International Journal of Approximate Reasoning 69 (2016) 91–105

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

The role of local partial independence in learning of Bayesian networks

Johan Pensar [a,*], Henrik Nyman [a], Jarno Lintusaari [b], Jukka Corander [b]

[a] Department of Mathematics and Statistics, Åbo Akademi University, 20500 Turku, Finland
[b] Department of Mathematics and Statistics, University of Helsinki, 00014 Helsinki, Finland

ARTICLE INFO

Article history:
Received 19 May 2015
Received in revised form 16 November 2015
Accepted 20 November 2015
Available online 27 November 2015

Keywords:
Bayesian network
Model learning
Local structure
Partial independence
Bayesian score

ABSTRACT

Bayesian networks are one of the most widely used tools for modeling multivariate systems. It has been demonstrated that more expressive models, which can capture additional structure in each conditional probability table (CPT), may enjoy improved predictive performance over traditional Bayesian networks despite having fewer parameters. Here we investigate this phenomenon for models of various degree of expressiveness on both extensive synthetic and real data. To characterize the regularities within CPTs in terms of independence relations, we introduce the notion of partial conditional independence (PCI) as a generalization of the well-known concept of context-specific independence (CSI). To model the structure of the CPTs, we use different graph-based representations which are convenient from a learning perspective. In addition to the previously studied decision trees and graphs, we introduce the concept of PCI-trees as a natural extension of the CSI-based trees. To identify plausible models we use the Bayesian score in combination with a greedy search algorithm. A comparison against ordinary Bayesian networks shows that models with local structures in general enjoy parametric sparsity and improved out-of-sample predictive performance, however, often it is necessary to regulate the model fit with an appropriate model structure prior to avoid overfitting in the learning process. The tree structures, in particular, lead to high quality models and suggest considerable potential for further exploration.

1. Introduction

Since its introduction nearly three decades ago, the class of Bayesian networks has become a well-established tool for many purposes of probabilistic inference. Conditional independence (CI), which is the central probabilistic concept of Bayesian networks, allows a modular and low-dimensional parametrization of a potentially very high-dimensional multivariate system, in terms of conditional probabilities usually specified by conditional probability tables (CPTs). Still, the CI-based dependence structure may in certain situations be unnecessarily coarse, resulting in an inefficient representation of the underlying distribution in the sense that unnecessary parameters need to be learned and retained. This phenomenon has been recognized by numerous authors [3,8,10,12,21,22] who have introduced local structures in order to capture regularities within the CPTs of a Bayesian network. Additionally, the related model classes Bayesian multinets [14], probabilistic

[*] Corresponding author.
E-mail address: johan.pensar@abo.fi (J. Pensar).

http://dx.doi.org/10.1016/j.ijar.2015.11.008
0888-613X/© 2015 Elsevier Inc. All rights reserved.

decision graphs [16], and chain event graphs [1,11,25] have also been developed in the pursuit of modeling asymmetric dependence structures that cannot be explicitly captured by traditional Bayesian networks.

One of the most widely known local restrictions is based on the notion of context-specific independence (CSI) which was formalized by Boutilier et al. [3]. Compact representations of such constraints have been obtained through CSI-trees [3,12], parent contexts [22], and labeled directed acyclic graphs [21]. In this work, we further generalize CSI to partial conditional independence (PCI) in order to obtain sound local restrictions that lead to more expressive CPT-structures which still follow an underlying structure. Due to computational advantages, we focus here on tree-based structures. In particular, we show that CSI-trees can easily be extended to also capture certain PCIs in an efficient way.

The role of structured CPTs in model learning has previously been investigated by numerous authors [5,8,10,12,21] and the results suggest it can have a positive effect on the predictive performance of the inferred models. The reason for this is that the increased flexibility, gained by modeling the CPTs separately, allows a model to better emulate an underlying distribution without inducing redundant parameters. As a result, a reduced number of parameters leads to a more stable estimation of the model distribution. Still, a negative aspect is a vastly increased model space making the already difficult task of model learning even more challenging. One of the key reasons to why structured CPTs are such a convenient add-on from a learning perspective is that popular model selection criteria such as the Bayesian score can still be evaluated in closed form [8,12]. For this choice of score, it is not uncommon to assume a uniform prior over the model space, however, Pensar et al. [21] noticed that this may result in dense models with poor out-of-sample performance when the data are generated from an actual Bayesian network. Here we investigate this particular phenomenon as well as the overall effect of structured CPTs of different degrees of expressiveness in extensive numerical experiments covering both synthetic and real data.

The remainder of this article is structured as follows. In the next section we define various forms of local independence and consider their representation in terms of local graphs. Section 3 presents a Bayesian scoring method for networks with structured CPTs. In addition, the section introduces a greedy search algorithm for identifying high-scoring models that use PCI- and CSI-based trees as well as decision graphs. The penultimate section presents results from extensive numerical experiments with both real and synthetic data. The last section provides further remarks on our findings and suggests several venues for further research on learning and use of local partial independence.

2. Local partial independence in Bayesian networks

We begin by introducing the following notation. We consider a set of discrete stochastic variables $X = \{X_1, \dots, X_d\}$ indexed by $V = \{1, \dots, d\}$. Each variable X_j takes values from a finite set of outcomes $\mathcal{X}_j$. For a subset $S \subseteq V$, we denote the associated set of variables by $X_S = \{X_j\}_{j \in S}$ and the joint outcome space is given by the Cartesian product $\mathcal{X}_S = \times_{j \in S} \mathcal{X}_j$. The cardinality of an outcome space $\mathcal{X}_S$ is denoted by $|\mathcal{X}_S|$. We use a lowercase letter x_S to indicate that the corresponding variables have been assigned a specific joint configuration in $\mathcal{X}_S$. Accordingly, we use $p(X_S)$ to denote a probability distribution over X_S whereas $p(x_S)$ denotes the probability of the variables being assigned a specific configuration.

2.1. Partial independence

We now proceed to formally define and discuss different classes of independences that can explain interactions, or the lack thereof, between variables. The most fundamental statement of independence is described by conditional independence.

Definition 1 *(Conditional Independence (CI)).* Let A, B, C be three disjoint subsets of V. We say that X_A is conditionally independent of X_B given X_C if

$$p(x_A \mid x_B, x_C) = p(x_A \mid x_C)$$

holds for all $(x_A, x_B, x_C) \in \mathcal{X}_A \times \mathcal{X}_B \times \mathcal{X}_C$ whenever $p(x_B, x_C) > 0$. This is denoted by

$$X_A \perp X_B \mid X_C.$$

If $C = \varnothing$, then $X_A \perp X_B$ is reduced to marginal independence (MI) between the two sets of variables.

A CI statement is global in the sense that it holds throughout the outcome space of the involved variables. In practice, it implies that the knowledge of X_C renders the information given by X_B irrelevant when considering the conditional distribution of X_A given X_B and X_C. The concept of Bayesian networks, or graphical models in general, is based on the notion of CI since it allows a decomposition of a network into smaller more manageable parts. Still, there are situations where CI alone can be unnecessarily restrictive for a model to efficiently capture the relationships between variables in real-world phenomena. For this reason, numerous authors [3,8,10,14,21,22,25] have introduced independence statements that only hold in parts of the outcome space, or in a certain domain. In an attempt to cover all such approaches, we now introduce the notion of partial conditional independence as an extension of the now well-established concept of CSI.

Definition 2 *(Partial Conditional Independence (PCI)).* Let A, B, C be three disjoint subsets of V. We say that X_A is partially conditionally independent of X_B in the domain $\mathcal{D}_B \subseteq \mathcal{X}_B$ given the context $X_C = x_C$ if