INCORPORATING TEXTUAL INFORMATION INTO BAYESIAN NETWORKS

András MILLINGHOFFER
Advisors: Tadeusz DOBROWIECKI, Péter ANTAL

I. Introduction

Today, Bayesian networks (BNs) are one of the most popular tools for representing and handling uncertainty. This is primarily due to that (1) they can expressively represent human knowledge and (2) can normatively combine it with observations. However, the learning of Bayesian networks from data is often inapplicable because of the lack or the high cost of data, and the evaluation of the resulting posterior distribution also can be problematic. To come around this difficulty, the paper proposes a text-mining method through which we can construct the structure of the domain model, which can be used directly or as a starting point (prior) for further refinement.

We also propose an extension to Bayesian networks, through which we gain an annotated knowledge representation method able to handle queries containing textual information concerning the domain.

II. Bayesian networks

Bayesian networks (see [1]) model the relevant quantities of the world as probabilistic variables. Our knowledge about the domain is represented by their joint distribution. A BN consists of two parts: (1) the directed acyclic graph G represents the variables with its nodes and direct probabilistic dependencies with its edges; (2) the local *conditional probability distributions* associated to each node, describe the node's distribution conditional on its parents in the graph. The joint distribution can be computed by multiplying the local models: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Parents(X_i))$.

A BN can be interpreted as (1) an effective representation of the joint distribution, (2) the map of probabilistic dependencies of the domain, or (3) the causal description of the domain, with edges representing direct cause-effect connections.

III. Text mining with Bayesian networks

Since the data needed for learning are often inaccessible or very expensive and expert provided information is difficult to handle for a knowledge engineer, a third source of information, namely scientific publications would be worth considering. An ambitious goal could be to build models based on textual data that represent a background knowledge comparable to those provided by the experts, i.e. to extract knowledge encoded in articles and papers, or even to determine the direction of further research into the domain (see [2]).

A. Causality appearing in Bayesian networks

The basic ideas about the connection between publishing and domain mechanisms are the followings. As the exploration of a research area proceeds, the considerations of the publications change. The main phases are: (1) settling the relevant factors (variables), (2) exploring associative or causal connections, and (3) determining numeric parameters.

We are mainly interested in processing papers of the second class: since these consider connections between variables, we await that those variables will appear together in a paper which depend on each other. This suggests that the dependency structure of the distribution describing the co-appearance of the concepts will be similar to that of the real-world domain. Hence if we can learn the generative

models of publications (w.r.t. what variables are mentioned together), then these models will fit the corresponding real-world area as well.

B. Learning domain models based on textual data

The main steps of learning are, as follows:

- Input data are: free-text articles and the set of short "kernel" descriptions of the variables.
- These are converted into binary vectors representing which words are contained in them.
- The relevance of a keyword to an article is determined by the similarity of the vector of its kernel description and the one of the article. Based on this, we assign to each article a binary vector representing which variables are relevant to it (for details see [3]).
- The model structure can now be learned by any standard learning algorithm, see e.g. [4].

C. Comparing results with expert knowledge

The structure of the network encode qualitative relations of the domain. To evaluate a model w.r.t. an other, we may consider how many of the pairwise causal relations of the nodes remained in it w.r.t. the model provided by the expert. The possible relations of two nodes are (in weakening order): (1) there is an edge between the variables, (2) there is a directed path between the variables, (3) the two variables have a common ancestor, and (4) none of the previous. The ideas of this section were discussed in details in [5].

IV. Annotated Bayesian networks

As we have seen above, prior knowledge provided by experts can take an important role in model construction. The basic idea of annotated Bayesian networks is to extend BNs by assigning textual descriptions to nodes and/or structures. Using these annotations, we can formulate expressions like: $\forall X_1, X_2$*: the annotations of* X_1 *and* X_2 *contain* $string_1$ *and* $string_2$*, and* $X_1 = X_2$ or regarding the structure $\forall X_1, X_2$*: if the annotations of* X_1 *and* X_2 *contain* $string_1$ *and* $string_2$*, then there is a directed path between* X_1 *and* X_2.

Since a BN defines a distribution over atomic events (the full instantiations of the nodes), and there exists a posterior distribution over structures conditional on observations ($P(G|D)$), through the equation $P(expr|D) = \sum_{G:\ expr\ is\ true\ in\ G} P(G|D)$ we can compute the probability of any such expressions being true, assuming that only non-textual objects are quantified.

Annotated Bayesian networks provide a first-order, yet finite extension of BNs, capable of incorporating textual information concerning the variables (like kernel descriptions) or even the domain itself (annotations of networks structures). Hence, they provide a computationally tractable, free-text-based first-order knowledge representation language, able to coherently deal with uncertainty through probabilities. For details, see [6].

References

[1] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Francisco, CA, 1988.

[2] C. Yoo, V. Thorsson, and G. Cooper, "Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational dna microarray data," in *Proceedings of Pacific Symposium on Biocomputing*, vol. 7, pp. 498–509, 2002.

[3] P. Antal, G. Fannes, Y. Moreau, D. Timmerman, and B. D. Moor, "Using literature and data to learn Bayesian networks as clinical models of ovarian tumors," *Artificial Intelligence in Medicine*, 30, 2004.

[4] D. Heckerman, D. Geiger, and D. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, 20:197–243, 1995.

[5] P. Antal and A. Millinghoffer, "Learning causal Bayesian networks from literature data," in *Proceedings of the 3rd International Conference on Global Research and Education, Inter-Academia'04*, 2004.

[6] P. Antal and A. Millinghoffer, "A probabilistic knowledge base using annotated Bayesian network features," in *Proc. of the 6th Int. Symp. of Hungarian Researchers on Computational Intelligence*, pp. 366–377, 2005.