Samsung Research

Synthetic Datasets and Evaluation Tools for Inductive Neural Reasoning

Cristina Cornelio - Samsung AI Center - Cambridge

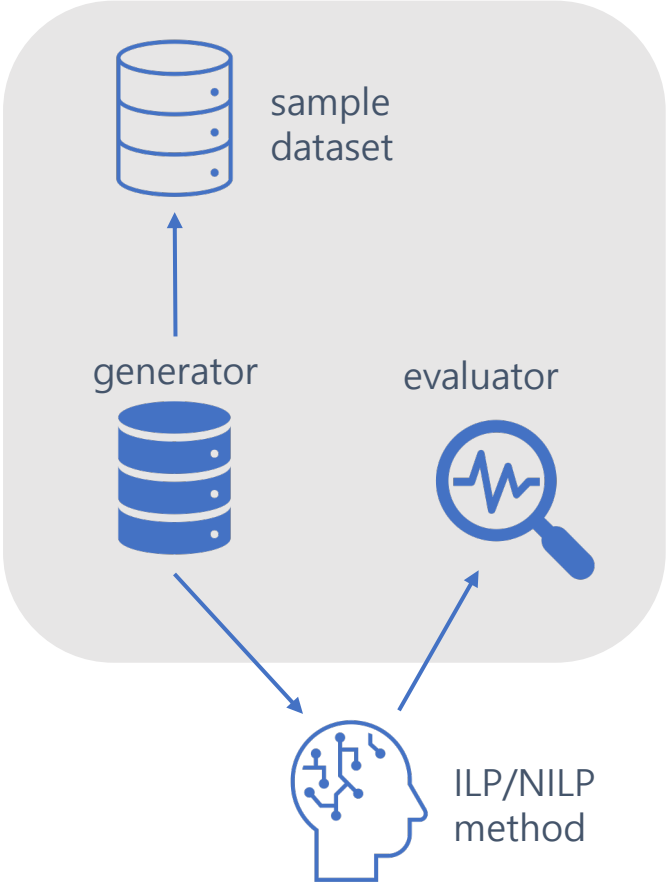
Veronika Thost - IBM Research/MIT-IBM Watson AI Lab

| 26 Oct. 2021 | ILP @ IJCLR 2021

Logic datasets for neural-ILP systems development

Samsung Research

- **Logical rules** are a popular and compact knowledge representation language in many domains
- **Learning rules automatically (ILP)** is a very active research field and, more recently, extended to **neural systems (NILP)**
- **NILP** research area is **missing adequate datasets** and evaluation approaches:
  - only toy dataset
  - not cover the various kinds of dependencies between rules
  - not allow for testing scalability
- **RuDaS** (**S**ynthetic **Da**tasets for **Ru**le Learning):
  - **logic generator** for synthetic datasets containing both facts and rules
    - including a pre-generated sample dataset
    - datalog expressivity
  - **performance evaluator** for NILP/ILP systems



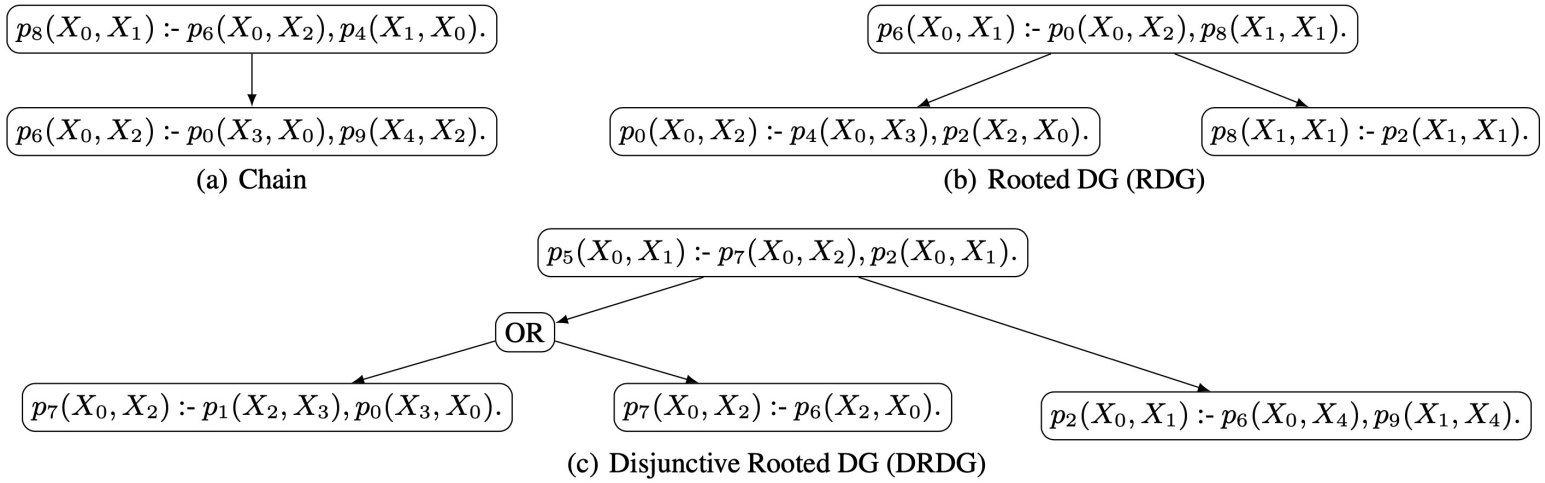
RuDaS is highly parameterizable

RuDaS is parameterizable in many dimensions:

1. maximal number of predicates, constants and facts
2. maximal arity of predicates
3. number and maximal length of rules
4. consequences of rules (i.e., completeness)
5. amount of noise (e.g., wrong or missing support facts)
6. evaluation metrics: novel and classic measures
7. type of dependencies between rules:
   Chain, R-DG, DR-DG, Mixed
8. maximal depth of rule graphs
9. minimal and maximal number of DGs in the rule set

Example of dataset

Rules.
p3(X0,X1) :- p7(X1,X0).
p7(X0,X2) :- p6(X0,X1), p6(X1,X2).
p7(X1,X0) :- p9(X3,X1), p9(X1,X0).
Facts.
p9(c127,c381).
p6(c324,c291).
p3(c363,c354). p7(c61,c96).
...



Sample Dataset: RuDaS.v0

#	Rule type	Size	Depth	#Rules			#Facts			#Pred			#Const		
				min	avg	max	min	avg	max	min	avg	max	min	avg	max
10	CHAIN	S	2	2	2	2	51	74	95	5	7	9	31	47	71
10	CHAIN	S	3	3	3	3	49	70	97	7	8	9	31	43	64
10	CHAIN	M	2	2	2	2	168	447	908	9	10	11	97	259	460
10	CHAIN	M	3	3	3	3	120	508	958	8	10	11	52	230	374
22	RDG	S	2	3	3	3	49	84	122	6	9	11	28	50	84
12	RDG	S	3	4	5	6	56	104	172	8	10	11	41	55	75
22	RDG	M	2	3	3	3	200	646	1065	6	11	11	71	370	648
22	RDG	M	3	4	5	7	280	613	1107	10	11	11	149	297	612
22	DRDG	S	2	3	4	5	60	100	181	6	9	11	29	55	82
12	DRDG	S	3	4	7	11	58	144	573	8	10	11	34	58	89
22	DRDG	M	2	3	4	5	149	564	1027	10	11	11	88	327	621
22	DRDG	M	3	4	7	12	111	540	1126	10	11	11	70	284	680

Performance evaluator

RuDaS provides tools to evaluate the performance of ILP systems, that compute distances between logic programs.

Need to ground the programs

- **Accuracy**
- **Precision** (or standard confidence)
- **Recall**
- **F1-score**
- **Herbrand distance:** distance between Herbrand models
- **Herbrand accuracy:** Herbrand distance normalized on the Herbrand base
- **Herbrand score** or **H-score**

$$\text{H-score}(\mathcal{R}, \mathcal{R}', \mathcal{F}) := \frac{|I(\mathcal{R}, \mathcal{F}) \cap I(\mathcal{R}', \mathcal{F})|}{|I(\mathcal{R}, \mathcal{F}) \cup I(\mathcal{R}', \mathcal{F})|}$$

No need to ground the programs

- **Rule-score**: an efficient measure that considers only the induced rules and not the grounded atoms.

$$\text{R-score}(\mathcal{R}, \mathcal{R}') = 1 - \frac{1}{|\mathcal{R}|}\left(\sum_{r_1 \in \mathcal{R}} \min_{r_2 \in \mathcal{R}'[hp(r_1)]} d_R(r_1, r_2)\right)$$

Experiments setup

- GOAL: demonstrate the need for a portfolio of diverse datasets for evaluating rule learning systems.

We compared the following systems:

- **FOIL**: traditional ILP system
- **AMIE+**: rule mining system
- **Neural-LP**: neural approach
- **NTP** : neural approach

Evaluated on:

- Our sample dataset RuDaS.v0
- Manually created complete dataset: EVEN

```
even(X) :- even(Z), succ(Z,Y),succ(Y,X).

succ(0,1).      succ(5,6).      even(0).
succ(1,2).      succ(6,7).      even(2).
succ(2,3).      succ(7,8).      even(4).
succ(3,4).      succ(8,9).
succ(4,5).      succ(9,10).
```

Results (1/2)

	EVEN	Compl.	Incompl.	Incompl.+Noise
FOIL	1.0	0.4053	0.1919	0.0849
AMIE+	-	0.2021	0.2098	0.2075
Neural-LP	-	0.0633	0.0692	0.0649
NTP	1.0	0.0482	0.0617	0.0574

	CHAIN	RDG	DRDG
FOIL	0.2024	0.0877	0.1633
AMIE+	0.3395	0.2275	0.1293
Neural-LP	0.1291	0.1050	0.0734
NTP	0.1239	0.0538	0.0368

We studied the impact of:

- missing information and noise
- rule structure
- scalability (dataset size)

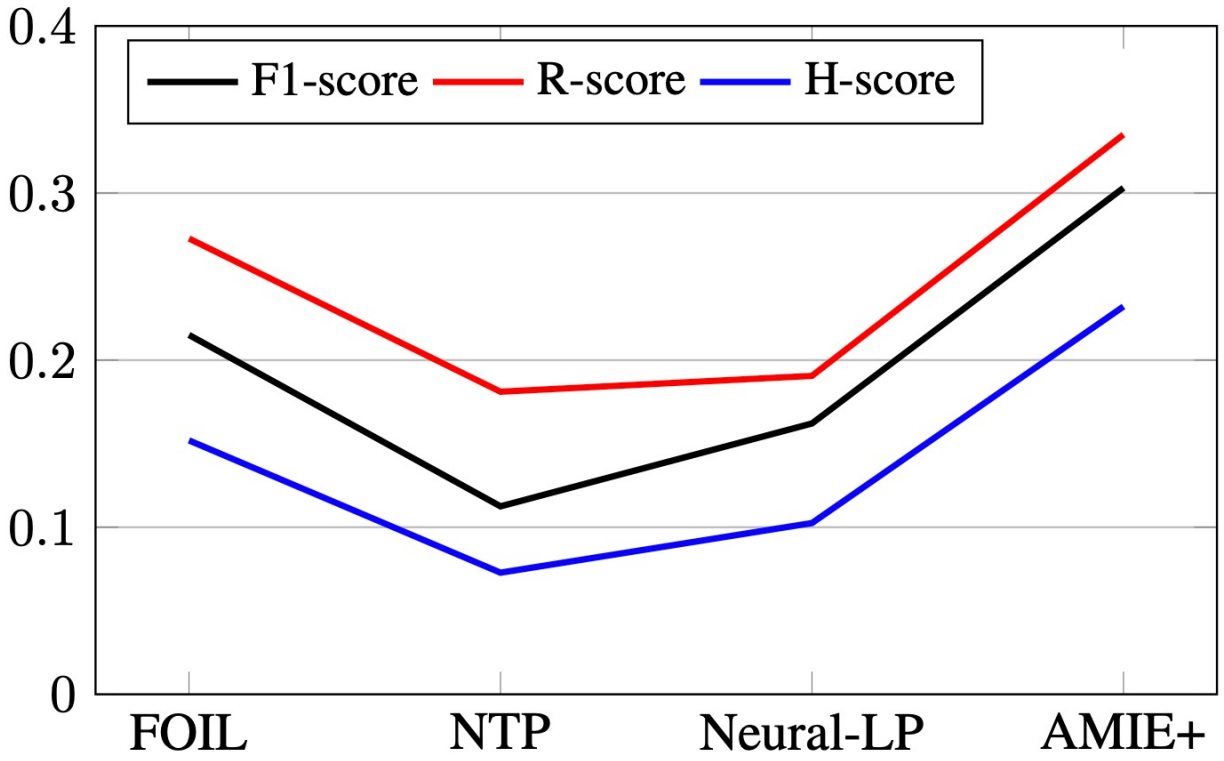
	S-2	S-3	M-2	M-3
FOIL	0.2815	0.2074	0.0356	0.0934
AMIE+	0.1449	0.1319	0.4392	0.2124
Neural-LP	0.1155	0.0673	0.1281	0.0992
NTP	0.1512	0.0432	0.0652	0.0374

All the results report the Herbrand score.

Results (2/2)

- Quality of evaluation metrics
- R-score: valid alternative with the advantage of computational efficiency

	FOIL	AMIE+	Neural-LP	NTP
H-accuracy	0.9873	0.8498	0.9850	0.9221
Accuracy	0.9872	0.8494	0.9849	0.9219
F1-score	0.2151	0.3031	0.1621	0.1125
H-score	0.1520	0.2321	0.1025	0.0728
Precision	0.5963	0.2982	0.1687	0.1021
Recall	0.2264	0.7311	0.2433	0.3921
R-score	0.2728	0.3350	0.1906	0.1811



Summary

We proposed **RuDaS**, a system that provides:

- a **logic generator** for different datasets types for rule learning
- a sample dataset **RuDaS.v0**
- an **evaluation tool** with measures that are more suitable for rule learning systems

Take-home message - It is important to:

- have differentiated data that consider several rules types
- test scalability and different type/amount of noise
- perform the evaluation using appropriate measures

Open-Source GIT repository:
github.com/IBM/RuDaS