Learning Models over Relational Databases

Dan Olteanu
Department of Computer Science, University of Oxford, Oxford, UK
dan.olteanu@cs.ox.ac.uk

Abstract

In this talk, I will make the case for a first-principles approach to machine learning over relational databases that exploits recent development in database systems and theory.

The input to learning classification and regression models is defined by feature extraction queries over relational databases. The mainstream approach to learning over relational data is to materialize the training dataset, export it out of the database, and then learn over it using statistical software packages. These three steps are expensive and unnecessary. Instead, one can cast the machine learning problem as a database problem by decomposing the learning task into a batch of aggregates over the feature extraction query and by computing this batch over the input database.

The performance of this database-centric approach benefits tremendously from structural properties of the relational data and of the feature extraction query; such properties may be algebraic (semi-ring), combinatorial (hypertree width), or statistical (sampling). It also benefits from database systems techniques such as factorized query evaluation and query compilation. For a variety of models, including factorization machines, decision trees, and support vector machines, this approach may come with lower computational complexity than the materialization of the training dataset used by the mainstream approach. Recent results show that this translates to several orders-of-magnitude speed-up over state-of-the-art systems such as TensorFlow, R, Scikit-learn, and mlpack.

While these initial results are promising, there is much more awaiting to be discovered.

2012 ACM Subject Classification Theory of computation → Database query processing and optimization (theory); Information systems → Database query processing; Computing methodologies → Supervised learning; Computing methodologies → Machine learning approaches

Keywords and phrases In-database analytics, Data complexity, Feature extraction queries, Database dependencies, Model reparameterization

Digital Object Identifier 10.4230/LIPIcs.ICDT.2019.1

Category Invited Talk

Funding This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 682588.

Acknowledgements This work is part of the FDB project (`https://fdbresearch.github.io`) and based on collaboration with Maximilian Schleich (Oxford), Mahmoud Abo-Khamis, Ryan Curtin, Hung Q. Ngo (RelationalAI), Ben Moseley (CMU), and XuanLong Nguyen (Michigan).

22nd International Conference on Database Theory (ICDT 2019).
Editors: Pablo Barcelo and Marco Calautti; Article No. 1; pp. 1:1–1:1
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany