Data & Knowledge Engineering 73 (2012) 73–88

Contents lists available at SciVerse ScienceDirect

Data & Knowledge Engineering

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

An ontology-based approach for constructing Bayesian networks

Stefan Fenz *

SBA Research, Favoritenstrasse 16, 2nd floor, 1040 Vienna, Austria
Vienna University of Technology, Favoritenstrasse 9-11/E188, 1040 Vienna, Austria

ARTICLE INFO

Article history:
Received 21 April 2011
Received in revised form 13 December 2011
Accepted 14 December 2011
Available online 22 December 2011

Keywords:
Bayesian network
Ontology
Knowledge re-use

ABSTRACT

Bayesian networks are commonly used for determining the probability of events that are influenced by various variables. Bayesian probabilities encode degrees of belief about certain events, and a dynamic knowledge body is used to strengthen, update, or weaken these assumptions. The creation of Bayesian networks requires at least three challenging tasks: (i) the determination of relevant variables (nodes), (ii) the determination of relationships between the identified variables (links), and (iii) the calculation of the conditional probability tables (CPTs) for each node in the Bayesian network. Based on existing domain ontologies, we propose a method for the ontology-based construction of Bayesian networks. The method supports (i) the construction of the graphical Bayesian network structure (nodes and links), (ii) the construction of CPTs that preserve semantic constraints of the ontology, and (iii) the incorporation of already existing knowledge facts (findings). The developed method enables the efficient construction and modification of Bayesian networks based on existing ontologies.

1. Introduction

Bayesian networks (also referred to as probabilistic networks or belief networks) have been applied to several real-world applications such as medical diagnosis and prognosis, spam analysis, information retrieval, and natural language processing (cf. [8,24,16,30,21]). According to Neapolitan [24], Bayesian networks are '*graphical structures for representing the probabilistic relationships among a large number of variables and doing probabilistic inference with those variables*'. According to Heckerman et al. [15], the Bayesian probability of an event *x* is a person's degree of belief in the very event. Therefore, the frequentistic probability class (e.g., the probability that a coin lands heads or tails, based on a large number of trials tending to infinity) is distinguished from the Bayesian (personal) probability, in which probabilities encode degrees of belief about certain events and data is used to strengthen, update, or weaken these assumptions (e.g., the degree of belief that the coin will land heads at the next throw) [27]. The following challenges arise at the construction of Bayesian networks:

- identification of variables that are relevant to the considered domain (nodes),
- identification of relationships between the identified variables (links), and
- creation of the conditional probability table for each variable.

The conditional probability table (CPT) is used to express how the potential states of the node's parents affect the posterior probability of the considered node. In general, two different methods or a combination of both are used to construct a Bayesian network: (i) automated construction of Bayesian networks from existing data (cf. [3]), and (ii) the domain expert-based construction of Bayesian networks covering complex knowledge domains with insufficient or non-existing empirical data regarding relevant variables. Both approaches have their specific shortcomings. While the domain expert-based approach is time-consuming and error-prone (cf. [8]), the automated construction of Bayesian networks from existing data faces a bias problem.

* Tel.: +43 1 5053688.
E-mail address: sfenz@sba-research.org.

0169-023X/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2011.12.001

The main problem is that the Bayesian network is constructed from a limited subpopulation (e.g., a medical database of a specific hospital covering the treated patients) to be applied to a large target population (e.g., the entire population of a country or continent). Therefore, automated construction methods based on existing data are often insufficient in practice [9].

To address these problems, Druzdzel et al. emphasize in [8] the need for approaches aimed at reducing the number of probabilities to be assessed and tools for supporting the quantification task in the construction of Bayesian networks.

According to the ontology definition given by Neches, ontologies are a potential solution to support the construction of Bayesian networks: *an ontology defines the basic terms and properties comprising the vocabulary of a topic area as well as the rules for combining terms and properties to define extensions to the vocabulary* [25]. An ontology's classes, properties, and individuals are widely used to capture and represent a given knowledge domain. Since Bayesian networks use nodes and links to represent a given domain, we can use the formally defined semantics of ontology classes, individuals, and properties to automatically construct the graphical structure (nodes and links) of Bayesian networks. Furthermore, ontologies can be used to automatically construct CPTs based on expert knowledge modeled in the ontology. Although experts would still be required to transfer their knowledge into a machine-readable format, the ontology enables them to model the knowledge in a way that is more aligned to the human way of thinking and reasoning engines can be used to infer new knowledge from already known facts. Furthermore, we may be able to reduce the time required to assess the necessary knowledge by using already existing domain ontologies. As previously noted, it is important to stress that the concepts of ontologies and Bayesian networks are not equivalent. While ontologies describe domain concepts and their interrelationships, Bayesian networks describe probability distributions and their interrelationships. For this reason, fully automated 1:1 mapping is not possible. However, real-world scenarios such as the use case described in this article have shown that ontologies already describing a certain domain can be valuable in the construction of Bayesian networks to represent causalities in that same domain. This work does not aim at constructing ontologies solely for the construction of Bayesian networks. Instead, it provides an approach that efficiently reuse knowledge that has already been modeled in an ontology in the construction of Bayesian networks. In such a setting, ontology tools can be used to manage the actual knowledge and therefore make the collaborative editing of the Bayesian network structures (e.g., the introduction of new variables) easier. To exploit the capabilities of ontologies for the efficient construction of Bayesian networks, a method for the ontology-based construction of Bayesian networks is required.

Most approaches in the field of ontology-based construction of Bayesian networks do not provide a method for using existing ontologies without specific extensions for the Bayesian network construction (cf. [6,28,34,5,35]). Approaches that do not require these extensions do not support the automated incorporation of findings (concrete knowledge) and conditional probability tables (CPTs) (cf. [17]). While general approaches for constructing CPTs exist (cf. [33,7,29,10]), existing ontology-based approaches require substantial ontology modifications to enable automated CPT computation (see Section 2 for further details). Therefore, the research questions are:

- RQ1: How can existing ontologies be used to construct the graphical structure of Bayesian networks?
- RQ2: How can existing ontologies be used to enrich Bayesian networks with CPTs and concrete findings while preserving the semantic constraints of the ontology?

Using existing domain ontologies, this article presents a generic method for the ontology-based construction of Bayesian networks by (i) using ontology classes/individuals to create the nodes of the Bayesian network, (ii) using ontology properties to link the Bayesian network nodes, (iii) utilizing the ontological knowledge base to support the conditional probability table calculation for each node, and (iv) enriching the Bayesian network with concrete findings from existing domain knowledge. The developed method enables the semi-automatic construction and modification of Bayesian networks based on existing ontologies. The method is demonstrated on the example of threat probability determination, which uses a security ontology as its underlying formal knowledge base.

This paper is organized as follows: The first part reviews related work and introduces the fundamentals of ontologies and Bayesian networks. The second part describes the proposed ontology-based construction of Bayesian networks in detail. The paper concludes with a prototype implementation of the proposed approach and a proof of concept conducted within the information security domain.

2. Related work

Before describing existing ontology-based Bayesian network approaches, we will provide a brief overview of approaches that can be used to learn Bayesian networks from data. Both fields share the same goal but differ substantially in the way they achieve that goal.

Lam and Bacchus [22] developed a Minimal Description Length (MDL) principle-based approach to learn Bayesian networks from raw data. The approach does not require any prior assumptions about the distributions and allows a trade-off between accuracy and complexity in the learned model. Larrañaga et al. [23] proposed a Bayesian network learning approach that uses a database of cases as input and generic algorithms to search for the best ordering of the system variables. Friedman and Koller [13] presented a Bayesian approach to structure discovery in Bayesian networks. Based on a given data set the approach expresses posterior distributions over Bayesian network structures and evaluates the posterior probability of important structural features of the distribution. Hruschka et al. [20] proposed a low computational complexity method based on feature ranking algorithms that can be used to define efficient variables ordering in the Bayesian network learning context.

The described approaches are just a small subset of available approaches to learn Bayesian networks from data. However, this brief overview shows that this type of approach relies on raw data to construct valid Bayesian networks. In some cases there is no raw data but concepts and relations between these concepts have already been defined. Ontologies are an example of such a case.