ARTICLE IN PRESS
IJA:7945

International Journal of Approximate Reasoning ••• (••••) •••–•••

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A two-phase method for extracting explanatory arguments from Bayesian networks

Sjoerd T. Timmer [a,*], John-Jules Ch. Meyer [a], Henry Prakken [a,b], Silja Renooij [a], Bart Verheij [c]

[a] *Utrecht University, Department of Information and Computing Sciences, The Netherlands*
[b] *University of Groningen, Faculty of Law, The Netherlands*
[c] *University of Groningen, Artificial Intelligence Institute, The Netherlands*

A R T I C L E   I N F O

Article history:
Received 6 November 2015
Received in revised form 5 September 2016
Accepted 6 September 2016
Available online xxxx

Keywords:
Bayesian networks
Argumentation
Probabilistic reasoning
Explanation
Inference
Uncertainty

A B S T R A C T

Errors in reasoning about probabilistic evidence can have severe consequences. In the legal domain a number of recent miscarriages of justice emphasises how severe these consequences can be. These cases, in which forensic evidence was misinterpreted, have ignited a scientific debate on how and when probabilistic reasoning can be incorporated in (legal) argumentation. One promising approach is to use Bayesian networks (BNs), which are well-known scientific models for probabilistic reasoning. For non-statistical experts, however, Bayesian networks may be hard to interpret. Especially since the inner workings of Bayesian networks are complicated, they may appear as black box models. Argumentation models, on the contrary, can be used to show how certain results are derived in a way that naturally corresponds to everyday reasoning. In this paper we propose to explain the inner workings of a BN in terms of arguments.

We formalise a two-phase method for extracting probabilistically supported arguments from a Bayesian network. First, from a Bayesian network we construct a *support graph*, and, second, given a set of observations we build arguments from that support graph. Such arguments can facilitate the correct interpretation and explanation of the relation between hypotheses and evidence that is modelled in the Bayesian network.

1. Introduction

Bayesian networks (BNs), which model probability distributions, have proven value in several domains, including medical and legal applications [1,2]. However, the interpretation and explanation of Bayesian networks is a difficult task, especially for domain experts who are not trained in probabilistic reasoning [3]. Legal experts, for example, such as lawyers and judges, may be more accustomed to argumentation-based models of proof because probabilistic reasoning is often considered a difficult task [4,5]. Recently, a scientific interest in combining argumentation-based models of proof with probabilities has arisen [6–10]. One possible combination is the use of argumentation to explain probabilistic reasoning. Argumentation is a well studied topic in the field of artificial intelligence (see chapter 11 of [11] for an overview). Argumentation theory provides models that describe how conclusions can be justified. These models closely follow the reasoning patterns present in human reasoning. This makes argumentation an intuitive and versatile model for common sense reasoning tasks.

* Corresponding author.
E-mail address: s.t.timmer@uu.nl (S.T. Timmer).

http://dx.doi.org/10.1016/j.ijar.2016.09.002

Please cite this article in press as: S.T. Timmer et al., A two-phase method for extracting explanatory arguments from Bayesian networks, Int. J. Approx. Reason. (2016), http://dx.doi.org/10.1016/j.ijar.2016.09.002

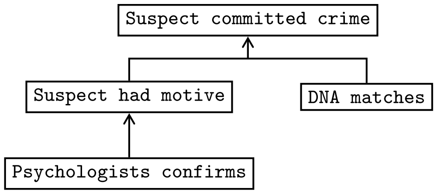

Fig. 1. An example of a complex argument. Every box represents one argument and the arrows show how subarguments support conclusions.

Argumentative explanations of Bayesian reasoning may prove helpful to interpret probabilistic reasoning in legal cases. Existing explanation methods for BNs can broadly be divided in two categories. First, the model itself can be explained. See, for instance, the work of Lacave and Dièz or Koiter [12,13]. Secondly, the evidence can be explained by calculating the so-called *most probable explanation* (MPE) or *maximum a-posteriori probability* (MAP) which is the most likely configuration of a (sub)set of non-evidence variables [14]. A MAP/MPE helps to explain the evidence, but does not explain why the posterior probabilities of variables of interest are high or low nor do they explain the reasoning steps between evidence and hypotheses. In this paper we take a third approach to explaining, which is to explain the derivation of probabilities resulting from the calculations in the BN and explain those using *reasoning chains* that have a clear argumentative interpretation. This resembles the work of Suermondt [15] although that does not apply argumentation, and the work of Schum [16] which is an informal approach to explaining Bayesian networks in argumentative terms. We formalise a method for extracting arguments from a BN, in which we first extract an intermediate support structure, which subsequently guides the argument construction process. This results in numerically backed arguments based on probabilistic information modelled in a BN. We apply our method to a legal example but the approach does not depend on this domain and can also be applied to other fields where BNs are used. Our method thus serves as a general explanation method for BNs.

In earlier work [17] we introduced the notions of probabilistic rules and arguments and a simple algorithm to extract those from a BN. For larger networks, however, this algorithm, which exhaustively enumerates every possible probabilistic rule and argument, is computationally infeasible because it examined inferences between all combinations of variable assignments. We improve on this by searching for explanations in nearby nodes only. Moreover, the algorithm from [17] does unnecessary work because many of the enumerated antecedents will never be met, resulting in irrelevant rules. Similarly, many arguments constructed in this way are superfluous because they argue for irrelevant conclusions from which no further inference is possible. Improving on this work, we proposed a new method that addresses these issues [18]. In this method, the process of argument generation is split into two phases: from the BN, first, a *support graph* is constructed for a variable of interest, from which arguments can be generated in a second phase. This eliminates the aforementioned problem of unnecessarily enumerating irrelevant rules and arguments. As a side-effect this also has the advantage that the support graph is independent of the evidence. When observations are added to the BN, only the resulting argumentation changes. In [18] we introduced an algorithm for the first phase but the second phase was only described informally. In [19] we further formalised the support graph generation phase and we proved a number of properties of this formalism. The current paper further extends [19]. Extensions include the addition of a more elegant and intuitive definition of support graphs and a proof that our algorithm correctly computes such a graph. We have also added a more detailed discussion of the support graph and argument construction method using small examples. Furthermore, we have formalised the second (argument generation) phase and added a case study using an example BN from the literature.

In Section 2 we will present backgrounds on argumentation and BNs. In Section 3 we formally define and discuss support graphs. Using the notion of a support graph we introduce a formalisation of argument construction in Section 4. We apply this method in a case study in Section 5.

2. Preliminaries

2.1. Argumentation

In argumentation theory, one possibility to deal with uncertainty is the use of defeasible inferences. A defeasible rule (as opposed to a strict, or deductive, inference rule) can have exceptions. In a defeasible rule the antecedents do not conclusively imply the consequence but rather create a presumptive belief in it. Using (possibly defeasible) rules, arguments can be constructed. Fig. 1, for instance, shows an argument graph with three nested arguments connected by two rules. From a psychological report it is derived that the suspect had a motive and together with a DNA match this is reason to believe that the suspect committed the alleged crime.

Argumentation can be used to model conflicting or contradictory information. This is modelled by attack between arguments. Undercutting and rebutting attacks between arguments with defeasible rules have been distinguished [20]. A rebuttal attacks the conclusion of an argument, whereas an undercutter directly attacks the inference. An undercutter exploits the fact that a rule is not strict by posing one of the exceptional circumstances under which it does not apply. In this paper we do not use undercutting and *undermining*, which is the third form of attack that can be present in the general case of ASPIC+. The attack relation between arguments can be analysed and from it the acceptability of arguments can be determined.

Please cite this article in press as: S.T. Timmer et al., A two-phase method for extracting explanatory arguments from Bayesian networks, Int. J. Approx. Reason. (2016), http://dx.doi.org/10.1016/j.ijar.2016.09.002