**MRLBase: Multi-Relational Learning with Multi-Relational Databases**

**Zhensong Qian**

Several well-developed software packages implement standard machine learning algorithms by assuming a flat data representation with a single data table or matrix (e.g., R, Weka). The single-table representation is appropriate when the data points are independent of those of others. However, many real-word enterprise datasets which are often maintained in a relational database have a more complex structure, with different classes of entities (customers, products, factories etc.), that have different attributes, and often interrelated in multiple ways. The statistical analysis of such multi-relational data requires new system capabilities and novel structured machine learning models. We show how SQL can be leveraged to support multi-relational learning.

We present MRLBase for “Multi-relational Learning Base”, a framework for building the system capabilities required for multi-relational learning that go beyond what is required for single-table learning. Statistical system tasks include accessing data accesses, constructing, storing, querying, and transforming parameter estimates and model structures. MRLBase follows a client-server paradigm, where the client is a machine learning application for multi-relational data, and the server is an RDBMS that supports a machine learning application. The RDBMS is used not only to store data, but also to store structured objects for statistical analysis as first-class citizens in the database. By separating system tasks from statistical issues, MRLBase facilitates extending single-table applications to multi-relational data. Our argument is that relational algebra can play the same role for multi-relational machine learning that linear algebra does for single-table machine learning: a unified language for both representing and computing with objects that support statistical analysis.

The main contributions of the framework may be summarized as follows.

1. Identifying new system requirements for multi-relational machine learning that go beyond traditional single table machine learning.

2. An integrated set of SQL-based solutions for providing these system capabilities, including

(a) Defining a default set of relational random variables, and extracting meta-information about them from the RDBMS system catalog.

(b) Computing contingency tables that store multi-relational sufficient statistics as database tables.

(c) Storing and scoring probabilistic models.

A case study on six benchmark databases shows how our system supports a challenging and important machine learning application, namely learning a Bayesian network model for an entire database. Our implementation shows how our SQL constructs in MRLBase facilitate fast, modular, and reliable program development. Empirical evidence indicates that leveraging the RDBMS capabilities achieves scalable learning and fast model testing.