Explications analogiques

Henri Prade Gilles Richard

IRIT

CNRS & Université Paul Sabatier

Toulouse

présenté à l'Atelier "Explain'AI" de la conférence EGC, Blois, 25 Jan. 2022

L'explication est une thématique ancienne en IA

- Une "intelligence", fut-elle artificielle, doit pouvoir *expliquer ses conclusions*
- Déjà, le succès des *systèmes experts*, à base de *règles*, il y a un peu plus de 30 ans avait conduit à des travaux pour développer des systèmes capables d'expliquer leurs conclusions
- Malgré une idée encore répandue, pas de divorce entre *logique* et *réseaux de neurones*

Explications et logique

- Explications (en termes de sensibilité) dans des systèmes de règles *incertaines*
  (dans le cadre de la *théorie des possibilités*)
  sous forme de cascades de produits min-max de matrices (Farreny et Prade, 1989-1990)
- repris, *généralisé*, mis sous forme de réseaux de neurones, et exploité à des fins d'explications (Baaj, 2021)
- Une analyse logique de *classifieurs* en termes d'explications *abductives* (réponse à une question de type "pourquoi ?") ou *contrastives* (type "pourquoi pas ?") (Marques-Silva et al., 2021)

Explications et logique

- Explications (en termes de sensibilité) dans des systèmes de règles *incertaines*
  (dans le cadre de la *théorie des possibilités*)
  sous forme de cascades de produits min-max de matrices (Farreny et Prade, 1989-1990)
- repris, *généralisé*, mis sous forme de réseaux de neurones, et exploité à des fins d'explications (Baaj, 2021)

- Une analyse logique de *classifieurs* en termes d'explications *abductives* (réponse à une question de type "pourquoi ?") ou *contrastives* (type "pourquoi pas ?") (Marques-Silva et al., 2021)

Explication abductive

- $\mathcal{A}$ un ensemble de n attributs $i = 1, \cdots, n$
  x_i une valeur de l'attribut i
  v_i une constante de $\mathcal{D}_i$, domaine de l'attribut i
  $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_n$
  et cl une fonction de classification
- Etant donné $cl(v) = c_0$ pour $v = (v_1, \cdots, v_n)$, une explication *abductive* (par implicant premier) est constituée par n'importe quel sous-ensemble minimal $\mathcal{X} \subseteq \mathcal{A}$ tel que
  $\forall x \in \mathcal{D}.[\bigwedge_{i \in \mathcal{X}}(x_i = v_i)] \rightarrow (cl(x) = c_0)$
- Cela suffit de fixer les valeurs x_i des attributs de $\mathcal{X}$ à v_i pour assurer que $cl(x) = c_0$

Explication contrastive

- Etant donné $cl(v) = c_0$, une explication *contrastive* est constituée par n'importe quel sous-ensemble minimal $\mathcal{Y} \subseteq \mathcal{A}$ tel que

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

- On peut trouver un x, en dehors de c_0, qui coincide avec v sur un sous-ensemble maximal d'attributs, i.e., on peut effectuer un changement minimal sur v afin que x ne soit plus dans c_0
- Cela correspond à une réponse à la question "Pourquoi pas $cl(v) \neq c_0$?", i.e., on identifie les attributs dont la valeur doit être changée pour cela

Modélisation booléenne

- analogie : mise en parallèle de 2 situations établir une correspondance entre les éléments des 2 situations et les relations qui les lient
- proportion analogique :"a est à b ce que c est à d" "le veau est à la vache ce que le poulain est à la jument"
- $a : b :: c : d =$ $((a \wedge \neg b) \equiv (c \wedge \neg d)) \wedge ((\neg a \wedge b) \equiv (\neg c \wedge d))$
  - $0 : 0 :: 0 : 0$
  - $1 : 1 :: 1 : 1$
  - $0 : 1 :: 0 : 1$
  - $1 : 0 :: 1 : 0$
  - $0 : 0 :: 1 : 1$
  - $1 : 1 :: 0 : 0$

Exemple et propriétés

- items a, b, c; d : *vecteurs* de valeurs de n attributs
  $a : b :: c : d$ ssi $\forall i \in \{1, \cdots, n\}, a_i : b_i :: c_i : d_i$

	mammifère	carnivore	bovidé	équidé	jeune	adulte
veau	1	0	1	0	1	0
vache	1	0	1	0	0	1
poulain	1	0	0	1	1	0
jument	1	0	0	1	0	1

- $a : b :: c : d \Rightarrow a : c :: b : d$ permutation centrale
  $a : b :: c : d \Rightarrow c : d :: a : b$ symétrie
  $a : b :: c : d$ et $c : d :: e : f \Rightarrow a : b :: e : f$ transitivité
  $a : b :: c : d \Rightarrow \neg a : \neg b :: \neg c : \neg d$ indépendance au codage
  on peut recalculer d à partir de a, b, c

Exemple et propriétés

- items a, b, c; d : *vecteurs* de valeurs de n attributs
  $a : b :: c : d$ ssi $\forall i \in \{1, \cdots, n\}, a_i : b_i :: c_i : d_i$

	mammifère	carnivore	bovidé	équidé	jeune	adulte
veau	1	0	1	0	1	0
vache	1	0	1	0	0	1
poulain	1	0	0	1	1	0
jument	1	0	0	1	0	1

- $a : b :: c : d \Rightarrow a : c :: b : d$ permutation centrale
  $a : b :: c : d \Rightarrow c : d :: a : b$ symétrie

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

$a : b :: c : d \Rightarrow \neg a : \neg b :: \neg c : \neg d$ indépendance au codage

on peut recalculer d à partir de a, b, c

Une lecture des données tournée vers l'explication

	$A_1 \ldots A_{i-1}$	$A_i \ldots A_{j-1}$	$A_j \ldots A_{k-1}$	$A_k \ldots A_{r-1}$	$A_r \ldots A_{s-1}$	$A_s \ldots A_n$	C
a	1	0	1	0	1	0	p
b	1	0	1	0	0	1	q
c	1	0	0	1	1	0	r
d	1	0	0	1	0	1	s

- $p : q :: r : s$ vrai ssi $p = q$ et $r = s$, ou $p = r$ et $q = s$
  - cas $p = r$ et $q = s$ ($p \neq q$) Le basculement de valeur de C de p à q entre a et b et entre c et d ne peut être expliqué, au vu des attributs considérés que par le changement de valeurs des attributs de A_r à A_n (qui est le même pour la paire (a, b) et la paire (c, d))
  - voir ces paires comme des instances d'une règle exprimant que le changement sur les attributs de A_r à A_n détermine le changement sur C quelque soit le contexte

Une lecture des données tournée vers l'explication

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

- $p : q :: r : s$ vrai ssi $p = q$ et $r = s$, ou $p = r$ et $q = s$
- cas $p = r$ et $q = s$ ($p \neq q$) Le basculement de valeur de $\mathcal{C}$ de p à q entre a et b et entre c et d ne peut être expliqué, au vu des attributs considérés que par le changement de valeurs des attributs de $\mathcal{A}_r$ à $\mathcal{A}_n$ (qui est le même pour la paire (a, b) et la paire (c, d))
  - voir ces paires comme des instances d'une règle exprimant que le changement sur les attributs de $\mathcal{A}_r$ à $\mathcal{A}_n$ détermine le changement sur $\mathcal{C}$ quelque soit le contexte

Une lecture des données tournée vers l'explication

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

- $p:q::r:s$ vrai ssi $p=q$ et $r=s$, ou $p=r$ et $q=s$
- cas $p=r$ et $q=s$ $(p\neq q)$ Le basculement de valeur de $\mathcal{C}$ de p à q entre a et b et entre c et d ne peut être expliqué, au vu des attributs considérés que par le changement de valeurs des attributs de $\mathcal{A}_r$ à $\mathcal{A}_n$ (qui est le même pour la paire (a, b) et la paire (c, d))
- voir ces paires comme des instances d'une règle exprimant que le changement sur les attributs de $\mathcal{A}_r$ à $\mathcal{A}_n$ détermine le changement sur $\mathcal{C}$ quelque soit le contexte

Petite illustration

cas	*situation*	*contre – indication* (*c. i*)	*decision*	*option* 1	*option* 2
a	sit_1	*oui*	δ	0	0
b	sit_1	*non*	δ	1	0
c	sit_2	*oui*	δ	0	1
d	sit_2	*non*	$\boldsymbol{\delta}$	**1**	**1**

- décision : peut-on servir un café avec ou sans sucre (option 1), avec ou sans lait (option 2) à une personne dans un établissement médicalisé
  Que peut-on faire en sit_2 sans *c. i.* ?
- **question** “pourquoi du lait et du sucre pour *d*?”
  réponse (le lait) “parce qu’on est en sit_2 (et pas en sit_1)”
  “parce qu’il n’y pas de *c. i.*” pour le sucre
  question “pourquoi pas de lait pour *b*?”,
  réponse “parce que l’on est en sit_1 (et pas en

Petite illustration

cas	*situation*	*contre – indication* (*c. i*)	*decision*	*option* 1	*option* 2
a	sit_1	*oui*	δ	0	0
b	sit_1	*non*	δ	1	0
c	sit_2	*oui*	δ	0	1
d	sit_2	*non*	$\boldsymbol{\delta}$	**1**	**1**

- décision : peut-on servir un café avec ou sans sucre (option 1), avec ou sans lait (option 2) à une personne dans un établissement médicalisé
  Que peut-on faire en sit_2 sans *c. i.* ?
- **question** “pourquoi du lait et du sucre pour *d*?”

réponse (le lait) “parce qu’on est en sit_2 (et pas en sit_1)”
“parce qu’il n’y pas de *c. i.*” pour le sucre

question “pourquoi pas de lait pour *b*?”,
réponse “parce que l’on est en sit_1 (et pas en

Petite illustration

cas	*situation*	*contre – indication* (*c. i*)	*decision*	*option* 1	*option* 2
a	sit_1	*oui*	δ	0	0
b	sit_1	*non*	δ	1	0
c	sit_2	*oui*	δ	0	1
d	sit_2	*non*	$\boldsymbol{\delta}$	**1**	**1**

- décision : peut-on servir un café avec ou sans sucre (option 1), avec ou sans lait (option 2) à une personne dans un établissement médicalisé
  Que peut-on faire en sit_2 sans *c. i.* ?
- **question** “pourquoi du lait et du sucre pour *d*?”

réponse (le lait) “parce qu’on est en sit_2 (et pas en sit_1)”
“parce qu’il n’y pas de *c. i.*” pour le sucre
question “pourquoi pas de lait pour *b*?”,
réponse “parce que l’on est en sit_1 (et pas en

Analogie et explication contrastive

cas	*contexte*	*changement*	*cl*
a	sit_1	*oui*	*p*
b	sit_1	*non*	*q*
c	sit_2	*oui*	*p*
d	sit_2	*non*	*q*

Table: Situation schématique de l'explication analogique

- la réponse à la question "pourquoi *d* n'est pas en classe *p*?" est dans les valeurs prises par *d* pour les attributs de *change*. Quand *c* est un proche voisin de *d*, le nombre d'attributs dans *change* est petit. On est proche d'une explication contrastive : $\exists x = c \in \mathcal{S}.[\bigwedge_{j \in \mathcal{A} \setminus change}(x_j = c_j = d_j)] \wedge (cl(x) \neq q)$
- explication contrastive $\exists x \in \mathcal{D}.[Desaccord(x, v) = \mathcal{Y} \wedge (cl(x) \neq c_0)]$

Analogie et explication abductive

- L'explication est ici plus riche, on connait au moins une autre paire (ici (a, b)), correspondant à un autre *contexte* où le même changement de valeur d'attributs conduit au même changement de classe, ce qui suggère la possibilité des règles $\forall$ *sit*,

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

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

la règle permet une lecture de la Table avec un parfum d'explication abductive, qui dit pourquoi the item is in class p (or in class q).

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

- MAIS exception si $\exists\ (a', b')$ t. q. $a' = (sit', oui)$, $b' = (sit', non)$ with $cl(a') = cl(b') = p$

Pour résumer

- les proportions analogiques ont un important potentiel explicatif à partir de données
- on peut répondre à des questions de types "*pourquoi*" et "*pourquoi pas*".
- approche apparentée, à celle du système LORE (Local Rule-Based explanations) (Guidotti et al. 2018)
- rôle des exemples "*adverses*" dans l'explication
- une proportion analogique a une valeur explicative "*Star Wars* (1977) est à *Raiders of the Lost Ark* (1981) comme *Return of the Jedi* (1983) est à *Indiana Jones and the Last Crusade* (1989)"

Pour résumer

- les proportions analogiques ont un important potentiel explicatif à partir de données
- on peut répondre à des questions de types "*pourquoi*" et "*pourquoi pas*".
- approche apparentée, à celle du système LORE (Local Rule-Based explanations) (Guidotti et al. 2018)
- rôle des exemples "*adverses*" dans l'explication
- une proportion analogique a une valeur explicative "*Star Wars* (1977) est à *Raiders of the Lost Ark* (1981) comme *Return of the Jedi* (1983) est à *Indiana Jones and the Last Crusade* (1989)"

Lien avec l'argumentation

- discutant d : situation S_2 est comme situation S_1
  ce qui s'est passé dans S_1 arrivera aussi dans S_2
  Le discutant adverse d' :

il y a une caractéristique (importante) où ils *diffèrent*
ce qui s'est passé dans S_1 n'arrivera pas en S_2

- d peut produire une autre paire de situations (S_3, S_4) où *la même différence* peut être observée ***sans affecter la conclusion*** de d pour S_2.
- d' peut contre-argumenter s'il connaît une autre paire de situations (S'_3, S'_4) où la même différence conduit à une *conclusion différente*
- analysable avec des proportions analogiques

Remarques de conclusion - 1

- Hüllermeier (2020) : usage explicatif des proportions analogiques en apprentissage
- classification et apprentissage de préférences
- inférence analogique conclut de $a : b :: c : d$ et de "a est préféré à b", que "c est préféré à d"
  L' explication analogique *s'appliquerait aussi*.
- plusieurs scénarios : 1) expliquer une prédiction concernant / la valeur d'un attribut de d, sur la base de triplets a, b, c de l'ensemble des données 2) extraire des proportions analogiques d'un ensemble de données

Remarques de conclusion - 1

- Hüllermeier (2020) : usage explicatif des proportions analogiques en apprentissage
- classification et apprentissage de préférences
- inférence analogique conclut de $a : b :: c : d$ et de "a est préféré à b", que "c est préféré à d"
  L' explication analogique *s'appliquerait aussi*.
- plusieurs scénarios : 1) expliquer une prédiction concernant / la valeur d'un attribut de d, sur la base de triplets a, b, c de l'ensemble des données
  2) extraire des proportions analogiques d'un ensemble de données

Remarques de conclusion - 2

- intéressant de *précompiler* l'ensemble des données sous forme de paires
  en repérant où les items sont égaux et où et comment ils diffèrent
  pour faciliter une analyse analogique des données

commencer par déterminer les attributs pertinents
confiance, *support* des règles associées aux paires
vient d'être d'implémenté

- une 2nde sorte de PA où *a* et *c* d'une part et *b* et *d* d'autre part appartiennent à 2 univers différents:
  "ce médicament est au rhume ce que l'aspirine est au mal de tête" (il est assez efficace et pas cher)

Remarques de conclusion - 2

- intéressant de *précompiler* l'ensemble des données sous forme de paires
  en repérant où les items sont égaux et où et comment ils diffèrent
  pour faciliter une analyse analogique des données

  commencer par déterminer les attributs pertinents
  confiance, *support* des règles associées aux paires

  vient d'être d'implémenté
- une 2nde sorte de PA où *a* et *c* d'une part et *b* et *d* d'autre part appartiennent à 2 univers différents: "ce médicament est au rhume ce que l'aspirine est au mal de tête" (il est assez efficace et pas cher)

Remarques de conclusion - 2

- intéressant de *précompiler* l'ensemble des données sous forme de paires
  en repérant où les items sont égaux et où et comment ils diffèrent
  pour faciliter une analyse analogique des données

  commencer par déterminer les attributs pertinents
  confiance, *support* des règles associées aux paires

  vient d'être d'implémenté
- une 2nde sorte de PA où *a* et *c* d'une part et *b* et *d* d'autre part appartiennent à 2 univers différents: "ce médicament est au rhume ce que l'aspirine est au mal de tête" (il est assez efficace et pas cher)