ConfMaster@IJCAI2019 printed for Veronika Thost (thost) at 2019-04-15 14:37:31

Reviews

Paper# 2728: RuDaS: Synthetic Datasets for Rule Learning

Authors Cristina Cornelio
Veronika Thost (Contact Author)

Abstract

Logical rules are a popular knowledge representation language in many domains, they represent background knowledge and encode information that can be derived from given facts in a compact form. However, rule formulation is a complex process that requires deep domain expertise, and is further challenged by today's often large, heterogeneous, and incomplete knowledge graphs. Several approaches for learning rules automatically, given a set of input example facts, have been proposed over time, including, more recently, neural systems. Yet, the area is missing adequate evaluation approaches: existing datasets (i.e., facts and rules to be learned) resemble toy examples that neither cover the various kinds of dependencies between rules nor allow for testing scalability. We present a tool for generating different kinds of datasets and for evaluating rule learning systems.

Paper Type	Full Paper
Keywords	[Special Track on Understanding Intelligence and Human-level AI in the New Machine Learning era] Learning knowledge representations (Special Track on Human AI and Machine Learning), [Special Track on Understanding Intelligence and Human-level AI in the New Machine Learning era] Neural Methods in Inductive Logic Programming (Special Track on Human AI and Machine Learning), [Special Track on Understanding Intelligence and Human-level AI in the New Machine Learning era] Neuro-Symbolic methods (Special Track on Human AI and Machine Learning)
The authors commit to…	yes
The list of authors is c…	yes
Authors have read and…	yes
The authors agree that…	Yes
The first author is a st…	[]

Reviews

Review #210721 New

Comments to Authors.

The paper presents a tool for generating different kinds of datasets to be used for learning and most of all for evaluating rule learning systems. The goal is, from one hand to have a widely recognized test best (to be possibly used also for checking the scalability performances of rule learning systems), on the other hand to have a tool that can be used for checking the quality of the learned rules in different scenarios and/or under different conditions such as heterogeneous and incomplete knowledge, checking for/learning specific types of dependencies etc.

The paper focuses on an important problem and it could be potentially an important contribution for the scientific community. Overall the idea appears promising, motivations for it are well presented but the contribution per se results rather preliminary. In the following more detailed comments are reported.

The paper is overall well motivated. An exception is provided by section 2.2, where motivations are a bit weak and vague.

Looking at table 1, the generated datasets appear rather small and this contradicts the arguments concerning KGs and scalability.

What is meant for noisy data should be defined in a more formal way, as well as how noisy data/facts are generated. Some intuitions and information at this regards are briefly reported only at the end of section 4. A more formal definition (and discussion) should be provided before this section.

It would be preferable to report an algorithm for the rule generation (plus its description). Also the rationale for the assigned probability during the rule generation process should be given. The provided description for the rule generation refers only to rules belonging to category (1). Nothing is said about rules pertaining to the to other (more complex) categories. Similar considerations apply to the fact generation process.

As for section 5, metrics for the evaluation of learned rules taking into account the OWA appear to be missing. At this regards, [Galarraga et al., 2015] reports a rather extensive discussion and a related discussion is provided in "Duc Minh Tran, Claudia d'Amato, Binh Thanh Nguyen, Andrea G. B. Tettamanzi. Comparing Rule Evaluation Metrics for the Evolutionary Discovery of Multi-relational Association Rules in the Semantic Web. EuroGP 2018: 289-305." On the contrary this aspect appears to be almost missing in this paper. A detailed analysis of the impact of the OWA in the evaluation of rules returned by rule learning systems working in OWA should be reported.

Details concerning the experiments presented in section 6.1 are missing, particularly regarding the cardinality of the adopted datasets, the amount of noise and missing consequences. Considering the information provided in 4, these datasets should be rather small. As such, considerations concerning the scalable performances of AMIE, that is well know at the state of the art for its scalability on rather large RDF collections such as DBPedia and YAGO, appear not very well positioned.

A large part of the motivations for this work is given by the necessity of checking the quality of the learned rules in different scenario and/or under different conditions such as heterogeneous and incomplete knowledge, learning specific types of dependencies. However, these aspects are only partially touched with some brief considerations reported at the end of section 6.2 and it does not seem to be enough for corroborating the paper assertions and motivations.

Review #250839 New

Comments to Authors.

The paper presents a system for generating datasets for evaluating rule learners. The authors argue that many existing datasets for such evaluations are “toy examples” and lack of rules neural rule learning approaches. Thus, a method is proposed to first generate rules (as part of the datasets) based on three types of rule dependency graphs, and then populate the facts based on the generated rules to form datasets of various ranges of sizes. Finally, the authors also introduce a set of quality measures to evaluated learned rules and present experiment results on several existing rule learners.

Rule learning is hot research area, and several approaches as well as scalable systems have been developed. Yet a well accepted set of benchmarking datasets and a set of widely agreed performance measures are still missing, thus a work on this direction is very relevant to relational learning and knowledge representation communities. A major concern of mine, however, is the usefulness of the generated datasets and the evaluation tools.

First, the authors claim existing datasets are “toy examples” due to their limited complexity and sizes, which is not necessarily true. Indeed, the fact that existing rule learners often learn rules of restricted forms does not necessarily mean complex dependencies do not exist in existing datasets, as such restrictions are adopted by rule learners to enhance their scalability. On the other hand, it is unclear what sort of complexity is desired and why. In particular, the proposed approach uses three types of rule dependencies to generated rules, but it is left unexplained why these dependencies are beneficial and wether they are missing in the rules learned by existing systems. Also, little is mentioned about the shapes of the rules generated, regarding the predicate arities, rule lengths or variable permutations, all of which are important aspect to measure rule complexity.

Furthermore, existing datasets extracted from popular knowledge graphs are indeed of huge sizes. For instance, the smallest dataset used in [Galarraga et al., 2015], YAGO2, has over 900K facts. Compared to this, the largest datasets generated by the proposed system, the XL datasets, have only up to 500K facts. Indeed, looking at the sizes of the datasets (in Table 1) used for the experiments, the largest one has only fewer than 500 facts and only 3 rules. This makes the evaluation much less convincing.

More importantly, although the authors claim the generated datasets "may model different practical scenarios", it is unclear how this is (or can be) achieved in the proposed approach. Indeed, the generated rules have artificial and somewhat random structures which might not reflect patterns or data dependencies in real-life scenarios. Also, I doubt how "instantiating the rule graphs multiple times" would generate datasets that represent practical ones. Hence, it is doubtful whether conclusions from evaluating on such generated datasets can be generalised to practical scenarios.

Finally, the system uses a set of evaluation measures that are different from existing ones, but their relation with the existing ones is unclear. What makes them more suitable for rule quality evaluation? While the experiments compare several systems on the generated datasets (from Table 1) against the proposed evaluation measures, the meaning of such a comparison is questionable.

Detailed comments:

- p1, "still maintaining their processing exhaustive", what does it mean?

- p2, the citation [ilp,] is broken.

- p2, the purpose of discussion on simple rules for WordNet is unclear. Does it suggest one cannot or should not learn complex rules over WordNet?

- p3, "Category Chain" —> "Category: Chain", same for the other two categories. Also, is not Chain a special case of Rooted DG? Or do you require a minimum branching factor for the latter?

- p3, the use of "open world degree" is confusing. I guess you meant "incompleteness degree"; also, what is the depth of a graph in this case?

- p4, it mentions rule generation requires a “target predicate”. The use of such target predicates is unexplained.

Review #253430 New

Comments to Authors.

**** Summary ****
The authors argue that there is a need for more datasets to evaluate rule learning systems. They introduce the RuDaS tool, that allows to generate synthetic datasets of different sizes, complexity (in term of the shape of rules) and quality (in terms of completeness of the ground facts and of noisiness). They describe the dataset generation process implemented in RuDaS and introduce several measures (one of them new) to assess the strengths and weaknesses of the rules learned by the evaluated systems. They conduct two experiments on four rule-learning systems and compare them on different RuDaS-generated datasets for two of the described measures.

**** Overall Evaluation ****
This paper offers a solution to the lack of benchmarks in KG-oriented rule learning systems. Although the paper itself has flaws, and some of them can't be corrected just by improving the paper (weak experiments) I believe the tool offered is of interest to the AI research community so I would be fine with it being accepted.

**** Rational for the scores ****

***** Relevance *****
The paper is within the scope of the conference. The existence of a tool able to generate synthetic datasets for rule-learning systems is useful not only to the ILP research domain but also to the KB research domain so it should be of interest to a substantial number of AI researchers.

***** Significance *****
Although this paper is not by itself a significant advance in the state of the art, it has the potential to ease the work of many others in rule-based learning so in that sense it could have a lasting impact.

***** Originality *****
This paper proposes an original answer to the lack of standard benchmarks for rule-learning systems.

The new evaluation metric proposed, Herbrand accuracy (ha), is interesting because it measures not only the correctness of the obtained rules but also their completeness, where the standard confidence (sc) only measures the correctness of the rules. This is also the case with the Herbrand distance (hd) but the value returned by ha is normalised, which makes it more convenient to use.

***** Technical Quality *****

The technical specifications of the hardware on which the experiments were run are missing and should be added in Sect. 6 of the paper. This is important to ensure that the experiments are reproducible.

I was not able to replicate the experiments despite the authors sending me an archive with their code, data and scripts (see Details section). Since this is apparently due to a technical trivia on my system, I am willing to try again if the authors have a quick fix to the problem I encountered or if they provide a new archive (see Details section of this review).

However, I was able to generate new KGs from their generator (despite a lack of documentation).

There is no discussion about the limits of RuDaS. It seems to me that this tool is designed specifically to address the need for benchmarks in rule-learning over KG. Whether this impression is correct or not, I believe this point should be discussed in the paper because other applications of rule-learning systems may need different kinds of benchmarks (see my comment on Scholarship).

At the end of sect. 6.1, it is written that Neural-LP is robust to noise and incomplete data but the XS-1 incomplete+noise benchmark doesn't seem to confirm this.

This paper has not convinced me that the categories of rules described (CHAIN, RDG, DRDG) are really relevant. The discrepancies of results between the *-S-2 and *-S-3 results in the experiment of section 6.2 hint that the results observed may be due to intrinsic characteristics of the generated KG that are unrelated to their categories. An analysis over a statistically significant number of KG of each type and size would be more convincing.

From the code, I saw that all generated KG should contain only binary predicates. I couldn't find this information in the paper. Generally speaking, the parameters used for generating the KG used in the experiments should be written somewhere in the paper.

I also have some questions regarding the technical quality of the paper that I wrote in the "Details" section of this review.

***** Clarity and quality of writing *****

The paper is globally well-written. I found a few typos and some sentences should be rewritten in a clearer way (see details) but there is nothing critical there in my opinion.

I am not convinced that the S, M, L... labels are suitable names for denoting the benchmark categories. This puts a defacto standard size for datasets (M) that is much smaller than what existing KB such as Yago or DBpedia contain. Also, imagining that this tool becomes a standard for benchmarking Rule learning systems and that rule systems become more and more efficient, the categories would become less and less readable (XXXXXL vs XXXXXXL is really annoying to read). Why not simply number them following the order of magnitude of the KB: XS -> 1, S -> 2,... XL -> 5?

***** Scholarship *****
I am surprised that all the state-of-the-art paragraphs center on tools for rule extraction instead of describing existing approaches to obtain KG and rule learning benchmarks, such as KG extraction (https://link.springer.com/chapter/10.1007/978-3-642-41335-3_34), or the ILP competition (2016, http://ilp16.doc.ic.ac.uk/competition).

**** Details ****

***** Mistake in the standard confidence formula *****
sc(R,R',F) is not equal to 1 - (hd(R,R',F)/|I(R',F)| as written in section 5. As defined by the authors, hd (the Herbrand distance) is the number of facts that differ between the two minimal Herbrand models of the two programs.
Thus hd can be formally defined as hd(R,R',F) = |I(R,F)-I(R',F)| + |I(R',F)-I(R,F)|.
By definition, sc = |I(R,F)\intersect I(R',F)|/|I(R',F)| = |I(R',F) - (I(R',F)-I(R,F))|/|I(R',F)| = |I(R',F)|/|I(R',F)| - |I(R',F)-I(R,F)|/|I(R',F)|.
Hence one part of hd is missing, namely the terms that are generated by R but not by R', for the equality to hold.

***** Attempt to reproduce the experiments on a linux system (Debian 9) *****
minor details:
- I needed to use chmod u+x on setup/setup_systems.sh and setup/neurallp.sh
- In RuDaS-code/experiments, I had to create the systems/amiep directory by hand or setup/setup_systems.sh wouldn't work
- The setup_experiments.py file is not located in RuDaS-code/experiments as written is the README but in RuDaS-code/experiments/code
main problem:
- The call

python RuDaS-code/experiments/code/setup_experiments.py

crashes with:

```
Traceback (most recent call last):
  File "./experiments/code/setup_experiments.py", line 16, in <module>
    import futils
  File "<my-path>/RuDaS-code/experiments/code/futils.py", line 16, in <module>
    import src.logic as log
ImportError: No module named src.logic
```

When moving setup_experiments.py to the experiments folder, only the first error remains. This error occurs both using python2 and python3

***** Questions *****
 - Instead of ha, another measure that could have been adopted to complement sc is the fraction of correctly inferred facts w.r.t. all facts that

can be inferred using the correct rules (i.e. replacing the |I(R',F)| in the denomination of sc by |I(R,F)|. Is there a reason to prefer ha to the measure I just described?
- How were the numbers (30%, 20%) given in the last paragraph of sect. 3 chosen?
- In Sect. 4, why is it only possible to control the maximal sizes of parameters and not their minimal size (it appears that in the code, it is at least also possible to control the minimal arity of the predicates) or to have finer control on them? Why constrain the missing and noisy facts to the same ratio n_Noise?
- Also in Sect. 4, I find the last sentence of the "Rule generation" paragraph really confusing. Apparently the probabilities described originate from the constraints previously set (at least this is what I infer from the fractions in parentheses). If that is really the case, it would make more sense to give these constraints directly (and to motivate them). If not, these probabilities should still be motivated or at least the fractions should be explained.
- How does your fact generation method scale?
- Although this is not the heart of the paper, I am surprised by the choice of the ILP tool used to run the experiments (FOIL), which is rather old. Given that there are a few other more recent ILP tools like Aleph or Metagol that are also available online and could have been more appropriate for an up-to-date comparison, why choose FOIL?

***** Typos
Abstract:
- "[...] in many domains, they represent [background knowledge...]" -> "[...] in many domains. They represent [...]"

Introduction
- (p1, 1st column) "... with the application of deep learning." Do you mean "apparition"?
- (p1, 2d column) premature -> immature

Sect. 2.1
- (p2, 2d column) [ilp,-] citation would be better as a footnote, given that there is apparently no related publication.

Sect. 3
- (Fig. 1b) an arrow from the p5 headed rule to the p0 headed rule appears to be missing if I understand this representation correctly.
- (Fig. 1c) the p6 predicate at the root should be a p4 (or the p4 in the two other nodes should be p6)
- (Rooted DG description) name the rule in "Furthermore, for each rule *R* there may be several..." and use that name in "... in the body of the former rule..." (-> "... in the body of the rule *R*...")
- (Disjuctive RDG description) "... but each child nodes contains different a predicate ..." -> "... but each child *node* contains *a* different predicate ..."

Sect. 4
- choose one of R-DG and RDG and one of DR-DG and DRDG and be consistent throughout the paper (tables and figures included).

Sect. 5
- add a comma after "The Herbrand distance hd between two logic programs"
- "[...all the facts in F] are considered corrected predictions..." -> "... are considered correct predictions..." ?

Sect. 6.2
- "..., since the good performance,..." Do you mean "thanks to the good performance" ?
- (forelast paragraph) a dot is missing between "...these datasets satisfy this condition" and "The latter condition..."
- (last paragraph) "different rules types" -> "different rule types"

Sect. 7
- (first paragraph) "several rules types" -> "sereval rule types"
- (same) "take in account" -> "take into account"

Review #255326 New

Comments to Authors.

The paper contributes:

1. a new tool (RuDaS) for generating synthetic datasets for evaluating rule learning systems.

2. an empirical evaluation of four rule learning systems on a synthesised dataset

I think that the paper does not make a sufficient contribution to warrant acceptance to IJCAI. Regarding the first contribution, although the described tool may be useful to rule learning researchers, I do not see how it will interest a broad AI audience. Even in the general area of rule learning, the tool will only appeal to a small number of researchers.

Because the first contribution is minor, the second contribution, the evaluation of existing rule learning systems, must therefore be significant to warrant acceptance. However, I think that the 'experiments' need to be substantially improved.

Foremost, there is no clear motivation for the experiments, i.e. there is no clear experimental hypothesis or question. What is the point of the evaluation?

Are you trying to compare state-of-the-art rule learning systems, i.e. do you claim that one system is better than the other on the synthetic datasets? If so, then there are several problems with this claim. For instance, the choice of rule learning systems is inappropriate, e.g. FOIL is not a state-of-the-art ILP system. Why compare systems from 2018 with one from 1990? Likewise, the experimental methodology is not clearly explained. Did you provide all the systems with the same computational restrictions, i.e. same CPU, memory, etc? Without such details it is difficult to accept the results.

Or is the point of the experiments to show show that the RuDaS generates difficult synthetic datasets? Or that it generates realistic datasets? If so, what is the justification for these claims?

I suggest that the authors make the experimental hypothesis or question clear so that the appropriate contribution can be clearly identified.

Moreover, I think the authors can improve experiments, especially the description and presentation of the results. For instance, in footnote 6 the authors state that missing values are due to the computational time limit for the evaluation, but the authors do not state what this limit is. I assume the limit is the same for all systems, but the authors should make this clear. In general, there is no clear experimental methodology.

Regarding the results, Tables 2 and 3 are difficult to interpret. Do you need to present the results to 4 decimal places? Would two decimal places be sufficient and more readable? Moreover, what is the statistical significance of the results? An advantage of using synthesised data is that you can repeat the experiments, which should allow you to calculate the statistical significance of the results. Overall it is hard to appreciate the significance of the results as they are currently presented.

Overall, the contributions of the paper are minor and are unlikely to make a significant contribution to IJCAI. The paper is more suitable for a workshop aimed at a more precise rule learning audience.