Justifications derived from inconsistent case bases using authoritativeness

ArgXAI 2022

Joeri Peters
Utrecht University & Netherlands National Police
j.g.t.peters@uu.nl

Floris Bex
Utrecht University & Tilburg University
f.j.bex@uu.nl

Henry Prakken
Utrecht University & University of Groningen
h.prakken@uu.nl

Contents

AF-CBA

- Post hoc analysis
- Model agnostic
- Access to training data
- Local explanations (justifications)

- An explanation takes the form of an argument game for grounded semantics played between a proponent and opponent of an outcome, in which they take turns to attack the other's last argument.

General idea

- Case base CB is the set of cases (precedents).
- A case c = ($F(c)$, $outcome(c)$) is a fact situation and an outcome (e.g. s).
- A fact situation contains dimensions (features) with values ($v(d, c) = x$)
- Focus case: current case and its predicted outcome

- *A fortiori* assumption:

  The focus case should have the same outcome as a precedent case if the differences between these cases only serve to add further support for that same outcome.

Comparing cases

- Given two fact situations F and F', $F \leq_s F'$ iff $v \leq_s v'$ for all $(d, v) \in F$ and $(d, v') \in F'$.
- Given case base CB and fact situation F, deciding F for s is forced iff CB contains a case $c = (F', s)$ such that $F' \leq_s F$.
- Any value assignment in the focus case that is not at least as favourable for the outcome as in the precedent is a relevant difference:

$$D(c, f) = \{(d, v) \in F(c) \mid v(d, c) \not\leq_s v(d, f)\}$$

- A best precedent to cite is one with the same outcome as the focus case and with the fewest relevant differences.

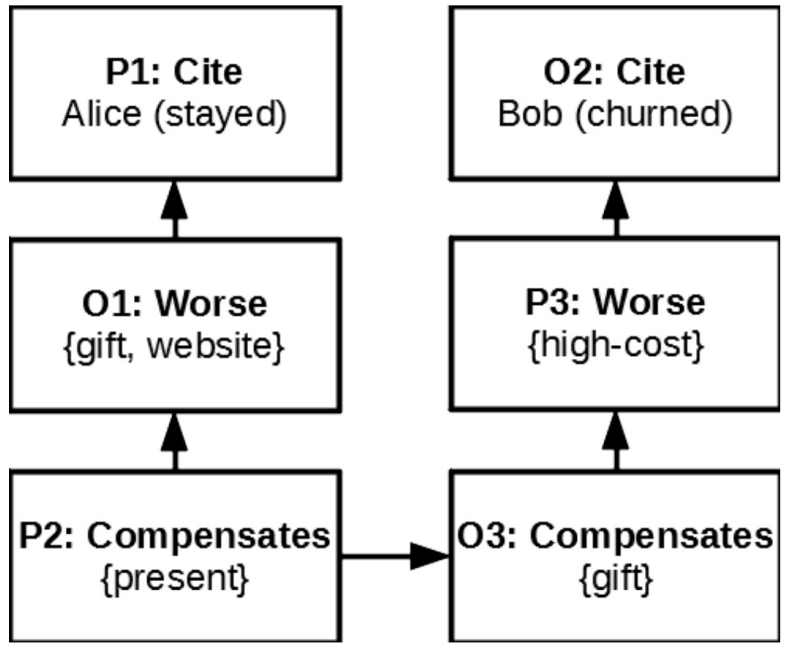
Arguing over cases

- Proponent argues in favour of the predicted outcome, opponent argues against this by:
  - Citing a precedent $c \in CB$
  - Distinguishing a cited precedent from focus case $f \notin CB$ by:
    - $Worse(c, x)$: f is on some dimensions x worse than c for $outcome(c)$.
    - $Compensates(c, x, y)$: the dimensions x on which f is not at least as good as c for $outcome(c)$ are compensated by dimensions y on which f is better for $outcome(c)$ than c.
    - $Transformed(c, c')$: initial citation can be transformed by the distinguishing moves into a case for which $D(c, f) = \emptyset$ and which can therefore attack the counterexample.

Example

Dimension	Name	Description
$d_1^{\downarrow}$	Gift	Whether the customer has received a gift from the provider
$d_2^{\downarrow}$	Present	Whether the customer was present during the last organised event
$d_3^{\downarrow}$	Website	The number of times the customer logged into their a profile
$d_4^{\uparrow}$	High cost	Whether the customer is in a high-cost category

Customer	$d_1^{\downarrow}$	$d_2^{\downarrow}$	$d_3^{\downarrow}$	$d_4^{\uparrow}$	Label (churn)
Alice	1	0	5	0	0
Bob	1	1	3	1	1
Charlie (focus)	0	1	3	0	?



Inconsistency

- A case base CB is consistent iff it does not contain two cases with opposite outcomes such that $F \leq_s F'$.

- Case bases can be inconsistent because:
  - Annotators/decision makers can make mistakes or disagree.
  - Feature vectors can lack information to distinguish data.

- AF-CBA does not strictly require that the CB be consistent, but inconsistencies are often due to exceptional cases (with a surprising outcome) and these can be problematic for the explanation due to the focus case being forced for both outcomes.

- Explanations containing inconsistent forcings essentially explain that a decision cannot be justified without acknowledging the inconsistency of the CB, which weakens the value of those explanations.

- The larger the number of inconsistent forcings (N_{inc}), the larger the number of explanations where this problem occurs.

Authoritativeness

- In earlier work, $N_{inc} = 0$ was achieved through case deletions. We would prefer to leave the case base (training data) intact and take the inconsistency of cases into account, preferably in a way which is intuitive.

- Previously: "Cases like this *always* receive outcome o"
- With this modification: "Cases like this *usually* receive outcome o"

- Cite the best precedent which has the highest value for authoritativeness $\alpha(c)$.

Alternative expressions

$n_a(c)$ = | $\{c' \in CB \mid outcome(c') = outcome(c)$ and $D(c, c') = \emptyset\}$ |
The number of cases with the same outcome without any relevant differences.

$$\alpha(c) = \frac{n_a(c)}{n_a(c) + n_d(c)} \qquad \alpha(c) = \frac{n_a(c)}{|\, CB \,|}$$

Both expressions capture some of the intuitive understanding of how authoritative a case is. We could therefore combine the two, e.g. as a product or harmonic mean.

This currently gives us four expressions for authoritativeness.

Example

Customer	$d_1^{\downarrow}$	$d_2^{\downarrow}$	$d_3^{\downarrow}$	$d_4^{\uparrow}$	$outcome$
c_1	1	1	0	0	s
c_2	1	1	0	0	s
c_3	1	1	0	0	s
c_4	1	1	2	0	s
c_5	1	1	2	0	s
c_6	1	1	2	0	$\overline{s}$
c_7	1	1	15	0	s

$\alpha(c_1) = 3/(3 + 0) = 1$
or
$\alpha(c_1) = 3/7 \approx 0.429$

Evaluation

	Base	Relative (1)	Absolute (2)	Product (3)	Harmonic ($\beta = 1$) (4)
Admission	$\mu = 105.67$ $N_{inc} = 496$	$\mu = 112.1$ $N_{inc} = 0$	$\mu = 105.95$ $N_{inc} = 0$	$\mu = 106.0$ $N_{inc} = 0$	$\mu = 105.97$ $N_{inc} = 0$
Churn	$\mu = 82.15$ $N_{inc} = 38012$	$\mu = 148.81$ $N_{inc} = 2$	$\mu = 94.68$ $N_{inc} = 42$	$\mu = 94.76$ $N_{inc} = 0$	$\mu = 94.75$ $N_{inc} = 0$
Mushroom	$\mu = 70.25$ $N_{inc} = 620$	$\mu = 72.37$ $N_{inc} = 0$	$\mu = 84.66$ $N_{inc} = 0$	$\mu = 86.75$ $N_{inc} = 0$	$\mu = 84.83$ $N_{inc} = 0$

Discussion and future work

- Why not use a simpler model instead?
  - There are problems for which the only satisfactory solutions are too opaque for practitioners.
- These metrics are only proxies, usability studies are required to study interpretability.
- Additional criteria for authoritativeness might turn out to be important, rendering our current expressions obsolete.
- Additional modifications of AF-CBA:
  - Other criteria for ranking precedents
  - Incorporating complex arguments
  - Accounting for highly dependent dimensions
  - Allow for dimensions with complex tendencies

Conclusion

- Modification of AF-CBA so as to include the reality of inconsistent case bases, expressed through various alternative formulations.

- The authors would like to thank the anonymous reviewers for their feedback and suggestions.

- H. Prakken, R. Ratsma, A top-level model of case-based argumentation for explanation: formalisation and experiments, Argument & Computation Preprint (2021) 1–36. doi:10.3233/AAC-210009, publisher: IOS Press.

Utrecht University

DISCLAIMER

The information in this presentation has been compiled with the utmost care,
but no rights can be derived from its contents.