Imperial College London

Beneficial and harmful explanatory machine learning

Lun Ai, Imperial College London

L. Ai, S.H. Muggleton, C. Hocquette, M. Gromowski, and U. Schmid. Beneficial and harmful explanatory machine learning. Machine Learning, 110:695-721, 2021.

Imperial College London

Challenges and issues

- Under-specified and ambiguous definitions
- A lack of empirical data
- Explanation effects are “subjective”
- Limited references to social science literature
- Little accounting for humans' perspective
- Focus on exclusively on beneficiality

Imperial College London

Contributions

1) Operational measure of machine learning explanatory effect
2) Cognitive window framework
3) Demonstration of beneficial/harmful explanatory effect on human comprehension

Imperial College London

Human comprehension

Human out-of-sample predictive accuracy after studying training materials

U. Schmid, C. Zeller, T. Besold, A. Tamaddoni-Nezhad, and S.H. Muggleton. How does predicate invention affect human comprehensibility?. In Alessandra Russo and James Cussens, editors, Proceedings of the 26th International Conference on Inductive Logic Programming, pages 52-67, Berlin, 2017.

Imperial College London

Explanatory effectiveness

Effect = machine-aided comprehension - self-learning comprehension

Beneficial = positive effect

Harmful = negative effect

Otherwise, no observable effect

Imperial College London

Cognitive window

Bound on hypothesis space size (**learning**):

Effect is negative when $|S| > B(M(E),H)$

Kolmogorov complexity bound on execution cost (**runtime**):

Effect is non-positive when $Cog(M(E),x) \geq CogP(E,\overline{M},\phi,x)$

Imperial College London

Materials

Noughts and Crosses isomorphism:

- Avoid ceiling effect
- Same complexity
- 2-ply initialisation

Easiest

Hardest

Last move to win as Player X

(win_1)

X		X
O	X	
O		O

Two moves to win as Player X

(win_2)

X		
O	X	
		O

Three moves to win as Player X

(win_3)

O	X	

Noughts and Crosses

X to move

Win - player has three pieces in a line

X		X
O	X	
O		O

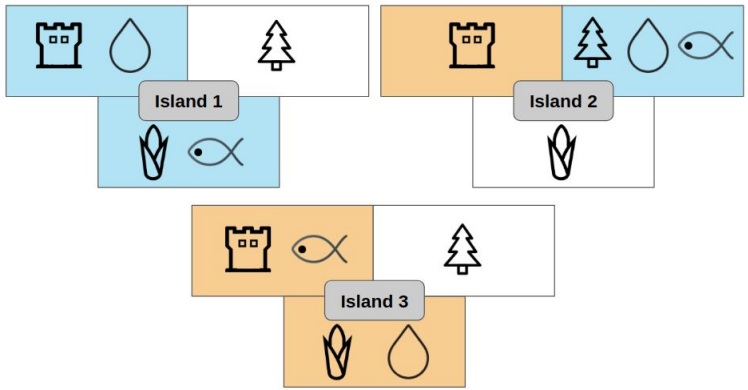
isomorphic

The Island Game

Blue to move

Win - 1) player has three cells on the same island

2) player has three of the same resource



Imperial College London

Two MIL systems

MIGO:

Sufficient and necessary BK

Positive examples only

Learns minimax algorithm

S.H. Muggleton and C. Hocquette. Machine discovery of comprehensible strategies for simple games using meta-interpretive learning. New Generation Computing, 37:203-217, 2019.

Imperial College London

Two MIL systems

MIPlain (adapted MIGO with Metaopt):

Additional BK

Positive and negative examples

Learns smaller program with less inferential cost

A. Cropper and S.H. Muggleton. Learning efficient logic programs. Machine Learning, 108:1063-1083, 2019.

Imperial College London

MIPlain's learned program

Depth	Rules
1	`win_1(A,B):-move(A,B),won(B)`
2	`win_2(A,B):-move(A,B),win_2_1(B)` `win_2_1(A):-number_of_pairs(A,x,2), number_of_pairs(A,o,0)`

3	`win_3(A,B):-move(A,B),win_3_1(B)`
	`win_3_1(A):-number_of_pairs(A,x,1),win_3_2(A)`
	`win_3_2(A):-move(A,B),win_3_3(B)`
	`win_3_3(A):-number_of_pairs(A,x,0),win_3_4(A)`
	`win_3_4(A):-win_2(A,B),win_2_1(B)`

Imperial College London

Comprehensibility test

Control group :

pre test => training with examples => post test

Treatment group:

pre test => training with examples & explanations => post test

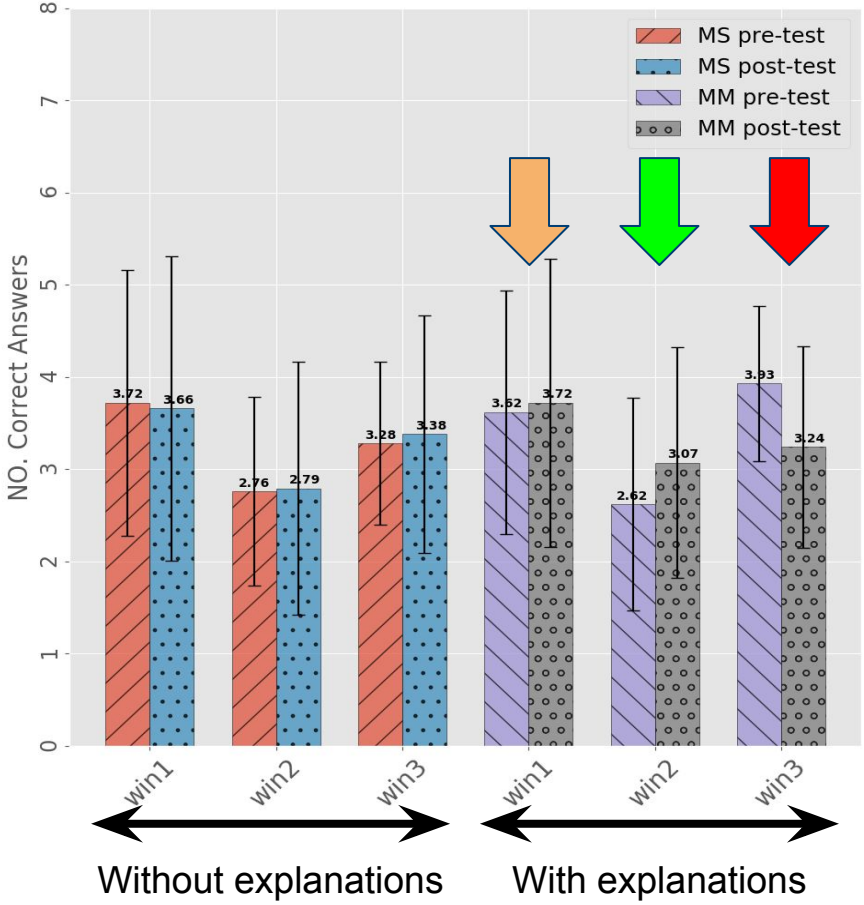
Imperial College London

Explanations in training

English translation of MIPlain's learned theory

Visualisation of game states

Contrastive consequences of good and bad moves



	Orange	Green	Red
Effect	No observable	Beneficial	Harmful

Imperial College London

Empirical results summary

1) Bound on the program size learnable
2) Short-term comprehension improved by heuristics
3) Evidence for a *cognitive window*

Imperial College London

Future work

How does the ordering of concepts affect comprehension?

How does machine-aided knowledge corrections affect comprehension?

Imperial College London

Thank you

Imperial College London

Cognitive window

Bound on hypothesis space size (**learning**):

Effect is negative when $|S| > B(M(E),H)$

Kolmogorov complexity bound on execution cost (**runtime**):

Effect is non-positive when $Cog(M(E),x) \geq CogP(E,\overline{M},\phi,x)$

Imperial College London

Human window

A balance between memory and computational complexity

D. Michie. “Experiments on the Mechanization of Game-Learning. 2-Rule-Based Learning and the Human Window.” Comput. J., pages 105-113, 1982.

Imperial College London

Ultra-strong machine learning

More training examples improves predictive accuracy

Symbolic output

Humans achieve better performance via teaching

D. Michie.Machine learning in the next five years. In Proceedings of the third European working session on learning, pages 107–122,1988.