Learning from Multirelational Data through Multiple Views

A Dissertation Presented
by
HONGYU GUO
B.Eng. Shanghai Jiao Tong University
M.Sc. University of Ottawa

Submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in Department of Computer Science

University of Ottawa, Canada
January 2008
Abstract

LEARNING FROM MULTIRELATIONAL DATA THROUGH MULTIPLE VIEWS

HONGYU GUO

COMPUTER SCIENCE, UNIVERSITY OF OTTAWA

Directed by: Professor Herna L. Viktor

Since their first release in the 1970s, relational databases have been routinely used to collect and organize real-world data—from financial transactions, marketing surveys, to health informatics observations. Traditional data mining methods expect data in the form of a single table, thus resulting in an inability to deal with such relational repositories. Multirelational Data Mining, on the other hand, aims to discover useful patterns across multiple inter-connected relations in a relational database. To this end, this work focuses on how to build classification models for relational databases through multiple views (feature sets).

This study developed four multiple view strategies for mining multirelational data. The thesis firstly introduces the Multi-View Relational Classification (MRC) framework, for constructing hypotheses from sets of attributes of the presented data. The MRC strategy distinguishes itself from existing multirelational mining algorithms by excluding the need to either transform multiple relations into a universal single table or to devise new techniques for direct relational learning. The MRC algorithm offers both predictive performance and efficiency gains over current relational models, when mining diverse relational databases.

Secondly, the MRC-IM method extends the MRC approach in order to deal with skew-class multirelational data. Here, the number of examples from one class is much higher than the others and correctly classifying the underrepresented examples is of prime importance. The MRC-IM method offers performance gains over
current relational model not only against majority class instances, but also against underrepresented examples. While the MRC and MRC-IM methods construct an individual view using features within a sole relation, the third multi-view strategy formulated by this work, namely the MRC-Cross approach, enables the search and collection of relevant attributes across multiple relations when constructing individual views. Finally, we present the SESP technique for pre-pruning uninteresting relations of complex relational databases. Through identifying uninteresting views from the MRC framework, our SESP method creates a pruned structure, while minimizing predictive performance loss on the final classification model.

The results of this study thus suggest that learning from multiple views sets a new direction for efficiently mining data in many relational forms, including relational databases, graphs and social networks.