Admissible generalizations of examples as rules

Philippe Besnard[1] — *Thomas Guyet*[3] — *Véronique Masson*[2,3]

[1] *CNRS-IRIT* [2] *Université Rennes-1* [3] *IRISA/LACODAM*

Presented by Sébastien Ferré (Université Rennes-1/SemLIS).

AIMLAI workshop @EGC – January 22nd, 2019

And now . . .

Attribute-value rule learning

	(C) *Price*	(A_1) *Area*	(A_2) *Rooms*	(A_3) *Energy*	(A_4) *Town*	(A_5) *District*	(A_6) *Exposure*
1	low-priced	70	2	D	Toulouse	Minimes	
2	low-priced	75	4	D	Toulouse	Rangueil	
3	expensive	65	3		Toulouse	Downtown	
4	low-priced	32	2	D	Toulouse		SE
5	mid-priced	65	2	D	Rennes		SO
6	expensive	100	5	C	Rennes	Downtown	
7	low-priced	40	2	D	Betton		S

- Task: induce rules to predict the value of the class attribute (C)
- Rules extracted by Algorithm CN2

$$\begin{aligned}
\pi_1^{CN2} &: A_5 = \text{Downtown} \Rightarrow C = \text{expensive} \\
\pi_2^{CN2} &: A_2 < 2.50 \wedge A_4 = \text{Toulouse} \Rightarrow C = \text{low-priced} \\
\pi_3^{CN2} &: A_1 > 36.00 \wedge A_3 = D \Rightarrow C = \text{low-priced}
\end{aligned}$$

Interpretability of rules and rulesets

- The logical structure of a rule can be easily interpreted by users

  `IF` *conditions* `THEN` *class-label*

- Rule learning algorithms generate rules according to implicit or explicit principles[1]
  but ..
  - are the generated rules the *interpretable* ones?
  - would it be possible to have different rulesets?
  - why a ruleset would be better than another one from the interpretability point of view?

$\Rightarrow$ We need ways to analyze the interpretability of the outputs of rule learning algorithms

[1] principles mainly based on statistical properties!

Analyzing the interpretability of rules

Analyzing the interpretativeness of ruleset

- Objective criteria on ruleset syntax [CZV13, BS15]
  - size of the rule (number of attributes)
  - size of the ruleset
- Intuitiveness of rules through the effects of cognitive biases [KBF18]

$\Rightarrow$ Our approach formalizes rule learning and formalizes some expected properties on rules to shed light on properties of some extracted ruleset

Rule learning at a glance

Rule learning is formalized by two main functions

- ϕ: selects possible subsets of data
- f: generalizes examples as a rule (LearnOneRule process [Mit82])

$\xrightarrow{\phi} S \xrightarrow{f} \pi$

$\xrightarrow{\phi} S' \xrightarrow{f} \pi'$

$\xrightarrow{\phi} S'' \xrightarrow{f} \pi''$

$\dots \qquad \dots$

$\xrightarrow{\phi} S''' \xrightarrow{f} \pi'''$

(C) *Price*	(A_1) *Area*	(A_2) *#Rooms*	(A_3) *Energy*	(A_4) *Town*	(A_5) *District*	(A_6) *Exposure*
low-priced	70	2	D	Toulouse	Minimes	
low-priced	75	4	D	Toulouse	Rangueil	
expensive	65	3		Toulouse	Downtown	
low-priced	32	2	D	Toulouse		SE
mid-priced	65	2	D	Rennes		SW
expensive	100	5	C	Rennes	Downtown	
low-priced	40	2	D	Betton		S

⇒ We focus on the generalisation of examples as rule

Toward the notion of admissibility

	(C) Price	(A_1) Area
1	low-priced	70
2	low-priced	75
4	low-priced	32
7	low-priced	40

- Rote learning of a rule
  $A_1 = \{70, 75, 32, 40\} \Rightarrow C = \text{low-priced}$
- Most generalizing rule
  $A_1 = [32 : 75] \Rightarrow C = \text{low-priced}$
- Would the following rule be better?

$$A_1 = [32 : 40] \cup [70 : 75] \Rightarrow C = \text{low-priced}$$

$\Rightarrow$ this is the question of admissibility!

The notion of admissibility has to capture an intuitive notion of generalization

- Admissible rules are rules less likely to be counter-intuitive
- Admissibility is an elementary notion for interpretability

Eliciting a rule (f function)

	(A_0) *Price*	(A_1) *Area*	(A_2) *Rooms*
1	low-priced	70	2
2	low-priced	75	4
4	low-priced	32	2
7	low-priced	40	2

$S_0 = \{\text{low-priced}\}$ $\quad S_2 = \{2, 4\}$

$S_1 = \{32, 40, 70, 75\}$

$S_0 \xrightarrow{f} \widehat{S}_0 \qquad S_1 \to \widehat{S}_1 \qquad S_2 \xrightarrow{f} \widehat{S}_2$

- For every attribute A_i, S_i is the set of values of A_i in items of S
- Each superset of S_i is, theoretically speaking, a generalization of S_i
- The generalisation process thus consists in selecting one of these supersets:
  - f choice function that is given as input a collection of supersets of S_i and picks one

We are looking for an appropriate $\widehat{\cdot}$ for (§) i.e.

$$A_1(x) \in \widehat{S}_1 \wedge \cdots \wedge A_n(x) \in \widehat{S}_n \to C(x) \in \widehat{S}_0 \qquad (\S)$$

Generalization of S_i : $\widehat{S}_i = f(\{Y \mid S_i \subseteq Y \subseteq \mathrm{Rng}\, A_i\})$

Notion of admissibility: propositions

Generalization of S_i : $\widehat{S_i} = f(\{Y \mid S_i \subseteq Y \subseteq \mathrm{Rng}\, A_i\})$

What collection $\mathcal{X} = \{\widehat{S_i} \mid S_i \subseteq \mathrm{Rng}\, A_i\}$ would do?

(i) $\mathrm{Rng}\, A_i \in \mathcal{X}$

(ii) if X and Y are in $\mathcal{X}$ then so $X \cap Y$.

- $\mathcal{X}$ is a closure system upon $\mathrm{Rng}\, A_i$.
- $\widehat{\cdot}$ is an operation enjoying weaker properties than closure operators; alternatives looked at:
  - pre-closure operator
  - capping operator

What choice function(s) can in practice capture these expected algebraic properties?

- Proposal for some classes of choice functions generating specific types of operators
- Concrete examples of such functions for numerical rules

Class of choice functions satisfying pre-closure

Theorem. *Given a set Z, let $f : 2^{2^Z} \to 2^Z$ be a function st for every upward closed $\mathcal{X} \subseteq 2^Z$ and every $\mathcal{Y} \subseteq 2^Z$:*

1. $f(2^Z) = \emptyset$
2. $f(\mathcal{X}) \in \mathcal{X}$
3. $f(\mathcal{X} \cap \mathcal{Y}) = f(\mathcal{X}) \cup f(\mathcal{Y})$
   whenever $\bigcup \min(\mathcal{X} \cap \mathcal{Y}) = \bigcup \min \mathcal{X} \cup \bigcup \min \mathcal{Y}$

Then, $\widehat{\cdot} : 2^Z \to 2^Z$ *as defined by*

$$\widehat{X} \stackrel{\mathrm{def}}{=} f(\{Y \mid X \subseteq Y \subseteq Z\})$$

is a pre-closure operator upon Z.

Intuition: Z is $\mathrm{Rng}\, A_i$
$\mathcal{X}$ (and $\mathcal{Y}$, too) is a collection of intervals over $\mathrm{Rng}\, A_i$
moreover, $\mathcal{X}$ is a collection containing all super-intervals of an interval belonging to the collection

Class of choice functions satisfying pre-closure

Theorem. *Given a set Z, let $f : 2^{2^Z} \to 2^Z$ be a function st for every upward closed $\mathcal{X} \subseteq 2^Z$ and every $\mathcal{Y} \subseteq 2^Z$:*

1. $f(2^Z) = \emptyset$
2. $f(\mathcal{X}) \in \mathcal{X}$
3. $f(\mathcal{X} \cap \mathcal{Y}) = f(\mathcal{X}) \cup f(\mathcal{Y})$
   whenever $\bigcup \min(\mathcal{X} \cap \mathcal{Y}) = \bigcup \min \mathcal{X} \cup \bigcup \min \mathcal{Y}$

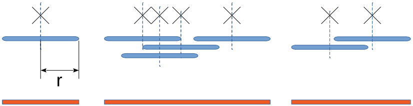
Then, $\widehat{\cdot} : 2^Z \to 2^Z$ *as defined by*

$$\widehat{X} \stackrel{\text{def}}{=} f(\{Y \mid X \subseteq Y \subseteq Z\})$$

is a pre-closure operator upon Z.

<u>Numerical attributes:</u> principle of single point (u) interpolation

$$A_i(x) \in [u - r : u + r] \to C(x) = c.$$



Example of framework usage

Dataset 1

	C	A
1	A	70
2	A	75
3	B	32
4	B	40

Dataset 2

	C	A
1	A	70
2	A	74
3	A	72
4	A	75
5	B	32
6	B	40

- CN2 chooses the boundary to be the middle of the bounds in between the two classes $\Rightarrow$ same rules in both cases
- CN2 is insensitive to examples density

Is it good or is it bad to be sensitive to examples density?

$\rightarrow$ it depends on the notion of admissibility!

- some admissible generalizations enjoying capping are sensitive to examples density, in contrast to CN2 generalizations!
- CN2 generalizations form an admissible class of generalizations enjoying *cumulation*!

Conclusion

- The logical structure of rules makes them easy to read but ...
- The interpretability of rules learned from examples requires, in particular, to take care of the way examples are generalized
- Qualifying the interpretable nature of rule learning outputs is challenging

- Our work contributes by giving a way to do such analysis
  - A proposal of a general framework for rule learning
  - A topological study of *admissible generalisations* of examples
- Perspectives: study the characteristics of extracted rulesets (set of rules)

Bibliography

Fernando Benites and Elena Sapozhnikova, *Hierarchical interestingness measures for association rules with generalization on both antecedent and consequent sides*, Pattern Recognition Letters **65** (2015), 197–203.

Alberto Cano, Amelia Zafra, and Sebastián Ventura, *An interpretable classification rule mining algorithm*, Information Sciences **240** (2013), 1–20.

Tomás Kliegr, Stepán Bahník, and Johannes Fürnkranz, *A review of possible effects of cognitive biases on interpretation of rule-based machine learning models*, CoRR **abs/1804.02969** (2018).

Kazimierz Kuratowski, *Topology*, vol. 1, Elsevier, 2014.

Tom M Mitchell, *Generalization as search*, Artificial Intelligence **18** (1982), 203–226.

Weakening closure operators

- List of Kuratowski's axioms [Kur14] (closure system):

$$\widehat{\emptyset} = \emptyset$$
$$S \subseteq \widehat{S} \subseteq \mathrm{Rng}\, A_i$$
$$\widehat{\widehat{S}} = \widehat{S} \qquad \textit{[to be dropped for pre-closure]}$$
$$\widehat{S \cup S'} = \widehat{S} \cup \widehat{S'}$$

- Actually, we downgrade Kuratowski's axioms as follows

$$\widehat{S} \subseteq \widehat{S'} \text{ whenever } S \subseteq S' \qquad \text{(closure)}$$
$$\widehat{S} = \widehat{S'} \text{ whenever } S \subseteq S' \subseteq \widehat{S} \qquad \text{(cumulation)}$$
$$\widehat{S \cup S'} \subseteq \widehat{S} \text{ whenever } S' \subseteq \widehat{S} \qquad \text{(capping)}$$

Lemma: Kuratowksi $\Rightarrow$ closure $\Rightarrow$ cumulation $\Rightarrow$ capping

Class of choice functions satisfying capping

Theorem. *Given a set Z, let $f : 2^{2^Z} \to 2^Z$ be a function st for every $\mathcal{X} \subseteq 2^Z$ such that $\bigcap \mathcal{X} \in \mathcal{X}$ and for every $\mathcal{Y} \subseteq 2^Z$*

1. $f(\mathcal{X}) \in \mathcal{X}$
2. *if* $\mathcal{Y} \subseteq \mathcal{X}$ *and* $\exists W \in \mathcal{Y}$, $W \subseteq f(\mathcal{X})$ *then* $f(\mathcal{Y}) \subseteq f(\mathcal{X})$

Then, $\widehat{\cdot} : 2^Z \to 2^Z$ *as defined by*

$$\widehat{X} \stackrel{\text{def}}{=} f(\{Y \mid X \subseteq Y \subseteq Z\})$$

is a capping operator upon Z.

Intuition: Z is $\mathrm{Rng}\, A_i$
$\mathcal{X}$ (and $\mathcal{Y}$, too) is a collection of intervals over $\mathrm{Rng}\, A_i$
moreover, $\mathcal{X}$ is a collection whose intersection
is itself a member of the collection

Class of choice functions satisfying capping

Theorem. *Given a set Z, let $f : 2^{2^Z} \to 2^Z$ be a function st for every $\mathcal{X} \subseteq 2^Z$ such that $\bigcap \mathcal{X} \in \mathcal{X}$ and for every $\mathcal{Y} \subseteq 2^Z$*

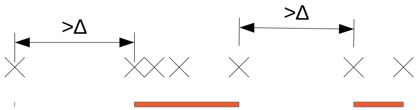
1. $f(\mathcal{X}) \in \mathcal{X}$
2. *if* $\mathcal{Y} \subseteq \mathcal{X}$ *and* $\exists W \in \mathcal{Y}$, $W \subseteq f(\mathcal{X})$ *then* $f(\mathcal{Y}) \subseteq f(\mathcal{X})$

Then, $\widehat{\cdot} : 2^Z \to 2^Z$ *as defined by*

$$\widehat{X} \stackrel{\text{def}}{=} f(\{Y \mid X \subseteq Y \subseteq Z\})$$

is a capping operator upon Z.

Numerical attributes: principle of pairwise point interpolation



Eliciting a rule

- S being a square is supposed to capture a rule π requires that every item of S satisfies π
  - $\rightarrow$ generalisation does not capture the statistical representativeness of dataset, but only elicits a rule generalizing all its items

(C) *Price*	(A_1) *Area*	(A_2) *#Rooms*	(A_3) *Energy*	(A_4) *Town*	(A_5) *District*	(A_6) *Exposure*
low-priced	70	2	D	Toulouse	Minimes	
low-priced	75	4	D	Toulouse	Rangueil	
expensive	65	3		Toulouse	Downtown	
low-priced	32	2	D	Toulouse		SE
mid-priced	65	2	D	Rennes		SW
expensive	100	5	C	Rennes	Downtown	
low-priced	40	2	D	Betton		S

$$\downarrow f$$

$$A_0 = 2 \wedge A_4 = \text{Toulouse} \Rightarrow C = \text{low-priced}$$

(C) *Price*	(A_1) *Area*	(A_2) *#Rooms*	(A_3) *Energy*	(A_4) *Town*	(A_5) *District*	(A_6) *Exposure*
low-priced	70	2	D	Toulouse	Minimes	
low-priced	75	4	D	Toulouse	Rangueil	
expensive	65	3		Toulouse	Downtown	
low-priced	32	2	D	Toulouse		SE
mid-priced	65	2	D	Rennes		SW
expensive	100	5	C	Rennes	Downtown	
low-priced	40	2	D	Betton		S

$$\downarrow f$$

$$A_0 \in [2, 4] \Rightarrow C \in \{\text{low-priced}, \text{expensive}\}$$