Learning Models over Relational Data Using Sparse Tensors and Functional Dependencies

MAHMOUD ABO KHAMIS and HUNG Q. NGO, RelationalAI, Inc.
XUANLONG NGUYEN, University of Michigan
DAN OLTEANU and MAXIMILIAN SCHLEICH, University of Oxford

Integrated solutions for analytics over relational databases are of great practical importance as they avoid the costly repeated loop data scientists have to deal with on a daily basis: select features from data residing in relational databases using feature extraction queries involving joins, projections, and aggregations; export the training dataset defined by such queries; convert this dataset into the format of an external learning tool; and train the desired model using this tool. These integrated solutions are also a fertile ground of theoretically fundamental and challenging problems at the intersection of relational and statistical data models.

This article introduces a unified framework for training and evaluating a class of statistical learning models over relational databases. This class includes ridge linear regression, polynomial regression, factorization machines, and principal component analysis. We show that, by synergizing key tools from database theory such as schema information, query structure, functional dependencies, recent advances in query evaluation algorithms, and from linear algebra such as tensor and matrix operations, one can formulate relational analytics problems and design efficient (query and data) structure-aware algorithms to solve them.

This theoretical development informed the design and implementation of the AC/DC system for structure-aware learning. We benchmark the performance of AC/DC against R, MADlib, libFM, and TensorFlow. For typical retail forecasting and advertisement planning applications, AC/DC can learn polynomial regression models and factorization machines with at least the same accuracy as its competitors and up to three orders of magnitude faster than its competitors whenever they do not run out of memory, exceed 24-hour timeout, or encounter internal design limitations.

CCS Concepts: • **Information systems → Database management system engines**; • **Theory of computation → Database query processing and optimization (theory)**; • **Computing methodologies → Supervised learning**;

Additional Key Words and Phrases: In-database analytics, functional aggregate queries, functional dependencies, model reparameterization, tensors

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 682588. X.N. is supported in part by Grants No. NSF CAREER DMS-1351362, No. NSF CNS-1409303, and the Margaret and Herman Sokol Faculty Award.

Authors' addresses: M. Abo Khamis and H. Q. Ngo, RelationalAI, Inc., 2120 University Ave, Berkeley, CA 94704; emails: {mahmoud.abokhamis, hung.ngo}@relational.ai; X. Nguyen, University of Michigan, 461 West Hall, 1085 South University, Ann Arbor, MI 48109-1107; email: xuanlong@umich.edu; D. Olteanu and M. Schleich, University of Oxford, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK; emails: {dan.olteanu, max.schleich}@cs.ox.ac.uk.

0362-5915/2020/06-ART7 $15.00
https://doi.org/10.1145/3375661