Journal of Computational and Applied Mathematics 287 (2015) 48–62

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

Structure space of Bayesian networks is dramatically reduced by subdividing it in sub-networks

Heni Bouhamed [a,c,*], Afif Masmoudi [b], Thierry Lecroq [c], Ahmed Rebaï [a]

[a] Bioinformatics Unit, Centre of Biotechnologie of Sfax, Sfax University, Tunisia
[b] Laboratory of Statistics and Probability, Sfax University, Tunisia
[c] LITIS EA 4108, University of Rouen, France

ARTICLE INFO

Article history:
Received 8 March 2012
Received in revised form 23 February 2015

Keywords:
Bayesian network
Structure learning
Modeling
Algorithmic complexity

ABSTRACT

Currently, Bayesian Networks (BNs) have become one of the most complete, self-sustained and coherent formalisms used for knowledge acquisition, representation and application through computer systems. However, learning of BNs structures from data has been shown to be an NP-hard problem. It has turned out to be one of the most exciting challenges in machine learning. In this context, the present work's major objective lies in setting up a further solution conceived to be a remedy for the intricate algorithmic complexity imposed during the learning of BN-structure with a massively-huge data backlog.

1. Introduction

The huge amounts of data, made recently available pertaining to the various research fields, have made it crucially critical for the learning techniques to be efficient, in so far as the processing of complex data dependencies is concerned. Owing to their flexibility and easily-recognizable mathematical formulations, BNs [1] are most often the basic model opted for in a wide-array of application-fields whether astronomic, textual, bioinformatics and web-mining applications. Yet, with an incredibly huge number of variables, the learning of BNs structure from data remains a big challenge to be retained and considered in terms of calculation power, algorithmic complexity and execution time [2]. Recently, however, various algorithms have been developed with respect to the BNs learning structures from databases [3–5]. A considerable class of these algorithms rests on the metric-scoring methods, excessively compared and exhaustively applied as approaches [6,7]. Nevertheless, these algorithms and scoring methods remain insufficient with regard to those cases in which the number of variables exceeds hundreds of thousands [8]. As far as our work is concerned, these algorithms and metric scores are not going to be dealt with or questioned. Rather, we seek to further enrich them through a new heuristic method based on clustering pertaining to structure learning, that aims to further reduce the algorithmic complexity as well as the execution time; this will allow modeling some previously non-modelizeable information systems, using, exclusively, the underway available algorithms.

In fact, another type of algorithms has been devised, they use the Hierarchical Class-Latent Models (HCLM) [2], along with the double-layer BNs [9]. In fact, these algorithms are very promising in their capability to gradually reduce the scope of data dimensions despite their inability to treat a quite large number of variables exceeding one thousand [8]. In [10] an approach is proposed that enable to process a considerable amount of data (6000 variables) by reducing the HCLM's research space to binary trees expanded by the possible connections among brothers. Still, the required restrictions imposed by this approach

* Corresponding author at: Bioinformatics Unit, Centre of Biotechnologie of Sfax, Sfax University, Tunisia.
E-mail address: heni_bouhamed@yahoo.fr (H. Bouhamed).

http://dx.doi.org/10.1016/j.cam.2015.02.055
0377-0427/

are likely to deviate from this model to diverge away from reality [11]. Noteworthy, in this respect, the common point binding our heuristic and the already-mentioned methods lies in the tendency towards reducing the algorithmic complexity in a bid to cope with the huge number of variables (to be modeled). It is worth noting, however, that our conceived heuristic method does not target to reduce data in such a way as to noticeably engender a considerable loss of information. Rather, it tends to divide and dislocate the information system into sub-parts by separately treating each part's learning structure, prior to assembling them as a whole afterwards.

Our present work is critically important for a number of various reasons. First, we have managed to prove, throughout its scope, that by wholly subdividing an information system into sub-sets, we tend to dramatically reduce the number of possible structures necessary for learning the whole BNs structure. Second, a special heuristic method has been devised and proposed whereby this reduction that could be exploited without engendering any significant loss of data. Ultimately, while combining our proper heuristic method with the existing prevailing structure-learning algorithms, the extent of algorithmic complexity as well as the learning of BNs structure from data execution time can be reduced considerably, in such a way that even a large number of non-modelizeable variables could be treated or processed.

The remainder of this paper was arranged as follows. Section 2 deals with the BNs and their structure learning problems. In Section 3, we put forward a new heuristic method which was tested upon a car-diagnosis and Lymphography diagnosis data bases. Finally, the main conclusions of this study were drawn before suggesting some potential perspectives relevant for future research.

2. Structure learning of BN from data

It is worth highlighting that knowledge representation and the related reasoning, therefore, have given birth to numerous models. The graphic probability model, namely, BN, introduced by Judea Pearl in the 1980s, has been the source of some practical tools useful for the representation of uncertain knowledge and reasoning process from incomplete information.

We denote by $B = (G, \Theta)$ a BN such that:

- $G = (X, E)$ is a graph managed without circuit summits of whose associates are a set of random variables $X = \{X_1, \ldots, X_n\}$.
- $\Theta = \{P(X_i \,|P_a(X_i))\}$ is a probability set of every knot X_i conditional upon the state of its parents relatives $P_a(Xi)$ in G.

Hence, the BN graphic representation indicates the dependencies (or independencies) between variables and provides a visual knowledge representation tool, that turns out to be more easily understood by its users. Furthermore, the use of probability allows to take into account the uncertainty, through quantifying the dependencies between variables. These two properties have been at the origin of the first terms allocated, initially, to the BN, "probabilistic expert systems", where the graph used to be compared with some set of rules pertaining to a classic expert system, and conditional probability presented as a quantification measurement of the uncertainty related to these rules.

In this respect, Ref. [12] has shown that the Bayesian networks have allowed to represent, the joint probability distribution relevant to all variables, in a compact way:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \,|P_a(X_i)) \tag{1}$$

The number of all BN possible structures has been shown to ascend sharply as a super-exponential on the number of variables. Indeed, Ref. [13] derived the following recursive formula for the number of Directed Acyclic Graph (DAG) with n variables:

$$r(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-1)} r(n-i) = n^{2^{O(n)}} \tag{2}$$

which gives: $r(1) = 1, r(2) = 3, r(3) = 25, r(5) = 29\,281, r(10) = 4, 2 \times 10^{18}$.

This means that, it is impossible to perform an exhaustive search of all structures in a reasonable time in case the number of nodes exceeds seven. In fact, most structure-learning methods use heuristics to search the DAGs space.

3. A new clustering-based heuristic: theoretical framework and methodology

The idea lying behind our conceived procedure lies in rapid super-exponential surge of algorithmic complexity of learning BN structure from data with respect to the rise in the number of variables. To remediate this problem, our idea consists in subdividing the variables into subsets (or clusters), by learning structure of each cluster separately, while looking for a convenient procedure by which the different structures could be assembled into a final structure. In this regard, it has been noticed that in numerous information systems, so as not to say in most of them, there exists, at least, one single central variable of a global interest constituting the basis of the system modelization. In this respect, we intend to execute the processing of each cluster learning structure with the central interest variable, then, proceed by assembling the different various structures around this central variable as a next step.