Neurocomputing 137 (2014) 142–149

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Discrete exponential Bayesian networks: Definition, learning and application for density estimation ☆

Aida Jarraya [a,b,*], Philippe Leray [b], Afif Masmoudi [a]

[a] Laboratory of Probability and Statistics, Faculty of Sciences of Sfax, University of Sfax, Tunisia

[b] LINA Computer Science Lab UMR 6241, Knowledge and Decision Team, University of Nantes, France

ARTICLE INFO

Article history:
Received 20 November 2012
Received in revised form
15 May 2013
Accepted 18 May 2013
Available online 15 February 2014

Keywords:
Exponential family
Bayesian network
Density estimation

ABSTRACT

Our work aims at developing or expliciting bridges between Bayesian networks (BNs) and Natural Exponential Families, by proposing discrete exponential Bayesian networks as a generalization of usual discrete ones. We introduce a family of prior distributions which generalizes the Dirichlet prior applied on discrete Bayesian networks, and then we determine the overall posterior distribution. Subsequently, we develop the Bayesian estimators of the parameters, and a new score function that extends the Bayesian Dirichlet score for BN structure learning. Our goal is to determine empirically in which contexts some of our discrete exponential BNs (Poisson deBNs) can be an effective alternative to usual BNs for density estimation.

1. Introduction

Probabilistic graphical models, specifically Bayesian networks are tools for knowledge representation under uncertainty. Under usual hypotheses, the joint probability distribution associated to a BN is decomposed in the product of local conditional probability distribution of each variable given its parents in the graph.

Some Bayesian estimation methods can be used in order to estimate the parameters of each conditional probability distribution, given one dataset and one a priori distribution over the parameters. The computation of the posterior distribution is useful for two purposes: to estimate the probability of a graph given the data and then to find the better graph (structure) that fits the data [6,17,16].

Reviewing the literature, we find abundant research and works dealing with discrete Bayesian network, where the conditional distribution of each variable, given its parents, is a multinomial distribution [28].

In this paper, we are interested in extending the distribution of variables to the natural exponential family (NEF) which represents a very important class of distributions in probability and statistical theory [2,24]. This idea has previously been developed by Beal and Ghahramani [3] (conjugate–exponential models) for Bayesian networks with latent variables. They concentrated their work on variational Expectation Maximization (EM) estimation needed because of latent variables, but they did not clarify the Bayesian estimators used, hence restricting their experiments to the usual multinomial distributions. Wainwright and Jordan [30] also proposed an interesting study of graphical models as exponential families, showing that very specific structures of directed or undirected probabilistic graphical models can be interpreted as exponential distributions. Our work pursues with the same general idea, developing bridges between BNs and NEFs, dealing with discrete exponential BNs instead of usual multinomial ones in order to explore a wider range of probabilistic models. We formally introduce a family of prior distributions valid for any NEF distribution and show that this family of priors generalizes the Dirichlet prior used in discrete multinomial BNs. Then, we are able to express the global posterior distribution and a new score function for learning the structure of any discrete exponential BN, and the Bayesian estimators for parameters of such BNs. Our work aims to determine empirically in which contexts one class of these discrete exponential Bayesian networks (Poisson deBNs) can be an effective alternative to usual Bayesian networks for density estimation.

The present paper is structured as follows. In Section 2, we review existing results for discrete multinomial BN learning. After that, in Section 3, we review some of the properties of natural exponential families with quadratic variance functions. In Section 4, we propose our results concerning structure and parameters learning for discrete exponential BNs. Section 5 aims at expliciting our experimental protocol, evaluation criteria, and gives an interpretations of results. Finally, we conclude with perspectives for a future work.

☆This paper is an extended version of theoretical and experimental results initially presented in [20,21].

* Corresponding author at: LINA Computer Science Lab UMR 6241, Knowledge and Decision Team, University of Nantes, France.

E-mail addresses: aidajarraya@yahoo.fr (A. Jarraya), philippe.leray@univ-nantes.fr (P. Leray), Afif.Masmoudi@fss.rnu.tn (A. Masmoudi).

http://dx.doi.org/10.1016/j.neucom.2013.05.061
0925-2312

2. Discrete Bayesian network background

2.1. General definition

Consider a finite set $X = \{X_1, \ldots, X_n\}$ of random variables. A Bayesian network (BN) is a directed acyclic graph G and a set of conditional probability distributions which represent a joint probability distribution

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | pa_i) \tag{1}$$

where pa_i represent the parents of node X_i in G.

Bayesian networks allow to represent distributions compactly and to construct efficient inference and learning algorithms. Distinct directed acyclic graphs may sometimes encode the same set of independence relations and hence, may be considered as Markov equivalent graphs.

In what follows, we assume that the local distributions depend on a finite set of parameters $\mu = (\mu_{ij})_{1 \le i \le n, 1 \le j \le q_i}$, where q_i is the number of parent configurations for variable X_i.

Usually, as described in [28], discrete Bayesian networks studied in the literature assume that local conditional distributions $P(X_i|pa_i)$ are multinomial distributions $Mult(1, \mu_{ij1}, \ldots, \mu_{ijr_i})$, i.e. multivariate Bernoulli distributions with parameters $(\mu_{ij1}, \ldots, \mu_{ijr_i})$.

2.2. Structure learning

Learning Bayesian network structure from dataset is a NP-hard problem [5] and is still one of the most exciting challenges in machine learning.

There are three classical approaches often used for BN structure learning. First, constraint-based methods consist in detecting (in) dependencies between the variables by performing conditional independence tests on data. Second, score-and-search based approaches use a score function to evaluate the quality of a given structure and a heuristic to search in the solution space. The third hybrid approach merges the two previous approaches.

In previous works, Acid et al. [1], Brown et al. [4], Fu [11] and Tsamardinos et al. [29] provide comparisons of Bayesian network learning algorithms. Recently, Daly et al. [9] states that score-and-search approaches are some of the most successful strategies for learning Bayesian networks. Even a simple heuristic such as greedy search [6] can produce good results. The choice of the scoring function is also an open question. This function, often related to the marginal likelihood can be approximated in several ways [7]. Daly et al. [9] proposes some elements in order to compare the two main scoring functions, the large scale approximation (BIC) [13] and the Bayesian Dirichlet approximation described below.

Lemma 2.1. *Suppose that we have one dataset $d = \{x^{(1)}, \ldots, x^{(M)}\}$ of size M where $x^{(h)} = (x_1^{(h)}, \ldots, x_n^{(h)})'$ is the hth sample and $x_i^{(h)}$ is the value of the variable X_i for this sample. Therefore, the distribution of the dataset d given $\mu = (\mu_1, \ldots, \mu_n)$ and the structure G is*

$$P(d|\mu, G) = \prod_{i=1}^{n} \prod_{h=1}^{M} P(x_i^{(h)} | pa_i^{(h)}, \mu_i, G) \tag{2}$$

where $pa_i^{(h)}$ contains the values of the parents of x_i in the hth sample.

Given a structure G, we denote by N_{ijk} the number of samples in d where X_i is in its kth state and its parents are in their jth configuration, by $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ the number of samples in d where pa_i is in the jth configuration, r_i denotes the number of states of X_i, q_i the number of parents configurations of X_i. We suppose that the conditional distribution of X_i given pa_i is a multinomial distribution $Mult(1, \mu_{ij1}, \ldots, \mu_{ijr_i})$ which corresponds to the multivariate Bernoulli distribution with parameters $(\mu_{ij1}, \ldots, \mu_{ijr_i})$. Starting from a prior distribution about the possible structures $P(G)$, the objective is to express the posterior probability of all possible structures conditional on a dataset d. After some calculation and using a Dirichlet prior, we get the Bayesian Dirichlet (BD) scoring function:

$$BD(d, G) = P(d, G) = P(G) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij} + \alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \tag{3}$$

where α_{ijk} are the parameters of the Dirichlet prior distribution.

Heckerman et al. [17] propose a constraint on Dirichlet coefficients to be used in order to obey to the Markov equivalence:

$$\alpha_{ijk} = N' P(X_i = k, pa_i = j | G_c) \tag{4}$$

where G_c denotes the completely connected graph and N' is an equivalent sample size defined by the user.

2.3. Parameter learning

The following theorem is introduced in the context of Bayesian networks under the name of global posterior independent parameters.

Theorem 2.2. *Under the same conditions of Lemma 2.1, we suppose that $\mu = (\mu_i)_{1 \le i \le n}$ are mutually independent given a dataset d, thus*

$$P(\mu | d, G) = \prod_{i=1}^{n} P(\mu_i | d, G) \tag{5}$$

In the case of discrete BNs where the joint distribution is a multinomial one and the prior is a Dirichlet distribution, the EAP (expected a posteriori) estimator of $\mu_{ij} = (\mu_{ijk})_{1 \le k \le r_i}$ is given by

$$\widehat{\mu}_{ijk}^{EAP} = \frac{N_{ijk} + \alpha_{ijk}}{\sum_{l=1}^{r_i} (N_{ijl} + \alpha_{ijl})}, k = 1, \ldots, r_i. \tag{6}$$

Note that the MAP (maximum a posteriori) estimator is

$$\widehat{\mu}_{ijk}^{MAP} = \frac{N_{ijk} + \alpha_{ijk} - 1}{\sum_{l=1}^{r_i} (N_{ijl} + \alpha_{ijl} - 1)}, \quad k = 1, \ldots, r_i. \tag{7}$$

3. Natural exponential families

3.1. Introduction and definitions

The exponential family is a widely used family of distributions. The natural exponential families (NEFs) is a subclass of this family, with interesting estimation properties [2]. In the last decades, several classifications of natural exponential families based on the form of the variance function have appeared in literature [25,15]. There are six basic natural exponential families having quadratic variance functions distributions: Normal, Poisson, Gamma, Binomial, Negative Binomial and the NEF generated by the generalized hyperbolic secant distributions. For an accurate presentation of the simple quadratic natural exponential families, let us begin with some traditional definitions and notations in Statistics. For more details, we refer to [24]. Let $(\theta, x) \longmapsto \langle \theta, x \rangle$ be the canonical scalar product on $\mathbb{R}^d \times \mathbb{R}^d$. For a positive measure ν on $\mathbb{R}^d$, we denote

$$L_\nu : \mathbb{R}^d \longrightarrow [0, +\infty[$$
$$\theta \longrightarrow \int_{\mathbb{R}^d} \exp \langle \theta, x \rangle \nu(dx)$$

$\Theta(\nu) = int\{\theta \in \mathbb{R}^d ; L_\nu(\theta) < +\infty\}$, $k_\nu = \log\ L_\nu$.

L_ν and k_ν are, respectively, the Laplace transform and the cumulant generating function of ν. The set $\mathcal{M}(\mathbb{R}^d)$ is now defined as the set of positive measures ν such that ν is not concentrated on