Explanation of Pseudo-Boolean Functions using Cooperative Game Theory and Prime Implicants

C. Labreuche [1,2]

[1] **THALES Research & Technology**, Palaiseau, France
[2]**SINCLAIR AI Lab**, Palaiseau, France
email: christophe.labreuche@thalesgroup.com

Outline

Two different explanations of a function f applied on an instance x

Formal approaches – Sufficient Explanations

- Find the caracteristics in x that are *sufficient* to get the outcome $f(x)$
- *Process of generalizing x (removing values on attributes) while keeping the same outcome $f(x)$*

CONS

- Restricted to Boolean (discrete) output

PROS

- Clear meaning
- Actionable explanation

Illustration with 2 features: $x = (true, true)$

Is the subset *SUFFICIENT*?	1 alone	2 alone	1, 2 together
$f = AND$	NO	NO	YES
$f = OR$	YES	YES	YES

Two different explanations of a function f applied on an instance x

Heuristics – Feature attribution

- Allocate a *contribution level* of each attribute of x in $f(x)$

CONS

- What to do with these numbers?
- Cannot represent the idea of *sufficiency*

PROS

- Highlights the most important features
- Model agnostic

Illustration with 2 features

Cannot distinguish between AND and OR operators!

Aim

Aim of the work

Define a *feature attribution* approach representing *sufficiency*.

- If a single feature is sufficient, it is enough to select it!

Outline

Outline

Setting

- $N = \{1, \ldots, n\}$: index set of attributes/features.
- We assume Boolean variables/features.
- $D = \{0,1\}^N$: set of alternatives/instances.

Boolean Function (BF)

A *BF* is a function $f : D \to \{0,1\}$.

0–1 Game

A 0–1 *game* is a set function $v : 2^N \to \{0,1\}$.

Pseudo-Boolean Function (PBF)

A *PBF* is a function $f : D \to \mathbb{R}$.

Game

A *game* is a set function $v : 2^N \to \mathbb{R}$.

- $f \mapsto v_f$ defined by $v_f(S) = f(1_S, 0_{N\setminus S})$.
- v (resp. f) is assumed to be monotone.

Sufficient Explanation: prime implicants & winning coalitions

$\mathcal{I}_f$: Implicants of f

An *implicant* is a conjuction of literals 1_S s.t. $f(1_S, x_{N\setminus S}) = 1$ for all x.

$\mathcal{W}_v$: Winning Coalitions

A *winning* coalition is a subset S s.t. $v(S) = 1$.

$\mathcal{PI}_f$: Prime Implicants of f

A *prime implicant* is a minimal implicant.

$\mathcal{MW}_v$: Minimal Winning Coalitions

Minimal Winning Coalitions w.r.t. $\subseteq$.

Irrelevant / mandatory coalition

A variable is *null* if changing the value on this variable never modifies the output v.
A variable is a *veto*, if all winning coalitions include this variable.

$f(x) = x_1 \wedge (x_2 \vee x_3)$ on $N = \{1, 2, 3, 4\}$

$\mathcal{I}_f = \{1_{\{1,2\}}, 1_{\{1,3\}}, 1_{\{1,2,3\}}, 1_{\{1,2,4\}}, 1_{\{1,3,4\}}, 1_{\{1,2,3,4\}}\}$
and $\mathcal{PI}_f = \{1_{\{1,2\}}, 1_{\{1,3\}}\}$.
Feature 4 is irrelevant and 1 is mandatory.

$v(S) = 1$ iff $(1 \in S) \wedge [(2 \in S) \vee (3 \in S)]$

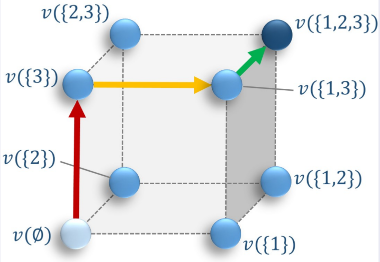
$\mathcal{W}_v =$
$\{\{1, 2\}, \{1, 3\}, \{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{1, 2, 3, 4\}\}$
and $\mathcal{MW}_v = \{\{1, 2\}, \{1, 3\}\}$.

Heuristic Explanation: feature attribution

How to distribute the total worth $v(N)$ among the players?

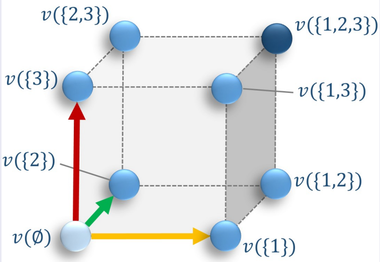
Shapley value

$$\phi_i^{\mathrm{Sh}}(N,v) = \sum_{S\subseteq N\setminus i} \frac{(n-|S|-1)!|S|!}{n!}\left[v(S\cup\{i\}) - v(S)\right]$$



Proportional Division

$$\phi_i^{\mathrm{PD}}(N,v) = \frac{v(\{i\})}{\sum_{j\in N} v(\{j\})} v(N)$$



Outline

Values cannot represent sufficiency

Illustration with $N = \{1, 2\}$

$$v_\wedge(S) = 1 \text{ iff } (1 \in S) \wedge (2 \in S) \quad \text{and} \quad v_\vee(S) = 1 \text{ iff } (1 \in S) \vee (2 \in S).$$

Prime Implicants	Game Theory
$\mathcal{MW}_{v_\wedge} = \{\{1, 2\}\}$	$\phi_1(N, v_\wedge) = \phi_2(N, v_\wedge) = {}^1\!/_2$
$\mathcal{MW}_{v_\vee} = \{\{1\}, \{2\}\}$	$\phi_1(N, v_\vee) = \phi_2(N, v_\vee) = {}^1\!/_2$

Values cannot represent sufficiency

Sufficient Feature Contribution

A value $\sigma^{0\text{–}1}$ on BFs is *sufficient* if

(i) if i is *null* (i.e. i is in no $\mathcal{MV}_v$), then $\sigma_i^{0\text{–}1}(N, v) = 0$,

(ii)$_a$ If $\{i\} \in \mathcal{MV}_v$ then $\sigma_i^{0\text{–}1}(N, v) = 1$,

(ii)$_b$ If i is a *veto* (i.e. i is in all $\mathcal{MV}_v$), then its influence cannot be smaller than that of any other player,

(iii) For $i, j \in N$: If for all $S \in \mathcal{MW}_v$ with $i \in S$, there exists $T \in \mathcal{MV}_v$ with $j \in T$ and $|S| \geq |T|$, then $\sigma_i^{0\text{–}1}(N, v) \leq \sigma_j^{0\text{–}1}(N, v)$.

How to define sufficient values on BFs?

Definition

$$\sigma_i^{0\text{–}1}(N,v) := \max_{S\in\mathcal{MW}_v\,:\,S\ni i} \frac{1}{|S|}.$$

Illustration

$$v_\wedge(S) = 1 \text{ iff } (1\in S)\wedge(2\in S) \quad \text{and} \quad v_\vee(S) = 1 \text{ iff } (1\in S)\vee(2\in S).$$

Prime Implicants	Game Theory
$\mathcal{MW}_{v_\wedge} = \{\{1,2\}\}$	$\sigma_1^{0\text{–}1}(N,v_\wedge) = \sigma_2^{0\text{–}1}(N,v_\wedge) = {}^1\!/_2$
$\mathcal{MW}_{v_\vee} = \{\{1\},\{2\}\}$	$\sigma_1^{0\text{–}1}(N,v_\vee) = \sigma_2^{0\text{–}1}(N,v_\vee) = 1$

Lemma

Value $\sigma^{0\text{–}1}$ is *sufficient.*

Outline

Outline

1. Context and Motivation
2. Case of Boolean Functions
   - Setting and definitions
   - Motivation and Proposal
3. Case of Pseudo-Boolean Functions
   - Definition & Properties
   - Construction of the optimal coalition

How to define sufficient values on PBFs?

How to extend σ^{0-1} to PBFs?

- Symmetry: players are no more symmetric in a $\mathcal{MW}_v$.
  ≫ Replace $\frac{1}{|S|}$ by $\phi_i(S, v_{|S})$.
- $\mathcal{MW}_v$: no more defined.
  ≫ Replace the min over elements of $\mathcal{MW}_v$ to any coalition.

Definition 0-1 games

$\sigma_i^{0-1}(N, v) := \max_{S \in \mathcal{MW}_v \,:\, S \ni i} \frac{1}{|S|}$

Definition on general games

$\sigma_i^{\phi}(N, v) := \max_{S \ni i} \phi_i(S, v_{|S})$

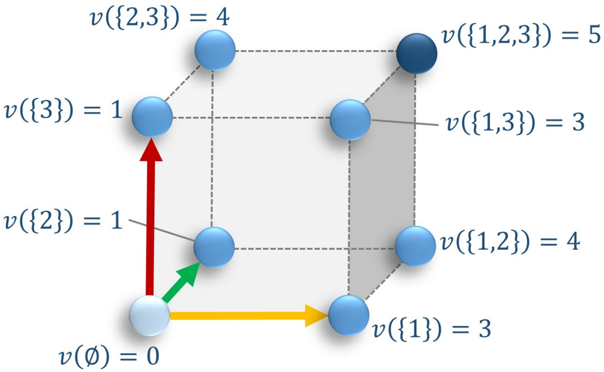
Lemma

For any 0-1 game v, we have $\sigma_i^{\phi^{\mathrm{PD}}}(N, v) = \sigma_i^{0-1}(N, v)$,

But $\sigma_i^{\phi^{\mathrm{Sh}}}(N, v) \neq \sigma_i^{0-1}(N, v)$.

How to define sufficient values on PBFs?

Illustration



	ϕ_1^{PD}	ϕ_2^{PD}	ϕ_3^{PD}
For $\{1,2,3\}$	**3**	1	1
For $\{1,2\}$	**3**	1	$\times$
For $\{1,3\}$	$^9/_4$	$\times$	$^3/_4$
For $\{2,3\}$	$\times$	**2**	**2**
For $\{1\}$	**3**	$\times$	$\times$
For $\{2\}$	$\times$	1	$\times$
For $\{3\}$	$\times$	$\times$	1
$\sigma^{\phi^{\text{PD}}} = \max \cdots$	**3**	**2**	**2**

Properties

Null Player (NP)

$\phi_i(N,v) = 0$ whenever i is *null* for v (i.e. $v(S \cup \{i\}) = v(S)$ for all $S \subseteq N \setminus \{i\}$).

Lemma

If ϕ satisfies $\mathbf{NP}$, so does σ^ϕ

Efficiency (E)

$\sum_{i\in N} \phi_i(N,v) = v(N)$.

Super Efficiency (SE)

$\sum_{i\in N} \phi_i(N,v) \geq v(N)$.

Lemma

If ϕ satisfies $\mathbf{E}$, then σ^ϕ satisfies $\mathbf{SE}$.

Essential Singleton (ES)

$\phi_i(N,v) = v(N)$ whenever $v(\{i\}) = v(N)$.

Lemma

If ϕ satisfies $\mathbf{E}$, then σ^ϕ satisfies $\mathbf{ES}$

Properties

Equal Treatment Property (ETP)

$\phi_i(N, v) = \phi_j(N, v)$ whenever $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subseteq N \setminus \{i, j\}$.

Lemma

If ϕ satisfies $\mathbf{ETP}$, so does σ^ϕ

Subset Dominance (SD)

$\phi_i(S, v) \geq \phi_i(S', v)$ for all $S' \subseteq S$.

Lemma

σ^ϕ satisfies $\mathbf{SD}$

Outline

A priori identification of the coalition realizing the max σ^ϕ

Problem statement:

How to identify a coalition realizing the maximum of the max in $I(N,v) := \sigma^\phi(N,v)$ without knowing explicitly ϕ?

Definition:

$$\mathcal{S}_i(N,v) = \Big\{ S \ni i \text{ such that } \phi_i(S, v_{|S}) \geq \phi_i(T, v_{|T}) \ \ \forall T \ni i \Big\}.$$

- $\mathcal{R}^{i,T} : \mathcal{G}(N) \to \mathcal{G}(N)$ defined for $T \subseteq N$ with $T \ni i$.
- $\mathcal{T}_i(N,v) = \{ T \ni i \text{ s.t. } I_i \left(N, \mathcal{R}^{i,T}(v)\right) = I_i(N,v) \}$
- $\underline{\mathcal{T}}_i(N,v)$: minimal elements of $\mathcal{T}_i(N,v)$ in the sense of $\subseteq$.

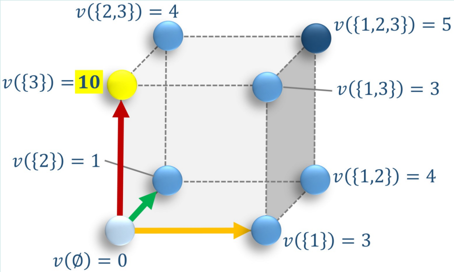
A priori identification of the coalition realizing the max σ^{ϕ}

Idea of $\mathcal{R}^{i,T}$: Modify v outside T so that $\max_{T \supseteq S} \phi_i(S, v_{|S})$ is very small.

$\mathcal{R}^{i,S}$: Case of ϕ^{PD}

$$v'(T) = \begin{cases} \vartheta & \text{if } T \subseteq N \setminus S \text{ and } |T| = 1 \\ v(T) & \text{otherwise} \end{cases}$$

Illustration on $\mathcal{R}^{1,\{1,2\}}$



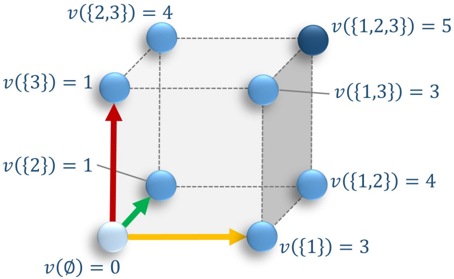
	ϕ_1^{PD}
For $\{1,2,3\}$	$^{15}/_{14}$
For $\{1,2\}$	3
For $\{1,3\}$	$^{12}/_{13}$
For $\{1\}$	3
$I_1(N, \mathcal{R}^{1,\{1,2\}}(v))$	3

$$\frac{3}{3+1+1} \times 5 \Longrightarrow \frac{3}{3+1+10} \times 5$$

$$\frac{3}{3+1} \times 4 \Longrightarrow \frac{3}{3+10} \times 4$$

A priori identification of the coalition realizing the max σ^ϕ

Illustration



T	$I_i(N, \mathcal{R}^{i,T}(v))$		
	$i = 1$	$i = 2$	$i = 3$
$\{1,2,3\}$	**3**	**2**	**2**
$\{1,2\}$	**3**	1	$\times$
$\{1,3\}$	$9/4$	$\times$	1
$\{2,3\}$	$\times$	**2**	**2**
$\{1\}$	**3**	$\times$	$\times$
$\{2\}$	$\times$	1	$\times$
$\{3\}$	$\times$	$\times$	1

- For $i = 1$: $\mathcal{T}_1(N, v) = \Big\{\{1,2,3\}, \{1,2\}, \{1\}\Big\}$ and $\underline{\mathcal{T}}_1(N, v) = \Big\{\{1\}\Big\}$
- For $i = 2$: $\mathcal{T}_2(N, v) = \Big\{\{1,2,3\}, \{2,3\}\Big\}$ and $\underline{\mathcal{T}}_2(N, v) = \Big\{\{2,3\}\Big\}$
- For $i = 3$: $\mathcal{T}_3(N, v) = \Big\{\{1,2,3\}, \{2,3\}\Big\}$ and $\underline{\mathcal{T}}_3(N, v) = \Big\{\{2,3\}\Big\}$

Are these axioms sufficient to derive I?

Lemma]

$\mathcal{I}_i(N, v) \subseteq \mathcal{S}_i(N, v) \subseteq \mathcal{T}_i(N, v)$.

Conclusion

Synthesis

- *Values* do not represent the idea of *sufficient explanation*
- $\sigma^{0\text{–}1}$: *sufficient* value restricted to 0–1 games
- σ^{ϕ}: *sufficient* value for general games
  - It uses a standard value ϕ

Extensions

- Non-Boolean variables
- Other baseline values