Using Literature and Data to Annotate and Learn Bayesian Networks

Peter Antal [a] Geert Fannes [a] Yves Moreau [a]
Bart De Moor [a]

[a] Electrical Eng. Dept., SCD, KULeuven, Kasteelpark Arenberg 10, 3001 Leuven

Abstract

The increasing availability of electronic literature opens up the possibility of using it as prior knowledge when dealing with complex statistical models where data is scarce or high levels of noise are present. This raises the question of how to actually perform the integration of domain literature with statistical data. In this paper, we assume that our textual information consists of short free-text descriptions of the domain variables and, optionally, of a large repository of related domain literature while the statistical data are classical vectors of observations. Because of their explicit semantics and powerful computational methods, Bayesian networks are suitable candidates for the integration of prior domain knowledge and statistical data. To connect our computational model with textual domain knowledge, we define an extended representation of Bayesian networks called the Annotated Bayesian Network. We introduce a text-based prior for the evaluation of Bayesian network substructures. We derive a text-based prior distribution over the space of Bayesian network structures and update this to a posterior with data. We evaluate our methodology in the ovarian cancer domain. In the case where only a small number of samples is available (which is a common situation in medical and biological applications), the prior derived from the literature increases the classification performance.

1 Introduction

The availability of increasingly large quantities of electronic literature begs the question of how to use both domain knowledge and data efficiently in knowledge engineering, machine learning, and decision support — theoretically as well as practically. This challenge is particularly acute in the complex and rapidly changing domains of medicine and genomics. Here, because of the inherent complexity, much of the voluminous domain knowledge is only available as free-text, scattered throughout the literature. The efficient access and incorporation of domain literature requires richer representations and beter integrated techniques than those used for the integration of the human expertise and data. Similarly, statistical models enriched with textual domain knowledge offer new types of priors for machine learning and new possibilities for decision support — for example, to generate enriched explanations.

In this paper we describe an extension of Bayesian networks called the *Annotated Bayesian network* (ABN) and introduce a language for information retrieval with ABNs, which links the information retrieval with the building of Bayesian networks. Furthermore, based on the free-text annotations of the variables, we define a prior probability for the Bayesian network structure and present results about the effect of text-based priors on structure learning in a classification task.

The paper is organized as follows. Section 2 introduces the Annotated Bayesian network and Section 3 describes an ABN in the ovarian cancer domain. Section 4 introduces the annotation-based priors for Bayesian network structures. Section 5 reports results about the effect of text-based priors on structure learning in the ovarian cancer domain. Section 6 contain the discussion and conclusion.

2 Annotated Bayesian networks

A Bayesian network represents a joint probability distribution over a set of variables by exploiting conditional independence relations. The model decomposes into a graphical part (a directed acyclic graph) and a numerical part (local dependency models). The vertices of the directed acyclic graph represent the random variables X_i and the edges define the dependency relations (each variable is independent of its non-descendants given its parents [6]). Additionally, there is a dependency model for every vertex (i.e., for the corresponding variable) to describe its dependency on the parent vertices (i.e., on the corresponding variables).

An extension of the Bayesian network representation for probability distributions, the Annotated Bayesian Networks (ABN), allows the definition of a hierarchy of classes over the domain variables and the attachment of free-text to the objects in the representation, such as values, variables, edges, dependency models and classes [2]. This ABN representation and the corresponding software environment provided the formal framework and the experimental environment for our investigations.

3 ABNs in the ovarian cancer domain

We perform the investigations in the ovarian cancer domain in the context of the International Ovarian Tumor Analysis (IOTA)[1], which is a multicentre study on the preoperative characterization of ovarian tumors based on artificial intelligence models. This study includes the multicentre collection of patient data and the corresponding data collection protocols. These descriptions provide a consistent source of domain literature for our experiments.

[1]http://www.iota-group.org

3.1 Sources of annotations

To ensure objective annotations, we used the IOTA protocols without modification as primary sources. A corresponding thesis [8] and the Merck Manual provided an extension for the IOTA descriptions. Additionally, we allowed references to the On-line Medical Dictionary[2], the CancerNet Dictionary[3], and the MEDLINE collection of abstracts of the US National Library of Medicine[4]. Concatenated together, these compose the free-text annotation, on average a hundred-word description for each of the domain variables. In the experiments we used a total of twenty-six variables which had been previously evaluated as the most relevant domain variables, such as parity, drug treatment for infertility, use of oral contraceptives, family history of breast and ovarian cancer, age, bilaterality of the tumor, pain, descriptors of the morphology, echogenicity and vascularization of the mass or the level of CA125 tumor marker.

3.2 Representations of annotations

According to our assumption, there is a free-text annotation for each domain variable, possibly containing references to standard document collections. The algebraic representation called the *vector space model* encodes a document in a D-dimensional space where each component represents a corresponding word, neglecting the grammatical structure of the text. We used the Porter stemmer to canonize the words [2] and the essential domain specific phrases and synonyms are processed appropriately. Based on these, a domain vocabulary was constructed containing more than one million words (D=1,135,017), which is used to generate a controlled index. The weights for the vector model were computed using the *term frequency–inverse document frequency* (tf-idf) term weighting scheme [7].

3.3 The ovarian cancer data

In addition to the prior background information, data has been collected in the framework of the IOTA project, consisting of 68 parameters for 1150 tumor masses at the moment of writing. In our experiments to evaluate the potential and use of text-based priors, we used only the completely observed cases with respect to the selected variables (318 cases).

4 Text-based priors and data for ABNs

We assume that the stochastic variables are discrete, the cases in the data are complete and independent and there is no a priori preference among parametrizations, a closed-form Bayesian formula was derived for the probability of a Bayesian network structure B_S given a dataset $\underline{D}$ (for details, see [4]). This formula consists

[2]http://www.graylab.ac.uk/omd/index.html
[3]http://thymoma.de/meddict.htm
[4]http://www.ncbi.nlm.nih.gov/PubMed/

of a product over the domain variables

$$P(B_S|\underline{D}) \propto P(B_S|\xi) \prod_{i=1}^{n} P(\pi_i \to X_i|\underline{D})$$

where the i-th factor $P(\pi_i \to X_i|\underline{D})$ is the probability of variable X_i having the parental set $\pi_i = \{\pi_{i1}, \ldots, \pi_{iL_i}\}$ according to the data, L_i is the number of parents for variable i and ξ represents the background information.

Common usage is to specify a uniform prior distribution for $P(B_S|\xi)$ rating all possible Bayesian network structures equally. But despite the decomposition of the learning to the selection of appropriate parental sets, the amount of data needed for statistically significant identification of networks is frequently unsufficient, particularly in recent high-dimensional, data scarce problems such as gene expression data analysis. Since the specification of a prior distribution over the Bayesian network space can be difficult for human experts in rapidly developing domains, such as the medical or genomic domain, we would like to automate the usage of the textual domain knowledge for this.

4.1 A probabilistic similarity metric for annotations

Our main assumption is that the vector representations of the annotations attached to the stochastic variables in the ABN convey useful information about these relations and that they can be used to define a prior $P(B_S|\xi)$ for Bayesian networks. The potential of the vector representation of text were investigated intensively in the context of information retrieval [7]. A principal IR goal is the definition of similarity metrics among the documents (or sets of documents), which expresses their semantic, information theoretic relation. A standard similarity metric for a pair of documents (X, Y) is the cosine of the angle between their corresponding normalized tf-idf vector representation

$$\mathrm{sim}(X,Y) = \cos(X,Y) = \frac{\sum_{i=1}^{D} f_{X,i}\mathrm{idf}_i \times f_{Y,i}\mathrm{idf}_i}{\|X\|\|Y\|}$$

where $f_{X,i}$ is the frequency of term i in document X and idf_i is the inverse document frequency expressing the discriminative power of term i as

$$\mathrm{idf}_i = \log\frac{N}{n_i}.$$

In the latter formula, N is the total number of documents and n_i is the number of documents containing term i.

Since this metric expressses the similarity between two documents describing two random variables, it also represents some connection between the two random variables. Next we observe that both the frequency and inverse document frequency factors are always positive, resulting in a similarity metric value between 0 and 1. These two elements allow us to directly interprete this value as the probability of the presence of an arc between the Bayesian network variables X and Y.

4.2 Text-based priors for Bayesian network structures

After summarizing the tf-idf-based relevance metric, we continue the derivation of an informative a priori distribution for $P(B_S|\xi)$. Assuming the independence of $(\pi_i \rightarrow X_i)$ and $(\pi_j \rightarrow X_j)$, our goal is to derive an estimate for $P(\pi_i \rightarrow X_i|\xi)$ based on the introduced relevance metric. Our further assumption of the independence of $X_k \in \pi_i$ and $X_l \in \pi_i$ (i.e., the independence of presence of the edges in the graph at a vertex) results into the decomposition of $P(\pi_i \rightarrow X_i|\xi)$, where $\pi_i = \{\pi_{i1}, \ldots, \pi_{iL_i}\}$:

$$P(\pi_i \rightarrow X_i|\xi) \quad = \quad \prod_{j=1}^{L_i} P(\pi_{ij} \rightarrow X_i|\xi).$$

Based on the annotations of the random variable in the ABN and the probabilistic similarity measure, $P(\pi_{ij} \rightarrow X_i|\xi)$ can be defined as follows (for simplicity, we use the same notation for the random variable and its annotation in the ABN)

$$P(\pi_{ij} \rightarrow X_i|\xi) \quad \propto \quad \mathrm{sim}(\pi_{ij}, X_i). \tag{1}$$

On the left of Fig. 1, we show all pairwise relevance scores. Without the edge independence assumption, $P(\pi_i \rightarrow X_i|\xi)$ can be defined using a general, non-pairwise relevance measure, which characterizes the global relevance of set π_i to X_i. Because of space limitations we confine ourselves only to the previous case.

5 Results

To evaluate the effect of the text-based a priori distribution, we report results for a classification task in the ovarian cancer domain. (For previous results about the application of Bayesian belief networks and multilayer perceptrons to classify ovarian masses, see [1].) Here, we report the classification performance in term of the area under the ROC curve [5] of two structure learning algorithms in combination of either our informative text-based a priori distribution for the Bayesian network structures (indicated with I) or a noninformative (uniform) prior distribution (indicated with N). We used the domain variables and the data described in Section 3. The first algorithm searches in the space of augmented naïve-Bayes belief networks (indicated with AnB), see [3] while the second algorithm is a general Bayesian network structure learning algorithm (indicated with G) based on randomly selected node orderings where the search for parental sets is exhaustive up to size three, after this a greedy algorithm is applied.

We divided the data set for various proportions of train and test set in an iterative scheme (30 repetitions). The right part of Fig. 1 shows the effect of the prior incorporation in the small sample range (percentage of data/100 used for training on the X-axis, the area under the ROC curve on the Y-axis).

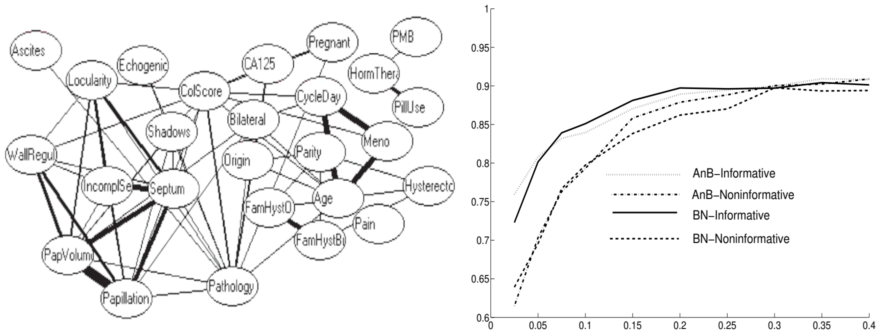

Figure 1: The visualisation of text-based relevance scores for the pairs of domain variables (left, thicker edges indicate higher scores). The learning curves of the BN-AnB-I, BN-AnB-N, BN-G-I and BN-G-N models using the area under the ROC curve for varying proportions (0% – 40%) of samples used in the training set. (right)

	$\widehat{E}[.]$	$\widehat{Std}[.]$	$\widehat{E}[.]$	$\widehat{Std}[.]$
train/test	10%/90%		40%/60%	
BN-AnB-N	0.793	0.036	0.908	0.007
BN-AnB-I	0.839	0.029	0.908	0.009
BN-G-N	0.797	0.026	0.894	0.015
BN-G-I	0.851	0.040	0.901	0.010

Table 1: Mean and standard deviation of the Area Under the ROC curve values at the 10%/90% and 40%/60% training/test proportion for the BN-AnB-I, BN-AnB-N, BN-G-I and BN-G-N models.

6 Conclusion

In the classification task, the text-based prior distribution for Bayesian network structures is beneficial in the small sample range, while it is not restrictive and vanishes in the middle and large sample range. However, the classification task is not an optimal evaluation for the effect of the text-based structure prior in the small sample range: the Bayesian networks structures proposed by the prior could be very good, while this does not show from the classification performance since there are simply not enough data points in the very small training set to train the parameters appropriately. To overcome this problem, we currently investigate the usage of model likelihoods and manual investigations to evaluate the text-based priors. Additionally, a prior distribution over the *parameters* for all Bayesian network structures is needed to achieve a better classification performance in the small sample range, together with the relaxation of the independent edge probability assumption. The former can be achieved by transforming the prior over the parameters for a *fixed structure* Bayesian network, specified by an expert, to all other structures (the parameter-prior specification for all possible Bayesian network structures is clearly a Sisyphus torment for the expert). The latter necessitates the usage of arbitrary probabilities for parental sets derived from the annotations.

As more and more domain knowledge becomes available as electronic literature, knowledge engineering increasingly needs representations and methods that integrate textual domain knowledge directly into knowledge modeling to assist the elicitation, learning, application and maintenance of complex models. The introduced *Annotated Belief Network* is a response to this challenge.

In this paper, we assumed to have short free-text descriptions for the domain variables and a huge repository of related domain literature. We derived a text-based prior over the Bayesian network structures and reported its effect in a medical task. These positive results suggest that the development of integrated text and data analysis methodologies, closely coupled representations and models for domain literature and statistical data is a promising direction.

The current results and evaluations can be extended in many directions. The current text-based priors, that is $\mathrm{sim}(.,.)$, currently relies on an algebraic representation of the text, while the annotations are already structured into various fields. A *structured* text representation with a corresponding text-based prior exploiting this structural information may have many advantages. A related research topic would be a more refined linguistic analysis, including better phrase identification or shallow parsing for a better representation of the content yielding an improved $\mathrm{sim}(.,.)$, enhancing also the text-based priors.

Acknowledgements

Dr. Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Peter Antal are Research Assistants with the K.U.Leuven. Geert Fannes is a Research Assistant with the F.W.O. Vlaanderen. Yves Moreau is a postdoctoral researcher with the F.W.O. Vlaanderen. Our research is supported by grants from: Research Council KUL: Concerted Research Action GOA-Mefisto 666, IDO (IOTA Oncology, Genetic networks); Flemish Government: Fund for Scientific Research Flanders (projects G.0256.97, G.0115.01, G.0240.99, G.0197.02, G.0407.02, research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s, STWW-Genprom, GBOU-McKnow, Eureka-Impact, Eureka-FLiTE); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006): Dynamical Systems and Control: Computation, Identification & Modelling), Program Sustainable Development PODO-II (CP-TR-18)

References

[1] P. Antal, G. Fannes, Y. Moreau, B. De Moor, J Vandewalle, and D. Timmerman, *Extended Bayesian regression models: a symbiotic application of belief networks and multilayer perceptrons for the classification of ovarian tumors*, Proc. of the 8th European Conference on Artificial Intelligence in Medicine (AIME01), 2000, Cascais, Portugal, pp. 177–187.

[2] P. Antal, T. Meszaros, B. De Moor, and T. Dobrowiecki, *Annotated Bayesian networks: a tool to integrate textual and probabilistic medical knowledge*, Proc. of the 13th IEEE Symp. on Comp.-Based Med.Sys. (CBMS01), 2001, Bethesda, MD, pp. 177–182.

[3] Jie Cheng and Russell Greiner, *Learning Bayesian belief network classifiers: Algorithms and system*, Proc. of the Canadian Conference on Artificial Intelligence (CSCSI01), 2001, pp. 141–151.

[4] G. F. Cooper and E. Herskovits, *A Bayesian method for the induction of probabilistic networks from data*, Machine Learning **9** (1992), 309–347.

[5] J. A. Hanley, *The meaning and use of the area under receiver operating characteristic (roc) curve.*, Radiology **143** (1982), 29–36.

[6] J. Pearl, *Probabilistic reasoning in intelligent systems*, Morgan Kaufmann, 1988.

[7] B. Ribeiro-Neto R. Baeza-Yates, *Modern information retrieval*, ACM Press, 1999.

[8] D. Timmerman, *Ultrasonography in the assessment of ovarian and tamoxifen-associated endometrial pathology*, Ph.D. dissertation, Leuven University Press, D/1997/1869/70, 1997.