al. 2024a). We formally define the problem of providing recourse under MM, and we briefly describe the six desirable properties that we argue that the solution should satisfy. We then introduce *argumentative ensembling*, a novel technique rooted in computational argumentation (see (Atkinson et al. 2017) for an overview). We briefly demonstrate how argumentative ensembling is able to solve the recourse problem effectively, while naturally incorporating user preferences over meta-evaluation aspects of the models, like fairness, robustness, and interpretability.

Recourse under Model Multiplicity

Given a set of classification *labels* $\mathcal{L}$, a *model* is a mapping $M : \mathbb{R}^n \rightarrow \mathcal{L}$; we denote that M classifies an *input* $\mathbf{x} \in \mathbb{R}^n$, consisting of n features, as ℓ iff $M(\mathbf{x}) = \ell$. Then, a *counterfactual explanation* (CE) for $\mathbf{x}$, given M, is some $\mathbf{c} \in \mathbb{R}^n \setminus \{\mathbf{x}\}$ such that $M(\mathbf{c}) \neq M(\mathbf{x})$, which may be optimised by some distance metric between the inputs.

Consider a non-empty set of models $\mathcal{M} = \{M_1, \ldots, M_m\}$ and, for an input $\mathbf{x}$, assume a set $\mathcal{C}(\mathbf{x}) = \{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$ where each $\mathbf{c}_i \in \mathcal{C}(\mathbf{x})$ is a CE for $\mathbf{x}$, given M_i. We assume each CE is *valid* on its associated model, i.e. $M_i(\mathbf{c}_i) \neq M_j(\mathbf{x})$, and we say a CE $\mathbf{c}_i$ is *valid* on model M_j iff $M_j(\mathbf{c}_i) \neq M_j(\mathbf{x})$. Our aim is to solve:

Problem: Recourse-Aware Ensembling (RAE)
Input: input $\mathbf{x}$, set $\mathcal{M}$ of models, set $\mathcal{C}$ of CEs
Output: "optimal" set $S \subseteq \mathcal{M} \cup \mathcal{C}$.

To characterise optimality, we propose six desirable properties for the outputs (solutions) of ensembling methods. *Non-emptiness* requires the solution satisfies $S \cap \mathcal{M} \neq \emptyset$ and $S \cap \mathcal{C} \neq \emptyset$, ensuring that the RAE method returns some models and some CEs. *Non-triviality* states that S should contain more than one model, such that the prediction result for $\mathbf{x}$ is jointly decided by multiple models and thus more robust. Then, S is said to satisfy *model agreement* if the models included all agree on the prediction label for $\mathbf{x}$, meaning no prediction-related conflict exists. *Majority vote* requires that the prediction result determined by S is the same as majority voting when considering all models in $\mathcal{M}$. *Counterfactual validity* states that each of the included CEs should be valid on each of the included models, i.e., there are no counterfactual validity-related conflicts. Finally, *counterfactual coherence* guarantees that a model is included in the solution set iff its associated CE is also in this set . Then, we theoretically prove that two ensembling methods naturally extended from naive ensembling do not satisfy non-emptiness, counterfactual validity, and counterfactual coherence.

Argumentative Ensembling

As can be seen from the problem definition and the desirable properties, prediction-related and CE validity-related conflicts are at the core of solving the RAE problem. *Computational argumentation*, a set of formalisms for dealing with

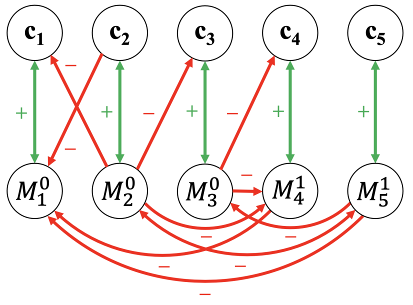

Figure 1: An example BAF constructed by argumentative ensembling where: models' predictions for the input $\mathbf{x}$ are given as superscripts, e.g. $M_1(\mathbf{x}) = 0$ but $M_4(\mathbf{x}) = 1$; reciprocal supports are represented by dual-headed green arrows labelled with + and standard (reciprocal) attacks are represented by single-headed (dual-headed, respectively) red arrows labelled with −.

conflicting information (Atkinson et al. 2017), is therefore identified as the ideal tool for obtaining the optimal solution sets. We propose a novel argumentative ensembling method to solve the RAE problem, which works in two steps.

First, we model the conflicts in $\mathcal{M} \cup \mathcal{C}$ in a bipolar argumentation framework (BAF) (Cayrol and Lagasquie-Schiex 2005), specifying the attack and support relations between models and models, and models and CEs, based on whether or not there exist conflicts between them, and whether a model is preferred over another in terms of some pre-defined model meta-evaluation preference rules. Figure 1 shows an example BAF for an input with five competing models under MM. In this example, bi-directional supports are established between each model and its associated CE. We then draw attacks between each of $\{M_1, M_2, M_3\}$ and each of $\{M_4, M_5\}$ because these two subsets of models predict different labels for the input. Additionally, attacks are drawn between CEs and models if the CE is not valid on the model. The directions of the attacks depend on the model preferences. Here, we assume the preferences over models are $M_2 \simeq M_5 > M_3 > M_4 > M_1$, where $\simeq$ and $>$ represent *is equally preferred to* and *is preferred to* respectively.

Then, by computing the *s-preferred* extension of the BAF (Cayrol and Lagasquie-Schiex 2005), we obtain a subset of $\mathcal{M} \cup \mathcal{C}$ satisfying certain argumentative properties, which we use as the solution set of the RAE problem. In our example (Figure 1), the solution identified by our method is $\{M_4, M_5, \mathbf{c}_4, \mathbf{c}_5\}$. Though it is not the set of models that produces the majority prediction (class 0 by $\{M_1, M_2, M_3\}$), it contains no CE validity-related conflicts.

We perform rigorous theoretical analyses linking our way of formulating the BAF, the argumentative properties of an s-preferred extension, and the desirable properties of RAE solutions, through which we prove that our argumentative ensembling method satisfies all the desirable properties except for majority vote, with an additional benefit of supporting user preferences on models. In our experiments, we instantiate RAE problems with model set sizes of $\{10, 20, 30\}$ on three datasets, and we demonstrate the superior performances of our method over baselines extended from naive ensembling on metrics quantifying the desirable properties.

Future Work

This paper opens up several interesting directions for future work. First, it would be interesting to examine whether considering attacks to or from *sets* of arguments (e.g. (Dvorák et al. 2022)), rather than single arguments, may help in MM. Further, *extended argumentation frameworks* (Modgil 2009) and *value-based argumentation frameworks* (Bench-Capon 2002) may provide useful alternative ways to account for preferences. Moreover, in order to support experiments with a high number of models (beyond the 30 we considered), large-scale argumentation solvers would be highly desirable. Finally, it would be interesting to assess the effect which MM has on users' evaluations of CEs.

Acknowledgements

Jiang, Rago and Toni were partially funded by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme. Leofante was supported by an Imperial College Research Fellowship grant. Rago and Toni were partially funded by the ERC (grant agreement No. 101020934).

References

Atkinson, K.; Baroni, P.; Giacomin, M.; Hunter, A.; Prakken, H.; Reed, C.; Simari, G. R.; Thimm, M.; and Villata, S. 2017. Towards artificial argumentation. *AI Magazine* 38(3):25–36.

Bench-Capon, T. J. M. 2002. Value-based argumentation frameworks. In *NMR 2002*, 443–454.

Black, E.; Leino, K.; and Fredrikson, M. 2022. Selective ensembles for consistent predictions. In *ICLR 2022*.

Black, E.; Raghavan, M.; and Barocas, S. 2022. Model multiplicity: Opportunities, concerns, and solutions. In *FAccT 2022*, 850–863.

Breiman, L. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16(3):199–231.

Cayrol, C., and Lagasquie-Schiex, M. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *ECSQARU 2005*, 378–389.

Dvorák, W.; König, M.; Ulbricht, M.; and Woltran, S. 2022. Rediscovering argumentation principles utilizing collective attacks. In *KR 2022*, 122–131.

Guidotti, R. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* 1–55.

Jiang, J.; Leofante, F.; Rago, A.; and Toni, F. 2024a. Recourse under model multiplicity via argumentative ensembling. In *AAMAS 2024*, 954–963.

Jiang, J.; Leofante, F.; Rago, A.; and Toni, F. 2024b. Robust counterfactual explanations in machine learning: A survey. In *IJCAI 2024*.

Marx, C. T.; Calmon, F. P.; and Ustun, B. 2020. Predictive multiplicity in classification. In *ICML 2020*, 6765–6774.

Modgil, S. 2009. Reasoning about preferences in argumentation frameworks. *Artif. Intell.* 173(9-10):901–934.