SQL for SRL: Structure Learning Inside a Database System

Oliver Schulte and Zhensong Qian, Simon Fraser University, Canada

Introduction
- Statistical-Relational Learning: Learn a joint statistical model for all tables in the input database.
- New approach to SRL system building:
  - The RDBMS stores structured objects for statistical analysis as first-class citizens in the database.
  - SQL is used to build and transform statistical objects:
    - Structured Model (Bayesian network, Markov Logic Network).
    - Parameter Estimates.
    - Sufficient Statistics.
- Empirical evaluation: leveraging the RDBMS capabilities achieves scalable learning and fast model testing.
- All code and datasets are available online [1].

Contributions
- Identifying new system requirements for multi-relational machine learning that go beyond single table machine learning.
- An integrated set of SQL-based solutions for providing these system capabilities.

Related Works
- BayesStore [3]: all statistical objects are first-class citizens in a relational database. Inference, no learning.
- Madlib [5]: leverages SQL for single-relational data table analysis.
- Tuffy [7]: reliable and scalable inference and parameter learning for Markov Logic Networks with an RDBMS. No structure learning.

System Overview

The Parameter Manager
- Goal: Learn Bayesian Network Parameters
  - Stored in Conditional Probability (CP) table.
  - Maximum Likelihood Estimate is easy to compute from database counts.

The Random Variable Database
- Meta data about random variables stored in database tables.
  - Domain of possible values.
  - Pointer to corresponding data table/column.

The Random Variable Database
- Meta data about random variables stored in database tables.
  - Domain of possible values.
  - Pointer to corresponding data table/column.

The Random Variable Database
- Meta data about random variables stored in database tables.
  - Domain of possible values.
  - Pointer to corresponding data table/column.

The Model Manager
- Goal: Learn First-Order Bayesian Network [2].
- Bayesian Network Structure Learning [6].
- Nodes = Random Variables
- Edges are stored in Database tables.
- Model selection scores are also stored, not shown (BIC, AIC, BDDeu).

The Model Manager
- Goal: Learn First-Order Bayesian Network [2].
- Bayesian Network Structure Learning [6].
- Nodes = Random Variables
- Edges are stored in Database tables.
- Model selection scores are also stored, not shown (BIC, AIC, BDDeu).

References