Learning the Causal Structure of Overlapping Variable Sets

David Danks

Institute for Human and Machine Cognition
University of West Florida
40 S. Alcaniz St.
Pensacola, FL 32501 U.S.A.
ddanks@ai.uwf.edu

Abstract. In many real-world applications of machine learning and data mining techniques, one finds that one must separate the variables under consideration into multiple subsets (perhaps to reduce computational complexity, or because of a shift in focus during data collection and analysis). In this paper, we use the framework of Bayesian networks to examine the problem of integrating the learning outputs for multiple overlapping datasets. In particular, we provide rules for extracting causal information about the true (unknown) Bayesian network from the previously learned (partial) Bayesian networks. We also provide the SLPR algorithm, which efficiently uses these previously learned Bayesian networks to guide learning of the full structure. A complexity analysis of the “worst-case” scenario for the SLPR algorithm reveals that the algorithm is always less complex than a comparable “reference” algorithm (though no absolute optimality proof is known). Although no “expected-case” analysis is given, the complexity analysis suggests that (given the currently available set of algorithms) one should always use the SLPR algorithm, regardless of the underlying generating structure. The results provided in this paper point to a wide range of open questions, which are briefly discussed.

1 The “Big Picture” Problem

Modern data collection has advanced to the point that the size and complexity of our datasets regularly exceed the computational limits of our algorithms (on modern machines). As a result, analysis is often rendered computationally tractable only when we consider proper subsets of the variables we have measured. In addition, the variables thought to be relevant often change over the course of an investigation, both in the data collection and analysis phases. For example, an unexpected correlation might suggest the need to find an unmeasured common cause. When these changes occur, we want to use as much information as possible from earlier analyses to minimize duplication of effort. Thus, in both of these situations, we must address the distinctive problems (such as integration of outputs and efficient use of prior learning in subsequent learning) that arise for learning on multiple overlapping sets of variables.
There is a further, more practical, motivation. Many social science datasets have overlapping variables but the datapoints are unlabelled (for privacy reasons), so that there is no possible way to create a “complete” dataset. For example, we might have a census dataset and an unemployment dataset, both of which have Income as a variable, but neither of which contains identifiers to be used for creation of a single, integrated dataset. Hence, in these domains where there are substantial practical barriers to creating a unified dataset, the problem of integrating the learning outputs becomes particularly salient.

At the same time as we face these difficulties, we want our analysis techniques to reveal causal relationships. Causal information allows for predictions about the (probabilistic) outcomes of interventions, and is more easily understood by human users of machine learning techniques. Hence, if possible, we want to use a representation that allows for causal inference and prediction.

To better understand these problems, we can try to express them more formally. Let $V$ be the full set of variables under consideration. We assume that the variables are either all discrete or all continuous, though in the former case they need not have the same number of values. Let $S_1, \ldots, S_n$ be (nonempty) subsets of $V$ such that $S_1 \cap \ldots \cap S_n \neq \emptyset$. We further assume that, throughout all stages of learning, there is some stationary generating process producing the data, and that we have sufficient data that the sample statistics are the same as the population statistics. In this paper, we will be concerned with the following two questions:

1. If we do not have joint data over $V$, but we do have the outputs of some reliable, correct learning process (e.g., a machine learning algorithm) on $S_1, \ldots, S_n$, what can we learn about the relationships among $V$ as a whole? That is, if we can only learn the causal structure of the subsets (because of lack of data), what can we learn (if anything) about the full structure underlying $V$?
2. If we do have joint data over $V$, as well as the learning outputs, how can we efficiently learn the full structure for $V$? That is, how (if at all) can we use learning results over subsets to guide the learning for $V$ as a whole?

For example, we might have three datasets (drawn from the same population) over the following variables:

1. $S_1 = \{Education, ParentalEducation, Income\}$;
2. $S_2 = \{Education, Housing, Income\}$; and
3. $S_3 = \{Education, Age, NumberOfChildren, Income\}$.

In this example, the above two questions correspond to: (1) Given just these three datasets, what can be learned about the interrelationships among the six variables (including pairs, like Housing and Age, that do not appear in the same dataset)? and (2) How could we efficiently learn the full causal structure if we actually had a complete dataset?

On one level, it will be surprising if the answer to question 1 is anything other than “nothing.” Any positive answer implies that we can determine something about the relationship between two variables, despite the fact that we have no