Dissertations

Leveraging Relational Representations for Causal Discovery

Matthew John Hale Rattigan, *University of Massachusetts - Amherst*

Date of Award
9-2012

Document Type
Open Access Dissertation

Degree Name
Doctor of Philosophy (PhD)

Degree Program
Computer Science

First Advisor
David Jensen

Second Advisor
Andrew Barto

Third Advisor
Andrea Foulkes

Keywords
causality, relational learning, social network analysis

Subject Categories
Computer Sciences

Abstract
This thesis represents a synthesis of relational learning and causal discovery, two subjects at the frontier of machine learning research. Relational learning investigates algorithms for constructing statistical models of data drawn from of multiple types of interrelated entities, and causal discovery investigates algorithms for constructing causal models from observational data. My work demonstrates that there exists a natural, methodological synergy between these two areas of study, and that despite the sometimes onerous nature of each, their combination (perhaps counterintuitively) can provide advances in the state of the art for both.

Traditionally, propositional (or "flat") data representations have dominated the statistical sciences. These representations assume that data consist of independent and identically distributed (iid) entities which can be represented by a single data table. More recently, data scientists have increasingly focused on "relational" data sets that consist of interrelated, heterogeneous entities. However, relational learning and

causal discovery are rarely combined. Relational representations are wholly absent from the literature where causality is discussed explicitly. Instead, the literature on causality that uses the framework of graphical models assumes that data are independent and identically distributed.

This unexplored topical intersection represents an opportunity for advancement --- by combining relational learning with causal reasoning, we can provide insight into the challenges found in each subject area. By adopting a causal viewpoint, we can clarify the mechanisms that produce previously identified pathologies in relational learning. Analogously, we can utilize relational data to establish and strengthen causal claims in ways that are impossible using only propositional representations.