International Journal of Approximate Reasoning 69 (2016) 18–34

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Learning from incomplete data in Bayesian networks with qualitative influences

Andrés R. Masegosa [a,*,1], Ad J. Feelders [b], Linda C. van der Gaag [b]

[a] Department of Computer and Information Science, The Norwegian University of Science and Technology, Norway
[b] Institute of Information and Computing Sciences, Utrecht University, The Netherlands

ARTICLE INFO

Article history:
Received 27 July 2015
Received in revised form 4 November 2015
Accepted 6 November 2015
Available online 11 November 2015

Keywords:
Bayesian networks
Qualitative influences
Parameter learning
Missing data
EM algorithm
Isotonic regression

ABSTRACT

Domain experts can often quite reliably specify the sign of influences between variables in a Bayesian network. If we exploit this prior knowledge in estimating the probabilities of the network, it is more likely to accepted by its users and may in fact be better calibrated with reality. We present two algorithms that exploit prior knowledge of qualitative influences in learning the parameters of a Bayesian network from incomplete data. The isotonic regression EM, or irEM, algorithm adds an isotonic regression step to standard EM in each iteration, to obtain parameter estimates that satisfy the given qualitative influences. In an attempt to reduce the computational burden involved, we further define the qirEM algorithm that enforces the constraints imposed by the qualitative influences only once, after convergence of standard EM. We evaluate the performance of both algorithms through experiments. Our results demonstrate that exploitation of the qualitative influences improves the parameter estimates over standard EM, and more so if the proportion of missing data is relatively large. The results also show that the qirEM algorithm performs just as well as its computationally more expensive counterpart irEM.

1. Introduction

While domain experts often have difficulties in coming up with numerical probability assessments, experience shows that they feel more comfortable with providing qualitative knowledge about the probabilistic influences between variables in a Bayesian network [10,19]. The qualitative knowledge provided by the experts, moreover, tends to be more robust than their numerical assessments. For example, experts can quite reliably state knowledge of the type: *smoking increases the probability of lung cancer*, but have much more problems with specifying the exact probabilities.

Previous work [2,27,14] has shown that the exploitation of prior knowledge about qualitative influences can improve parameter learning in Bayesian networks. This improvement can be observed in particular when training data is scarce, because in that case the order constraints on the parameters (resulting from the specified qualitative influences) may be violated in the training sample due to sampling variability. If the specified influences are correct and the training sample is large, the constraints tend to be satisfied anyway, and are therefore less useful. In practical learning problems we are often confronted with missing values in the training data, and it stands to reason that knowledge of qualitative influences can be

* Corresponding author.
E-mail addresses: andresmasegosa@ntnu.no (A.R. Masegosa), A.J.Feelders@uu.nl (A.J. Feelders), L.C.vanderGaag@uu.nl (L.C. van der Gaag).

[1] This research was conducted when the author was affiliated at the Department of Computer Science and Artificial Intelligence, University of Granada, Spain.

http://dx.doi.org/10.1016/j.ijar.2015.11.004
0888-613X/

particularly helpful in such cases. Therefore, in this paper we focus on parameter learning from incomplete data, assuming there are order restrictions on the parameters resulting from qualitative influences specified by a domain expert.

Earlier work on parameter learning with order constraints from complete data [14] proposed to use the *isotonic regression* [24] for this purpose. In case the network variables are binary, the isotonic regression produces the constrained maximum likelihood estimates, and in the more general case of ordinal variables its performance was shown to be indistinguishable from that of constrained maximum likelihood estimation [12]. In this paper we restrict our attention to networks with binary variables. We propose to augment the maximization step of the EM algorithm with the isotonic regression in order to obtain parameter estimates that satisfy the order constraints from incomplete data. This algorithm, called irEM, requires the application of the isotonic regression in each iteration of EM. In an attempt to speed up the learning process, we also propose qirEM, which applies the isotonic regression only once, namely after the standard EM algorithm has converged. Experiments show that qirEM produces parameter estimates of the same quality as the computationally more expensive irEM algorithm.

This work is related to [8,22] which address the problem of parameter learning from incomplete data with a more general class of constraints that contains the order constraints resulting from qualitative influences as a special case. To handle this more general class of constraints, they require the application of convex programming in the M-step of EM, thereby adding another iterative numerical optimization step. By limiting ourselves to qualitative influences we can compute the exact constrained maximum in polynomial time in the M step. Moreover, qirEM passes over the constrained optimization step altogether. It is applied only once, after the convergence of normal EM.

This paper is organized as follows. In the next section we introduce the basic concepts and notation required for the remainder of the paper. In section 4 we present the irEM algorithm and its faster alternative qirEM. This section also provides an analysis of the properties of these algorithms. Section 5 presents an experimental evaluation of the algorithms on three well-known Bayesian networks from the literature and five network structures that were learned from data. Finally, section 6 concludes.

2. Preliminaries

We introduce our notational conventions and briefly review the EM algorithm and the isotonic regression for parameter estimation under incomplete data and monotonicity constraints, respectively.

2.1. Notations

We consider a set $\mathbf{X}$ of binary random variables. Each variable $X \in \mathbf{X}$ takes the values 0 and 1; we will use x to denote $X = 1$ and $\bar{x}$ to denote $X = 0$. A joint value assignment to a (sub-)set of variables $\mathbf{Y} \subseteq \mathbf{X}$ is written as a vector $\mathbf{y}$. The set of all joint assignments to $\mathbf{Y}$ is denoted by $\mathrm{Val}(\mathbf{Y})$.

A Bayesian network B is a probabilistic graphical model defining a joint probability distribution Pr over $\mathbf{X}$. This probabilistic model is composed of an acyclic digraph G, with nodes for the random variables and directed arcs to capture the independency structure over them, and an associated parameter vector $\boldsymbol{\theta}$. We use $\Pi_X \subset \mathbf{X}$ to denote the set of parents of the variable X in G. A joint value assignment to this set Π_X will be termed a parent configuration for X. We associate with the set $\mathrm{Val}(\Pi_X)$ of all parent configurations of X a partial order on its elements, which will be denoted as $\preceq_X$. The parameter vector $\boldsymbol{\theta}$ now includes for each variable $X \in \mathbf{X}$, the elements $\theta_{x\pi} = \Pr(x \mid \pi)$ for all parent configurations π; we use θ_X to denote the set of all elements specified for the variable X. The parameter vector $\boldsymbol{\theta}$ fully describes the joint distribution Pr given the independency structure from G. In the remainder of the paper, we consider the digraph G of a network B to be fixed, and address the estimation of its parameter vector $\boldsymbol{\theta}$.

A qualitative influence between two variables X and Y connected by an arc $X \to Y$ describes how observing a value for the one variable affects the probability distribution of the other variable. A positive qualitative influence of X on Y expresses that observing x increases the probability of observing y (i.e. $Y = 1$), assuming that the values of the other parents of Y remain the same, that is,

$$\Pr(y \mid x, \mathbf{s}) \geq \Pr(y \mid \bar{x}, \mathbf{s}) \tag{1}$$

for any combination of values $\mathbf{s}$ for the set of parents of Y other than X; a negative influence between X and Y expresses that

$$\Pr(y \mid x, \mathbf{s}) \leq \Pr(y \mid \bar{x}, \mathbf{s}) \tag{2}$$

for any such combination $\mathbf{s}$.

We further consider a multiset $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$, of data samples that were independently drawn from the distribution Pr over $\mathbf{X}$. Each sample $\mathbf{x}^{(i)}$ from $\mathcal{D}$ is a partially observed vector which may include one or more missing values. We use $\mathbf{O}^{(i)}$ and $\mathbf{o}^{(i)}$ to denote the set of observed variables of the sample $\mathbf{x}^{(i)}$ and its associated joint value assignment, respectively. We use $\mathbf{U}^{(i)}$ for the set of unobserved variables of $\mathbf{x}^{(i)}$; note that the set of unobserved variables may differ among samples. We further consider the possible completions of a data sample $\mathbf{x}^{(i)}$ and write $\mathcal{U}^{(i)}$ to denote the set of all possible joint value assignments $\mathbf{u}^{(i)}$ to the sample's unobserved variables. The pair $(\mathcal{U}, \mathcal{D})$ with $\mathcal{U} = \times_{i=1,\ldots,m}\, \mathcal{U}^{(i)}$ thus defines the possible completions of the dataset $\mathcal{D}$; we will write $\mathbf{u} \in \mathcal{U}$ to denote such a completion.