Annotated Bayesian Networks: a Tool to Integrate Textual and Probabilistic Medical Knowledge

Péter Antal, M.Sc.[1], Tamás Mészáros, M.Sc..[2]

Bart De Moor M.Eng. Ph.D. [1], Tadeusz Dobrowiecki M.Eng. Ph.D. [2]

[1]*Department of Electrical Engineering*
Katholieke Universiteit Leuven
{antal, demoor}@esat.kuleuven.ac.be [2]

Dept. of Measurement and Inf. Systems,
Budapest Univ. of Tech. and Economics
{meszaros, tade}@mit.bme.hu

Abstract

In previous publications we have reported on the development of Bayesian Network models for the preoperative discrimination between malignant and benign ovarian mass. The models incorporated both medical background knowledge and patient data, which required the traceability of the incorporated prior medical knowledge. For this purpose, we followed a particular annotation method for Bayesian Networks using a dedicated representation. In this paper we present the resulting Annotated Bayesian Network (ABN) representation that consists of a regular Bayesian Network with standard probabilistic semantics and a corresponding semantic network, to which the textual information sources are attached. We demonstrate the applicability of such dual model to represent both the rigorous probabilistic and the unconstrained textual medical knowledge. We describe methods on how these Annotated Bayesian Network models can be used: (1) as a domain model to arrange the personal textual information of the clinician according to the semantics of the domain, (2) in decision support to provide detailed, even personalized explanation, and (3) to enhance the information retrieval to find new textual information more efficiently.

1. Introduction

Bayesian Networks have become a prevalent method to build probabilistic medical knowledge bases due to the corresponding clear probabilistic semantics, dual qualitative-quantitative components, the capability to serve as a universal probabilistic predictive model for decision support and their capability to integrate both the available background knowledge and the observations. We have developed a Bayesian Network model to assist clinicians discriminating between benign and malignant ovarian tumors [2]. Though we have obtained good performance comparable to the human experts and to other statistical models, a particular limit of this technique has been detected, specifically that the documenting the incorporated huge amount of prior background information from studies, papers and experts to the Bayesian Network model is a problem. Consequently, we investigated methods to attach informal medical information to the formal probabilistic model and utilize this in the model-building phase and in decision support in general.

To solve this problem at first we have defined a heuristic annotation method for Bayesian Networks and modified our modeling environment accordingly. The annotation method allows the attachment of natural language comments and references at various levels of a Bayesian Network, such as model, groups, variables, values of the variables, the probabilistic dependency relation between the variables and the corresponding probability distributions of the variables. In the next step, to fully exploit the potential of the probabilistic model enriched with textual information, we have defined a formal framework for this annotation by introducing an adjoined semantic network to which the textual information is attached. Based on the Bayesian Network model with its strict probabilistic semantics, a semantic network can be derived describing explicitly the entities and their relations in the corresponding domain (i.e. the domain ontology).

> John F. Sowa, Knowledge Representation:
> Logical, Philosophical, and Computational
> Foundations, Brooks Cole Publishing Co.,
> Pacific Grove, CA, 2000

The most immediate application of such semantic domain model or ontology is the possibility to organize the textual information of a clinician (relevant papers, interesting patient records, important observations), i.e. to attach this textual information to various parts of the derived semantic network. An Annotated Bayesian Network can also provide a more detailed and even personalized explanation for the derived conclusion. The underlying probabilistic model can serve as a decision theoretic computational engine to select optimal decisions and to generate plausible explanations, while the corresponding semantic network with its natural language comments and references can enrich and personalize such explanations. Finally we have investigated the applicability of this model in an increasingly important task, namely in helping either knowledge engineers or clinicians in finding new textual information either on Intra- or Internet. We shortly describe methods that use the derived semantic network and the attached textual information to make the information retrieval more efficient. This is accomplished by expanding the query of the user, by influencing the relevance evaluation function and by ordering and/or filtering the resulting documents.

The paper is organized as follows. Section 2 introduces the Annotated Bayesian Network model: the standard Bayesian Network model and the corresponding semantic network model with attached textual information. In Section 3 we sketch a method to organize textual information items according to the semantics of the domain and using an Annotated Bayesian Network model. Section 4 describes a method to enrich explanations of predictions with the attached textual information. Section 5 describes the implemented system architecture, using Annotated Bayesian Network models in information retrieval. We conclude the paper with a summary about the advantages of this dual model.

2. Annotated Bayesian Networks

A Bayesian Network represents a joint probability distribution over a set of variables. The model consists of a qualitative part (a directed graph) and quantitative parts (dependency models). For a particular domain, the vertices of the graph represent the domain variables and the directed edges describe the probabilistic dependency-independency relations among the variables in accordance with the joint probability distribution over the domain variables. There is a dependency model for every vertex (i.e., for the corresponding variable) to describe its probabilistic dependency on the parents (i.e., on the corresponding variables) (see e.g. [1]). These dependency models can be considered as input-output probabilistic models defined by a

parametric family of distributions and a corresponding parameterization. If the variables are discrete, a common dependency model is the table model, which contains the conditional distribution of the child variable conditioned on its parent.

In practical Bayesian Network models, the previous - theoretically complete - representation of a probability distribution over a set of domain variables is augmented with:

- ?? textual names for the domain variables and for the values of the discrete variables,
- ?? discretization thresholds for the values of the discrete variables,
- ?? a partition of the variables grouping them in accordance with the domain semantics.

In the paper we assume two further extensions: the existence of variable types (e.g. discretization schemes with value names) and the hierarchy of groups over the domain variables. The following description illustrates the representation of the extended Bayesian Network model shown on the left side in Fig. 1.
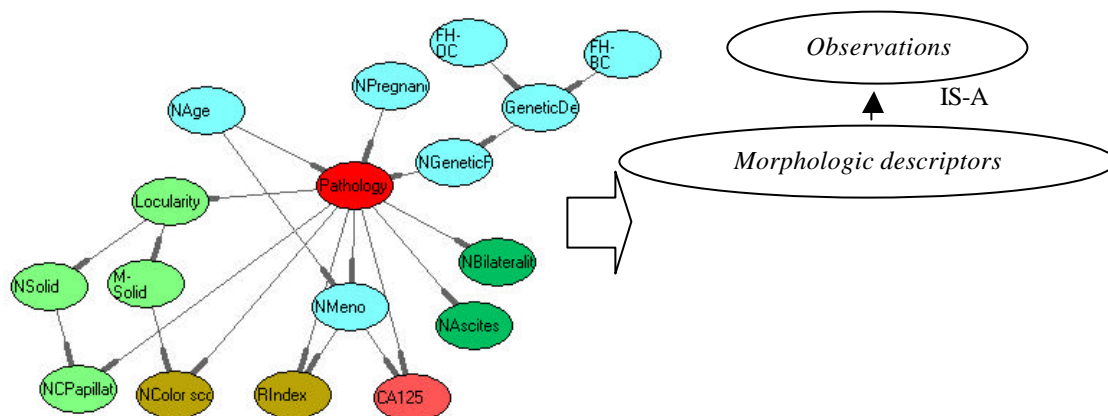
1. Variables{values(bins)}: Pathology{benign,malignant}, CA125{normal(0,55), high[55,)},…
2. Graph (directed edges): (Pathology,Papillation), (Pathology,Locularity), …
3. Dependency models: Pathology: Table{0.7,0.3,…},…
4. Groups: Observation={Morphologic,…}, Morphologic={Papillation, Locularity}, …

ezek hol latszanak az abran ??

While such a representation is fully functional both for the knowledge engineers and the users of a Bayesian Network, we have noticed that the attachment of informal resources, such as natural language documents, comments or figures, could help significantly the development and the maintenance of the model, e.g. by documenting the sources of various information in the model [2]. Consequently, we further augmented our representation of Bayesian Networks by allowing attachment of a free text to the objects in the representation (i.e. to values, variables, edges, dependency models and groups).

The introduced Annotated Bayesian Network proved particularly useful with respect to the heterogeneous sources of information usually incorporated in the Bayesian Network used for ovarian cancer classification. The initial heuristic extension of the original representations has resulted in a hybrid representation: on one hand the Bayesian Network representation with its strictly defined, but implicit probabilistic semantics (e.g. the (in)dependency model) and the augmentations with their domain specific, explicit semantics (e.g. the groups and their hierarchy), on the other hand …. ?????

A possible solution is to represent the Bayesian Network and partly its semantics explicitly in a representation that is capable to incorporate the newly introduced domain specific augmentations. Fig.1 illustrates the 'Semantic Network representation for Annotated Bayesian Networks' and the list below describes the transformation of a Bayesian Network and partly its semantics into the Semantic Network representation. Informally, the standard Bayesian Network is incorporated automatically in the foundation of the domain ontology and the higher-level concepts are built on the top of it.
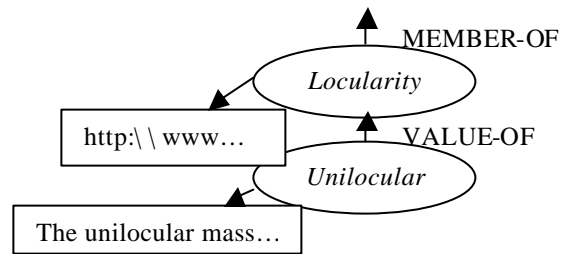
Figure 1. Semantic network representation of Annotated Bayesian Networks (vedd le a szineket, mert a szurke arnyalatoktol nem olvashato).

1. Value ?   Value node and value-of edge
2. Variable ?   Variable node and is-a edge
3. Group ?   Group node and is-a edge
4. Dependency model ?   Dependency model node and input-of, output-of edges
5. Annotation ?   Annotation node and reference edge

For simplicity, in the paper in the following the 'Annotated Bayesian Network' (ABN) denotes the 'semantic network representation for Annotated Bayesian Networks'.

3. Annotated Bayesian Networks in information management

The Annotated Bayesian Network incorporates partly the standard Bayesian Network semantics, which already offers hierarchical and lateral relations between the concepts (e.g. variable-value and parent variable-child variable). The hierarchy of the groups, however, makes it possible to represent the ontology for a particular domain. Since the domain ontology represents the relevant concepts and relations in the domain, such representation is an ideal candidate for organizing the documents (i.e. electronic articles) of the knowledge engineer, who develops the model/ontology or of the user of the finished model/ontology. In the case of a knowledge engineer the information sources used to develop the model can be attached to the corresponding objects in the model, while a researcher or a clinician can organize, store and access her/his collection of relevant articles, patient records, etc. according to the domain semantics. In our experimental environment we allow references to local files and documents on the network, and provide a browsing functionality for Annotated Bayesian Networks.

4. Annotated Bayesian Networks in decision support

In developing anAnnotated Bayesian Network representation our central goal was to link a strictly defined computational domain model with a more flexible, mainly textual domain model. Our goal was on one hand to help the manual development of the computational model as described above and on the other hand to help the usage of the model in decision support as follows.

Bayesian Networks or the related Influence Diagrams used in decision support are frequently used as input-output models, predicting the conditional distribution of an 'output' variable conditioned on the observed values of the 'input' variables. For example, in the ovarian cancer classification problem the developed Bayesian Network is used to predict the conditional

probability of benigness/malignancy given the observations about the morphologic, the vascularization and other properties of the pelvic mass. Though the performance of the model based on the predicted conditional class probability is comparable to that of human experts [2,6], an explanation of the computed prediction is frequently necessary. Following [4], we assume that the 'predictive explanation' in Bayesian Network based models consists of the most influential observed variables and the most influential paths connecting these variables to the output variable. Since this 'explanation set' and the paths can be identified automatically, the attached annotations can be used either directly by offering them for browsing or indirectly by analyzing the attached textual information. The analysis can be a deep linguistic analysis to provide natural language excerpts or summaries for the variables involved in the explanation. Another approach followed by us uses a shallow statistical analysis. We experimented with generating vector representation [5] for the annotated objects in the Annotated Bayesian Network, which contains relevance factors for each textual term occurring in the annotation. Based on the vector representation, the explanation derived from the computational model can be enriched, providing for example the average of the relevance factors for the variables in the explanation set or in certain path.

5. Information retrieval using Annotated Bayesian Networks

Information retrieval (IR) deals with methods for indexing, searching, and recalling data, particularly text and other unstructured data forms. Two major trends leading to increase the efficiency of the IR process are to utilize user specific information and domain specific information [5]. The Annotated Bayesian Network model can serve both as the user profile and as the domain ontology. Fig. 2 shows the standard model of the IR process (left) and modified model based on ABN (right) with the following function support:
1. ABN as domain ontology
   a. Deriving a domain specific controlled vocabulary from an ABN.
   b. Using the ABN as domain ontology for generalizing/specializing query terms.
2. ABN as user profile
   a. Expanding/transforming the query according to the ABN.
   b. Modifying the query matching formula to require that the relevant documents should be close to the aggregate ABN document (e.g. concatenation of all the annotation).
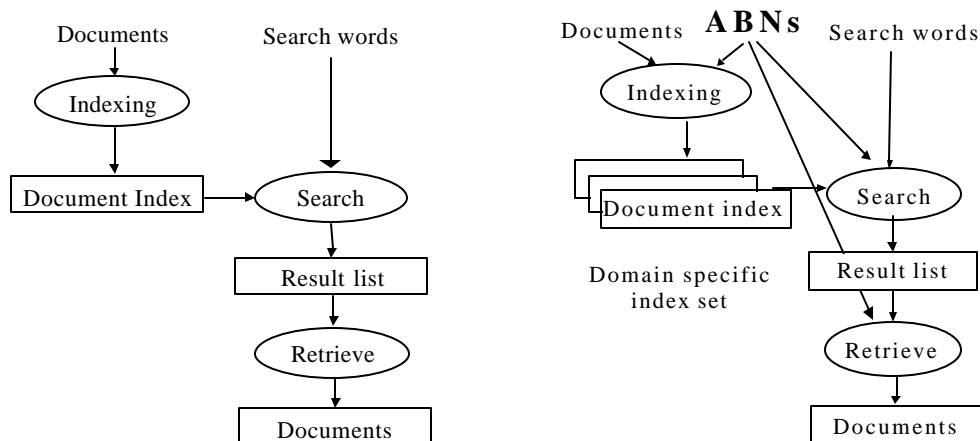   c. Filtering the returned documents with respect to the ABN.

Figure 2. IR models (general and ABN based).

Our main purpose was to increase the efficiency of the model building by helping the medical experts and knowledge engineers to find the necessary information. It means that we have attempted to integrate our Bayesian Network modeling system and an IR system. The goal was to create an integrated environment in which the computational-statistical modeling is accessible to the IR system as a kind of context (user profile) to increase the efficiency of the retrieval. The core of this system is the ABN. The availability of a computational-statistical domain model with textual annotations provides the link to integrate the function supports available in these contexts (e.g. probabilistic reasoning/learning versus information retrieval). We have investigated methods for points 2a, 2b, 2c and 1a in this environment.

In the IR subsystem the documents are represented by a vector representation [5]. To improve the performance of the system we used a controlled index to represent the documents, in which the domain vocabulary was derived from the ABN.

To be able to use the ABN to characterize the interest of the knowledge engineer more precisely, we identified phases in the model building, such as finding new domain variables, defining variables, checking conditional pair-wise (in)dependencies, quantifying dependency models, etc. Appropriately, the query is automatically modified according to the actual model-building phase, so the document matching criteria is always biased towards the actual goal of the model building process. Finally, the returned documents are filtered to eliminate the redundancy.

The integrated system consists of two components: the statistical-computational modeling system to work with Bayesian Networks/Annotated Bayesian Networks and a client-server document storage, analysis and retrieval system.

6. Summary

We have shown a methodology to create a link between probabilistic modeling and the corresponding textual information. The annotation of the extended Bayesian Networks seems to be particularly useful in the model-building phase to enhance the information retrieval, to support the documentation and maintenance of the model. Additionally it can provide a more detailed explanation for decision support and can be used to organize electronic documents according to the domain semantics. The proposed semantic network representation for Annotated Bayesian Networks provides semantic clarity and facilitates the usage of the practical extensions of the Bayesian Network representation.

Acknowledgement

References

[1] Pearl, J.: Probabilistic reasonning in intelligent systems, MIT Press, 1992

[2] P. Antal, H. Verrelst, D. Timmerman, S. Van Huffel, B. De Moor, I. Vergote, Y. Moreau: Bayesian networks in ovarian cancer diagnosis: potentials and limitations, 13th IEEE SYMPOSIUM ON COMPUTER-BASED MEDICAL SYSTEMS, June 23-24, 2000, Texas Medical Center, Houston, Texas, pp. 103-108

[3] M.R. Quillian: Semantic Memory, in (ed.)M. Minsky, Semantic Information Processing, MIT Press, 1968

[4] Urszula Chajewska, D. L. Draper: Explaining Predictions in Bayesian Networks and Influence Diagrams, Interactive and Mixed-Initiative Decision-Theoretic Systems, AAAI 1998 Spring Symposium Series, Stanford University, California March 23-25, 1998

[5] R. Baeza-Yates, B.Ribeiro-Neto: Modern Information Retrieval, ACM Press, 1999

[6] D. Timmerman et al., Artificial neural network models for the pre-operative discrimination between malignant and benign adnexal masses. Ultrasound Obstet Gynecol 1999; 13: 17-25.