Amalur: Data Integration Meets Machine Learning

Rihan Hai
TU Delft

1 Abstract

The accuracy of an ML model heavily depends on the training data. In real world applications, often the data is not stored in a central database or file system, but spread over different data silos. Take, for instance, drug risk prediction: the features can reside in datasets collected from clinics, hospitals, pharmacies, and laboratories distributed geographically [1]. With data privacy and security constraints, data often cannot leave the premises of data silos, hence model training should proceed in a decentralized manner. Data integration (DI) systems enable interoperability among multiple, heterogeneous sources, and provide a unified view for users. Notably, they allow us to describe data sources and their relationships [3]: i) mappings between different source schemata, i.e., schema matching and mapping [7, 4] and ii) linkages between data instances, i.e., data matching (also known as record linkage or entity resolution) [2]. Yet, a data integration system's goal is to facilitate query answering or data transformation over silos, and not to directly support machine learning applications.

In this talk, I will present the vision of our new research line in Delft, which aims to bridge the traditional data integration techniques with the requirements of modern machine learning [8]. We explore the possibilities of utilizing metadata obtained from data integration processes for improving the effectiveness and efficiency of ML models. During the talk, I will give a brief introduction on traditional DI systems, recent advances of in-database machine learning, and federated learning. I will analyze two common use cases over data silos, feature augmentation and federated learning. Then I will explain our recent works on ML-aware data augmentation [5], model zoo metadata management [6], and cost estimation for factorized learning. Bringing data integration and machine learning together, I will highlight new research opportunities from the aspects of systems, intermediate representations, factorized learning and federated learning.

References

[1] J. M. Bos, G. A. Kalkman, H. Groenewoud, P. M. van den Bemt, P. A. De Smet, J. E. Nagtegaal, A. Wieringa, G. J. van der Wilt, and C. Kramers. Prediction of clinically relevant adverse drug events in surgical patients. *PloS one*, 13(8):e0201645, 2018.

[2] D. G. Brizan and A. U. Tansel. A. survey of entity resolution and record linkage methodologies. *Communications of the IIMA*, 6(3):5, 2006.

[3] A. Doan, A. Halevy, and Z. Ives. *Principles of data integration*. Elsevier, 2012.

[4] R. Fagin, L. M. Haas, M. Hernández, R. J. Miller, L. Popa, and Y. Velegrakis. Clio: Schema mapping creation and data exchange. In *ER*, pages 198–236. Springer, 2009.

[5] A. Ionescu, R. Hai, M. Fragkoulis, and A. Katsifodimos. Join path-based data augmentation for decision trees. In *2022 IEEE 38th International Conference on Data Engineering Workshops (ICDEW)*, pages 84–88. IEEE, 2022.

[6] Z. Li, R. Hai, A. Bozzon, and A. Katsifodimos. Metadata representations for queryable ML model zoos. 2022.

[7] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *the VLDB Journal*, 10(4):334–350, 2001.

[8] A. I. a. A. K. Rihan Hai, Christos Koutras. Amalur: Next-generation data integration in data lakes.