Explaining Predictions from Data Argumentatively

Explain AI@Imperial Workshop

Ken Satoh[1] Oana Cocarascu Kristijonas Čyras Francesca Toni

April 25, 2018

Department of Computing, Imperial College London, UK
[1]National Institute of Informatics, Tokyo, Japan

Problem

- Examples/instances/cases $DB = \{e_1, \ldots, e_n\}$
  Example $e = (F, \mathsf{o}) \in DB$ consists of:
  - (set of) features/attribute-value pairs/factors
    $F = \{\mathsf{f}_1, \ldots, \mathsf{f}_m\} \subseteq \mathbb{F}$
  - label/class/outcome
    $\mathsf{o} \in \mathbb{L} = \{\varphi, \overline{\varphi}\}$
- New example $(N, ?)$
  - features $N \subseteq \mathbb{F}$
  - unknown label ?
- Prediction: determine whether $? = \varphi$ or $? = \overline{\varphi}$
- *Explain* why

(Some) Existing Approaches

- To predict labels, could use
  - Case-Based Reasoning (CBR) [Richter and Weber, 2013]
  - Artificial Neural Networks (ANNs) [LeCun et al., 2015]
  - etc.
- But may be hard to explain predictions [Andrews et al., 1995, Sørmo et al., 2005]
  - hard to define formally
  - showing similar examples need not suffice
  - transparent/interpretable $\neq$ explanatory
- May also be data-hungry
  - e.g. large DB needed

Our Approach

- Abstract Argumentation (AA) [Dung, 1995]
  - deals with conflicting information
- AA-CBR [Čyras et al., 2016a]: AA-driven CBR
  - models and deals with conflicting examples
- AA-CBR Explanations [Čyras et al., 2016b]
  - debates explaining predictions
- ANNs with AA-CBR
  - ANNs for feature selection
  - AA-CBR predictions and explanations
  - rule-based predictions and explanations

Feature Selection (ANN)

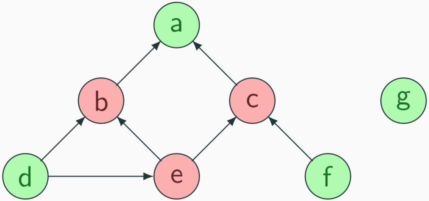
- Start with a *training set* $\mathcal{E}$ of examples (Y, o)
  - features (of e.g. mushrooms[2])
    $Y \subseteq \mathbb{F}_{\mathcal{E}} = \{\ldots, \mathit{white}, \mathit{pink}, \mathit{red}, \mathit{crimson}, \mathit{maroon}, \ldots\}$
  - label $\mathrm{o} \in \mathbb{L} = \{\mathit{edible}\ (\varphi),\ \mathit{poisonous}\ (\overline{\varphi})\}$
- Use autoencoder to get a *trimmed dataset* DB of examples
  - $\{\ldots, \mathit{white}, \mathit{red}, \ldots\} = \mathbb{F} \subseteq \mathbb{F}_{\mathcal{E}}$
  - $DB = \{(Y, \mathrm{o})\ :\ (X, \mathrm{o}) \in \mathcal{E},\ Y = X \cap \mathbb{F}\}$
- Ensure $\mathbb{F}$ leads to *coherent* DB
  - $\forall\ (X, \mathrm{o}_X), (Y, \mathrm{o}_Y) \in DB$, if $X = Y$, then $\mathrm{o}_X = \mathrm{o}_Y$
  - DB is 'rational'

[2] archive.ics.uci.edu/ml/datasets/Mushroom[Dheeru and Karra Taniskidou, 2017]

Abstract Argumentation (AA)

AA is used to create a *model* of *DB*.

- An AA framework is a graph $(Args, \rightsquigarrow)$
  - Nodes: arguments *Args* represent information
  - Edges: attacks $\rightsquigarrow$ represent conflicts
- Semantics determine 'good' arguments
  - E.g. grounded extension (set of arguments)



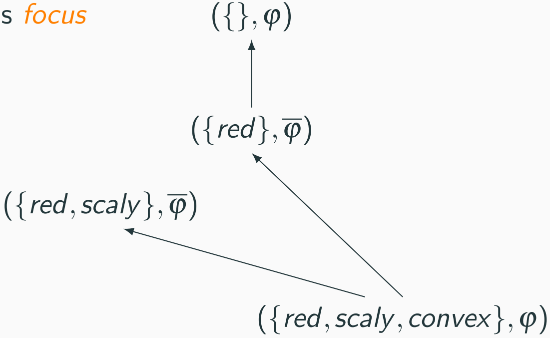
AA-CBR Model

From DB and φ construct $(Args, \rightsquigarrow)$ with:

- $Args = DB \cup \{(\{\}, \varphi)\}$;
  - examples are arguments
  - $(\{\}, \varphi)$ (being *edible*) is *focus argument*
- for $(X, o_X), (Y, o_Y) \in DB \cup \{(\{\}, \varphi)\}$,
  it holds that $(X, o_X) \rightsquigarrow (Y, o_Y)$ iff
  1. $o_X \neq o_Y$, and (different outcomes)
  2. $Y \subsetneq X$, and (specificity)
  3. $\nexists (Z, o_X) \in CB$ with $Y \subsetneq Z \subsetneq X$. (concision)

AA-CBR Model Graph (Mushrooms)

φ is *focus*



AA-CBR Prediction

From DB, focus φ and $(N, ?)$ construct $(Args_N, \rightsquigarrow_N)$ with:

- $Args_N = Args \cup \{(N, ?)\}$;
- $\rightsquigarrow_N = \rightsquigarrow \cup \{((N, ?), (Y, o_Y)) : (Y, o_Y) \in Args$ and $Y \nsubseteq N\}$.
  - $(Args_N, \rightsquigarrow_N)$ extends $(Args, \rightsquigarrow)$ with $(N, ?)$ attacking 'irrelevant' examples

Let $\mathbb{G}$ be the grounded extension of $(Args_N, \rightsquigarrow_N)$.

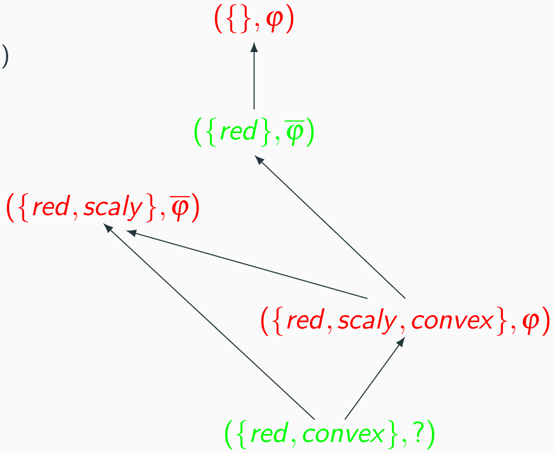
The *AA-CBR prediction* of $(N, ?)$ is:

- φ, if $(\{\}, \varphi) \in \mathbb{G}$;
  - edible if focus argument is good
- $\overline{\varphi}$, otherwise, if $(\{\}, \varphi) \notin \mathbb{G}$.
  - poisonous otherwise

AA-CBR Prediction Graph (Mushrooms)

φ is *focus*

$(N, ?) = (\{red, convex\}, ?)$



$\mathbb{G} = \{(\{red, convex\}, ?), (\{red\}, \overline{\varphi})\}$.

$(\{\}, \varphi) \notin \mathbb{G}$. So prediction is poisonous ($\overline{\varphi}$).

AA-CBR Explanations

Explanations of predictions are *disputes* between a proponent P (arguing for focus) and an opponent O (arguing against).

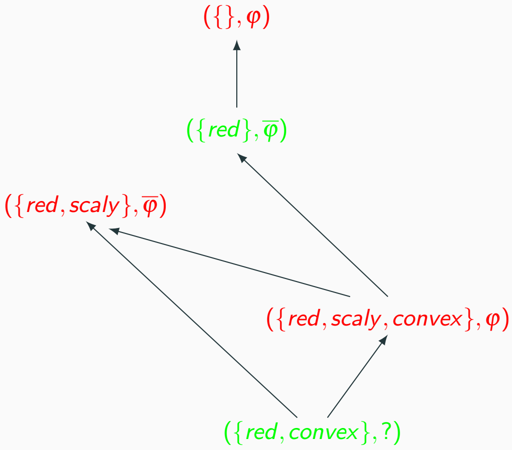
Disputes as sub-graphs of $(Args_N, \rightsquigarrow_N)$:

- Prediction is φ – an explanation is any *admissible dispute tree* $\mathcal{T}$ for the focus argument $(\{\}, \varphi)$
  - every O node has a child
  - no argument labels both P and O
- Prediction is $\overline{\varphi}$ – an explanation is any *maximal dispute tree* $\mathcal{T}$ for the focus argument $(\{\}, \varphi)$
  - every O leaf is unattacked in $(Args_N, \rightsquigarrow_N)$

Explanation for Poisonous

$(Args_N, \rightsquigarrow_N)$

φ is *focus*



$\mathcal{T}$

$$\begin{tikzcd}
{[P : (\{\}, \varphi)]} \\
{[O : (\{red\}, \overline{\varphi})]} \arrow[u] \\
{[P : (\{red, scaly, convex\}, \varphi)]} \arrow[u] \\
{[O : (\{red, convex\}, ?)]} \arrow[u]
\end{tikzcd}$$

Explanation for Edible

φ is *focus*

$$
\begin{tikzcd}
 & (\{\}, \varphi) & \\
 & (\{red\}, \overline{\varphi}) \arrow[u] & \\
(\{red, scaly\}, \overline{\varphi}) & & \\
 & & (\{red, scaly, convex\}, \varphi) \arrow[uul] \arrow[ull] \\
 & (\{red, scaly, convex, smooth\}, ?) &
\end{tikzcd}
$$

$$
\begin{tikzcd}
[P : (\{\}, \varphi)] \\
[O : (\{red\}, \overline{\varphi})] \arrow[u] \\
[P : (\{red, scaly, convex\}, \varphi)] \arrow[u]
\end{tikzcd}
$$

Rules

Logic programming rules from $(Args, \rightsquigarrow)$

- Alternative description of the model of DB
- Rule predictions coincide with AA-CBR predictions
- Alternative explanations of predictions

Logic program $\mathcal{P}$:

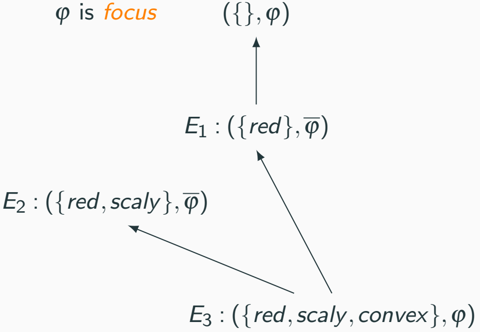
- For $E : (\{\mathrm{f}_1, \ldots, \mathrm{f}_m\}, \mathrm{o}) \in Args$, create a rule
  $$acc(E) \leftarrow \mathrm{f}_1, \ldots, \mathrm{f}_m, \text{not } acc(E_1), \ldots, \text{not } acc(E_k).$$
  stating that E is accepted
  - if all features $\mathrm{f}_1, \ldots, \mathrm{f}_m$ apply,
  - unless any of the attackers $E_1, \ldots, E_k$ of E are accepted;
- Repeat for each attacker and its attackers in turn;

For rule prediction, add features from N as facts to get $\mathcal{P}_N$.

Rules (Mushrooms)

$(Args, \rightsquigarrow)$

φ is *focus*



$N = \{red, convex\}$

$\mathcal{P}$:

$$acc(focus) \leftarrow \text{not } acc(E_1).$$
$$acc(E_1) \leftarrow red, \text{not } acc(E_3).$$
$$acc(E_2) \leftarrow red, scaly, \text{not } acc(E_3).$$
$$acc(E_3) \leftarrow red, scaly, convex.$$

$\mathcal{P}_N$ is $\mathcal{P}$ with

$$red \leftarrow \top.$$
$$convex \leftarrow \top.$$

Ongoing Work

- Datasets
- ANNs
- Categorical rather than binary features
- Multiple labels
- Rule simplification
- Related (argumentation-based) explanation concepts, e.g. [García et al., 2013, Fan and Toni, 2015, Schulz and Toni, 2016]
- Related (rule-based) explanation concepts, e.g. (neural) decision trees, inductive logic programming

Summary

- ML for feature selection within data
- Argumentation for
  - model creation
  - predictions
  - rules
  - *dialectical and logical* explanations

References i

Andrews, R., Diederich, J., and Tickle, A. B. (1995).
Survey and critique of techniques for extracting rules from trained artificial neural networks.
Knowledge-Based Systems, 8(6):373–389.

Čyras, K., Satoh, K., and Toni, F. (2016a).
Abstract Argumentation for Case-Based Reasoning.
In Baral, C., Delgrande, J. P., and Wolter, F., editors, *Principles of Knowledge Representation and Reasoning, 15th International Conference*, pages 549–552, Cape Town. AAAI Press.

Čyras, K., Satoh, K., and Toni, F. (2016b).
Explanation for Case-Based Reasoning via Abstract Argumentation.
In *6th International Conference on Computational Models of Argument*, pages 243–254, Potsdam. IOS Press.

Dheeru, D. and Karra Taniskidou, E. (2017).
UCI Machine Learning Repository.

Dung, P. M. (1995).
On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-person Games.
Artificial Intelligence, 77:321–357.

Fan, X. and Toni, F. (2015).
On Computing Explanations in Argumentation.
In Bonet, B. and Koenig, S., editors, *29th AAAI Conference on Artificial Intelligence*, pages 1496–1502, Austin, Texas. AAAI Press.

References ii

García, A. J., Chesñevar, C., Rotstein, N., and Simari, G. R. (2013).
Formalizing Dialectical Explanation Support for Argument-Based Reasoning in Knowledge-Based Systems.
Expert Systems with Applications, 40:3233–3247.

LeCun, Y., Bengio, Y., and Hinton, G. (2015).
Deep learning.
Nature, 521(7553):436–444.

Richter, M. and Weber, R. (2013).
Case-Based Reasoning.
Springer.

Schulz, C. and Toni, F. (2016).
Justifying Answer Sets Using Argumentation.
Theory and Practice of Logic Programming, 16(1):59–110.

Sørmo, F., Cassens, J., and Aamodt, A. (2005).
Explanation in Case-Based Reasoning–Perspectives and Goals.
Artificial Intelligence Review, 24(2):109–143.