Explainable Shapley-Based Allocation

Meir Nizri[1,2], **Amos Azaria**[1], **Noam Hazon**[1]
[1]Department of Computer Science
Data Science and Artificial Intelligence Center
Ariel University, Israel
meir.nizri@msmail.ariel.ac.il, amos.azaria@ariel.ac.il, noamh@ariel.ac.il
[2] Address: 4 Tzofnat, Bracha 4483500, Israel
Phone: +972 50-3377-637

Abstract

The Shapley value is one of the most important normative division scheme in cooperative game theory, satisfying basic axioms. However, some allocation according to the Shapley value may seem unfair to humans. In this paper, we develop an automatic method that generates intuitive explanations for a Shapley-based payoff allocation, which utilizes the basic axioms. Given a coalitional game, our method decomposes it to sub-games, for which it is easy to generate verbal explanations, and shows that the given game is composed of the sub-games. Since the payoff allocation for each sub-game is perceived as fair, the Shapley-based payoff allocation for the given game should seem fair as well. We run an experiment with 210 human participants and show that when applying our method, humans perceive Shapley-based payoff allocation as significantly more fair than when using a general standard explanation.

1 Introduction

The Shapley value (Shapley 1953), which has been termed the most important normative division scheme in cooperative game theory (Winter 2002), is based on the idea that the payoff of the game should be divided such that each agent's share is proportional to its contribution to the payoff. Indeed, the Shapley value is considered fair since it is the only payoff allocation that satisfies the following four desirable axioms: efficiency, symmetry, null player property and additivity (Hart 1989).

While the axioms satisfied by the Shapley value seem necessary, humans presented with an allocation according to the Shapley value may sometimes not observe it as fair. For example, consider the following game with three agents: r, l_1, and l_2, which is also known as the classical "glove game". Agents l_1 and l_2 have a left-glove and agent r has a right-glove. A pair of left and right gloves is worth \$12, but a single glove is worth nothing. If all agents collaborate, the Shapley value allocates \$8 to agent r and only \$2 to l_1 and \$2 to l_2. While it seems plausible that agent r should receive a higher payoff, a right-glove alone is worth nothing and thus, it may seem unfair that the payoff for this agent is 4-times more than each of the other agents. However, any other allocation would violate at least one of the axioms. It is thus desirable to increase human acceptance of the allocation according to the Shapley value, which can be achieved by providing explanations. In this paper, we develop an automatic method that generates intuitive explanations for a Shapley-based payoff allocation.

Now, the essence of our explanation is that any game is decomposed into several sub-games that their Shapley allocation is easier to perceive as fair. Specifically, any sub-games is built such that all the agents are either null players or equivalent to one another, and the values are either all non-negative or all non-positive. According to the null player axiom each agent who is a null player should receive a payoff of 0, and according to the symmetry and efficiency axioms all other agents should equally share the total outcome, and thus the Shapley allocation in each sub-game is intuitively fair. For example, the "glove game" can be decomposed into few sub-games; in one of the sub-games, agent r obtains a value of \$12 when collaborating with l_1, but not when collaborating with l_2. When all three agents collaborate, they obtain a value of \$12. In this sub-game l_2 is a null player, and agents r and l_1 are equivalent. Thus, the Shapley allocation of \$6 to agent r, \$6 to agent l_1 and \$0 to agent l_2 is intuitively fair. Finally, following the additivity axiom, since the Shapley allocation of every sub-game is intuitively fair, and the sum of the Shapley allocations in each sub-game is equal to the Shapley allocation in the original game, then the latter is easier to perceive as fair. We note that this process follows the arguments in the proof of the uniqueness of the Shapley value (Shapley 1953). Practically, we do not directly present the axioms to the users. Instead, our algorithm, which we termed *X-SHAP*, decomposes any coalitional game into several sub-games, and automatically generates a brief verbal explanation that accompanies each sub-game.

We run an experiment with 210 human participants and show that the explanations that were generated by X-SHAP achieved significantly higher fairness rating compared to the general explanation in all the games examined. This indicates that humans perceive the Shapley payoff allocation fairer if they receive X-SHAP's explanations.

To summarize, the main contribution of this paper is that it provides the first successful automatic method that generates customized explanations of the Shapley allocation for any given coalitional game.

Related Work

Our work belongs to the field of Explainable AI (XAI) (Gunning et al. 2019). The work that is closest to ours is the paper by Cailloux and Endriss (Cailloux and Endriss 2016). They develop an algorithm that automatically derives a justification for any outcome of the Borda rule. The algorithm's main idea is to decompose the preference profile into a sequence of sub-profiles, and use one of six axioms for providing explanations for the sub-profiles and for their combinations. Our approach for explaining the Shapley allocation is also based on axioms, and we also decompose the given coalitional game into a set of sub-games, which together compose an explanation for the given coalitional game.

Spliddit (Goldman and Procaccia 2015) is a website implementing algorithms for various division tasks (e.g., rent division), which also explains how the outcomes satisfy certain fairness requisites. While the website enables users to compute the Shapley value in a ride-sharing context, it provides only a general explanation that states the benefits of the Shapley value. Our work can thus serve as an extension for Spliddit by providing customized explanations for the Shapley value.

2 X-SHAP

In this section we propose the *X-SHAP* algorithm, which given any coalitional game, automatically decomposes the coalitional game into a number of sub-games.

The *X-SHAP* algorithm works as follows. It receives a coalitional game (N, v) as an input and provides a set X of characteristic functions that maintains the following two properties:

1. Each coalitional game (N, x), where $x \in X$, is easy-to-explain.
2. The sum of all the characteristic functions in X equals v. That is, $\sum_{x \in X} x = v$.

Note that since the Shapley value satisfies the additivity axiom, the sum of Shapley value payoffs assigned to each agent $i \in N$ in each characteristic function in X is equal to the Shapley value payoff for i in (N, v). That is, $\forall i \in N, \sum_{x \in X} Sh_i(N, x) = Sh_i(N, v)$. Once the set X is generated, we generate explanations for each of the sub-games.

Algorithm 1 describes the pseudo-code for *X-SHAP*. The algorithm iterates over all subsets $S \subseteq N$ in ascending order according to $|S|$. It maintains a characteristic function $accum$ that accumulates all the characteristic functions it builds in each iteration. For each subset S whose value in v is different from its value in $accum$, X-SHAP adds the following characteristic function x to X. For each subset of N, T, that contains S, $x(T)$ is set to the difference between $v(S)$ and $accum(S)$.

Algorithm 1: X-SHAP

Input : A coalitional game (N, v).
Output: A set of characteristic functions X, along with their explanations.

$X \leftarrow \emptyset$
Let $accum, x$ be characteristic functions on N
Initialize $accum$ to 0 for any subset
for $i \leftarrow 1$ *to* $|N|$ **do**
  for *every* $S \subseteq N$*, such that* $|S| = i$ **do**
    Initialize x to 0 for any subset
    if $v(S) \neq accum(S)$ **then**
      for *every* $T \supseteq S$ **do**
        $x(T) \leftarrow v(S) - accum(S)$
      $X \leftarrow X \cup \{x\}$
      $accum \leftarrow accum + x$
Generate an explanation for each $x \in X$
return X along with the explanations

3 Experimental Evaluation

In order to evaluate the performance of X-SHAP, we conducted a survey with human participants. The survey examined six coalitional games, representing a variety of scenarios. Each of the coalitional games was presented to the participants along with its Shapley payoff allocation as a suggestion for dividing the payoff among the agents. Then, each participant was given either X-SHAP's explanation or a general explanation that states the benefits of the Shapley value, which served as a baseline. The participants were asked to rate the proposed allocation by indicating to what extent they agree or disagree that it is fair, using a seven-point Likert scale. Overall, 210 different people participated in the survey, each answering two different coalitional games.

The results were obtained by averaging over the 35 ratings of each of the two explanations in each of the six scenarios. The explanations that were generated by X-SHAP significantly outperformed the general explanation in terms of fairness rating in all the scenarios examined ($p < 0.0001$). That is, the human participants perceive the payoff allocation fairer if they receive the explanations that are generated by X-SHAP. Overall, the average fairness rating in scenarios in which the X-SHAP explanation was provided is 5.3, which is significantly higher than the rating of 4.4 obtained for scenarios accompanied by the general explanation.

References

Cailloux, O.; and Endriss, U. 2016. Arguing about Voting Rules. In *AAMAS*.

Goldman, J.; and Procaccia, A. D. 2015. Spliddit: Unleashing Fair Division Algorithms. *SIGecom Exch.*, 13(2): 41–46.

Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; and Yang, G.-Z. 2019. XAI—Explainable artificial intelligence. *Science Robotics*, 4(37).

Hart, S. 1989. Shapley value. In *Game Theory*, 210–216. Springer.

Shapley, L. S. 1953. A value for n-person games. *Contributions to the Theory of Games*, 2(28): 307–317.

Winter, E. 2002. The Shapley value. *Handbook of game theory with economic applications*, 3: 2025–2054.