Artificial Intelligence 218 (2015) 56–73

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Most frugal explanations in Bayesian networks

Johan Kwisthout *

Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, The Netherlands

A R T I C L E   I N F O

Article history:
Received 1 October 2013
Received in revised form 13 October 2014
Accepted 16 October 2014
Available online 22 October 2014

Keywords:
Bayesian abduction
Parameterized complexity
Approximation
Heuristics
Computational complexity

A B S T R A C T

Inferring the most probable explanation to a set of variables, given a partial observation of the remaining variables, is one of the canonical computational problems in Bayesian networks, with widespread applications in AI and beyond. This problem, known as MAP, is computationally intractable (NP-hard) and remains so even when only an approximate solution is sought. We propose a heuristic formulation of the MAP problem, denoted as Inference to the Most Frugal Explanation (MFE), based on the observation that many intermediate variables (that are neither observed nor to be explained) are irrelevant with respect to the outcome of the explanatory process. An explanation based on few samples (often even a singleton sample) from these irrelevant variables is typically almost as good as an explanation based on (the computationally costly) marginalization over these variables. We show that while MFE is computationally intractable in general (as is MAP), it can be tractably approximated under plausible situational constraints, and its inferences are fairly robust with respect to which intermediate variables are considered to be relevant.

1. Introduction

Abduction or inference to the best explanation refers to the process of finding a suitable explanation (the *explanans*) of observed data or phenomena (the *explananda*). In the last decades, Bayesian notions of abduction have emerged due to the widespread popularity of Bayesian or probabilistic techniques for representing and reasoning with knowledge [5,26,30,47,52]. They are used in decision support systems in a wide range of problem domains (e.g., [7,11,21,23,32,45,64]) and as computational models of economic, social, or cognitive processes [10,25,33,48,58,60]. The natural interpretation of 'best' in such models is 'most probable': the explanation that is the most probable one given the evidence, i.e., that has maximum posterior probability, is seen as the hypothesis that best explains the available evidence; this explanation is traditionally referred to as the MAP explanation and the computational problem of inferring this explanation as the MAP problem.[1]

However, computing or even approximating the MAP explanation is computationally costly (i.e., NP-hard), especially when there are many intermediate (neither observed nor to be explained) variables that may influence the explanation

* Correspondence to: Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, PO Box 9104, 6500HE Nijmegen, The Netherlands. Tel.: +31 0 24 3616288.

E-mail address: j.kwisthout@donders.ru.nl.

[1] Other relationships have been proposed that compete in providing 'sufficiently rational' relations between observed phenomena and their explanation that can be used to describe *why* we judge one explanation to be preferred over another [28,44]. Examples include *maximum likelihood* [29], which does not take the prior probabilities of the hypotheses into account, the *conservative Bayesian* approach [6], *generalized Bayes* factor [66], and various Bayesian formalisms of *coherence theory* [5,15,26,49]. While the posterior probability of such explanations is not the deciding criterion to prefer one explanation over another, it is typically so that explanations we consider to be good for other reasons also have a high posterior probability compared to alternative explanations [27,44].

http://dx.doi.org/10.1016/j.artint.2014.10.001

[1,4,51,56]. To compute the posterior probability distribution of the explanation variables, all these intermediate variables need to be marginalized over. One way of dealing with this intractability might be by assuming modularity of knowledge representations, i.e., by assuming that these representations are informationally encapsulated and do not have access to background knowledge. However, this is problematic as we cannot know beforehand which elements of background knowledge or observations may be relevant for determining the best explanation [17,19].

Fortunately, even when a full Bayesian computation may not be feasible in large networks, we need not constrain inferences only to small or disconnected knowledge structures. It is known that in general the posterior probability distribution of a (discrete) Bayesian network is skewed, i.e., few joint value assignments together cover most of the probability space [13], and that typically only a few of the variables in a network are relevant for a particular inference query [14]. We propose to utilize this property of Bayesian networks in order to make tractable (approximate) inferences to the best explanation over large and unencapsulated knowledge structures. We introduce a heuristic formulation of MAP, denoted as Inference to the Most Frugal Explanation (MFE), that is explicitly designed to reflect that only a few intermediate variables are typically relevant in real-world situations. In this formulation we partition the set of intermediate variables in the network into a set of 'relevant' intermediate variables that are marginalized over, and a set of 'irrelevant' intermediate variables that we sample from in order to estimate an explanation.

Note that in the MFE formalism there is marginalization over *some* of the intermediate variables (the variables that are considered to be relevant), but not over those intermediate variables that are not considered to be relevant. Thus, MFE can be seen as a 'compromise' between computing the explanation with maximum posterior probability, where one marginalizes over all intermediate variables, and the previously proposed Most Simple Explanation (MSE) formalism [35] where there is no marginalization at all, i.e., all intermediate variables are seen as irrelevant. We want to emphasize that the notions 'relevant' and 'irrelevant' in the problem definition denote *subjective* partitions of the intermediate variables; we will revisit this issue in Section 3.1.

We claim that this heuristic formalism of the MAP problem exhibits the following desirable properties:

1. The knowledge structures are *isotropic*, i.e., they are such that, potentially, everything can be relevant to the outcome of an inference process. They are also *Quinean*: candidate explanations are sensitive to the entire belief system [17,18].
2. The inferences are provably computationally tractable (either to compute exactly or to approximate) under realistic assumptions with respect to situational constraints [43,53].
3. The resulting explanations have an optimal or close-to-optimal posterior probability in many cases, i.e., MFE actually 'tracks truth' in the terms of Glass [28].

In the remainder of this paper, we will discuss some needed preliminaries in Section 2. In Section 3 we discuss MFE in more detail. We give a more formal definition, including a formal definition of relevance in the context of Bayesian networks, and show how MFE can be tractably approximated under realistic assumptions despite computational intractability of the problem in general. In Section 4 we show that MFE typically gives an explanation that has a close-to-optimal posterior probability, even if only a subset of the relevant variables is considered. We discuss how MFE performs under various scenarios (e.g., when there are few or many relevant variables, when there are many hypotheses that are almost equally likely, or when there is a misalignment between the *actual* relevant variables and the variables that are mistakenly presumed to be relevant). We conclude our paper in Section 5.

2. Preliminaries

In this section we will introduce some preliminaries from Bayesian networks, in particular the MAP problem as standard formalization of Bayesian abduction. We will discuss the ALARM network which we will use as a running example throughout this paper. Lastly, we introduce some needed concepts from complexity theory, in particular the complexity class PP, oracles, and fixed parameter tractability.

2.1. Bayesian networks and Bayesian abduction

A Bayesian or probabilistic network $\mathcal{B}$ is a graphical structure that models a set of stochastic variables, the conditional independences among these variables, and a joint probability distribution over these variables [52]. $\mathcal{B}$ includes a directed acyclic graph $\mathbf{G}_{\mathcal{B}} = (\mathbf{V}, \mathbf{A})$, modeling the variables and conditional independences in the network, and a set of parameter probabilities Pr in the form of conditional probability tables (CPTs), capturing the strengths of the relationships between the variables. The network models a joint probability distribution $\Pr(\mathbf{V}) = \prod_{i=1}^{n} \Pr(V_i \mid \pi(V_i))$ over its variables, where $\pi(V_i)$ denotes the parents of V_i in $\mathbf{G}_{\mathcal{B}}$. We will use upper case letters to denote individual nodes in the network, upper case bold letters to denote sets of nodes, lower case letters to denote value assignments to nodes, and lower case bold letters to denote joint value assignments to sets of nodes. We will sometimes write $\Pr(\mathbf{x} \mid \mathbf{y})$ as a shorthand for $\Pr(\mathbf{X} = \mathbf{x} \mid \mathbf{Y} = \mathbf{y})$ if no ambiguity can occur.

In a Bayesian abduction task there are three types of variables: the *evidence* variables, the *explanation* variables, and a set of variables called *intermediate* variables that are neither evidence nor explanation variables. The evidence variables are instantiated, i.e., have been assigned a value; the joint value assignment constitutes the explananda (what is to be explained,