

Learning Bayesian Networks from Incomplete Data: An Efficient Method for Generating Approximate Predictive Distributions

Carsten Riggelsen

Department of Information & Computing Sciences

Universiteit Utrecht

P.O. Box 80.098, 3508TB Utrecht, The Netherlands

carsten@cs.uu.nl

Abstract

We present an efficient method for learning Bayesian network models and parameters from incomplete data. With our approach an approximation is obtained of the predictive distribution. By way of this distribution any learning algorithm that works for complete data can be easily adapted to work for incomplete data as well. Our method exploits the dependence relations between the variables explicitly given by the Bayesian network model to predict missing values. Based on strength of influence and predictive quality, a subset of those predictor variables is selected, from which an approximate predictive distribution is generated by taking the observed part of the data into consideration. The approximate predictive distribution is obtained by traversing the data sample only twice and no iteration is required. Therefore our algorithm is more efficient than iterative algorithms such as EM and SEM. Our experiments show that the method performs well both for parameter learning and model learning compared to EM and SEM.

1 Introduction

Most methods for performing statistical data analysis require complete data samples in order to work or produce valid results. Unfortunately real-life databases are rarely complete. For doing statistical analysis of incomplete data, the standard tools for complete data often don't suffice anymore. Principled data analysis of incomplete data leads to analytical intractability and high computational complexity compared to the complete data scenario.

This paper is concerned with learning Bayesian networks (BN), a formalism built upon statistical principles. BNs are so-called directed graphical models which is a class of statistical models defined by a collection of conditional independences between variables represented by a graph. This graph offers an appealing way of structuring an otherwise confusing number of equations expressing the (in)dependences between variables.

BNs occupy a prominent position in decision support environments where they are used for diagnostic and prediction purposes. Also, in the context of data mining especially the graphical structure (model) of a Bayesian network is an appealing formalism for visualising the relationships between domain variables.

In this paper we show how to learn BNs from incomplete data when the missing data mechanism is *ignorable* as defined by Little & Rubin [15], which entails that data should be missing at random (MAR) or missing completely at random (MCAR). This essentially means that the probability that some entry is missing possibly depends on observed data, but is independent of unobserved data. In the typical MAR missing data mechanism, the probability of occurrence of a missing entry in a variable depends on fully observed covariates only. Without the ignorability assumption it is impossible to develop a fully automated procedure that produces statistically valid results. The reason is that the MAR assumption provides a minimal condition on which valid statistical analysis can be performed without modelling the underlying missing data mechanism. Under the MAR assumption, all information about the missing data, necessary for performing valid statistical analysis, is contained in the observed data, but structured in a way that complicates the analysis [3].

The method we develop has low computational cost compared to most existing incomplete data methods. The price we have to pay for this efficient algorithm is a certain degree of approximation.

Our missing data approach is not directly linked to model learning or parameter estimation *per se*. Instead we focus on the so-called *predictive distribution* which plays a crucial role when we want existing learning methods developed for complete data to work with incomplete data as well.

We proceed as follows: In section 2 we give a short review of previous research on learning Bayesian networks from incomplete data. Section 3 introduces the

