Using Analogical Proportions for Explanations

Suryani Lim[1] Henri Prade[2] Gilles Richard[2]

1. Federation University, Churchill, Australia

2. IRIT - CNRS & Université Paul Sabatier, Toulouse, France

Explanation is an old topic in AI

- We expect from an “intelligence”, even an artificial one, that it *can explain its conclusions*
- The success of *expert systems*, based on *rules*, a little over 30 years ago, had led to work to develop systems capable of explaining their conclusions

- The success of learning methods based on *neural networks* has renewed interest, over the last past years, in explanation, by raising the problem of explaining the outcome of “black box” methods

Explanations

- Explanation in *neural networks* is often seen as a problem of sensitivity analysis,
  In the *logical* view, we distinguish

abductive explanations for “why?” questions
contrastive explanations for “why not?” questions

- Both in expert systems and in machine learning, we have the knowledge about the process that led to the conclusion to be explained:
  we know the set of rules used and the classification function
- Such knowledge is no longer necessary in the approach proposed here

Explanations

- Explanation in *neural networks* is often seen as a problem of sensitivity analysis,
  In the *logical* view, we distinguish
  abductive explanations for "why?" questions
  contrastive explanations for "why not?" questions
- Both in expert systems and in machine learning, we have the knowledge about the process that led to the conclusion to be explained:
  we know the set of rules used and the classification function
- Such knowledge is no longer necessary in the approach proposed here

Abductive explanation

- $\mathcal{A}$ a set of n attributes $i = 1, \cdots, n$
  x_i a value of attribute i
  v_i a constant in $\mathcal{D}_i$, domain of attribute i
  $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_n$
  and cl a classification function
- Given $cl(v) = c_0$ for $v = (v_1, \cdots, v_n)$, an *abductive explanation* (by prime implicant) consists of any minimal subset $\mathcal{X} \subseteq \mathcal{A}$ such that
  $\forall x \in \mathcal{D}.[\bigwedge_{i \in \mathcal{X}}(x_i = v_i)] \to (cl(x) = c_0)$
- It is enough to fix the values x_i of attributes in $\mathcal{X}$ to v_i for insuring that $cl(x) = c_0$

Contrastive explanation

- Given $cl(v) = c_0$, a *contrastive* explanation consists of any minimal subset $\mathcal{Y} \subseteq \mathcal{A}$ such that

$$\exists x \in \mathcal{D}.[\bigwedge_{j \in \mathcal{A} \setminus \mathcal{Y}} (x_j = v_j)] \wedge (cl(x) \neq c_0)$$

- One can find an x, outside c_0, which coincides with v on a maximal subset of attributes, i.e., one can perform a minimal change on v so that x is no longer in c_0
- This corresponds to an answer to a question “Why not $cl(v) \neq c_0$?”, i.e., one identifies the attributes whose value should be changed for that

Boolean modeling

- analogical proportion :"a is to b as c is to d"
  "the calf is to the cow
  as the foal is to the mare"
- $a : b :: c : d =$
  $$((a \land \neg b) \equiv (c \land \neg d)) \land ((\neg a \land b) \equiv (\neg c \land d))$$
- $0 : 0 :: 0 : 0$
  $1 : 1 :: 1 : 1$
  $0 : 1 :: 0 : 1$
  $1 : 0 :: 1 : 0$
  $0 : 0 :: 1 : 1$
  $1 : 1 :: 0 : 0$
- nominal values
  $(a, b, c, d) \in \{(g, g, g, g), (g, h, g, h), (g, g, h, h)\}$

Example and properties

- items a, b, c; d : *vectors* de values of n attributes
  $a : b :: c : d$ ssi $\forall i \in \{1, \cdots, n\}, a_i : b_i :: c_i : d_i$

Table: AP: example with Boolean and nominal attributes

-

	mammal	*carnivore*	*young*	*adult*	*family*
`calf`	1	0	1	0	bovidae
`cow`	1	0	0	1	bovidae
`foal`	1	0	1	0	equidae
`mare`	1	0	0	1	equidae

- $a : b :: c : d \Rightarrow a : c :: b : d$ central permutation
  $a : b :: c : d \Rightarrow c : d :: a : b$ symmetry
  $a : b :: c : d$ et $c : d :: e : f \Rightarrow a : b :: e : f$ transitivity
  $a : b :: c : d \Rightarrow \neg a : \neg b :: \neg c : \neg d$
  code independence

Example and properties

- items a, b, c; d : *vectors* de values of n attributes
  $a : b :: c : d$ ssi $\forall i \in \{1, \cdots, n\}, a_i : b_i :: c_i : d_i$

Table: AP: example with Boolean and nominal attributes

-

	mammal	*carnivore*	*young*	*adult*	*family*
calf	1	0	1	0	bovidae
cow	1	0	0	1	bovidae
foal	1	0	1	0	equidae
mare	1	0	0	1	equidae

- $a : b :: c : d \Rightarrow a : c :: b : d$ central permutation
  $a : b :: c : d \Rightarrow c : d :: a : b$ symmetry

$a : b :: c : d$ et $c : d :: e : f \Rightarrow a : b :: e : f$ transitivity

$a : b :: c : d \Rightarrow \neg a : \neg b :: \neg c : \neg d$

code independence

A reading of data oriented towards explanation

	$\mathcal{A}_1 \ldots \mathcal{A}_{i-1}$	$\mathcal{A}_i \ldots \mathcal{A}_{j-1}$	$\mathcal{A}_j \ldots \mathcal{A}_{k-1}$	$\mathcal{A}_k \ldots \mathcal{A}_{r-1}$	$\mathcal{A}_r \ldots \mathcal{A}_{s-1}$	$\mathcal{A}_s .. \mathcal{A}_n$	$\mathcal{C}$
a	1	0	1	0	1	0	p
b	1	0	1	0	0	1	q
c	1	0	0	1	1	0	p
d	1	0	0	1	0	1	q

- ($p \neq q$) The change of value of $\mathcal{C}$ from p to q between a and b and between c and d can only be explained by, giving the data, the change of values of attributes from $\mathcal{A}_r$ to $\mathcal{A}_n$

(which is the same for the pair (a, b) and pair (c, d))

- see these pairs as instances of a rule

expressing that the change on attributes from $\mathcal{A}_r$ to $\mathcal{A}_n$ determines the change for $\mathcal{C}$ whatever the context

A reading of data oriented towards explanation

	$\mathcal{A}_1 \ldots \mathcal{A}_{i-1}$	$\mathcal{A}_i \ldots \mathcal{A}_{j-1}$	$\mathcal{A}_j \ldots \mathcal{A}_{k-1}$	$\mathcal{A}_k \ldots \mathcal{A}_{r-1}$	$\mathcal{A}_r \ldots \mathcal{A}_{s-1}$	$\mathcal{A}_s .. \mathcal{A}_n$	$\mathcal{C}$
a	1	0	1	0	1	0	p
b	1	0	1	0	0	1	q
c	1	0	0	1	1	0	p
d	1	0	0	1	0	1	q

- $(p \neq q)$ The change of value of $\mathcal{C}$ from p to q between a and b and between c and d can only be explained by, giving the data, the change of values of attributes from $\mathcal{A}_r$ to $\mathcal{A}_n$

(which is the same for the pair (a, b) and pair (c, d))

- see these pairs as instances of a rule

expressing that the change on attributes from $\mathcal{A}_r$ to $\mathcal{A}_n$ determines the change for $\mathcal{C}$ whatever the context

A reading of data oriented towards explanation

	$\mathcal{A}_1 \ldots \mathcal{A}_{i-1}$	$\mathcal{A}_i \ldots \mathcal{A}_{j-1}$	$\mathcal{A}_j \ldots \mathcal{A}_{k-1}$	$\mathcal{A}_k \ldots \mathcal{A}_{r-1}$	$\mathcal{A}_r \ldots \mathcal{A}_{s-1}$	$\mathcal{A}_s .. \mathcal{A}_n$	$\mathcal{C}$
a	1	0	1	0	1	0	p
b	1	0	1	0	0	1	q
c	1	0	0	1	1	0	p
d	1	0	0	1	0	1	q

- $(p \neq q)$ The change of value of $\mathcal{C}$ from p to q between a and b and between c and d can only be explained by, giving the data, the change of values of attributes from $\mathcal{A}_r$ to $\mathcal{A}_n$

(which is the same for the pair (a, b) and pair (c, d))

- see these pairs as instances of a rule

expressing that the change on attributes from $\mathcal{A}_r$ to $\mathcal{A}_n$ determines the change for $\mathcal{C}$ whatever the context

Illustrative example

case	situation	c. – i.	dec.	opt. 1	opt. 2
a	sit_1	yes	δ	0	0
b	sit_1	no	δ	1	0
c	sit_2	yes	δ	0	1
d	sit_2	no	δ	**1**	**1**

- decision: serve a coffee with or without sugar (option 1), with or without milk (option 2) to a person

  What to do in sit_2 if no *c. i.* ?

- **question** "why milk and sugar for *d*?" *answer* (for milk) "because we are in sit_2 (not in sit_1)" "because there is no *c. i.*" for sugar

  question "why no milk for *b*?", *answer* "because we are in sit_1 (not in sit_2)"

Illustrative example

case	situation	c. – i.	dec.	opt. 1	opt. 2
a	sit_1	yes	δ	0	0
b	sit_1	no	δ	1	0
c	sit_2	yes	δ	0	1
d	sit_2	no	δ	**1**	**1**

- decision: serve a coffee with or without sugar (option 1), with or without milk (option 2) to a person

  What to do in sit_2 if no *c. i.* ?
- **question** "why milk and sugar for *d*?"

answer (for milk) "because we are in sit_2 (not in sit_1)"
"because there is no *c. i.*" for sugar

question "why no milk for *b*?",
answer "because we are in sit_1 (not in sit_2)"

Illustrative example

case	situation	$c.-i.$	dec.	opt. 1	opt. 2
a	sit_1	yes	δ	0	0
b	sit_1	no	δ	1	0
c	sit_2	yes	δ	0	1
d	sit_2	no	δ	**1**	**1**

- decision: serve a coffee with or without sugar (option 1), with or without milk (option 2) to a person
  What to do in sit_2 if no $c.\ i.$?
- **question** "why milk and sugar for d?"
  answer (for milk) "because we are in sit_2 (not in sit_1)"
  "because there is no $c.\ i.$" for sugar
  question "why no milk for b?",
  answer "because we are in sit_1 (not in sit_2)"

Analogy and contrastive explanations

case	context	change	class
a	sit_1	yes	p
b	sit_1	no	q
c	sit_2	yes	p
d	sit_2	no	q

Table: Schematic situation of analogical explanation

- The answer to the question "why d is not in class p?" relies in the values taken d for the attributes in *change*. When c is a close neighbor of d, the number of attributes in *change* is small. We are close to a contrastive explanation :
  $\exists x = c \in \mathcal{S}.[\bigwedge_{j \in \mathcal{A} \setminus change}(x_j = c_j = d_j)] \wedge (cl(x) \neq q)$
- contrastive explanation
  $\exists x \in \mathcal{D}.[Disagree(x, v) = \mathcal{Y} \wedge (cl(x) \neq c_0)]$

Analogy and abductive explanation

- The explanation is richer here, one knows at least another pair (here (a, b)) that corresponds to another *context* where the same change of attribute values leads to the same change of classe, which suggests the possibility of rules $\forall$ *sit*,

$$(contexte = sit) \wedge (chang. = oui) \rightarrow cl((sit, non)) = p$$
$$(contexte = sit) \wedge (chang. = non) \rightarrow cl((sit, non)) = q$$

The rules enable a reading of the Table with an abductive explanation flavor, which says why the item is in class p (or in class q).

abductive explanation

$$\forall x.[(Acc.(x, v) = \mathcal{X}) \rightarrow (cl(x) = c_0)]$$

Confidence in explanations

case	context	change	class
$\vec{a}$	sit_1	yes	p
$\vec{b}$	sit_1	no	q
$\vec{c}$	sit_2	yes	p
$\vec{d}$	sit_2	no	q
$\vec{a'}$	sit'	yes	p
$\vec{b'}$	sit'	no	p

- BUT exception if $\exists\ (\vec{a'}, \vec{b'})$ s. t. $\vec{a'} = (sit', yes)$, $\vec{b'} = (sit', no)$ with $cl(\vec{a'}) = cl(\vec{b'}) = p$
- So we may calculate the confidence and support of the rule associated with pairs (a, b) and (c, d) in the data set

Concluding remarks - 1

- Explanatory use of analogical proportions in learning Hüllermeier (2020)
- Analogical proportions have great explanatory potential from data
- *"why"* and *"why not"* questions can be answered
- has been *implemented*
  - interesting to *precompile* the data set in pairs by identifying where items are *equal* and where and how they *differ* to facilitate an analogical analysis of the data
  - start by determining the relevant attributes,
- *confidence*, *support* of *rules* associated with pairs

Concluding remarks - 1

- Explanatory use of analogical proportions in learning Hüllermeier (2020)
- Analogical proportions have great explanatory potential from data
- *"why"* and *"why not"* questions can be answered
- has been *implemented*

  - interesting to *precompile* the data set in pairs
  by identifying where items are *equal*
  and where and how they *differ*
  to facilitate an analogical analysis of the data

  - start by determining the relevant attributes,
- *confidence*, *support* of *rules* associated with pairs

Concluding remarks - 2

- apply to preferences learning
  From $a : b :: c : d$ and "a is preferred to b"
  analogical inference concludes "c is preferred to d"
  Analogical explanation *would also apply*
- A 2nd kind of analogical proportion
  where a and c on the one hand and b and d on the other hand belong to 2 different universes:
  "*this drug is to colds what aspirin is to headache*"
  (it is quite effective and cheap)
- Analogical proportions have an explanatory value
  "*Star Wars* (1977) is to *Raiders of the Lost Ark* (1981) as *Return of the Jedi* (1983) is to *Indiana Jones and the Last Crusade* (1989)"

Concluding remarks - 2

- apply to preferences learning
  From $a : b :: c : d$ and "a is preferred to b"
analogical inference concludes "c is preferred to d"
  Analogical explanation *would also apply*
- A 2nd kind of analogical proportion
  where a and c on the one hand and b and d on the other hand belong to 2 different universes:
  "*this drug is to colds what aspirin is to headache*"
  (it is quite effective and cheap)
- Analogical proportions have an explanatory value
  "*Star Wars* (1977) is to *Raiders of the Lost Ark* (1981) as *Return of the Jedi* (1983) is to *Indiana Jones and the Last Crusade* (1989)"