Explainable Decisions under Incomplete Knowledge with Supports and Weights

Florence Dupin de St-Cyr, **Romain Guillaume** and Umer Mushtaq

IRIT, Université de Toulouse, Toulouse, France,
Universite Paris II - Pantheon Assas, Paris, France

March 23, 2020

Contents

1. Introduction
2. Possibility Theory and Defaults
3. Bipolar Framework with Supports and Weights
4. Interpreting BLFSW in Possibility Theory
5. Conclusion and prospects

Introduction

- Qualitative and Argumentative decision problem
- Under incomplete knowledge
- Bipolar evaluation
- Bipolar Layered Framework = visual representation
  - Connection between possible knowledge and goals,
  - Explication of a decision choice
- Aim of this proposal:
  - Take into account the notion of support and weight.
  - Make explicit the principles governing decision in possibilistic context

Possibility Theory

A possibility distribution π :

- $\pi(x) \leq \pi(x')$: x' *is the actual situation* is at least as plausible as for x
- $\pi(x) = 0$: *impossibility*
- $\pi(x) = 1$: x *is unsurprising* or *normal*

The possibility measure Π of a formula:

- $\Pi(\varphi) = max_{\omega \models \varphi} \pi(\omega)$: *how unsurprising the formula is*,
- $N(\varphi) = 1 - \Pi(\neg\varphi) = min_{\omega \models \neg\varphi}(1 - \pi(\omega))$: Dual Necessity Measure

Defaults

Definition

A default rule $a \rightsquigarrow b$ = compact way to express a general rule without mentioning its exceptions.

In Possibility theory $a \rightsquigarrow b$: constraint $\Pi(a \wedge b) > \Pi(a \wedge \neg b)$.

Note: $\Pi(a \wedge b) > \Pi(a \wedge \neg b)$ equivalent to $N(b|a) > 0$

Bipolar Layered Framework with Supports and Weights (BLFSW)

Definition (BLFSW)

A Bipolar Layered Framework with Supports and Weights is a tuple $(\mathcal{P}, \mathcal{I}, \mathcal{S}, pol, \preceq, w)$.

- $\mathcal{P}$: *set of decision principles:* $\mathcal{P} = \{\varphi \rightsquigarrow g | \varphi \in \mathscr{L}_F, g \in LIT_G\}$
- $\mathcal{I} \subseteq (\mathcal{L}_F \times \mathcal{P})$: *set of inhibitors*
- $\mathcal{S} \subseteq (\mathcal{L}_F \times \mathcal{P})$: *set of supports*
- $pol : \mathcal{V}_G \to \{\oplus, \ominus\}$: *polarity of a goal* $g \in \mathcal{V}_G$.
- LIT_G *is totally ordered by the relation* $\preceq$
- $w : \mathcal{I} \cup \mathcal{S} \to]0, 1]$: *weight function on inhibitors and supports.*

.

Example of a BLFSW: Finding an Hotel

An agent wants to find a hotel which is:

- not expensive (e)
- in which she/he can swim (s)
- she/he prefers to avoid crowded hotels (c)

Information about the following attributes:

- "to have a pool" (p)
- "to be a four star hotel" (f)
- "to be in a place where the weather is fine" (w)
- "to propose special offers" (o)

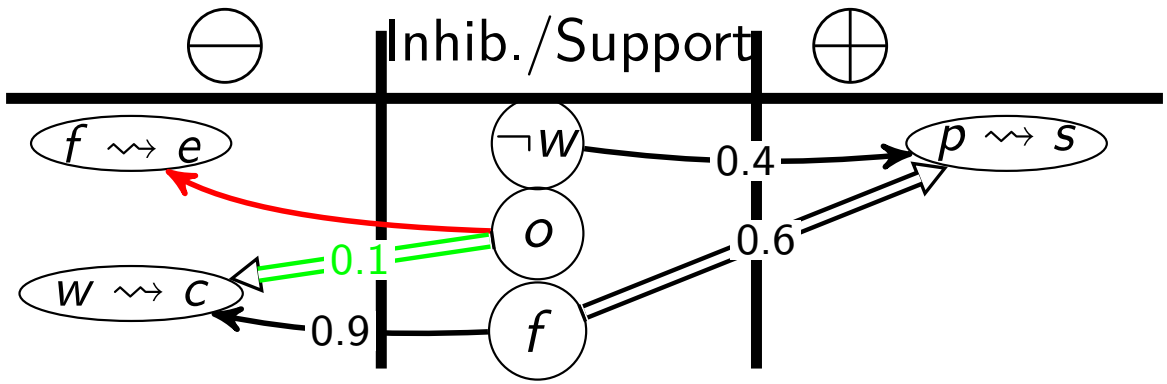
The agent may consider the following principles:

- “a priori when there is a pool the agent can swim” ($p \rightsquigarrow s$)
- “a priori if the hotel is four star then it is expensive” ($f \rightsquigarrow e$)
- “if the weather is fine in this area then the hotel is a priori crowded” ($w \rightsquigarrow c$)

Example of BLFSW: Find an Hotel

Principles can be supported or inhibited:

- A special offer (o) increases the certainty of $w \rightsquigarrow c$: to have a crowded hotel (c) when the weather is fine (w)
- Four star hotels are expensive: $f \rightsquigarrow e$ but a special offer may inhibit this deduction



Reasoning about goal achievements

First: Reasoning about goal achievements:

- a decision is based on the knowledge about the features
- reason wrt the argumentation graph: check the realized goals

K-BLFSW

Definition (K-BLFSW)

Given a consistent base K and a BLFSW $B = (\mathcal{P}, \mathcal{I}, \mathcal{S}, pol, \preceq, w)$
A K-BLFSW associated to B is $(\mathcal{P}_K, \mathcal{I}_K, \mathcal{S}_K, pol, \preceq, w_K)$ where

- $\mathcal{P}_K = \{(\varphi, g) \in \mathcal{P}, \text{ s.t. } K \models \varphi\}$: *set of valid DPs (wrt K)*
- $\mathcal{I}_K = \{(\varphi, p) \in \mathcal{I}, \text{ s.t. } K \models \varphi \text{ and } p \in \mathcal{P}_K\}$: *set of valid inhibitions*
- $\mathcal{S}_K = \{(\varphi, p) \in \mathcal{S}, \text{ s.t. } K \models \varphi \text{ and } p \in \mathcal{P}_K\}$: *set of valid supports*
- w_K: *restriction of w on $\mathcal{I}_K \cup \mathcal{S}_K$.*

Realized goals

Definition (activation level of a principle)

Given $(\mathcal{P}_K, \mathcal{I}_K, \mathcal{S}_K, pol, \preceq, w_K)$, *the activation level of* $p \in \mathcal{P}_K$ *is*

$$\alpha(p) = \sum_{s \in \mathcal{S}_K(p)} w_K(s,p) - \sum_{i \in \mathcal{I}_K(p)} w_K(i,p)$$

p *is* *inhibited* *iff* $\alpha(p) < 0$ $\quad$ p *is* *supported* *iff* $\alpha(p) > 0$

p *is* *unaffected* *iff* $\alpha(p) = 0$

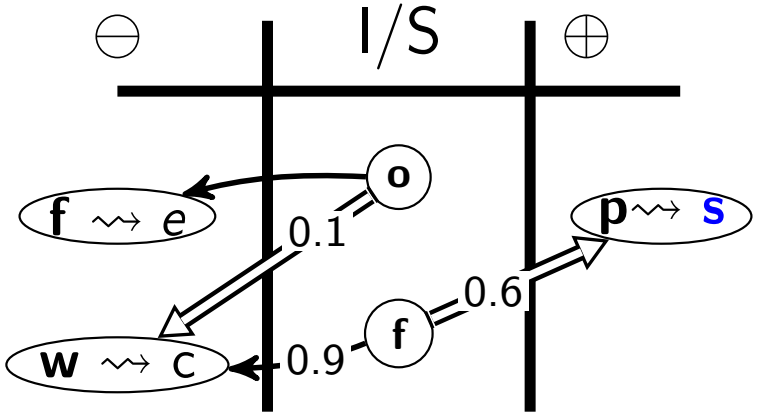
Definition (realized goals)

A goal g *is* *realized* *if there is a principle that concludes* g *and that is not* *inhibited*.

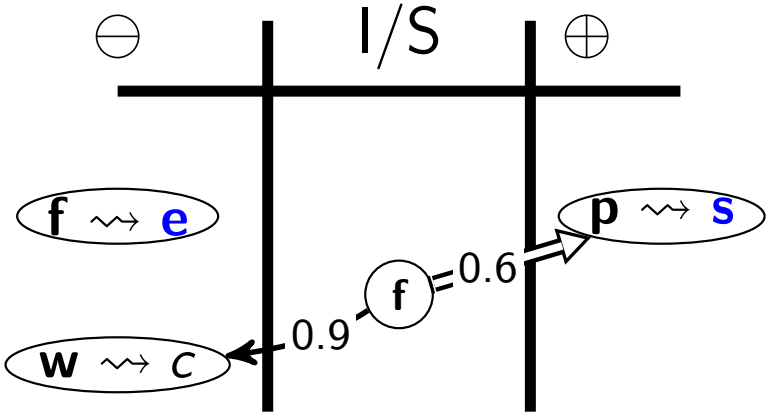
Example of K-BLFSW

The same BLFSW used with 2 Knowledge Bases K_1 and K_2:

$K_1 = \{p \wedge w \wedge f \wedge o\}$



$K_2 = \{p \wedge w \wedge f\}$



Realized goals:

$R_{K1} = \{s\}$

$R_{K2} = \{s, e\}$

Handling preferences

Decide if K is preferred to K': use Bipolar decision rules

Definition (BiLexi decision rule)

Given two alternatives K and K' with resp. realized goals R and R': The Bipolar Leximin dominance $\succeq_{BiLexi}$ (BiLexi-preferred to) is s.t.:

$$\begin{array}{lll} K \succ_{BiLexi} K' & \text{iff} & \left| \begin{array}{l} M \text{ exists and} \\ |\mathrm{R}_M^{\oplus}| \geq |\mathrm{R}'^{\oplus}_M| \text{ and } |\mathrm{R}_M^{\ominus}| \leq |\mathrm{R}'^{\ominus}_M| \end{array} \right. \\ K \simeq_{BiLexi} K' & \text{iff} & M \text{ does not exist} \end{array}$$

where $M = \max(\{g \in \mathrm{R} \cup \mathrm{R}' \text{ s.t. } |\mathrm{R}_g^{\oplus}| \neq |\mathrm{R}'^{\oplus}_g| \text{ or } |\mathrm{R}_g^{\ominus}| \neq |\mathrm{R}'^{\ominus}_g|\}, \preceq)$ *and* $\mathrm{R}_g^{\oplus}$ *=set of positive realized goals at same level as* g

Refining the classical rules using Certainty evaluation

Definition (Certainty of a realized goal)

Given a BLFSW $B = (\mathcal{P}, \mathcal{I}, \mathcal{S}, pol, \preceq, w)$ *and an alternative described by* K*, for all goal* g *realized in* B_K*, the certainty associated to* g *is:*

$$\alpha_K(g) = \max_{\varphi \in \mathcal{L}_F, p = \varphi \rightsquigarrow g \in \mathcal{P}_K} \alpha(p)$$

Example

$K_2 = \{p \wedge w \wedge f\}$, $K_3 = \{p \wedge \neg w \wedge f\}$ same realized goals: $\{e, s\}$
But $\alpha_{K_2}(s) = 0.6$ and $\alpha_{K_3}(s) = 0.2$

Interpreting a BLFSW in Possibility Theory

Given a possibility measure Π,

Definition (Π-DP)

A Π-DP $\varphi \rightsquigarrow g$ is s.t. $N(g|\varphi) > 0$

Definition (Π-Inhibitor)

ψ is a Π-Inhibitor of $\varphi \rightsquigarrow g$ if $N(g|\varphi \wedge \psi) = 0$

Definition (Π-Support)

ψ is a Π-Support of $\varphi \rightsquigarrow g$ if $N(g|\varphi \wedge \psi) > N(g|\varphi)$

Interpreting a BLFSW in Possibility Theory

Given a possibility measure Π

Definition (Π-weight)

w is a Π-weight function iff for all possible K-BLFSW ($\mathcal{P}_K$, $\mathcal{I}_K$, $\mathcal{S}_K$, pol, $\preceq$, w_K) where w_K is the restriction of w on $\mathcal{I}_K \cup \mathcal{S}_K$ and for all decision principles $p = \varphi \rightsquigarrow g$, $p' = \varphi' \rightsquigarrow g' \in \mathcal{P}_K$

- $\alpha(p) < 0$ *iff* $N(g \mid \varphi \bigwedge_{\psi \in \mathcal{I}_K(p) \cup \mathcal{S}_K(p)} \psi) = 0$
- $\alpha(p) \geq 0$ *iff* $N(g \mid \varphi \bigwedge_{\psi \in \mathcal{I}_K(p) \cup \mathcal{S}_K(p)} \psi) > 0$
- $\alpha(p) \geq \alpha(p') \geq 0$ *iff* $N(g \mid \varphi \bigwedge_{\psi \in \mathcal{I}_K(p) \cup \mathcal{S}_K(p)} \psi) \geq N(g' \mid \varphi' \bigwedge_{\psi \in \mathcal{I}_K(p') \cup \mathcal{S}_K(p')} \psi) > 0$

with $\alpha(p)$ defined from w_K

Interpreting a BLFSW in Possibility Theory

Proposition

Given a Π*-Bipolar Layered Framework with Supports and Weights* B*,built on a possibility distribution* π *and on a utility function* u *on* LIT_G*.*

B *is s.t. for all consistent knowledge bases* K, K' *and for any goals* g, g'*:*

- g *not realized wrt* B_K *iff* $N(g|K) = 0$ *or* $u(g) = 0$
- *if* g *realized wrt* B_K *and* g' *realized wrt* $B_{K'}$ *then* $\alpha_K(g) > \alpha_{K'}(g')$ *iff* $N(g|K) > N(g'|K')$

Conclusion

Conclusions

- The BLFSW is a visual tool made to help human decision makers in their tasks.
- BLFSW can justify a decision by giving the features that support and attack the goals.
- BLFSW can be built from a possibility distribution (help the decision maker to understand decision situations)

Prospects

Prospects

- Link between other criteria for qualitative decision under uncertainty and BLFSW
- Better understand the relation between polarity and optimism/pessimism of the decision maker.
- Learn a BLFSW from data with knowledge and utility or knowledge and ranking