Rule Learning over Knowledge Graphs with Genetic Logic Programming (Extended Abstract)

Lianlong Wu[1] , **Emanuel Sallinger**[1,2] , **Evgeny Sherkhonov**[1] , **Sahar Vahdati**[1] , **Georg Gottlob**[1,2]

[1]Department of Computer Science, University of Oxford
[2]Institute of Logic and Computation, TU Wien
{first.last}@cs.ox.ac.uk

Abstract

Declarative rules such as Prolog and Datalog rules are common formalisms to express expert knowledge and facts. They play an important role in Knowledge Graph (KG) construction and completion. Such rules not only encode the expert background knowledge and the relational patterns among the data, but also infer new knowledge and insights from them. Formalizing rules is often a laborious manual process, while learning them from data automatically can ease this process. Within the rule hypothesis space, current approaches resort to exhaustive search with a number of heuristics and syntactic restrictions on the rule language, which impacts the efficiency and quality of the outcome rules. In this paper, we extend the rule hypothesis space from usual path rules to general Datalog rule space by proposing a novel Genetic Logic Programming algorithm named Evoda. It is an iterative process to learn high-quality rules over large scale KG for a matter of seconds. We have performed experiments over multiple real-world KGs and various evaluation metrics to show its mining capabilities for higher quality rules and more precise predictions. We have applied it on the KG completion tasks to illustrate its competitiveness with several state-of-the-art embedding or neural-based models. The experiments demonstrate the feasibility, effectiveness and efficiency.

1 Introduction

While the use of rules is highly effective, the traditional process of acquiring them (often addressed as knowledge engineering) requires scarce expert knowledge and significant human effort. Nowadays, often manual creation of rules together with (semi-)automated processes are complementary, allowing engineered background knowledge to be used at the same time as rules learned from data. The integration of both manual and automated approaches is aimed toward jointly obtaining a higher-quality result compared with each of these processes used in isolation. The automatic process of acquiring rules is the main focus of this paper and employs learning techniques which also highly depend on the formal systems in which the rules are being defined.

Evolutionary Algorithms (EA) impose few restrictions on the rule language, and have the capability of efficient searching in complex, unrestricted rule language spaces like the rule-language Datalog. Specifically, Genetic Programming is based on a number of operators mirroring their biological counterparts.

In this paper we combine approaches from Evolutionary Algorithms (EA) and Knowledge Representation and Reasoning(KRR). The introduced algorithm, called Evoda[1], produces multiple near-optimal solutions efficiently, and, moreover, has fewer syntactic restrictions on the form of learned rules, where no rule templates are required and that produces general Datalog rules.

In our experiments, we compare Evoda with the top search-based rule learning methods namely AMIE (Galárraga et al. 2013) and RuDiK (Ortona, Meduri, and Papotti 2018), and show that Evoda can learn higher quality rules with comparable running time. As additional experiments, we apply Evoda to popular KG completion or link prediction task. In KG completion, given an entity and a relation, the task is to find the target entity. We show that Evoda outperforms existing state-of-the-art search based AMIE, RuDiK and AnyBURL, and embedding based methods such as ConvE (Dettmers et al. 2018) and ComplEx-N3 (Lacroix, Usunier, and Obozinski 2018).

The main contributions of this paper are:

- We first extend the rule learning hypothesis space to the more expressive Datalog language space instead of the path rules, with specific designed genetic logic programming algorithm Evoda.
- We show comprehensive experimental results on latest large scale Knowledge Graphs under Open World Assumption, evaluated with multiple metrics to show the effectiveness and efficiency.
- We apply Evoda to the KG completion tasks, and show significant improved precision compared to other state-of-the-art rule-based or embedding-based methods.

2 Preliminaries

In this paper we study the following problem. Given a Knowledge Graph $\mathcal{K}$ and a target predicate A, find a rule r such that A is in the head. For example, there is a small KG and a rule:

$$\mathcal{K}_1 = \{livesIn(\texttt{John}, \texttt{Oxford}), marriedTo(\texttt{John}, \texttt{Mary})\}$$

$$r_1 : livesIn(Y, Z) \leftarrow livesIn(X, Z), marriedTo(X, Y)$$

[1]Evoda, a portmanteau of "Evolutionary" and "Vadalog", a rule language extension of Datalog.

Such learned rules not only confirm existing knowledge in KGs, but also predict *new facts* that can be added to KGs. For example, rule r_1 predicts a new fact $livesIn(Marry, Oxford)$ over $\mathcal{K}_1$. Each rule has an intermediate representation in the form of an ordered tree.

We use internal evaluation includes support, standard confidence and Partial Completeness Assumption (PCA) confidence. And for external evaluation, we use the same manual evaluation method as AMIE and RuDiK.

3 Algorithm

The outstanding strategy in Evoda is the use of Genetic Logic Programming that efficiently searches huge rule space for discovering near-optimal rules. In particular, one rule or a set of rules can be treated as a chromosome, from which a new population of chromosomes are derived via mutation, crossover and selection operators. While performing these operations, a quality measure, called fitness function is computed to judge whether an obtained new generation of chromosomes is fit for continuing the search. PCA confidence is used as the fitness function, but other appropriate fitness functions can be used such as Closed World Assumption (CWA) confidence, or number of predictions.

Rule Transformation Operators that perform transformation of a given set of rules includes selection, mutation and crossover. In order to complete the link prediction evaluation, we also introduce Rule Covering Algorithm and Rule Aggregation Algorithm.

4 Experiments

In this section, we present our results from evaluation of the proposed algorithm on a set of benchmarks. We present a rather wide set of comparisons in different application areas of Evoda, to give an as broad as possible experimental evaluation of our system. It is noted that not *every* system mentioned will be comparable to any other system mentioned, as many of the state-of-the-art systems are designed to solve particular problems we are going to consider, but not be applicable to solve other such problems.

In particular, we are concentrating on executing time and memory footprint characteristics on KG datasets such as the YAGO, the DBpedia and Wikidata families, compared with many state-of-the-art systems such as AMIE, AMIE+, AMIE3, OP, and RuDiK. The rule learning average execution time for a single predicate ranges from 1s for small scale Kinship dataset to 30s for the large YAGO4-Wikidata2020 dataset. In addition, the single rule average *inference* speed is 0.110s for Kinship, 0.017s for YAGO2 Sample, 3.461s for YAGO3 and 1.428s for YAGO4. For memory usage, take YAGO2s, YAGO3 and Wikidata2014 for example, Evoda uses around 1GB and latest AMIE3 uses around 20GB memory.

We are concentrating on the quality of learned rules, and there the natural comparison made is to systems such as AMIE and RuDiK. On the YAGO2s dataset, we run Evoda for every predicate 30 times and pick the top rule from each run for each predicate. For AMIE, we pick top rule for each predicate from all the 247 rules in their published results. The average PCA value for Evoda is 0.78, which is much higher than 0.39 for AMIE. Furthermore, for 94.58% (35 out of 37) of all predicates in YAGO2s, Evoda has found rules of higher or equal PCA confidence than AMIE has. It shows the advantage of the large unrestricted rule search space and unlimited rule length.

In the other precision comparision experiments, Evoda founds 18 rules for YAGO2 and 16 rules for YAGO2s. The precision of AMIE rules are in the range of 30%-40% where all the Evoda rules are with more than 60% precision.

In the quality and coverage comparison, the rule distribution shows more Evoda rules are in the higher PCA range. There are more AMIE rules are making lots predictions, but the PCA confidences are in the low (< 0.1) range.

Complementarily, we are extending our system for Knowledge Graph Completion, and compared with other systems such as AnyBURL, ConvE, ComplEx, etc. Evoda leads all the Hits@1, Hits@10 and MRR metrics in both WN18 and YAGO3-10 datasets. Especially for the YAGO3-10, Evoda's results have increased top performance of 26.5 percentage points from 50.3 to 76.8 in Hits@1, and 22.4 percentage points from 71.2 to 93.6 in Hits@10. The test coverage of YAGO3-10 is 0.626.

5 Conclusion

In this work, we highlighted the problem of rule learning in KGs and provided a formal definition. The ultimate goal is to emphasize the importance of explicit rule learning. A novel evolutionary based algorithm Evoda is proposed for rule mining over large scale KGs. The experiments demonstrate the feasibility, effectiveness and efficiency of this algorithm. This work provides the foundation for the development of a comprehensive framework for rule learning. Our approach governs fewer language restrictions, so that a larger hypothesis space can be explored and provide higher quality rules. In the KG completion tasks, it performs significantly better than the state-of-the-art embedding models. In future work, we aim at addressing constants in rules and rule simplifications. The application area can also be extended to logical question answering, multi-modal KGs, reasoning under uncertainty with the use of rule and confidence values.

References

Dettmers, T.; Minervini, P.; Stenetorp, P.; and Riedel, S. 2018. Convolutional 2D knowledge graph embeddings. In *AAAI 2018*, 1811–1818.

Galárraga, L.; Teflioudi, C.; Hose, K.; and Suchanek, F. M. 2013. AMIE: Association rule mining under incomplete evidence in ontological knowledge bases. *Proceedings of WWW 2013* 413–422.

Lacroix, T.; Usunier, N.; and Obozinski, G. 2018. Canonical tensor decomposition for knowledge base completion. In *ICML 2018*, volume 7, 4475–4486.

Ortona, S.; Meduri, V. V.; and Papotti, P. 2018. RuDiK: Rule discovery in knowledge bases. *Proceedings of the VLDB Endowment* 11(12):1946–1949.