Expert Systems with Applications 41 (2014) 6755–6772

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A hybrid algorithm for Bayesian network structure learning with application to multi-label learning

Maxime Gasse, Alex Aussem *, Haytham Elghazel

Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France

ARTICLE INFO

Article history:
Available online 9 May 2014

Keywords:
Bayesian networks
Multi-label learning
Markov boundary
Feature subset selection

ABSTRACT

We present a novel hybrid algorithm for Bayesian network structure learning, called H2PC. It first reconstructs the skeleton of a Bayesian network and then performs a Bayesian-scoring greedy hill-climbing search to orient the edges. The algorithm is based on divide-and-conquer constraint-based subroutines to learn the local structure around a target variable. We conduct two series of experimental comparisons of H2PC against Max–Min Hill-Climbing (MMHC), which is currently the most powerful state-of-the-art algorithm for Bayesian network structure learning. First, we use eight well-known Bayesian network benchmarks with various data sizes to assess the quality of the learned structure returned by the algorithms. Our extensive experiments show that H2PC outperforms MMHC in terms of goodness of fit to new data and quality of the network structure with respect to the true dependence structure of the data. Second, we investigate H2PC's ability to solve the multi-label learning problem. We provide theoretical results to characterize and identify graphically the so-called minimal label powersets that appear as irreducible factors in the joint distribution under the faithfulness condition. The multi-label learning problem is then decomposed into a series of multi-class classification problems, where each multi-class variable encodes a label powerset. H2PC is shown to compare favorably to MMHC in terms of global classification accuracy over ten multi-label data sets covering different application domains. Overall, our experiments support the conclusions that local structural learning with H2PC in the form of local neighborhood induction is a theoretically well-motivated and empirically effective learning framework that is well suited to multi-label learning. The source code (in *R*) of H2PC as well as all data sets used for the empirical tests are publicly available.

1. Introduction

A Bayesian network (BN) is a probabilistic model formed by a structure and parameters. The structure of a BN is a directed acyclic graph (DAG), whilst its parameters are conditional probability distributions associated with the variables in the model. The problem of finding the DAG that encodes the conditional independencies present in the data attracted a great deal of interest over the last years (Gasse, Aussem, & Elghazel, 2012; Kojima, Perrier, Imoto, & Miyano, 2010; Peña, 2012; Perrier, Imoto, & Miyano, 2008; Rodrigues de Morais & Aussem, 2010a; Scutari, 2010; Scutari & Brogini, 2012; Villanueva & Maciel, 2012). The inferred DAG is very useful for many applications, including feature selection (Aliferis, Statnikov, Tsamardinos, Mani, & Koutsoukos, 2010; Peña, Nilsson, Björkegren, & Tegnér, 2007; Rodrigues de Morais & Aussem, 2010b), causal relationships inference from observational data (Aliferis et al., 2010; Aussem, Rodrigues de Morais, & Corbex, 2012, 2010; Brown & Tsamardinos, 2008; Cawley, 2008; Ellis & Wong, 2008; Prestat et al., 2013) and more recently multi-label learning (Dembczyski, Waegeman, Cheng, & Hüllermeier, 2012; Guo & Gu, 2011; Zhang & Zhang, 2010).

Ideally the DAG should coincide with the dependence structure of the global distribution, or it should at least identify a distribution as close as possible to the correct one in the probability space. This step, called structure learning, is similar in approaches and terminology to model selection procedures for classical statistical models. Basically, constraint-based (CB) learning methods systematically check the data for conditional independence relationships and use them as constraints to construct a partially oriented graph representative of a BN equivalence class, whilst search-and-score (SS) methods make use of a goodness-of-fit score function for evaluating graphical structures with regard to the data set. Hybrid methods attempt to get the best of both worlds: they learn a skeleton with a CB approach and constrain on the DAGs considered during the SS phase.

* Corresponding author. Tel.: +33 426234466.
E-mail address: aaussem@univ-lyon1.fr (A. Aussem).

http://dx.doi.org/10.1016/j.eswa.2014.04.032
0957-4174/

In this study, we present a novel hybrid algorithm for Bayesian network structure learning, called H2PC.[1] It first reconstructs the skeleton of a Bayesian network and then performs a Bayesian-scoring greedy hill-climbing search to orient the edges. The algorithm is based on divide-and-conquer constraint-based subroutines to learn the local structure around a target variable. HPC may be thought of as a way to compensate for the large number of false negatives at the output of the weak PC learner, by performing extra computations. As this may arise at the expense of the number of false positives, we control the expected proportion of false discoveries (i.e. false positive nodes) among all the discoveries made in $\mathbf{PC}_T$. We use a modification of the Incremental association Markov boundary algorithm (IAMB), initially developed by Tsamardinos et al. in Tsamardinos, Aliferis, and Statnikov (2003) and later modified by Peña in Peña (2008) to control the FDR of edges when learning Bayesian network models. HPC scales to thousands of variables and can deal with many fewer samples ($n < q$). To illustrate its performance by means of empirical evidence, we conduct two series of experimental comparisons of H2PC against Max–Min Hill-Climbing (MMHC), which is currently the most powerful state-of-the-art algorithm for BN structure learning (Tsamardinos, Brown, & Aliferis, 2006), using well-known BN benchmarks with various data sizes, to assess the goodness of fit to new data as well as the quality of the network structure with respect to the true dependence structure of the data.

We then address a real application of H2PC where the true dependence structure is unknown. More specifically, we investigate H2PC's ability to encode the joint distribution of the label set conditioned on the input features in the multi-label classification (MLC) problem. Many challenging applications, such as photo and video annotation and web page categorization, can benefit from being formulated as MLC tasks with large number of categories (Dembczyski et al., 2012; Kocev, Vens, Struyf, & Džeroski, 2007; Madjarov, Kocev, Gjorgjevikj, & Džeroski, 2012; Read, Pfahringer, Holmes, & Frank, 2009; Tsoumakas, Katakis, & Vlahavas, 2010b). Recent research in MLC focuses on the exploitation of the label conditional dependency in order to better predict the label combination for each example. We show that local BN structure discovery methods offer an elegant and powerful approach to solve this problem. We establish two theorems (Theorems 6 and 7) linking the concepts of marginal Markov boundaries, joint Markov boundaries and so-called label powersets under the faithfulness assumption. These Theorems offer a simple guideline to characterize graphically: (i) the minimal label powerset decomposition, (i.e. into minimal subsets $\mathbf{Y}_{LP} \subseteq \mathbf{Y}$ such that $\mathbf{Y}_{LP} \perp\!\!\!\perp \mathbf{Y} \setminus \mathbf{Y}_{LP} | \mathbf{X}$), and (ii) the minimal subset of features, w.r.t an Information Theory criterion, needed to predict each label powerset, thereby reducing the input space and the computational burden of the multi-label classification. To solve the MLC problem with BNs, the DAG obtained from the data plays a pivotal role. So in this second series of experiments, we assess the comparative ability of H2PC and MMHC to encode the label dependency structure by means of an indirect goodness of fit indicator, namely the 0/1 loss function, which makes sense in the MLC context.

The rest of the paper is organized as follows: In the Section 2, we review the theory of BN and discuss the main BN structure learning strategies. We then present the H2PC algorithm in details in Section 3. Section 4 evaluates our proposed method and shows results for several tasks involving artificial data sampled from known BNs. Then we report, in Section 5, on our experiments on real-world data sets in a multi-label learning context so as to provide empirical support for the proposed methodology. The main theoretical results appear formally as two theorems (Theorems 6 and 7) in Section 5. Their proofs are established in the Appendix. Finally, Section 6 raises several issues for future work and we conclude in Section 7 with a summary of our contribution.

[1] A first version of HP2C without FDR control has been discussed in a paper that appeared in the Proceedings of ECML-PKDD, pp. 58–73, 2012.

2. Preliminaries

We define next some key concepts used along the paper and state some results that will support our analysis. In this paper, upper-case letters in italics denote random variables (e.g., X, Y) and lower-case letters in italics denote their values (e.g., x, y). Upper-case bold letters denote random variable sets (e.g., $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$) and lower-case bold letters denote their values (e.g., $\mathbf{x}, \mathbf{y}$). We denote by $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$ the conditional independence between $\mathbf{X}$ and $\mathbf{Y}$ given the set of variables $\mathbf{Z}$. To keep the notation uncluttered, we use $p(\mathbf{y}|\mathbf{x})$ to denote $p(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x})$.

2.1. Bayesian networks

Formally, a BN is a tuple $\langle \mathcal{G}, P \rangle$, where $\mathcal{G} = \langle \mathbf{U}, \mathbf{E} \rangle$ is a directed acyclic graph (DAG) with nodes representing the random variables $\mathbf{U}$ and P a joint probability distribution in $\mathcal{U}$. In addition, $\mathcal{G}$ and P must satisfy the Markov condition: every variable, $X \in \mathbf{U}$, is independent of any subset of its non-descendant variables conditioned on the set of its parents, denoted by $\mathbf{Pa}_i^{\mathcal{G}}$. From the Markov condition, it is easy to prove (Neapolitan, 2004) that the joint probability distribution P on the variables in $\mathbf{U}$ can be factored as follows:

$$P(\mathcal{V}) = P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \mathbf{Pa}_i^{\mathcal{G}}) \tag{1}$$

Eq. (1) allows a parsimonious decomposition of the joint distribution P. It enables us to reduce the problem of determining a huge number of probability values to that of determining relatively few.

A BN structure $\mathcal{G}$ entails a set of conditional independence assumptions. They can all be identified by the *d-separation criterion* (Pearl, 1988). We use $X \perp_{\mathcal{G}} Y | \mathbf{Z}$ to denote the assertion that X is d-separated from Y given $\mathbf{Z}$ in $\mathcal{G}$. Formally, $X \perp_{\mathcal{G}} Y | \mathbf{Z}$ is true when for every undirected path in $\mathcal{G}$ between X and Y, there exists a node W in the path such that either (1) W does not have two parents in the path and $W \in \mathbf{Z}$, or (2) W has two parents in the path and neither W nor its descendants is in $\mathbf{Z}$. If $\langle \mathcal{G}, P \rangle$ is a BN, $X \perp_P Y | \mathbf{Z}$ if $X \perp_{\mathcal{G}} Y | \mathbf{Z}$. The converse does not necessarily hold. We say that $\langle \mathcal{G}, P \rangle$ satisfies the *faithfulness condition* if the d-separations in $\mathcal{G}$ identify *all and only* the conditional independencies in P, i.e., $X \perp_P Y | \mathbf{Z}$ iff $X \perp_{\mathcal{G}} Y | \mathbf{Z}$.

A Markov blanket $\mathbf{M}_T$ of T is any set of variables such that T is conditionally independent of all the remaining variables given $\mathbf{M}_T$. By extension, a Markov blanket of T in $\mathbf{V}$ guarantees that $\mathbf{M}_T \subseteq \mathbf{V}$, and that T is conditionally independent of the remaining variables in $\mathbf{V}$, given $\mathbf{M}_T$. A Markov boundary, $\mathbf{MB}_T$, of T is any Markov blanket such that none of its proper subsets is a Markov blanket of T.

We denote by $\mathbf{PC}_T^{\mathcal{G}}$, the set of parents and children of T in $\mathcal{G}$, and by $\mathbf{SP}_T^{\mathcal{G}}$, the set of *spouses* of T in $\mathcal{G}$. The *spouses* of T are the variables that have common children with T. These sets are unique for all $\mathcal{G}$, such that $< \mathcal{G}, P >$ satisfies the faithfulness condition and so we will drop the superscript $\mathcal{G}$. We denote by $\mathbf{dSep}(X)$, the set that d-separates X from the (implicit) target T.

Theorem 1. *Suppose $< \mathcal{G}, P >$ satisfies the faithfulness condition. Then X and Y are not adjacent in $\mathcal{G}$ iff $\exists \mathbf{Z} \in \mathbf{U} \setminus \{X, Y\}$ such that $X \perp\!\!\!\perp Y | \mathbf{Z}$. Moreover, $\mathbf{MB}_X = \mathbf{PC}_X \cup \mathbf{SP}_X$.*

A proof can be found for instance in Neapolitan (2004).