Legal Knowledge and Information Systems
G. Sileno et al. (Eds.)

doi:10.3233/FAIA230950

Learning Case Relevance in Case-Based Reasoning with Abstract Argumentation

Guilherme PAULINO-PASSOS [1] and Francesca TONI
Imperial College London, Department of Computing, London, United Kingdom
ORCiD ID: Guilherme Paulino-Passos https://orcid.org/0000-0003-3089-1660, Francesca Toni https://orcid.org/0000-0001-8194-1459

Abstract. Case-based reasoning is known to play an important role in several legal settings. We focus on a recent approach to case-based reasoning, supported by an instantiation of abstract argumentation whereby arguments represent cases and attack between arguments results from outcome disagreement between cases and a notion of relevance. We explore how relevance can be learnt automatically with the help of decision trees, and explore the combination of case-based reasoning with abstract argumentation ($AA\text{-}CBR$) and learning of case relevance for prediction in legal settings. Specifically, we show that, for two legal datasets, $AA\text{-}CBR$ with decision-tree-based learning of case relevance performs competitively in comparison with decision trees, and that $AA\text{-}CBR$ with decision-tree-based learning of case relevance results in a more compact representation than their decision tree counterparts, which could facilitate cognitively tractable explanations.

Keywords. case-based reasoning, argumentation, machine learning, explainable AI

1. Introduction

Case-based reasoning (CBR) is a methodology in which concrete past occasions are directly used as sources of knowledge and solutions for new situations. It has been studied in AI and Law since its inception [1]. This is a not a surprise, given the centrality of the use of cases in Common Law systems, although not exclusively [2].

In this paper we focus on recent approaches to CBR [3,4,5,6] using argumentation [7]. Argumentation itself has a long history in AI and Law, and its use to support CBR has been shown to pave the way towards novel forms of explanations for the outcomes of CBR, including via arbitrated dispute trees [8,9]. Specifically, we focus on the $AA\text{-}CBR$ approach [3,4,5], where arguments correspond to cases and attacks between arguments result from outcome disagreement between cases and *relevance* between cases, guided by a partial order over cases capturing some notion of specificity. Originally [3], $AA\text{-}CBR$ expects a representation of cases in terms of sets of *manually engineered binary* features and the partial order is defined via the subset relation. This expectation is a restriction for applicability. While previous work has generalised beyond

[1]Corresponding Author: Guilherme Paulino-Passos, gppassos@imperial.ac.uk.
We thank ERC (grant agreement No.101020934, ADIX); J.P. Morgan and the Royal Academy of Engineering, UK; and Capes (Brazil, Ph.D. 88881.174481/2018-01).

binary features in order to support different applications [4], a systematic generalisation to tabular datasets, including categorical and continuous data, is still missing. This is essential for applying $AA\text{-}CBR$ to realistic datasets, including legal ones, to realise the original inspiration from legal reasoning for $AA\text{-}CBR$. In this work we close this gap, focusing on applying $AA\text{-}CBR$ to possibly non-binary tabular data from legal settings.

Our first contribution is a general method for applying $AA\text{-}CBR$ to any tabular data by extracting binary features from decision trees when learning for the final task. Our second contribution is showing that this method is competitive with decision trees on two legal datasets: COMPAS [10] and a simulated legal dataset [11] for welfare benefit. Finally, as a third contribution, we show that our method creates smaller models, leading to potentially more cognitively tractable explanations.[2]

Background. We use the formulation of $AA\text{-}CBR_{\succeq}$ by [5]. We highlight that every case consists of a *characterisation* and an *outcome*, the set of characterisations is equipped with a partial order $\succeq$, and there is a particular *default case*. The partial order $\succeq$ defines a notion of *relevance* $\sim$ between characterisations, where $x_1 \sim x_2$ iff $x_2 \preceq x_1$. This notion and crucially *irrelevance* (defined as $\not\sim$) are used to compare new and past cases as well as two past cases (thus in $AA\text{-}CBR_{\succeq}$ relevance is not symmetric). The idea is that the partial order $\succeq$ captures *specificity* between cases, and that the outcome for a new case depends only on past cases than which the new case is more specific. Each case becomes an argument in an (abstract) argumentation framework (AF). We also apply $cAA\text{-}CBR_{\succeq}$ [5] and use arbitrated dispute trees (ADTs) for explanation [8,9].

2. Learning Relevance

Learning relevance in $AA\text{-}CBR_{\succeq}$ amounts to learning the partial order $\succeq$, which represents specificity. Here we use decision tree learning to extract characterisations suitable for $AA\text{-}CBR_{\supseteq}$ (i.e. $AA\text{-}CBR_{\succeq}$ with $\succeq{=}\supseteq$) from tabular data. Specifically, we use the CART algorithm for decision trees, in which decision nodes are greedily created choosing the feature and the split threshold which minimises a loss function; each split can then be seen as a binary feature, and each example can be characterised as a set of binary features. Thus specificity here is having all (binary) features of another case.

Example 1. Consider the dataset with the examples below and a decision tree trained on it (left of Figure 1):

$$\alpha = ((\text{age} = 20, \text{prior\_count} = 2), +), \qquad \gamma = ((\text{age} = 35, \text{prior\_count} = 7), +),$$

$$\beta = ((\text{age} = 30, \text{prior\_count} = 1), -), \qquad \epsilon = ((\text{age} = 19, \text{prior\_count} = 1), -),$$

$$\eta = ((\text{age} = 19, \text{prior\_count} = 10), +)$$

Assume further that the default outcome for $AA\text{-}CBR_{\succeq}$is $+$, reflecting the majority output `recid`. Then on the right of Figure 1 we show the AF mined from the processed dataset: each example is represented by the split tests for which it is evaluated `true`. The result is then used as a casebase for $AA\text{-}CBR_{\supseteq}$. The correspondence is one to many, since examples falling into the same leaf may correspond to different cases in the AF.

[2] An extended version of this paper is available at: www.github.com/GPPassos/learning-relevance-aacbr-technical-report

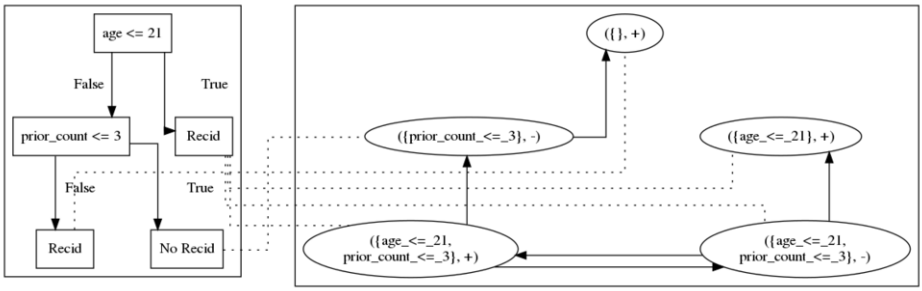

Figure 1. On the left, decision tree learnt from D in Example 1. On the right, the AF mined from D drawn from the splits in the decision tree. Dotted lines show correspondence between leaves (left) and cases (right).

Also, when multiple cases would have the same characterisation but different outcomes, either an incoherence is generated, or it is avoided via preprocessing.

3. Experiments

We train decision trees with pre-pruning, that is, limiting maximum depth and number of leaf nodes as regularisation, with values chosen by cross-validation. We consider: for maximum depth, varying from 3 to 13, in a step of 2; for maximum number of leaf nodes, from 4 to 512, in geometric progression of ratio 2. Nodes are created in a best-first search fashion, using Gini impurity. We evaluate three approaches for the problem of incoherence: 1. `keep`: to keep the incoherence and let each model deal with it in their own ways; 2. `removal`: to remove every incoherent pair of cases; 3. `majority`: for each characterisation in the resulting transformation, select the majority outcome.

COMPAS Dataset. This dataset contains predicted scores of recidivism and data of actual (measured) recidivism [10]. Our goal is simply to use this dataset as a way of evaluating our methodology in a legally relevant scenario. This should not be seen as results of a ready-to-deploy system or which allow conclusions from a criminal justice point of view. We use the two-year recidivism dataset and apply the original filtering strategy for missing data, resulting in 6172 entries. Each row corresponds to a defendant and contains personal information, information about the current charge, criminal history and whether the defendant has reoffended. We experimented with 4 different feature sets, each removing (A) no features; (B) `age_cat`; (C) `age_cat` and race; (D) `age_cat`, race and gender. We do so since `age_cat` is redundant with the age feature, while race and gender are protected features. When we do not specify the feature set, we mean C.

COMPAS Results. Comparing the strategies, `keep` is the weakest strategy even for $cAA\text{-}CBR_{\succeq}$, which deals with incoherence directly, while `majority` is the strongest. This is shown not only on the test set directly (Table 1) but also over almost all hyperparameter choices (Table 2). On Table 3 we directly compare performance. Under our method, $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$ show comparable performance with decision trees on COMPAS. Interestingly in most cases the optimal hyperparameter choice for $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$ resulted in both having the same behaviour on the test set, suggesting they may have learned the same decision function, despite different structures.

Table 1. Percentage accuracy of each $AA\text{-}CBR$ model and each strategy for incoherence in the casebase, aggregated over hyperparameter choice. Results on COMPAS test set, feature set A.

	$AA\text{-}CBR_{\succeq}$			$cAA\text{-}CBR_{\succeq}$		
	keep	removal	majority	keep	removal	majority
min	45.6	54.4	58.2	47.0	54.4	57.5
max	55.3	63.9	68.1	57.8	61.9	68.1
avg±stddev	49.1 ± 4.3	57.3 ± 2.2	64.1 ± 4.0	52.3 ± 3.0	57.4 ± 2.4	63.9 ± 4.2

Table 2. Difference in percentage accuracy between the `removal` or `keep` strategies and the `keep` strategy for incoherence, aggregated over hyperparameter choice. Aggregation is performed over the difference. Results on COMPAS test set, feature set A.

	$AA\text{-}CBR_{\succeq}$		$cAA\text{-}CBR_{\succeq}$	
	removal − keep	majority − keep	removal − keep	majority − keep
min	2.1	2.9	−0.1	1.8
max	13.0	22.5	10.4	21.00
mean±stddev	8.1 ± 4.1	15.0 ± 8.2	5.1 ± 3.5	11.6 ± 7.1

Table 3. Percentage accuracy for COMPAS, averaged over 5-fold cross validation, with standard deviation, and using hyperparameter optimisation by internal validation split.

	Feature set A	Feature set B	Feature set C	Feature set D
Decision tree	67.60 ± 1.31	67.60 ± 1.31	67.48 ± 1.56	67.00 ± 1.15
$AA\text{-}CBR_{\succeq}$	66.32 ± 1.20	66.32 ± 1.20	66.32 ± 1.20	66.41 ± 1.31
$cAA\text{-}CBR_{\succeq}$	66.32 ± 1.20	66.32 ± 1.20	66.32 ± 1.20	66.41 ± 1.31

Welfare Benefit Dataset. The welfare benefit domain was originally proposed in [12], with the goal of having a dataset that captures conditions typically found in law. Our goal is evaluating our method for learning relevance for $AA\text{-}CBR_{\succeq}$, so a thorough evaluation of rationales is outside our scope. We use the available `WelfareFailMany` dataset, containing contains 2000 cases, where 1000 are eligible cases and 1000 are ineligible.

Welfare Benefit Results. Table 4 shows that `majority` is the stronger strategy also for Welfare. Interestingly, for $AA\text{-}CBR_{\succeq}$ `keep` shows better performance than `removal`, that presents very high variance. By inspecting the learned models, this happened since many such learned models end up containing very few cases or even just the default case, due to the learned AFs having always incoherent cases for each or many characterisations. This also suggests a higher sensibility of $cAA\text{-}CBR_{\succeq}$ to noise. This is shown here by the high variance of the `removal` strategy. On the other hand, `majority` has not only a higher average, but also is more stable, with a smaller variance. Overall, the results confirm the ones seen for COMPAS, where `majority` is a better strategy in which both $AA\text{-}CBR$ approaches show performance on par with decision trees.

Explainability. Explanations come in two forms: global explanations, which explain the behaviour of entire model over all possible inputs; and local explanations, which explain the behaviour of (or around) a particular prediction. Given that both decision trees and $AA\text{-}CBR_{\succeq}$ are intrinsically interpretable models, the models themselves are subject

Table 4. Percentage accuracy for Welfare, for each strategy for incoherence, using hyperparameter search by internal validation split. Averages over 5-fold cross-validation, with standard deviation.

Decision Tree	$AA\text{-}CBR_{\succeq}$			$cAA\text{-}CBR_{\succeq}$		
	keep	removal	majority	keep	removal	majority
99.6 ± 0.1	99.3 ± 0.6	90.5 ± 18.0	99.5 ± 0.4	82.9 ± 18.8	90.5 ± 18.0	99.6 ± 0.2

to human inspection and can thus be evaluated as global explanations. As for local explanations, we use explanations tailored for each model. For decision trees, we consider the decision path traversed by the classified example. As for $AA\text{-}CBR_{\succeq}$, we use ADTs. We choose an ADT with minimum number of nodes by a minimax tree search algorithm. As there are no standard methodologies in the literature to evaluate explanations, we here use explanation size as a proxy for ease of interpretation. As the explanations that we use are all rooted graphs, they can be evaluated uniformly. We compare depth, number of nodes, and number of unique nodes (which all coincide for decision paths).

Explainability Results. As illustrated on Figure 1, a single leaf can become many nodes in $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$. While only half of the nodes of the decision tree are leaves, $AA\text{-}CBR$ could suffer from a combinatorial explosion of many features. However, this is not what we see empirically (Table 5). For COMPAS we see a 91.2% reduction in of the average size for $AA\text{-}CBR_{\succeq}$ and 94.2% for $cAA\text{-}CBR_{\succeq}$. This is subject to the high variance in decision tree size, but the $AA\text{-}CBR$ models show consistently smaller sizes. For Welfare there is a 29.1% reduction of the average size for $AA\text{-}CBR_{\succeq}$ and 58.8% for $cAA\text{-}CBR_{\succeq}$, with minimal variance for decision tree sizes. Thus, for comparable accuracy, $AA\text{-}CBR_{\succeq}$ and (specially) $cAA\text{-}CBR_{\succeq}$ can generate notably smaller models. Therefore, for scenarios where an interpretable graph form of the model is required, $AA\text{-}CBR_{\succeq}$ and $cAA\text{-}CBR_{\succeq}$ present a strong advantage over decision trees.

As for the local explanations (Table 6), ADTs show a larger number of nodes than decision paths. This is expected, since ADTs require multiple occurrences of sub-graphs of the original AF. ADTs for $cAA\text{-}CBR_{\succeq}$ show comparable number of nodes to decision paths in COMPAS, but are still larger in Welfare. The number of unique nodes is considerably larger than decision paths for $AA\text{-}CBR_{\succeq}$ and marginally so for $cAA\text{-}CBR_{\succeq}$. Furthermore, both $AA\text{-}CBR$ approaches result in a reduced depth as compared to decision paths. Thus, ADTs result in wider explanations, with multiple paths in the tree, but each path smaller than decision paths. Besides, an important difference between the $AA\text{-}CBR$ approaches and decision trees is that every node in $AA\text{-}CBR$ corresponds to at least one case in the casebase, as each node contains some counterfactual information (what would the outcome be for an input exactly equal to the past case, but not only [13]). Therefore the smaller representations also contain more information, despite requiring a more complex computation. This reflects into the size of the local explanation, with more nodes being required for it to be sufficient. The trade-off is favourable for $AA\text{-}CBR$, especially for $cAA\text{-}CBR_{\succeq}$, which has ADTs of similar size to decision paths.

4. Conclusions and Future Work

We presented an approach to learn case relevance for $AA\text{-}CBR$ from data on the case of COMPAS and Welfare Benefits, two tabular legal datasets. We show that binary splits of

Table 5. Size of models in number of nodes. Averages over 5-fold cross-validation, with standard deviation.

COMPAS			Welfare		
Decision Tree	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$	Decision Tree	$AA\text{-}CBR_{\succeq}$	$cAA\text{-}CBR_{\succeq}$
143.0 ± 184.9	12.6 ± 3.1	8.2 ± 1.6	11.0 ± 0.0	7.8 ± 4.3	4.6 ± 0.5

Table 6. Size of local explanations. Averages over 5-fold cross-validation, with standard deviation.

	COMPAS			Welfare		
	depth	# nodes	# unique	depth	# nodes	# unique
Decision Tree	6.2 ± 1.6	6.2 ± 1.6	6.2 ± 1.6	4.2 ± 0.1	4.2 ± 0.1	4.2 ± 0.1
$AA\text{-}CBR_{\succeq}$	5.6 ± 0.3	11.9 ± 1.7	7.9 ± 1.1	3.5 ± 0.0	8.1 ± 3.9	5.1 ± 1.1
$cAA\text{-}CBR_{\succeq}$	5.9 ± 0.4	6.1 ± 0.4	6.0 ± 0.3	3.9 ± 0.5	5.1 ± 0.4	4.5 ± 0.4

learned decision trees can be used as features for $AA\text{-}CBR$ and allow its instantiation as $AA\text{-}CBR_{\supseteq}$. Future work includes comparing with other forms of CBR for legal tasks [14,15,16,6], as well as learning case relevance for images and text [17].

References

[1] Rissland EL, Ashley KD, Branting K. Case-based reasoning and law. The Knowledge Engineering Review. 2005;20:293 298.

[2] Lewis S. Precedent and the Rule of Law. Oxford Journal of Legal Studies. 2021 Mar;41(4):873–898.

[3] Čyras K, Satoh K, Toni F. Abstract Argumentation for Case-Based Reasoning. In: KR; 2016. p. 549-52.

[4] Cocarascu O, Stylianou A, Čyras K, Toni F. Data-Empowered Argumentation for Dialectically Explainable Predictions. In: 24th ECAI; 2020. .

[5] Paulino-Passos G, Toni F. Monotonicity and Noise-Tolerance in Case-Based Reasoning with Abstract Argumentation. In: 18th KR; 2021. p. 508-18.

[6] Prakken H, Ratsma R. A top-level model of case-based argumentation for explanation: Formalisation and experiments. Argument Comput. 2022;13:159-94.

[7] Prakken H. Historical Overview of Formal Argumentation. College Publications; 2018. .

[8] Čyras K, Birch D, Guo Y, Toni F, Dulay R, Turvey S, et al. Explanations by arbitrated argumentative dispute. Expert Syst Appl. 2019;127:141-56.

[9] Čyras K, Rago A, Albini E, Baroni P, Toni F. Argumentative XAI: A Survey. In: IJCAI; 2021. p. 4392-9.

[10] Larson J, Mattu S, Kirchner L, Angwin J. How We Analyzed the COMPAS Recidivism Algorithm; 2016.

[11] Steging C, Renooij S, Verheij B, Bench-Capon TJM. Arguments, rules and cases in law: Resources for aligning learning and reasoning in structured domains. Argument Comput. 2023;14(2):235-43.

[12] Bench-Capon TJM. Neural Networks and Open Texture. In: ICAIL; 1993. p. 292-7. Available from: https://doi.org/10.1145/158976.159012.

[13] Paulino-Passos G, Toni F. On Monotonicity of Dispute Trees as Explanations for Case-Based Reasoning with Abstract Argumentation. In: ArgXAI@COMMA; 2022. .

[14] Horty JF, Bench-Capon TJM. A factor-based definition of precedential constraint. Artificial Intelligence and Law. 2012 May;20(2):181–214.

[15] Grabmair M. Modeling purposive legal argumentation and case outcome prediction using argument schemes in the value judgment formalism. University of Pittsburgh, USA; 2016.

[16] van Woerkom W, Grossi D, Prakken H, Verheij B. Landmarks in Case-Based Reasoning: From Theory to Data. In: HHAI 2022; 2022. p. 212-24.

[17] Mumford J, Atkinson K, Bench-Capon TJM. Reasoning with Legal Cases: A Hybrid ADF-ML Approach. In: 35th JURIX; 2022. .