Empowering Industry Professionals with Machine Learning through Knowledge Graphs

Antonis Klironomos[1,2], Gad-Elrab Mohamed[1] and Evgeny Kharlamov[1,3]

[1]Bosch Center for Artificial Intelligence, Germany
[2]University of Mannheim, Germany
[3]University of Oslo, Norway

Abstract

The application of machine learning (ML) has become increasingly prevalent in various industries, offering valuable insights and predictive capabilities. However, the adoption of ML by domain experts, who possess deep industry-specific knowledge but may lack technical expertise, presents unique challenges. This paper explores strategies for scaling out the usage of ML to industry professionals, enabling them to leverage the power of ML in their respective domains. We discuss a comprehensive user-friendly ML system with an interface for democratizing ML within industry domains. The system includes automatic feature engineering through ontologies, and simplifying ML pipeline creation using knowledge graphs (KGs). We also present real-world use cases supported by user study results.

Keywords

machine learning, knowledge graphs, industry

1. Introduction

The adoption of machine learning (ML) in various industries has revolutionized the way businesses operate, offering valuable insights and predictive capabilities that were previously unattainable. However, the prerequisites and complexity of developing ML pipelines pose a barrier for domain experts who want to use ML.

To overcome this barrier, the paper shows a comprehensive user-friendly system with a graphical user interface (GUI) designed to cater to industry professionals. This system includes automatic feature engineering through ontologies, which allows domain experts to leverage their industry-specific knowledge to create and extract relevant features for ML models. Additionally, the paper addresses the system's utilization of knowledge graphs (KGs) to simplify the creation of ML pipelines, making it easier for domain experts to build ML models without requiring extensive technical expertise.

To show the system's impact, the paper presents two real-world Bosch use cases that demonstrate the successful integration of ML into industry domains, supported by user study results. By showcasing the practical applications of ML in the industry, the paper aims to highlight the potential benefits and opportunities that ML can offer to domain experts. Ultimately, the goal is to democratize ML within industry domains, empowering domain experts to leverage ML for improved decision-making, enhanced productivity, and innovative solutions to industry-specific challenges.

2. Automated Feature Engineering using Ontologies

A crucial step in applying ML involves feature engineering, which typically necessitates domain knowledge repeatedly used for similar data. Integrating domain knowledge into ontologies can reduce repetition and help perform automatic ML algorithm selection. This section describes the phases and benefits of our Semantically-Enhanced Feature Engineering (SemFE) tool [1].

First International Workshop on Scaling Knowledge Graphs for Industry, co-located with 20th International Conference on Semantic Systems (SEMANTICS) - Amsterdam, Sept. 17–19, 2024

antonis.klironomos@de.bosch.com (A. Klironomos); mohamed.gad-elrab@de.bosch.com (G. Mohamed); evgeny.kharlamov@de.bosch.com (E. Kharlamov)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-3759/workshop5.pdf

Semantic Data Preparation. The process of semantically enhancing feature engineering involves a data preparation phase. This begins with integrating raw data sources, where the Domain Knowledge Annotator maps raw feature names to the terms of domain ontologies (DO) in a semi-automated fashion. The resulting Data-to-DO mapping serves as the foundation for subsequent automated processes. The Feature Group (FG) Annotator then utilizes reasoning to infer ML feature groups for each source from the Data-to-DO mapping, generating the DO-to-FG mapping. This automated step allows subsequent modules to abstract from the concrete features and generically work with feature groups.

Semantic Feature Processing. Following the data preparation phase, the process moves into semantic feature processing, which involves retrieving feature processing algorithms (FPAlg) for each feature group from the ML Ontology. The Feature Processing Algorithm Selector retrieves these algorithms, each with varying computational complexity, depending on the feature groups. The default algorithms are specified, and for specific feature groups, users can manually modify the default algorithms. The Processed Feature Groups Annotator then uses the FPAlg-to-FPG mapping to infer the feature processed groups (FPG) for the specified feature groups and chosen feature processing algorithms. This automated process generates names for new features and applies the feature processing algorithms to compute these new features.

ML Modeling and Implementation. The final step involves ML modeling, where the ML Algorithm Selector module selects ML algorithms with different feature settings, ML methods, and hyper-parameters based on the FPAlg-to-FPG mapping. After the ML model training and testing, the formal representation of domain knowledge, ML feature processing strategies, and the algorithms and their application order enable the execution of these workflows with minimal adjustments and adaptations to new datasets. SemFE has been implemented as an extension of the ML pipeline, incorporating several semantic modules and communicating with a triple store and reasoner to store ontologies and retrieve inference results.

3. Convenient ML Pipeline Creation via Knowledge Graphs

The aforementioned automated feature engineering tool (i.e. SemFE) is complemented by our Semantically-Enhanced Machine Learning (SemML) tool [2] used for conveniently creating ML pipelines. The structured knowledge representation provided by ML-related ontologies facilitates the efficient construction of executable KGs that represent ML pipelines [3, 4]. This section describes the relevant ontologies and the process of creating machine learning pipelines from the user's perspective.

Semantic Artifacts for Executable Knowledge Graphs. The tool includes various ontologies, such as the upper domain ontology, manufacturing ontology, and domain ontologies for specific manufacturing domains at Bosch. The upper domain ontology contains axioms, classes, object properties, and datatype properties to model the general knowledge of discrete manufacturing processes. Domain experts create the domain ontologies, which consist of sub-classes of the upper domain ontology's classes. The tool also includes the data science ontology, which formalizes the general knowledge of data science activities, and task ontologies for visualization, statistical analytics, and machine learning analytics. These task ontologies describe common methods, allowed data structures, and the organization of tasks in pipelines.

Executable Knowledge Graph Construction. The construction of pipelines as executable KGs in the tool can be done via GUI. All user's actions reflect changes in KGs and the task options are based on the KG structure. Users can create pipelines from scratch, and modify, or integrate existing ones. The creation of pipelines involves selecting input data and tasks (i.e. steps) based on the respective ontologies. The modification of pipelines can be done by adding, deleting, or changing tasks, while the integration of KGs involves combining the outputs of different pipelines. The translation of executable KGs into code is done with Python, which is used as the language for discussion.

4. Use cases

We showcase our system's benefits through two real-world use cases. We demonstrate SemFE's capabilities with a welding use case including two experiments and test SemML on real welding data.

Use Case 1: Professionals extend SemFE's domain ontologies. Users were tasked with using SemFE for domain ontology extension (Experiment 1) and data mapping between column names and ontology terms (Experiment 2) [5]. These tasks are important for domain experts to establish a common vocabulary. In Experiment 1, domain experts and data scientists created terms for resistance spot welding (RSW) and hot-staking (HS), with average correctness for applying a template at 93% and making choices of dependencies at 92%. The correlations between user performance and self-reported expertise indicated that domain expertise greatly increased the correctness and efficiency. In Experiment 2, most users correctly mapped column names to newly introduced terms, achieving 100% correctness, with average time spent for each term at about 50 seconds. Similar to Experiment 1, there was the same conclusion for correlations between user performance and self-reported expertise, while experience with mapping tools showed minimal effect on correctness and efficiency.

Use Case 2: Welding experts develop ML pipelines using SemML. SemML was deployed on welding data from Bosch to predict the quality of resistance spot welding [4]. A user study involving 28 experts from various fields, including ML, welding, and sensor engineering, was conducted to evaluate the tool's effectiveness. The study included a series of tasks for visualization, statistics, and ML. The users were asked to complete the tasks with and without using SemML. ML experts explained the tasks to non-ML experts, who then completed the tasks by using technical language or creating, modifying, and merging knowledge graphs through a GUI. The study measured the percentage of tasks completed, completion time, and correctness of answers, and compared the actions taken during the tasks with ground truth to measure correctness. The results showed that most participants had a high completion percentage and correctness when using our tool and needed less time to complete tasks. The tool also improved transparency, usability, and the coverage of tasks, making previously impossible tasks achievable for non-ML experts.

5. Conclusion

This paper describes SemFE and SemML, two tools that complement each other to form a semantic-based ML system. The automation of feature engineering by SemFE is followed by SemML's facilitation of ML pipeline creation. In this system, we harness the power of ontologies and KGs to allow experts in various domains to use ML in their work, without needing ML expertise. We prove the system's benefits, such as user-friendliness and efficiency, by presenting two real-world use cases. The system is capable of enabling non-ML professionals to define a common vocabulary through ontologies and conveniently create ML pipelines through knowledge graphs.

Acknowledgments

The Graph Massiviser (GA 101093202) EU project as well as Dome 4.0 (GA 953163) and enRichMyData (GA 101093202) partially supported this work.

References

[1] B. Zhou, Y. Svetashova, T. Pychynski, I. Baimuratov, A. Soylu, E. Kharlamov, SemFE: Facilitating ML Pipeline Development with Semantics, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, ACM, Virtual Event Ireland, 2020, pp. 3489–3492. doi:10.1145/3340531.3417436.

[2] B. Zhou, Y. Svetashova, A. Gusmao, A. Soylu, G. Cheng, R. Mikut, A. Waaler, E. Kharlamov, SemML: Facilitating development of ML models for condition monitoring with semantics, Journal of Web Semantics 71 (2021) 100664. doi:10.1016/j.websem.2021.100664.

[3] A. Klironomos, B. Zhou, Z. Tan, Z. Zheng, G.-E. Mohamed, H. Paulheim, E. Kharlamov, ExeKGLib: Knowledge Graphs-Empowered Machine Learning Analytics, in: C. Pesquita, H. Skaf-Molli, V. Efthymiou, S. Kirrane, A. Ngonga, D. Collarana, R. Cerqueira, M. Alam, C. Trojahn, S. Hertling (Eds.), The Semantic Web: ESWC 2023 Satellite Events, volume 13998, Springer Nature Switzerland, Cham, 2023, pp. 123–127. doi:10.1007/978-3-031-43458-7_23.

[4] Z. Zheng, B. Zhou, D. Zhou, X. Zheng, G. Cheng, A. Soylu, E. Kharlamov, Executable Knowledge Graphs for Machine Learning: A Bosch Case of Welding Monitoring, in: U. Sattler, A. Hogan, M. Keet, V. Presutti, J. P. A. Almeida, H. Takeda, P. Monnin, G. Pirrò, C. d'Amato (Eds.), The Semantic Web – ISWC 2022, volume 13489, Springer International Publishing, Cham, 2022, pp. 791–809. doi:10.1007/978-3-031-19433-7_45.

[5] Y. Svetashova, B. Zhou, T. Pychynski, S. Schmid, Y. Sure-Vetter, R. Mikut, E. Kharlamov, Ontology-Enhanced Machine Learning: A Bosch Use Case of Welding Quality Monitoring, 2020. doi:10.1007/978-3-030-62466-8_33.