Artificial Intelligence 174 (2010) 984–1006

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Understanding the scalability of Bayesian network inference using clique tree growth curves

Ole J. Mengshoel

Carnegie Mellon University, NASA Ames Research Center, Mail Stop 269-3, Moffett Field, CA 94035, United States

ARTICLE INFO

Article history:
Received 16 January 2009
Received in revised form 18 May 2010
Accepted 18 May 2010
Available online 25 May 2010

Keywords:
Probabilistic reasoning
Bayesian networks
Clique tree clustering
Clique tree growth
C/V-ratio
Continuous approximation
Gompertz growth curves
Controlled experiments
Regression

ABSTRACT

One of the main approaches to performing computation in Bayesian networks (BNs) is clique tree clustering and propagation. The clique tree approach consists of propagation in a clique tree compiled from a BN, and while it was introduced in the 1980s, there is still a lack of understanding of how clique tree computation time depends on variations in BN size and structure. In this article, we improve this understanding by developing an approach to characterizing clique tree growth as a function of parameters that can be computed in polynomial time from BNs, specifically: (i) the ratio of the number of a BN's non-root nodes to the number of root nodes, and (ii) the expected number of moral edges in their moral graphs. Analytically, we partition the set of cliques in a clique tree into different sets, and introduce a growth curve for the total size of each set. For the special case of bipartite BNs, there are two sets and two growth curves, a mixed clique growth curve and a root clique growth curve. In experiments, where random bipartite BNs generated using the BPART algorithm are studied, we systematically increase the out-degree of the root nodes in bipartite Bayesian networks, by increasing the number of leaf nodes. Surprisingly, root clique growth is well-approximated by Gompertz growth curves, an S-shaped family of curves that has previously been used to describe growth processes in biology, medicine, and neuroscience. We believe that this research improves the understanding of the scaling behavior of clique tree clustering for a certain class of Bayesian networks; presents an aid for trade-off studies of clique tree clustering using growth curves; and ultimately provides a foundation for benchmarking and developing improved BN inference and machine learning algorithms.

1. Introduction

Bayesian networks (BNs) play a central role in a wide range of automated reasoning applications, including in diagnosis, sensor validation, probabilistic risk analysis, information fusion, and decoding of error-correcting codes [64,6,59,38,37,60,43,58]. A crucial issue in reasoning using BNs, as well as in other forms of model-based reasoning, is that of computational scalability. Most BN inference problems are computationally hard in the general case [10,63,61,1], thus there may be reason to be concerned about scalability. One can make progress on the scalability question by studying classes of problem instances analytically and experimentally. Such problem instances may come from applications or they may be randomly generated. In the area of application BNs, both encouraging and discouraging scalability results have been reported. For example, a prominent bipartite BN for medical diagnosis is known to be intractable using current technology [64]. Decoding of error-correcting codes, which can be understood as BN inference, is also not tractable but has empirically been found to be solvable with high reliability using inexact BN inference [20,37]. On the other hand, it is well-known that BNs that are

E-mail address: Ole.Mengshoel@sv.cmu.edu.

0004-3702/$ – see front matter
doi:10.1016/j.artint.2010.05.007

tree-structured, including the so-called naive Bayes model, are solvable in polynomial time using exact inference algorithms. There are also encouraging empirical results for application BNs that are "close" to being tree-structured or more generally for application BNs that are not highly connected [26,43].

Clique tree clustering, where inference takes the form of propagation in a clique tree compiled from a BN, is currently among the most prominent BN inference algorithms [33,2,62]. The performance of tree clustering algorithms depends on a BN's treewidth or the optimal maximal clique size of a BN's induced clique tree [16,11,15]. The performance of other exact BN inference algorithms also depends on treewidth. A key research question is, then, how the size of a clique tree generated from a BN (and consequently, inference time) depends on structural measures of BNs. One way to investigate this is through the use of random generation from distributions of problem instances [66,5,11,52,23]. Taking this approach, and increasing the ratio C/V between the number of leaf nodes C and the number of root nodes V in bipartite BNs, an easy-hard-harder pattern along with approximately exponential growth have previously been observed for clique tree clustering for a certain class of BNs, namely BPART BNs [45].

In this article, we develop a more precise understanding of this easy-hard-harder pattern. This is done by formulating macroscopic and approximate models of clique tree growth by means of restricted growth curves, which we illustrate by using bipartite BNs created by the BPARTalgorithm [45]. For the sake of this work, we assume that a clique tree propagation algorithm, operating on a clique tree compiled from a BN, is executed in order to answer probabilistic queries of interest. We introduce a random variable for total clique tree size. This random variable is, for the case of bipartite BNs, the sum of two random variables, one for the size of root cliques and one for the size of mixed cliques. Reflecting the random variable for total clique tree size, we introduce a continuous growth curve for total clique tree size which is the sum of growth curves for the size of root cliques and mixed cliques. Of particular interest is the growth curve for root clique size, where Gompertz curves of the form $g(\infty)e^{-\zeta e^{-\gamma x}}$, where $g(\infty)$, ζ, and γ are parameters, turn out to be useful. A key finding is that Gompertz growth curves are justified on theoretical grounds and also fit very well to experimental data generated using the BPART algorithm [45]. While we emphasize bipartite BNs in this article, we also discuss how to generalize to arbitrary BNs, by using multiple growth curves or translating arbitrary BNs to bipartite BNs via factor graphs [32,70].

For experimentation, we sampled bipartite BNs using an implementation of the BPART algorithm. For the number of root nodes, V, we used $V = 20$ and $V = 30$. The number of leaf nodes was also varied, thereby creating BNs of varying hardness; 100 BNs per C/V-level were randomly generated. A clique tree inference system, employing the minimum fill-in weight heuristic, was used to generate clique trees for the sampled BNs. Let W be a random variable representing the number of moral edges in moral graphs induced by random BNs. In addition to $x = C/V$, we consider $x = E(W)$ as an independent variable. In experiments, we compared different growth curves and investigated $x = C/V$ versus $x = E(W)$ as independent variables for Gompertz growth curves. Linear regression was used to obtain values for the parameters ζ and γ based on a linear form of the Gompertz growth curve; values for $g(\infty)$ were obtained by analysis. Gompertz growth curves are common in biological, medical, and neuroscience research [4,35,17], but have not previously been used to characterize clique tree growth (except for in our earlier conference paper [41] which this article extends). We provide improved results compared to previous research, where an easy-hard-harder pattern and approximately exponential growth of upper bounds on optimal maximal clique size as a function of C/V-ratio were established [45].

We believe this research is significant for the following reasons. First, analytical growth curves improve the understanding of clique tree clustering's performance for a certain class of BNs, namely BPART BNs. Consider Kepler's three laws of planetary motion, developed using Brahe's observational data of planetary movement. There is a need to develop similar laws for clique tree clustering's performance, and in this article we obtain such laws in the form of Gompertz growth curves for BPART BNs [45]. While they admittedly have a strong empirical basis, these Gompertz growth curves give significantly better fit to the raw data than alternative curves. Consequently, they provide better insight into the underlying mechanisms of the clique tree clustering algorithm and may be used to approximately predict the performance of the algorithm. Since the performance of other exact BN inference algorithms — including conditioning [55,11] and elimination algorithms [34,71, 14] — also depends on optimal maximal clique size, our results may have significance for these algorithms as well. A second benefit of growth curves is that they can be used to summarize performance of different BN inference algorithms or different implementations of the same algorithm on benchmark sets of problem instances, and thereby aid in evaluations.[1] Suppose that the growth curves $g_1(x)$ and $g_2(x)$ were obtained by benchmarking slightly different clique tree algorithms. Compared to looking at and evaluating potentially large amounts of raw data, it may be easier to understand the performance difference between the two algorithms by studying their curves $g_1(x)$ and $g_2(x)$ or by comparing their respective Gompertz curve parameter values ζ_1 and γ_1 versus ζ_2 and γ_2. A third benefit is that growth curves provide estimates of resource consumption in terms of clique tree size, estimates that can easily be translated into requirements on memory size and inference time. Hence, this approach enables trade-off studies of resource consumption (or requirements) versus resource bounds, which is important in resource-bounded reasoners [48,40], and may also be of use if one wants to take into account, during BN structure learning, the computational resources needed for reasoning.

The rest of this article is organized as follows. After introducing notation and background concepts related to Bayesian networks and clique tree clustering in Section 2, we study, in the context of the BPART algorithm, the development and

[1] Such evaluations have been performed, for example, at recent UAI conferences, see http://ssli.ee.washington.edu/~bilmes/uai06InferenceEvaluation/ and http://graphmod.ics.uci.edu/uai08/ for details. Application BNs for benchmarking can be found at http://genie.sis.pitt.edu/networks.html and http://www.cs.huji.ac.il/labs/compbio/Repository/.