Inductive Programming
Lecture 7
Data Wrangling

Stephen Muggleton
Department of Computing
Imperial College, London and
University of Nanjing

24th October, 2024

Papers for this lecture

Paper7.1 L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo, F. Martinez-Plumed, M.J. Ramirez-Quintana General-purpose Declarative Inductive Programming with Domain-Specific Background Knowledge for Data Wrangling Automation. arXiv:1809.10054v1 [cs.AI] 26 Sep 2018.

Paper7.2: A. Cropper, A. Tamaddoni-Nezhad and S.H. Muggleton. Meta-interpretive learning of data transformation programs. In Proceedings of the 25th International Conference on Inductive Logic Programming, pages 46-59. Springer-Verlag, 2016.

Motivation

- Inductive Programming
- Data Wrangling - transform raw data for analysis
- Large amount of effort writing small, error-prone programs
- Applications in Business, Science, Medicine
- Microsoft Research Redmond - Inductive Programming
- Academic Research

Commercial Data Wrangling Video

- Microsoft Inductive Programming products - Sumit Gulwani
- YouTube - Data Wrangling using Programming by Examples
- https://www.youtube.com/watch?v=XWRsxy8SbzY

Inductive Functional Programming Data Wrangling [Paper7.1]

Id	Input	Outputs
1	25-03-74	25/03/74
2	29-03-86	29/03/86
3	11-02-96	11/02/96
4	11-17-98	17/11/98
5	17-05-17	17/05/17
6	25-08-05	25/08/05
7	30-06-75	30/06/75
8	...	...

Dates with desired output format

Materials and Method [Paper7.1]

id	Domain	#Ex.	Description
1	Freetext	12	Complete brackets (From [29])
2	...	...	...
6	Dates	26	Change the punctuation of a date (From [30])
7	...	...	...
14	Emails	24	Extract words after '@' (From [33])
15	...	...	...
30	Units	12	Extract the units of a value (From [32]
31	...	...	...

Table 6: Data wrangling repository.
http://dmip.webs.upv.es/datawrangling/

Method: MagicHaskeller

Results [Paper7.1]

id	Domain	default	freetext	dates	...	all
1	freetext	0.00	**1.00**	0.00	...	0.00
...	...	...	...	...	...	...
6	dates	0.00	**1.00**	**1.00**	...	0.00
...	...	...	...	...	...	...
14	emails	0.00	0.04	0.04	...	0.00
...	...	...	...	...	...	...
30	units	0.64	0.18	0.18	...	0.00
...	...	...	...	...	...	...

Table 7: Accuracy depending on set of primitives (DSBK). Demonstrates Background **Relevance** problem.

Inductive Logic Programming Data Wrangling
Extracting predation facts from Ecological papers.
[Paper7.2]

Input

Harpalus rufipes eats large prey such as Lepidoptera .
Bembidion lampros : In cereals the main food was Collembola .

Ecological data output [Paper7.2]

Output

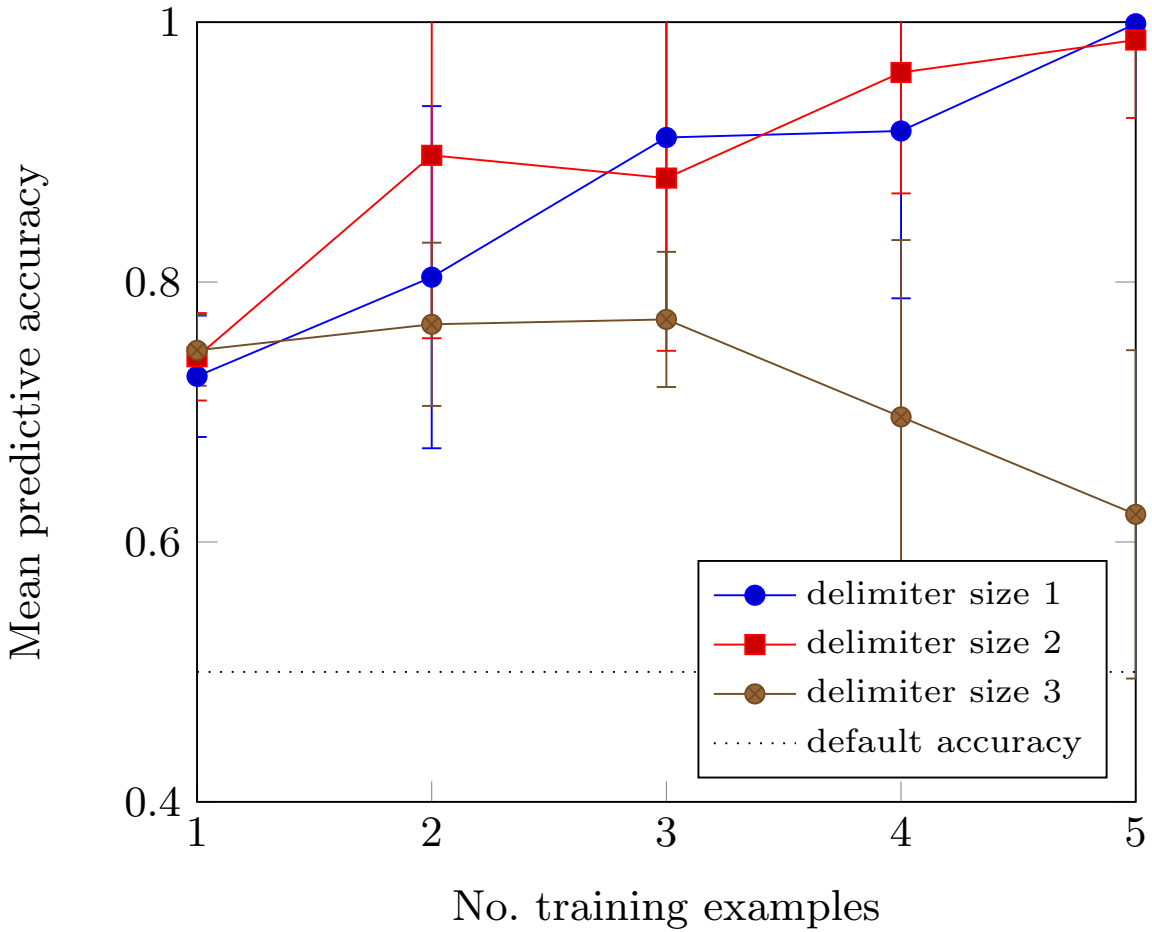
Harpalus rufipes	eats	Lepidoptera
Bembidion lampros	food	Collembola

Ecological data Background Knowledge (BK) [Paper7.2]

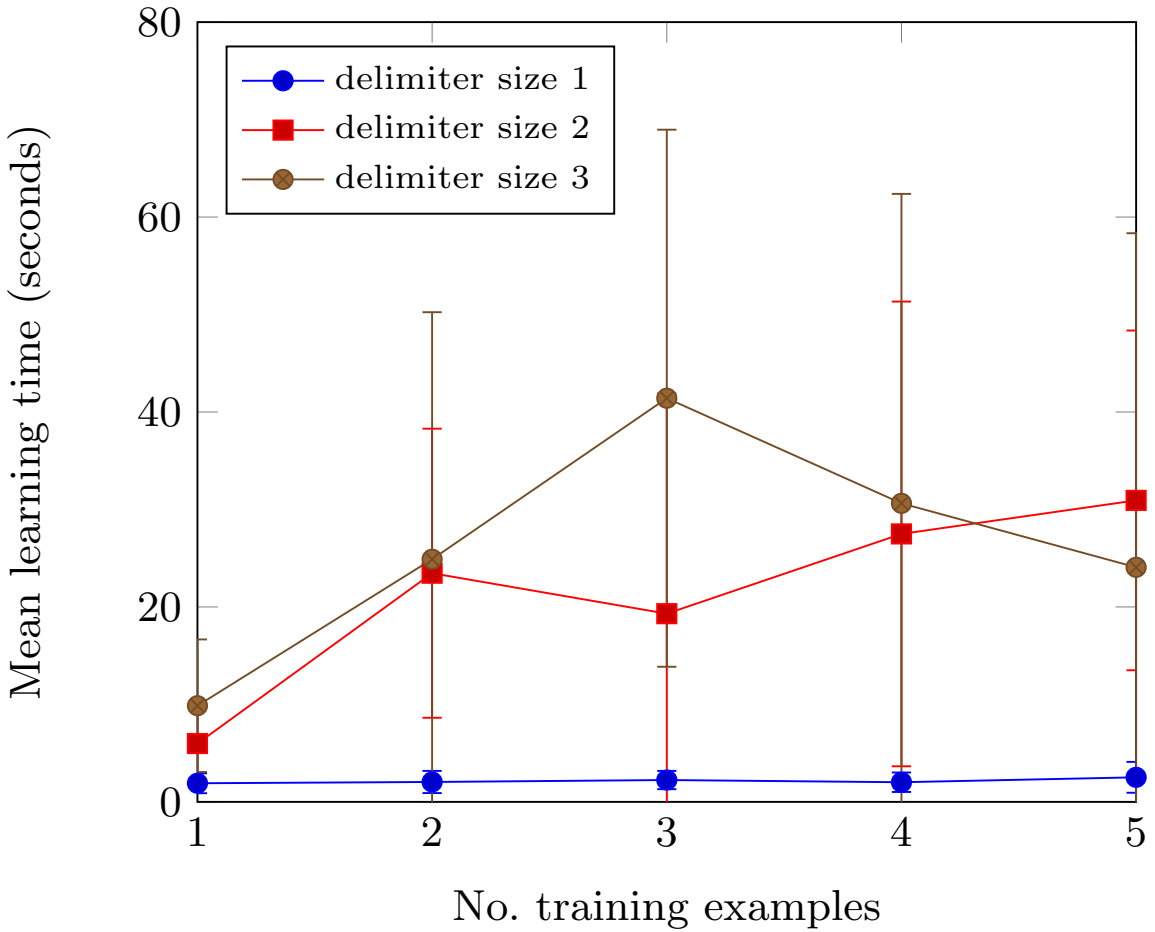
BK

```
find_species(A,B):-
   known_species(Species),
   find_sublist(A,B,Species).
known_species([L,o,r,i,c,e,r,a, ,p,i,l,i,c,o,r,n,i,s]).
known_species([H,a,r,p,a,l,u,s, ,r,u,f,i,p,e,s]).
```

Ecological predictive accuracies [Paper7.2]



Ecological training times [Paper7.2]



Ecological Induced Program[Paper7.2]

```
f(A,B):- f3(A,C), find_species(C,B).
f3(A,B):- find_species(A,C), f2(C,B).
f2(A,B):- closed_interval(A,B,[f,o],[o,d]).
f3(A,B):- find_species(A,C), f1(C,B).
f1(A,B):- closed_interval(A,B,[e,a],[t,s]).
```

Medical Data Wrangling input [Paper7.2]

P_001 67 year lung disease: n/a, Diagnosis: Unknown 80.78
P_003 56 Diagnosis: carcinoma, lung disease: unknown 20.78
P_013 70 Diagnosis: pneumonia 55.9

Output [Paper7.2]

Output

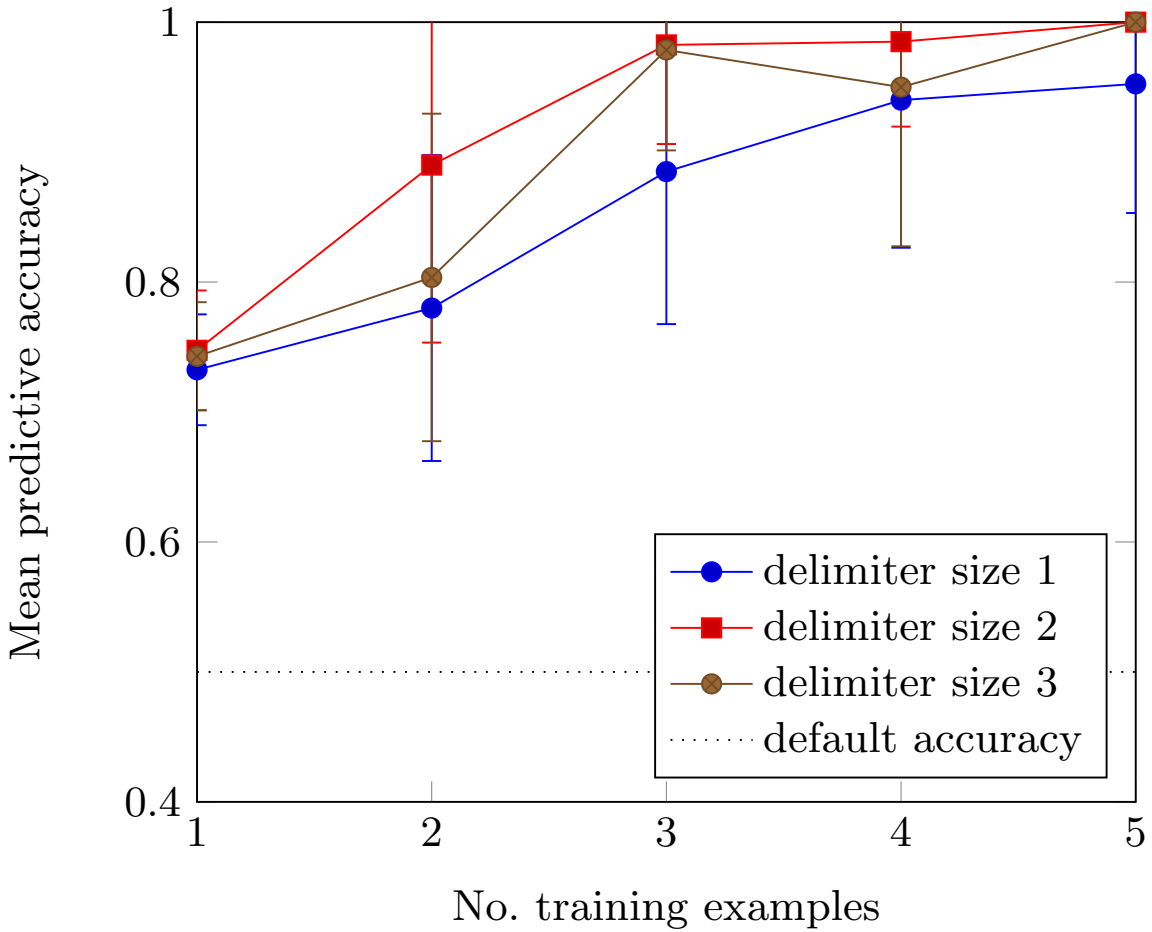
P_001	67	Unknown
P_003	56	carcinoma
P_013	70	pneumonia

Induced program [Paper7.2]

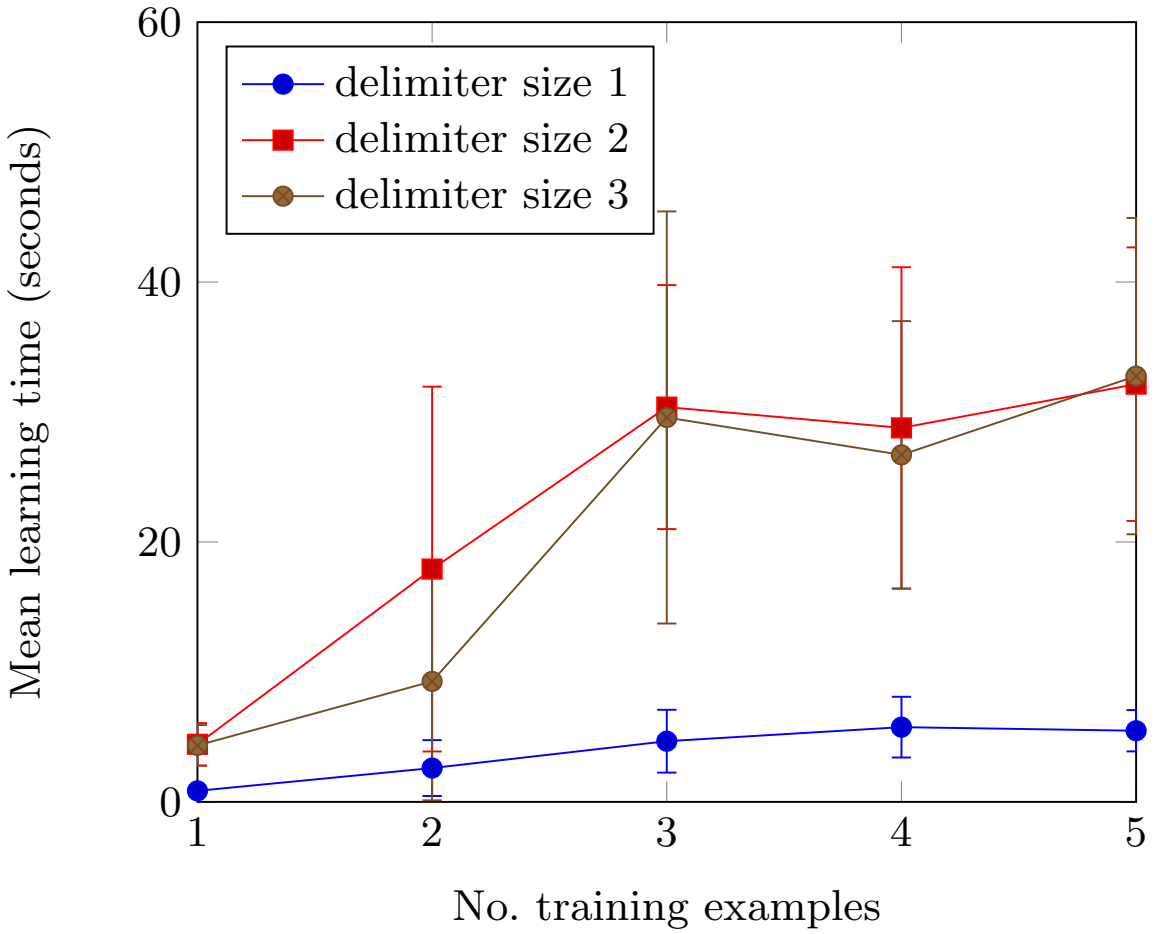
Induced program

```
f(A,B):- f2(A,C), f1(C,B).
f2(A,B):- find_patient_id(A,C), find_int(C,B).
f1(A,B):- open_interval(A,B,[':',' '],[';'n']).
f1(A,B):- open_interval(A,B,[':',' '],[',',' ']).
```

Medical predictive accuracies [Paper7.2]



Medical training times [Paper7.2]



Medical Induced Program[Paper7.2]

```
f(A,B):- f2(A,C), f2(C,B).
f2(A,B):- find_patient_id(A,C), find_int(C,B).
f2(A,B):- f1(A,C), find_float(C,B).
f1(A,B):- open_interval(A,B,[':',' '],[';'n']).
f1(A,B):- open_interval(A,B,[':',' '],[',',' ']).
```

Summary

- Data Wrangling - transform raw data for analysis
- First commercial mass market use of IP in 2013
- Microsoft Research Redmond - Inductive Programming
- Academic Research - results dependent on primitives used
- Medical records example - patientID and delimiters
- Ecological records example - speciesID and delimiters
- Induced programs readable - supports debugging