Justifications derived from inconsistent case bases using authoritativeness

Joeri G. T. Peters[1,2], Floris J. Bex[1,3] and Henry Prakken[1,4]

[1]Department of Information and Computing Sciences, Utrecht University, the Netherlands
[2]National Police Lab AI, Netherlands National Police, the Netherlands
[3]Tilburg Institute for Law, Technology and Society, Tilburg University, the Netherlands
[4]Faculty of Law, University of Groningen, the Netherlands

Abstract

Post hoc analyses are used to provide interpretable explanations for machine learning predictions made by an opaque model. We modify a top-level model (AF-CBA) that uses case-based argumentation as such a post hoc analysis. AF-CBA justifies model predictions on the basis of an argument graph constructed using precedents from a case base. The effectiveness of this approach is limited when faced with an inconsistent case base, which are frequently encountered in practice. Reducing an inconsistent case base to a consistent subset is possible but undesirable. By altering the approach's definition of best precedent to include an additional criterion based on an expression of authoritativeness, we allow AF-CBA to handle inconsistent case bases. We experiment with four different expressions of authoritativeness using three different data sets in order to evaluate their effect on the explanations generated in terms of the average number of precedents and the number of inconsistent *a fortiori* forcing relations.

1. Introduction

Both machine learning (ML) and rule-based classification approaches involve a trade off between accuracy and transparency, specifically the ability of end-users to understand decisions (class predictions) [1]. Deep neural networks in particular tend to produce predictions with a high degree of accuracy at the cost of transparency due to their technical complexity. However, the perceived complexity may vary according to a person's level of understanding, so even much simpler approaches might be thought of as relatively opaque by some people. Another reason for poor transparency can be proprietary protection of the approach, which can render even a relatively simple approach opaque. Regardless of the underlying reason, the term 'black box' is often used to refer to a particularly opaque approach [1, 2]. A black box model is more difficult to trust. It is harder to see its shortcomings, including biases and ethical concerns, which is why Explainable Artificial Intelligence (XAI) is aimed at increasing the transparency of black box models [3]. In the case of a binary classification problem in ML, this entails that we can explain why one class label was predicted by the model and not the other in a particular instance.

Methods for explaining ML decisions vary in a number of respects. A distinction can be made between methods that generate local explanations (explaining individual instances) and those

ArgXAI 2022: 1st International Workshop on Argumentation for eXplainable AI, September 12, 2022, Cardiff, Wales, UK.
j.g.t.peters@uu.nl (J. G. T. Peters); f.j.bex@uu.nl (F. J. Bex); h.prakken@uu.nl (H. Prakken)

CEUR Workshop Proceedings (CEUR-WS.org)

that generate global explanations (explaining a whole model). Some methods have access to the learnt model, while others are model agnostic. We use the term ‘justifications’ for explanations generated without model access to signify that such explanations do not explain exactly how a decision was reached, but instead they explain the assumptions under which the model’s decision can be justified. Justifications are thus not intended as separate predictions. This is related to the notion of *post hoc* analysis, which implies that an explanation is produced after the fact [1]. In this paper, we are concerned with local justifications produced by a model-agnostic, post hoc analysis for binary classification.

One way to justify a classifier’s prediction is to show a most similar case to one whose class is being predicted (the focus case). Pointing out a similar case constitutes an argument that requires no knowledge of ML to understand. To this end, Prakken & Ratsma [4] draw on AI & law research to propose a top-level model using case-based argumentation (CBA) to explain black-box predictions based on Horty’s model of *a fortiori* reasoning [5, 6] and inspired by CATO [7], hereafter referred to as ‘A Fortiori Case-Based Argumentation’ (AF-CBA). As ML classification is a supervised approach, there exists a training set used to train a classifier. AF-CBA requires that this set be accessible. AF-CBA produces a human-interpretable justification of that classifier’s binary prediction for a focus case by treating this training set as a case base (CB) and comparing precedents and their outcomes from that CB to the focus case. The underlying a fortiori assumption of AF-CBA is that 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 [4].

When asked to provide a justification, AF-CBA constructs an argument graph through a grounded argument game consisting of a fixed set of allowed moves. A proponent defends why the focus case should receive the same outcome as a best precedent (a most similar case) and the opponent argues against this. In doing so, they cite examples and counterexamples from the CB. There are distinguishing moves that set cases apart and moves to downplay these differences. When a precedent case has no relevant differences with the focus case, deciding for the focus case is said to be ‘forced’. The effectiveness of AF-CBA hinges in part on the distance measure between cases and any feature selection technique used to promote an interpretable argument graph [4].

CBs may contain inconsistencies. Because training sets are used as CBs, they constitute annotated data, i.e. data instances labelled by a person or process with the intention of allowing the ML model to learn to perform the same classification task. Annotators (people who label data to this end) produce a labelled data set specifically for the purpose of training a model, but may not necessarily be fully consistent when doing so [8]. Multiple annotators might disagree or an annotator can make an occasional mistake, thus leading to an inconsistent case base. Labels may also be produced by decision makers—people who produce labels as part of their role in some decision process, such as judges who decide on court cases, with their verdict being the label that is stored in a body of case law. This can also lead to contradictory classifications, as case law can contain conflicting opinions and interpretations. Finally, the feature vector itself may be a subset of all relevant details, thereby potentially lacking necessary data to discriminate between seemingly similar cases [9]. These sources of noise make the labelling seem inconsistent, since identical feature vectors might receive conflicting labels. Under the a fortiori assumption, this notion of inconsistency becomes even broader: a case which is at least as good as another yet receives the opposite outcome is a source of inconsistency. For these reasons, CB consistency is generally not a safe assumption in practice.

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. In experiments by Prakken & Ratsma [4], significant portions of a CB had to be ignored (by removing a minimal number of cases when instantiating the CB) in order to make them consistent—namely 0.32%, 11.35% and 3.20% for three different inconsistent data sets. We would preferably use the whole training set as a CB, without having to take this consistent subset to circumvent the problem. The problem is exasperated by feature selection techniques, which would otherwise benefit the simplicity of AF-CBA's explanations. In conclusion, CB consistency forms a problematic constraint for AF-CBA.

In this paper, we present a modification of AF-CBA that takes into account the degree (which we call 'authoritativeness') to which the CB is consistent in regards to a best precedent. This measure is used to prevent inconsistent forcing by modifying the selection of best precedents to cite, as it makes intuitive sense to cite cases with the highest authoritativeness. We investigate the desirability of this modification through exploratory experiments with several possible alternatives of quantifying authoritativeness, demonstrating it to have a beneficial effect on AF-CBA without adversely affecting its explanations. The rest of this paper is structured as follows. We describe AF-CBA and its background in Section 2. We consider how to address the problem of inconsistency in Section 3. We subsequently experiment with our proposed solution in Section 4. We discuss the results and future work in Section 5.

2. Case-Based Argumentation

In this section, we present the CBA framework by Prakken & Ratsma (with some differences in notation) for explanations with dimensions [4]. As our running example, we make use of the Telco Customer Churn data set [10], which describes the customers of a telecommunications provider and whether or not they churned (switched providers). Table 1 describes the dimensions ('features' in ML) used. The optional superscript arrow reflects the tendency of a dimension, i.e. whether a higher value promotes a result of 1 for the class label. Here, only the dimension of $high\ cost$ makes it likelier for a customer to churn; the other three dimensions make it less likely for a customer to do so.

Table 1
The dimensions used in the Churn 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

We take Table 2 as our example CB. Let us presume that Alice and Bob are previous customers and Charlie is a new customer whose predicted outcome we want to justify (the focus case).

Table 2
A fictional example based on the Churn data set with a CB consisting of only two cases and a new (focus) case.

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	?

Formally, we denote this as follows. Let o and o' be the two possible outcomes of a case in the CB. The variables s and $\bar{s}$ denote the two sides, meaning that $s = o$ if $\bar{s} = o'$ and vice versa. A dimension is defined as a tuple $d = (V, \leq_o, \leq_{o'})$, with value set V and two partial orderings on V, $\leq_o$ and $\leq_{o'}$, such that $v \leq_o v'$ iff $v' \leq_{o'} v$ for $v, v' \in V$. A value assignment is a pair (d, v). We denote the value x of dimension d as $v(d, c) = x$ for case $c \in CB$. Value assignments to all dimensions $d \in D$ (where D is nonempty) constitute a fact situation F. A case is defined as $c = (F, outcome(c))$ for such a fact situation and an $outcome(c) \in \{o, o'\}$. In this context, a case base CB specifically refers to the set of cases with value assignments for D. We denote the fact situation of a case c as $F(c)$. In the rest of this paper, we assume that any two fact situations assign values to the same set D.

Say we have a ML model that predicts Charlie will stay. An explanation of this outcome could be that it is forced by another case. We model Horty's [5] a fortiori reasoning using Definitions 1 and 2, meaning that the outcome of a focus case is forced if there is a precedent with the same outcome where all their differences make the focus case even stronger for that outcome.

Definition 1 (Preference relation for fact situations). *Given two fact situations F and F', $F \leq_s F'$ iff $v \leq_s v'$ for all $(d, v) \in F$ and $(d, v') \in F'$.*

Definition 2 (Precedential constraint). *Given case base CB and fact situation F, deciding F for s is forced iff CB contains a case $c = (F', s)$ such that $F' \leq_s F$.*

A fact situation could be forced for both s and $\bar{s}$, which brings us to the following definition of CB consistency:

Definition 3 (Case base consistency). *A case base CB is consistent iff it does not contain two cases $c = (F, s)$ and $c' = (F', \bar{s})$ such that $F \leq_s F'$. Otherwise it is inconsistent.*

An explanation takes the form of an argument game for grounded semantics [11] played between a proponent and opponent of an outcome, in which they take turns to attack the other's last argument. Since neither of the cases in Table 2 is identical to the focus case, it is not forced and the proponent and opponent argue about the outcome. An argument is justified if the proponent has a winning strategy, meaning the opponent runs out of moves. The proponent starts by citing a best precedent. This is a case which has the outcome for which the proponent is arguing and has a minimal subset of relevant differences with the focus case. Determining the relevant differences between two cases is defined according to Definition 4 (as presented in [4]).

Definition 4 (Differences between cases). *Let $c = (F(c), outcome(c))$ and $f = (F(f), outcome(f))$ be two cases. The set $D(c, f)$ of differences between c and f is:*

1. $D(c, f) = \{(d, v) \in F(c) \mid v(d, c) \not\leq_s v(d, f)\}$ *if* $outcome(c) = outcome(f) = s$.
2. $D(c, f) = \{(d, v) \in F(c) \mid v(d, c) \not\geq_{\bar{s}} v(d, f)\}$ *if* $outcome(c) \neq outcome(f)$ *and* $outcome(c) = s$.

So one relevant difference between Charlie (assuming he is predicted to stay) and Alice (who stayed) in Table 2 would be for dimension d_1, where Alice received a gift and Charlie did not, making her case better for staying.

Definition 5 (Best precedent)**.** *Let* $c = (F(c), outcome(c))$ *and* $f = (F(f), outcome(f))$ *be two cases, where* $c \in CB$ *and* $f \notin CB$. c *is a best precedent for* f *iff:*

- $outcome(c) = outcome(f)$ *and*
- *there is no* $c' \in CB$ *such that* $outcome(c') = outcome(c)$ *and* $D(c', f) \subset D(c, f)$.

Definition 5 defines a best precedent to cite. Multiple cases can meet these criteria. A lower number of best precedents is preferable, because of computational reasons and because one could say that a higher number of possible citations would make a single explanation somewhat arbitrary. This is why Prakken & Ratsma evaluated AF-CBA in part on the average number of best precedents found for three different data sets [4].

The opponent can reply to the initial citation by playing a distinguishing move or by citing a counterexample. The proponent can reply in turn with similar distinguishing moves. The distinguishing moves are $Worse(c, x)$ — the focus case is on some dimensions x worse than the precedent c for $outcome(c)$ — , $Compensates(c, x, y)$ — the dimensions x on which the focus case is not at least as good as the precedent c for $outcome(c)$ are compensated by dimensions y on which the focus case is better for $outcome(c)$ than c — and $Transformed(c, c')$ — the initial citation of a most similar case for $outcome(f)$ can be transformed by the distinguishing moves into a case for which $D(c, f) = \emptyset$ and which can therefore attack the counterexample. For the sake of brevity, see [4] for formal motivations of these moves and the need to allow the $Compensates$ move to be empty in order to state that the differences with the focus case do not matter.

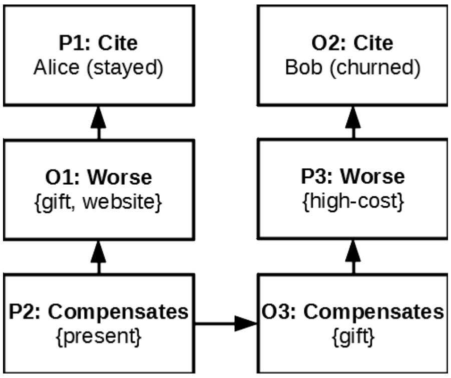

Figure 1: A fictional example of an explanation (dialogue between proponent and opponent).

Returning to our example, Figure 1 presents the resulting explanation as an argument game, which can be read as follows. P1: Alice stayed and her case is similar to Charlie's. O1: Charlie's scores for $d_1^{\downarrow}$ and $d_3^{\downarrow}$ make him worse for staying than Alice. P2: Charlie's score for $d_2^{\downarrow}$ compensates for O1. O2: Bob churned and his case is similar to Charlie's. P3: Charlie's score for $d_4^{\uparrow}$ makes him worse for churning than Bob. O3: Charlie's score for $d_1^{\downarrow}$ compensates for P3. P2: Charlie's score for $d_2^{\downarrow}$ compensates for O3. After this, the opponent has run out of possible moves to make and the proponent wins. The similarity to Alice's case has held up and acts as a explanation for the prediction that Charlie will stay as well.

Formalising this brings us to the following definition (after [4]) for the AF-CBA framework:

Definition 6 (Case-based argumentation framework)**.** *Given a finite case base CB, a focus case $f \notin CB$, and definitions of compensation dc, an abstract argumentation framework AAF is a pair $< \mathcal{A}, attack >$, where:*

- $\mathcal{A} = CB \cup M$,
  with $M = \{Worse(c, x) \mid c \in CB$, $x \neq \emptyset$ *and*
  $x = \{(d, v) \in F(f) \mid v(d, f) <_{outcome(f)} v(d, c)\}\} \cup$
  $\{Compensates(c, y, x) \mid c \in CB$, $y \subseteq \{(d, v) \in F(f) \mid v(d, c) <_{outcome(f)} v(d, f)\}$,
  $x = \{(d, v) \in F(f) \mid v(d, f) <_{outcome(f)} v(d, c)\}$ *and* y *compensates* x *according to* $dc\} \cup$
  $\{Transformed(c, c') \mid c \in CB$ *and* c *can be transformed into* c' *and* $D(c', f) = 0\}$
- *A attacks B iff:*
  - *$A, B \in CB$ and $outcome(A) \neq outcome(B)$ and $D(B, f) \not\subset D(A, f)$;*
  - *$B \in CB$ with $outcome(B) = outcome(f)$ and A is of the form $Worse(B, x)$;*
  - *B is of the form $Worse(c, x)$ and A is of the form $Compensates(c, y, x)$;*
  - *$B \in CB$ and $outcome(B) \neq outcome(f)$ and A is of the form $Transformed(c, c')$.*

In summary, AF-CBA provides justifications for individual binary classifications predicted by a ML model by presenting a winning strategy for a grounded argument game in favour of the predicted class label. This winning strategy represents a dialogue between a proponent and opponent on the basis of citations from the labelled training set (the case base) and shows how the opponent runs out of moves and the proponent thus wins the argument.

3. CB inconsistency

As we argued in Section 1, CB consistency is not always a safe assumption to make. 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.

Instead of mitigating the problem through case deletion [4], we explicitly take inconsistencies into account. Informally, one might say that when there is consistency, a precedential case has a strong backing when cited and should indeed immediately force the outcome; if there is inconsistency, it has less backing and thus should not. We therefore introduce the concept of 'authoritativeness', by which we mean that, given any case $c \in CB$, the authoritativeness $\alpha(c)$ numerically expresses (normalised between 0 and 1) the degree to which the rest of the CB supports the citing of c for $outcome(c)$. We subsequently use $\alpha(c)$ as an additional criterion in the selection of best precedents. The intuition behind authoritativeness is that whereas the a fortiori rule applied to a consistent CB can be expressed as the phrase 'cases like this always receive outcome o,' our idea of authoritativeness changes this phrase to 'cases like this *usually* receive outcome o'—where 'usually' has to be quantified in some manner which expresses the inconsistency of the CB with regards to the focus case. Since $\alpha(c)$ is a number, we can have a total ordering $\leq$ on the authoritativeness of cases.

Table 3 is another instance of our Churn example. Depending on how one chooses to define $\alpha(c)$, c_1 and c_2 should arguably receive a higher value for $\alpha(c)$ than c_3 due to c_4 having the opposite outcome.

Table 3
Example of a CB with two identical cases that are consistent with each other and two identical cases which contradict each other.

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	5	0	s
c_4	1	1	5	0	$\bar{s}$

First of all, the definition of best precedent has to be modified to reflect the additional criterion of maximising the authoritativeness:

Definition 7. *(Best authoritative precedent) Let CB be a case base and let $c = (F(c), outcome(c))$ and $f = (F(f), outcome(f))$ be two cases, where $c \in CB$ and $f \notin CB$. c is a best precedent for f iff:*

- *$outcome(c) = outcome(f)$,*
- *there is no $c' \in CB$ such that $outcome(c') = outcome(c)$ while $D(c', f) \subset D(c, f)$ and $\alpha(c') \geq \alpha(c)$.*

In order to quantify authoritativeness, we require expressions of agreement and disagreement between a precedent and the rest of the CB:

Definition 8. *(Agreement) Let CB be a case base. Given $c \in CB$, the agreement $n_a(c)$ is defined as:*
$n_a(c) = | \{c' \in CB \mid outcome(c') = outcome(c) \text{ and } D(c, c') = \emptyset\} |$

Definition 9. *(Disagreement) Let CB be a case base. Given $c \in CB$, the disagreement $n_d(c)$ is defined as:*
$n_d(c) = | \{c' \in CB \mid outcome(c') \neq outcome(c) \text{ and } D(c, c') = \emptyset\} |$

We understand $n_a(c)$ as the number of cases which have the same outcome as the precedent case and are at least as good for that outcome as c (thereby lending support to c). Similarly,

$n_d(c)$ is the number of cases which have the opposite outcome yet are at least as good for $outcome(c)$. The agreement $n_a(c)$ has at least a value of 1 due to c itself being a member of the CB. The disagreement $n_d(c)$ can have a value of 0.

Exactly how the level of agreement relates to authoritativeness is not self-evident, as various expressions may have equal merit. For example, given a case $c \in CB$, we could express the authoritativeness $\alpha(c)$ as the relative number of cases which lend further support to c (1). In Table 3, c_3 is supported by (other than itself) c_1 and c_2, but opposed by c_4. So in that situation, $\alpha(c_3) = 3/(3+1) = 0.75$.

$$\alpha(c) = \frac{n_a(c)}{n_a(c) + n_d(c)} \tag{1}$$

However, this overlooks any intuitive understanding of authoritativeness which stems from the absolute number of cases that can act as precedents (2). Intuitively, obscure cases are less authoritative than common ones. In Table 4, c_1 is supported by two other cases (again, other than itself), namely c_2 and c_3, while c_5 is supported by c_1 through c_4. We divide by $\mid CB \mid$ to normalise the expression between 0 and 1. So for example $\alpha(c_1) = 3/(3+0) = 1$ according to (1) but $\alpha(c_1) = 3/7 \approx 0.429$ according to (2).

$$\alpha(c) = \frac{n_a(c)}{\mid CB \mid} \tag{2}$$

Table 4
Example of an inconsistent CB showcasing different levels of support.

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	$\bar{s}$
c_7	1	1	15	0	s

Both (1) and (2) would appear to have some merit intuitively. Using a combination of the two seems even more intuitive. One option (3) is to take the product of (1) and (2), essentially using (1) as a weight factor for (2).

$$\alpha(c) = \frac{n_a(c)}{n_a(c) + n_d(c)} \cdot \frac{n_a(c)}{\mid CB \mid} \tag{3}$$

Alternatively, (1) and (2) can be combined as a harmonic mean (4). This introduces a parameter β, the relative importance of one expression over the other. The added advantage of this is that (1) could be considered twice as important than (2), for instance. At a value of $\beta = 1$, the two are equally important.

$$\alpha(c) = (1+\beta^2) \cdot \frac{\frac{n_a(c)}{n_a(c)+n_d(c)} \cdot \frac{n_a(c)}{|CB|}}{\frac{n_a(c)}{n_a(c)+n_d(c)} + \frac{n_a(c)}{|CB|}} \tag{4}$$

How desirable each expression is, is difficult to say. In the next section, we attempt to answer this question through experimentation. However, two observations can be made here regarding the four expressions. One is that $\alpha(c) = 1$ implies that the CB is consistent with regard to c, but only in case of (1) is this value obtained without the whole CB being in agreement with c.

Proposition 1. *Let CB be a case base and let $\alpha(c) = 1$ for a case $c \in CB$. Then CB must be consistent with regard to c.*

Proof. Recall that CB is consistent with regards to c if there exists no other case $c' \in CB$ with $outcome(c) \neq outcome(c')$ and $F(c) \leq_{outcome(c)} F(c')$. Recall also that this can be expressed as $n_d(c) = 0$ and that $n_a(c) \geq 1$ for any $c \in CB$ since c is in agreement with itself. Suppose that $\alpha(c) = 1$ according to (1). Then $n_a(c) = n_a(c) + n_d(c)$ and it must follow that $n_d(c) = 0$. Suppose now that $\alpha(c) = 1$ according to (2). Then $n_a(c) =\mid CB \mid$. Since c cannot count towards both $n_a(c)$ and $n_d(c)$, $n_d(c) = 0$. Suppose now that $\alpha(c) = 1$ according to (3). Since (3) is the product of (1) and (2), both expressions must have a value of 1 and therefore $n_a(c) = \mid CB \mid$ and $n_d(c) = 0$. Suppose now that $\alpha(c) = 1$ according to (4). Since (4) is the harmonic mean of (1) and (2), both expressions must have a value of 1 and therefore $n_a(c) = \mid CB \mid$ and $n_d(c) = 0$. □

The other observation is that $\alpha(c) = 0$ is not obtainable for any of the expressions for authoritativeness.

Proposition 2. *Let CB be a case base. Then $\alpha > 0$ for any $c \in CB$.*

Proof. Recall that $n_a(c)$ is the cardinality of the set of cases for which the condition holds that $outcome(c) = outcome(c')$ and $D(c, c') = \emptyset$ for $c' \in CB$. Suppose that $\alpha(c) = 0$. Then $n_a(c) = 0$ when evaluating $\alpha(c)$ according to (1), (2), (3) or (4). Since the condition for the cases counting towards $n_a(c)$ holds when $c = c'$, $n_a(c)$ can only be 0 when $c \notin CB$. □

These observations do not affect our evaluation in the next section and do not form a limitation of our current approach, but they do result in the following question regarding the intuitive understanding of authoritativeness: should the minimum value of 0 and maximum value of 1 for $\alpha(c)$ be of significance? If so, this would steer our choice for an expression for authoritativeness. We return to this point in Section 5.

4. Evaluation

Using authoritativeness as a criterion when selecting a best precedent is intended to improve the ability of AF-CBA to generate useful justifications in light of CB inconsistency. Since AF-CBA generates justifications for the same outcome as the ML model predicts, we cannot use fidelity (the agreement between an XAI approach and the ML model it explains) to assess the

Table 5

The results of our evaluation experiments for three different data sets and four different expressions of authoritativeness in addition to the base method where authoritativeness is not taken into account.

	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	$\mu = 70.25$ $N_{inc} = 620$	$\mu = 72.37$ $N_{inc} = 0$	$\mu = 84.66$ $N_{inc} = 0$	$\mu = 86.75$ $N_{inc} = 0$	$\mu = 84.83$ $N_{inc} = 0$

efficacy of our modification. Evaluation of our approach therefore requires a more investigative experimentation and interpretation of the results.

To this end, we follow a similar strategy to Prakken & Ratsma [4]. We also rely on the same data sets as they do, namely Graduate Admission [12], Telco Customer Churn [10] and Mushroom [13]. As an expression of how inconsistent each data set is, we determine the minimum number of case deletions required to make each CB consistent. The result is 26 (3.20%), 647 (9.20%) and 16 (0.32%) for the Admission, Churn and Mushroom data set, respectively. The tendencies of all dimensions are determined using the Pearson correlation coefficient. Prakken & Ratsma [4] attempt to gain insights into the feasibility of AF-CBA in terms of the justifications themselves and in the treatment of inconsistencies. As they explain, desirable characteristics for AF-CBA include fewer best precedents (reducing the solution space for citing a precedent, see Section 2). This is one of the metrics on which we compare our four alternative formulations of authoritativeness to the base method. We treat each case in the CB as a focus case and compute the number of best precedents for that case given the rest of the CB, reporting the mean number (μ). Changes in μ would depend on the average distribution of inconsistent cases per focus case. There is no well-motivated cut off point for when these numbers become too high, but it is worthwile to consider whether μ increases by orders of magnitude and whether there are surprising differences between the alternative formulations of authoritativeness.

We also report the number of inconsistent forcing relations (N_{inc}) given each experiment. As described in Section 3, this is the number of forcing relations between two cases that contradict each other on the outcome. $N_{inc} = 0$ for a consistent data set and our intention is to achieve this without having to take a consistent subset of the data. We therefore expect N_{inc} to drop to very low numbers (if not zero) for all experiments where we make use of authoritativeness.

We present the results of these experiments[1] in Table 5. A qualitative assessment of these results suggests that inconsistent forcing is indeed largely avoided by taking the authoritativeness of precedents into account, without having a costly impact on the best precedent distributions. The relative version of authoritativeness (1) raises μ the most for two of the three data sets and especially for the most inconsistent set (Churn), which suggests that this particular version of authoritativeness can complicate explanations slightly with inconsistent CBs. Relative authoritativeness (1) but especially absolute authoritativeness (2) does not reduce N_{inc} completely to zero for the Churn data set. The product (3) and harmonic (4) versions of

[1] https://github.com/JGTP/CBA-precedent.git

authoritativeness therefore appear to be the more desirable expressions. The results do not suggest any meaningful differences between (3) and (4).

5. Discussion and Future Work

Post hoc analyses often constitute classifiers themselves, although evidently worse than the actual models (or *they* would be used as the models instead). This is not the case with AF-CBA. One can still hold the view that a simpler albeit more transparent model is preferable to a post hoc analysis. However, it is our experience that this is often unfeasible, as there exist many problems for which the only satisfactory solutions are too opague—especially for people who are not researchers or data scientists. It is our belief that this warrants the use of post hoc analyses in many situations.

Our modification of AF-CBA relies on the intuition that one precedent can be more authoritative than another. We demonstrate its consequences in numerical terms (μ and N_{inc}), given that lower numbers indicate better explanations (as was argued in the original paper [4]). However, it could be argued that these metrics are but proxies for the capacity to justify ML predictions in an intuitive fashion. Testing this would require a usability study to evaluate the explanatory power and interpretability of various explanations. A variety of alternative modifications and additional metrics could then be compared to study their efficacy in a real-world setting.

None of our expressions for authoritativeness would ever reach a value of 0 for any case in the CB. This seems intuitive, since any case should have at least some authoritativeness simply due to its being a precedent. A value of $\alpha(c) = 1$ is only realistic when using our relative expression of authoritativeness 1. This would only be a problem if values due to different expressions of authoritativeness would have to be compared to each other, which is not the aim of our method. If multiple explanations are ever to be compared as part of some overarching approach, these (and possibly other) characteristics of alternative authoritativeness expressions would have to be taken into account.

Additional modifications to AF-CBA could include other criteria for ranking precedents, incorporating complex arguments in the explanations (AF-CBA is qualified as a 'top-level' model due to the possibility of providing it with a set of definitions as to why specific downplaying moves can be played) or accounting for dimensions which are highly dependent. Another possibility is an alteration that allows dimensions to have a more complex effect on predictions than the tendencies used in this paper. There exist binary classification tasks for which this would be desirable. For example, a dimension such as blood pressure could be a predictor for illness both at very low and very high values, with a value in the intermediate range being a predictor for the patient not being ill. We intend to include this in our future work.

Conclusion

In this paper, we have presented an extension of an earlier top-level model (AF-CBA) for case-based argumentation used to provide *post hoc* justifications for opaque machine learning predictions. We have modified its definition of best precedent to include a quantified expression of how authoritative that precedent is, thereby affecting which cases are likely to be cited. This is not strictly in conflict with the *a fortiori* assumption underpinning the approach. Instead, it recognises the limitations of that assumption in light of the inconsistency that can be expected from real-world case bases. We have experimented with multiple versions of this expression to study which appears to be the most fruitful regarding the handling of inconsistency without adversely affecting the explanations. Our evaluation suggests our two somewhat more elaborate expressions of authoritativeness are more suitable. Future work is to be aimed at other modifications and usability studies.

Acknowledgements

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

References

[1] Z. Lipton, The mythos of model interpretability, Communications of the ACM 61 (2016) 96–100. doi:10.1145/3233231.

[2] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 (2018) 93:1–93:42. doi:10.1145/3236009.

[3] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artificial Intelligence 267 (2019) 1–38. doi:10.1016/j.artint.2018.07.007.

[4] 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.

[5] J. Horty, Rules and reasons in the theory of precedent, Legal Theory 17 (2011) 1–34.

[6] J. Horty, Reasoning with dimensions and magnitudes, Artificial Intelligence and Law 27 (2019) 309–345. doi:10.1007/s10506-019-09245-0.

[7] V. Aleven, Teaching case-based argumentation through a model and examples, Ph.D. thesis, University of Pittsburgh, Pittsburgh, 1997.

[8] C. G. Northcutt, A. Athalye, J. Mueller, Pervasive label errors in test sets destabilize machine learning benchmarks, arXiv:2103.14749 [cs, stat] (2021). ArXiv: 2103.14749.

[9] B. Frenay, M. Verleysen, Classification in the presence of label noise: a survey, IEEE Transactions on Neural Networks and Learning Systems 25 (2014) 845–869. doi:10.1109/TNNLS.2013.2292894.

[10] IBM, Telco Customer Churn, 2018.

[11] S. Modgil, M. Caminada, Proof Theories and Algorithms for Abstract Argumentation Frameworks, in: G. Simari, I. Rahwan (Eds.), Argumentation in Artificial Intelligence, Springer US, Boston, MA, 2009, pp. 105–129. doi:10.1007/978-0-387-98197-0_6.

[12] M. Acharya, A. Armaan, A. Antony, A comparison of regression models for prediction of graduate admissions, in: 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1–5.

[13] D. Wagner, D. Heider, G. Hattab, Mushroom data creation, curation, and simulation to support classification tasks, Scientific Reports 11 (2021). doi:10.1038/s41598-021-87602-3, number: 1 Publisher: Nature Publishing Group.