1 Introduction

1.1 Resumen:

The paper introduces a novel graph query framework for relational learning, aiming to address the challenges in existing methods. Relational learning, which takes into account relationships between objects during the learning process, has shown prominence in various domains such as social networks and bioinformatics. The two main approaches to relational learning are the latent feature (connectionist) approach and the graph pattern-based (symbolic) approach. While the connectionist approach has been successful, the symbolic approach faces issues due to computational complexity from relational queries and lack of robust, general frameworks.

The proposed framework seeks to overcome these challenges by providing controlled complexity for graph pattern matching and stepwise pattern expansion using well-defined operations. This would enable automatic extraction of characteristic relational patterns from data in relational machine learning techniques.

However, the study does not focus on performance analysis or comparison with other methods. Instead, it aims to formalize an efficient graph query system that allows atomic operations for expanding queries in a partitioned manner and evaluating cyclic patterns in polynomial time. To the authors' knowledge, no existing approach meets these requirements.

The paper is organized as follows: Section 1 reviews related research; Section 2 introduces the proposed framework with its main definitions, properties, representative examples, and computational complexity analysis; Section 3 discusses the implementation of the framework for relational machine learning; and Section 4 concludes the study and suggests future research directions.

1.2 Evaluación:

Motivation: YES, The motivation is clearly explained in the introduction. The authors highlight the importance of relational learning methods and their applications in various domains, such as social networks, biology, and chemistry (lines 2-6). Improvement: While the motivation is strong, it could be further strengthened by providing more specific examples or statistics that demonstrate the significance of these problems.

Novelty: YES, The novelty of the proposed approach is clearly described in the introduction. The authors explain how their graph query framework addresses the limitations of existing relational query systems and provides a solution to the fundamental problems of computational complexity and lack of robust frameworks (lines 10-25). Improvement: To further emphasize the novelty, the authors could explicitly compare their approach with related work and highlight unique contributions.

Clarity: YES, The section is well-written and easy to understand. The authors use appropriate terminology and avoid ambiguity (lines 7-9, 10-25). Improvement: To improve clarity, the authors could consider defining technical terms like "relational learning" and "graph pattern matching" at their first mention.

Grammar and Style: YES, The section is free of grammatical errors and uses language appropriate for an academic setting (lines 1-25). Improvement: No specific improvements are suggested as the grammar and style are already excellent.

Typos and Errors: YES, The section appears to be free of typos and other errors (lines 1-25). Improvement: No specific corrections are suggested as there are no apparent errors in the text.

2 Related work

2.1 Resumen:

The related work section of this paper discusses the common approach to executing relational queries, which involves developing patterns in an abstract representation of data and searching for their occurrences in actual datasets. This method is part of Graph Pattern Matching, a field that has been extensively researched for over three decades. The authors highlight various distinctions in pattern matching methods, including structural, semantic, exact, inexact, optimal, and approximate. They also mention graph pattern matching based on isomorphisms, graph simulation, and bounded simulation.

The paper then distinguishes between two types of relational learning models: the latent feature approach and the graph-pattern based approach. The authors focus on the latter, which involves automatically extracting relational patterns from data.

Most pattern-based relational learning methods are derived from Inductive Logic Programming (ILP). However, ILP does not inherently offer relational classifiers and requires data relationships to be transformed into logical predicates. The authors discuss TILDE (Top-down Induction of Logical Decision Trees), an algorithm that can learn decision trees from a given set of examples but does not cater to relational learning, thus lacking certain operations for refining relational queries.

The paper also mentions Multi-relational decision tree learning (MRDTL) and Selection Graphs, which enable atomic operations to enhance queries. However, these methods cannot distinguish between query elements that constitute the query result and those that relate to objects that should or should not be linked to the query result.

Finally, the authors discuss Graph-Based Induction of Decision Trees (DT-GBI), a decision tree construction algorithm for learning graph classifiers using graph-based induction. This method generates attributes during the execution of the algorithm.

In conclusion, some pattern-based approaches can learn to classify complete graphs and construct node classifiers, while others, like the authors' proposal, support learning from general subgraphs as base cases. The proposed technique can also execute cyclic queries, allowing for the extraction of cyclic patterns from data during the learning process.

2.2 Evaluación:

Motivation: YES, The section clearly explains the significance and relevance of the study. It emphasizes the importance of graph pattern matching in executing relational queries and highlights the limitations of existing methods that prevent the evaluation of non-existence of elements.

Novelty: YES, The section effectively describes the novelty of the proposed approach by differentiating it from existing work. It explains how the proposal falls within the scope of semantic, exact, and optimal graph pattern matching implemented with an approach similar to simulations, while also addressing the limitations of current methods.

Clarity: YES, The section is well-written, easy to understand, and uses appropriate terminology. It avoids ambiguity by providing clear definitions and examples of various types of graph pattern matching methods.

Grammar and Style: YES, The section is free of grammatical errors and uses language appropriate for an academic setting.

Typos and Errors: YES, The section appears to be free of typos and other errors.

Improvement Suggestions:

Motivation: While the motivation is clear, it could be strengthened by providing more specific examples or data that highlight the problem's importance and its wider impacts.

Novelty: The section effectively emphasizes the novelty of the proposed approach. However, it could further enhance this by explicitly comparing with related work and highlighting unique contributions in a more detailed manner.

Clarity: The section is clear and easy to understand. To improve clarity further, consider defining some technical terms earlier in the text or providing illustrative examples for complex concepts.

Grammar and Style: The language used is appropriate for an academic setting and appears to be free of grammatical errors. No suggestions for improvement are needed.

Typos and Errors: No typos or other errors were found in the provided section.

3 Relational machine learning

3.1 Resumen:

The section on 'Relational machine learning' discusses the use of the presented framework to develop relational classifiers for graph data sets. The process begins with a labeled subset of subgraphs within a graph data set, and then employs a pattern search technique based on information gain to identify typical patterns for each subgraph class.

The 'Information-gain pattern mining' subsection explains that this pattern identification is achieved through top-down decision tree induction, which explores the pattern space using graph queries as test tools in the internal nodes of the trees. The best refinement sets are identified during the tree construction process, resulting in queries that define classes within the graph dataset.

The 'Relational tree learning examples' subsection provides practical instances to demonstrate this process. It starts with a simple social network where the goal is to classify nodes based on extracted patterns. The process results in a relational decision tree that accurately assigns types (User A, User B, or Item) to all nodes by exploiting relational information from the network. This tree also acquires distinctive patterns for each node type at its leaves, which can be used for future classifications.

The section further illustrates this process using a Star Wars character graph, where each character node and corresponding species property are used as a training dataset. The resulting relational decision tree accurately categorizes and explains each character's species in the graph.

3.2 Evaluación:

Motivation: YES, The section clearly explains the study's significance and relevance. It highlights the importance of acquiring relational classifiers on graph data sets and how the proposed framework can leverage this advantage. The problem's importance is justified through the use of real-world examples such as social networks and Star Wars toy graphs.

Novelty: YES, The section clearly describes the novelty or originality of the proposed approach. It differentiates itself from existing work by emphasizing the use of a pattern search technique founded on information gain to obtain typical patterns for each subgraph class.

Clarity: YES, The section is well-written and easy to understand. It uses appropriate terminology and avoids ambiguity. For instance, it clearly explains the concept of 'information gain' and how it can be used to identify characteristic patterns in graph data sets.

Grammar and Style: YES, The section is free of grammatical and stylistic errors. It uses language that is appropriate for an academic setting. For example, it effectively uses passive voice where necessary to maintain a formal tone ("A top-down decision tree induction will be conducted...").

Typos and Errors: YES, The section is free of typos and other errors. There are no apparent mistakes in the text or figures provided.

Improvement: Not Applicable, as all criteria have been fully met according to the evaluation.

4 Conclusions and future work

4.1 Resumen:

The paper introduces a new graph query framework that enables polynomial cyclic assessment of queries and refinements through atomic operations. The framework is consistent, supports subgraphs beyond individual nodes, allows controlled and automated query construction, and can be used for top-down learning techniques. Unlike graph isomorphism-based systems, this framework assesses the existence/non-existence of paths and nodes in a graph, making it capable of evaluating cyclic patterns in polynomial time.

The initial proof-of-concept implementation has shown promising results in relational learning procedures, extracting interesting patterns from relational data. The current query definition is limited to binary graph data sets but can be adapted for hypergraphs once their usage becomes more widespread.

While the paper provides a basic set of refinement operations, it suggests that more complex refinement families could be developed to prevent plateaus in pattern space and achieve faster learning algorithms. The framework's potential applications include characterizing subgraph categories, justifying decisions in sensitive applications, serving as features in other machine learning methods, and enhancing predictive power when used with ensemble methods like Random Forest.

Future work will focus on developing automated methods to generate refinement sets based on specific learning tasks and graph dataset characteristics. The paper concludes that this framework can establish effective techniques for matching graph patterns, learning symbolic relationships, exploring pattern space systematically, expressing queries highly, and keeping computational costs reasonable.

4.2 Evaluación:

Motivation: YES, The section clearly explains the significance of the study and its relevance in terms of relational learning and automatic feature extraction. It emphasizes that the results are significant for both explainable learning and automatic feature extraction tasks.

Novelty: YES, The section describes the novelty of the proposed approach by comparing it to graph isomorphism-based query systems, highlighting its ability to evaluate cyclic patterns in polynomial time.

Clarity: YES, The section is well-written, easy to understand, and uses appropriate terminology. It provides clear explanations of complex concepts such as relational learning processes, atomic operations, and top-down learning techniques.

Grammar and Style: YES, The section is free of grammatical errors and uses language appropriate for an academic setting. It employs precise language and technical terms to describe the framework's capabilities and limitations.

Typos and Errors: YES, The section appears to be free of typos and other errors.

Improvement Suggestions:

Motivation: No specific improvements are needed as the motivation is clearly stated. However, providing more concrete examples or data to illustrate the importance of the problem could further strengthen this section.

Novelty: The novelty is well-established, but it could be further emphasized by highlighting unique contributions in a bullet-point list or table for easy reference.

Clarity: While the section is clear overall, some technical terms (e.g., "atomic operations," "top-down learning techniques") could benefit from brief definitions or explanations to make them more accessible to non-experts.

Grammar and Style: The grammar and style are appropriate for an academic setting. However, using shorter sentences in places could improve readability.

Typos and Errors: No specific corrections are needed as the section appears to be free of typos and errors.

Acknowledgements

Proyecto PID2019-109152G financiado por MCIN/AEI/10.13039/501100011033

DISARM project - Grant n. PDC2021-121197, and the HORUS project - Grant n. PID2021-126359OB-I00 funded by MCIN/AEI/310.13039/501100011033 and by the "European Union NextGenerationEU/PRTR"

References

References

[1] Almagro-Blanco, P., & Sancho-Caparrini, F. (2017). Generalized graph pattern matching. *CoRR*, *abs/1708.03734*. URL: http://arxiv.org/abs/1708.03734. arXiv:1708.03734.

[2] Barceló, P., Libkin, L., & Reutter, J. L. (2011). Querying graph patterns. In *Proceedings of the Thirtieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* PODS '11 (pp. 199–210). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/1989284.1989307. doi:10.1145/1989284.1989307.

[3] Blockeel, H., & Raedt, L. D. (1998). Top-down induction of first-order logical decision trees. *Artificial Intelligence*, *101*, 285–297. URL: http://www.sciencedirect.com/science/article/pii/S0004370298000344. doi:10.1016/S0004-3702(98)00034-4.

[4] Bonifati, A. Fletcher, G.Voigt, H.Yakovets, N.Jagadish, H. V. (2018 Querying Graphs Morgan & Claypool Publishers

[5] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 2787–2795). Curran Associates, Inc. URL: http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf.

[6] Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? workforce implications. *Science*, *358*, 1530–1534.

[7] Camacho, R., Pereira, M., Costa, V. S., Fonseca, N. A., Adriano, C., Simões, C. J., & Brito, R. M. (2011). A relational learning approach to structure-activity relationships in drug design toxicity studies. *Journal of integrative bioinformatics*, *8*, 176–194.

[8] Chang, K.-W., Yih, S. W.-t., Yang, B., & Meek, C. (2014). Typed tensor decomposition of knowledge bases for relation extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. ACL – Association for Computational Linguistics. URL: https://www.microsoft.com/en-us/research/publication/typed-tensor-decomposition-of-knowledge-bases-for-relation-extraction/.

[9] Consens, M. P., & Mendelzon, A. O. (1990). Graphlog: A visual formalism for real life recursion. In *Proceedings of the Ninth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* PODS '90 (pp. 404–416). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/298514.298591. doi:10.1145/298514.298591.

[10] Cook, S. A. (1971). The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing* STOC '71 (pp. 151–158). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/800157.805047. doi:10.1145/800157.805047.

[11] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., & Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '14 (pp. 601–610). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/2623330.2623623. doi:10.1145/2623330.2623623.

[12] Fan, W., Li, J., Ma, S., Tang, N., Wu, Y., & Wu, Y. (2010). Graph pattern matching: From intractable to polynomial time. *Proc. VLDB Endow.*, *3*, 264–275. URL: http://dx.doi.org/10.14778/1920841.1920878. doi:10.14778/1920841.1920878.

[13] Gallagher, B. (2006). Matching structure and semantics: A survey on graph-based pattern matching. *AAAI FS*, *6*, 45–53.

[14] García-Jiménez, B., Pons, T., Sanchis, A., & Valencia, A. (2014). Predicting protein relationships to human pathways through a relational learning approach based on simple sequence features. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *11*, 753–765. doi:10.1109/TCBB.2014.2318730.

[15] Geamsakul, W., Matsuda, T., Yoshida, T., Motoda, H., & Washio, T. (2003). Classifier construction by graph-based induction for graph-structured data. In K.-Y. Whang, J. Jeon, K. Shim, & J. Srivastava (Eds.), *Advances in Knowledge Discovery and Data Mining: 7th Pacific-Asia Conference, PAKDD 2003, Seoul, Korea, April 30 – May 2, 2003 Proceedings* (pp. 52–62). Berlin, Heidelberg: Springer Berlin Heidelberg. URL: http://dx.doi.org/10.1007/3-540-36175-8_6. doi:10.1007/3-540-36175-8_6.

[16] Gupta, S. (2015). *Neo4j Essentials*. Community experience distilled. Packt Publishing. URL: https://books.google.es/books?id=WJ7NBgAAQBAJ.

[17] Luc De Raedt, Sebastijan Dumančić, Robin Manhaeve, Giuseppe Marra. *From Statistical Relational to Neural-Symbolic Artificial Intelligence*. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI'20)*, 2021. ISBN: 9780999241165. Article No.: 688. Pages: 8. Yokohama, Japan.

[18] Henzinger, M. R., Henzinger, T. A., & Kopke, P. W. (1995). Computing simulations on finite and infinite graphs. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on* (pp. 453–462). IEEE.

[19] Jacob, Y., Denoyer, L., & Gallinari, P. (2014). Learning latent representations of nodes for classifying in heterogeneous social networks. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* WSDM '14 (pp. 373–382). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/2556195.2556225. doi:10.1145/2556195.2556225.

[20] Jiang, J. Q. (2011). Learning protein functions from bi-relational graph of proteins and function annotations. In T. M. Przytycka, & M.-F. Sagot (Eds.), *Algorithms in Bioinformatics* (pp. 128–138). Berlin, Heidelberg: Springer Berlin Heidelberg.

[21] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*, 3.

[22] Karp, R. M. (1975). On the computational complexity of combinatorial problems. *Networks*, *5*, 45–68.

[23] Knobbe, A. J., Siebes, A., Wallen, D. V. D., & Syllogic B., V. (1999). Multi-relational decision tree induction. In *In Proceedings of PKDD' 99, Prague, Czech Republic, Septembre* (pp. 378–383). Springer.

[24] Latouche, P., & Rossi, F. (2015). Graphs in machine learning: an introduction.

[25] Leiva, H. A., Gadia, S., & Dobbs, D. (2002). Mrdtl: A multi-relational decision tree learning algorithm. In *Proceedings of the 13th International Conference on Inductive Logic Programming (ILP 2003* (pp. 38–56). Springer-Verlag.

[26] Milner, R. (1989). *Communication and Concurrency*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

[27] Nguyen, P. C., Ohara, K., Motoda, H., & Washio, T. (2005). Cl-gbi: A novel approach for extracting typical patterns from graph-structured data. In T. B. Ho, D. Cheung, & H. Liu (Eds.), *Advances in Knowledge Discovery and Data Mining: 9th Pacific-Asia Conference, PAKDD 2005, Hanoi, Vietnam, May 18-20, 2005. Proceedings* (pp. 639–649). Berlin, Heidelberg: Springer Berlin Heidelberg. URL: `http://dx.doi.org/10.1007/11430919_74`. doi:`10.1007/11430919_74`.

[28] Wenfei Fan. *Graph Pattern Matching Revised for Social Network Analysis*. In *Proceedings of the 15th International Conference on Database Theory (ICDT '12)*, 2012. ISBN: 9781450307918. Publisher: Association for Computing Machinery. Address: New York, NY, USA. Pages: 8–21. Num. Pages: 14. Location: Berlin, Germany. DOI: 10.1145/2274576.2274578.

[29] Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, Juncheng Liu, Bryan Hooi. *NodeAug: Semi-Supervised Node Classification with Data Augmentation*. En *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*, 2020, páginas 207–217. DOI: `https://doi.org/10.1145/3394486.3403063`.

[30] Seyed Mehran Kazemi and David Poole. *RelNN: A Deep Neural Model for Relational Learning*. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, University of British Columbia, Vancouver, Canada, 2018. Email: `smkazemi@cs.ubc.ca`, `poole@cs.ubc.ca`.

[31] Maria Leonor Pacheco and Dan Goldwasser. *Modeling Content and Context with Deep Relational Learning. Transactions of the Association for Computational Linguistics*, 2021; 9, 100–119. DOI: `https://doi.org/10.1162/tacl_a_00357`.

[32] K. Ahmed, A. Altaf, N.S.M. Jamail, F. Iqbal, R. Latif. *ADAL-NN: Anomaly Detection and Localization Using Deep Relational Learning in Distributed Systems. Applied Sciences*, 2023, 13, 7297. DOI: `https://doi.org/10.3390/app13127297`.

[33] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Maosong Sun. *Graph Neural Networks: A Review of Methods and Applications. AI open*, 2020, 1, 57-81.

[34] Lingfei Wu, Peng Cui, Jian Pei, Liang Zhao, Xiaojie Guo. *Graph Neural Networks: Foundation, Frontiers and Applications*. En *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, 2022, páginas 4840–4841. DOI: https://doi.org/10.1145/3534678.3542609.

[35] Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, *104*, 11–33.

[36] Namkyeong L., Dongmin H., Gyoung S. Na, Sungwon K., Junseok L. and Chanyoung P. (2023). Conditional Graph Information Bottleneck for Molecular Relational Learning.

[37] Plotkin, G. (1972). Automatic methods of inductive inference .

[38] van Rest, O., Hong, S., Kim, J., Meng, X., & Chafi, H. (2016). Pgql: a property graph query language. In *Proceedings of the Fourth International Workshop on Graph Data Management Experiences and Systems* (p. 7). ACM.

[39] Reutter, J. L. (2013). *Graph Patterns: Structure, Query Answering and Applications in Schema Mappings and Formal Language Theory*. Ph.D. thesis Laboratory for Foundations of Computer Science School of Informatics University of Edinburgh.

[40] Segaran, T., Evans, C., Taylor, J., Toby, S., Colin, E., & Jamie, T. (2009). *Programming the Semantic Web*. (1st ed.). O'Reilly Media, Inc.

[41] Tang, L., & Liu, H. (2009). Relational learning via latent social dimensions. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 817–826). ACM.

[42] Zou, L., Chen, L., & Özsu, M. T. (2009). Distance-join: Pattern match query in a large graph database. *Proc. VLDB Endow.*, *2*, 886–897. URL: http://dx.doi.org/10.14778/1687627.1687727. doi:10.14778/1687627.1687727.

[43] Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, Tianyu Wo. *Strong Simulation: Capturing Topology in Graph Pattern Matching. ACM Trans. Database Syst.*, January 2014, Volume 39, Number 1, Article No. 4. ISSN: 0362-5915. Publisher: Association for Computing Machinery. Address: New York, NY, USA. DOI: 10.1145/2528937. Keywords: dual simulation, graph simulation, data locality, Strong simulation, subgraph isomorphism.