Generating Explanations and Tutorial Problems from Bayesian Networks

Peter Haddawy, Ph.D., Joel Jacobson, Charles E. Kahn, Jr., M.D.
Department of Electrical Engineering and Computer Science, University of Wisconsin–Milwaukee;
Section of Information and Decision Sciences, Department of Radiology, Medical College of Wisconsin,
Milwaukee, Wisconsin

We present a system that generates explanations and tutorial problems from the probabilistic information contained in Bayesian belief networks. BANTER is a tool for high-level interaction with any Bayesian network whose nodes can be classified as hypotheses, observations, and diagnostic procedures. Users need no knowledge of Bayesian networks, only familiarity with the particular domain and an elementary understanding of probability. Users can query the knowledge base, identify optimal diagnostic procedures, and request explanations. We describe BANTER's algorithms and illustrate its application to an existing medical model.

INTRODUCTION

Bayesian networks have become the representation of choice for building decision-making systems in domains characterized by uncertainty, and have been applied to several medical domains [1-5]. The models currently available and under development provide a wealth of detailed knowledge that can be used for educational purposes as well as clinical decision support. Unfortunately, the information contained in these models is not easily intelligible; tools are needed to make this information comprehensible. The availability of shells for performing inferences over Bayesian network models [6,7] and the recent development of explanation generation algorithms [8,9] have made building such a tool possible.

This report presents BANTER (Bayesian Network Tutoring and Explanation), a generic Bayesian-network shell that provides decision support and tutors users in diagnosis and in selection of optimal diagnostic procedures. BANTER can be used with any Bayesian network containing nodes that can be classified as hypotheses, observations, and diagnostic procedures. The system is designed so that the user need know nothing about Bayesian networks in order to interact with it effectively. In fact, none of the system's dialogs with the user indicates that the system is using a Bayesian network to perform its reasoning. The user needs only some knowledge of the particular domain and an elementary understanding of probability.

BANTER computes the posterior probability of a diagnosis, determines the best diagnostic procedure to affirm ("rule in") or exclude ("rule out") a diagnosis, quizzes the user on the selection of optimal diagnostic procedures, and generates explanations of its reasoning. It can generate story problems and quiz the user on diagnoses and selection of optimal diagnostic procedures. Almost all of the system's reasoning is driven by the Bayesian network knowledge base; setting up the system for a new network requires minimal effort.

METHODS

System Environment

BANTER is implemented in C* and runs on top of the HUGIN Bayesian network inference system [6]. HUGIN performs all probability computations using a belief network specified in HUGIN's network definition format. The HUGIN interface consists of a set of functions from the HUGIN libraries, which are used to load and compile a belief network, instantiate and uninstantiate nodes, propagate changes in individual nodes throughout the network, and obtain probability values for nodes. BANTER's graphical interface is written using the Xaw graphics tool kit of the X11 public-domain windowing package; the widespread popularity of this package makes the interface highly portable.

System Configuration

BANTER is easily configured for new networks. To set up a new network model, BANTER requires a HUGIN network definition file, a BANTER definition file, and a story template file. For medical models, the BANTER definition file contains a list of nodes grouped as HISTORY, PHYSICAL FINDINGS,

* The software is available at *ftp//ftp.cs.uwm.edu/pub /tech_reports/ai/BANTER.tar.Z.*

0195-4210/94/$5.00 © 1994 AMIA, Inc.

DISEASES, and DIAGNOSTIC PROCEDURES. Each node is of the type FLOAT, INTEGER, STRING, or BOOLEAN.

Generating Tutorial Problems

The story template file is used to create the text for randomly generated story problems. The system generates a story problem by randomly choosing a set of values for the patient history and physical findings, randomly choosing a disease of interest, and expressing these choices by instantiating the story template. The template contains the following types of directives:

{*label*: *text*$_1$:*text*$_2$: ... :*text*$_{n+1}$}

Print the text for the corresponding state of node *label*: *text*$_1$ corresponds to the first state (as listed in the HUGIN definition file), *text*$_2$ for the second, and *text*$_{n+1}$ for the UNKNOWN state. A percent sign ("%") in the text stream indicates where the node's value will be inserted; for nodes of type FLOAT, one can specify the number of printed digits (e.g., "%.%%").

<BOOLEAN:*text*$_1$:*text*$_2$>

Pick a random boolean value, and print *text*$_1$ if TRUE or *text*$_2$ if FALSE.

[*class*]

Print the names of nodes of the specified *class* (for medical models: HISTORY, PHYSICAL-FINDINGS, or DISEASES), excluding those that have been selected already. If a node is BOOLEAN, its name will be included only if its current state is TRUE. For nodes of other types, the name and value will be displayed.

(*class*)

Print the names of nodes of the specified *class*, excluding those that have been selected already. If a node is of type BOOLEAN, its name will be included only if its current state is FALSE. For nodes of other types, the name and value will be displayed.

Determining the Best Test

The best test to rule in or rule out a hypothesis is determined by positively and negatively instantiating each test outcome and determining the posterior probability of the hypothesis given the test outcome and the patient's history and physical findings. The best test to rule in the hypothesis is the one that results in the highest post-test probability and the best test to rule out the hypothesis is the one that results in the lowest post-test probability.

Generating Explanations

Following Suermondt's INSITE method [8], BANTER generates explanations in two steps. The first step identifies the evidence that has the most influence on the given hypothesis. The second step identifies the strongest and most comprehensible paths linking the influential evidence with the hypothesis. Both algorithms are used to explain the current belief in a disease: we first identify those nodes among the specified history and physical findings that were most influential in producing the reported posterior probability of the disease and then find the paths along which that influence flows. To explain the selection of the best test, only the second algorithm is used: here we only need to find the paths of influence from each test outcome to the disease.

Identifying Influential Evidence. To identify the most influential pieces of evidence, we first determine the influence of each evidence node on a hypothesis by performing a sensitivity analysis. We remove all evidence from the network and then instantiate each evidence node individually and record the posterior probability of the hypothesis. We then filter out all evidence nodes that do not influence the hypothesis in the direction of its posterior probability given all the evidence. For the remaining nodes, the posterior probabilities are then normalized so that they fully span the range 0 to 1; call this the importance of each node. We define important nodes to be ones with an importance value greater than some threshold. The threshold is selectable by the user and is currently set to 0.7. We normalize the posterior probabilities since we are interested in identifying pieces of evidence with relatively strong influence on the probability of the hypothesis. This is not determined by the absolute value of the posterior probability but rather by the value relative to the prior probability of the hypothesis and the posteriors for the other pieces of evidence.

Our algorithm differs from that of Suermondt [8]. Rather than instantiating each piece of evidence individually, Suermondt removes each piece of evidence individually and computes the posterior probability of the hypothesis without that piece of evidence. An influential piece of evidence is one for which the posterior probability without the evidence is significantly lower than with the evidence. While our approach will identify each piece of evidence that is individually significant, Suermondt's approach will not flag a piece of evidence as significant if it does

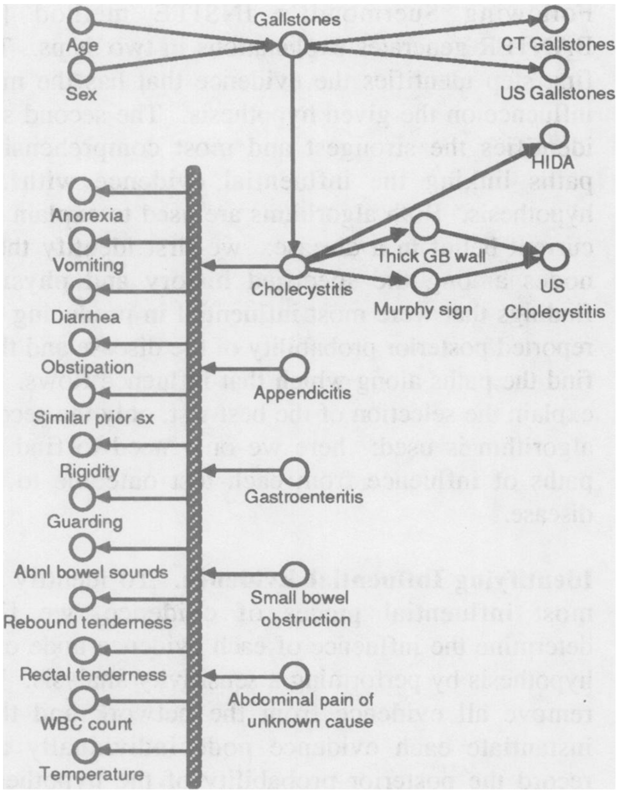


Figure 1. Bayesian network model of gallbladder disease. (The vertical bar simplifies the illustration: all input nodes influence all output nodes.)

not increase the probability of the hypothesis in the presence of other pieces of evidence. For example, if two pieces of evidence each raise the probability of the hypothesis to one, neither will be flagged as significant since when each is removed the probability of the hypothesis is still one. Suermondt further discusses using his technique on all possible subsets of the set of evidence in order to identify sets of evidence that may be collectively significant but for which no single element is individually relevant.

Identifying Paths of Influence. To determine the paths along which an evidence node influences a hypothesis node, we first identify all paths along which evidence can flow based on d-separation [10]. This set often will be too large for meaningful explanation, so we limit the explanation to five paths, ranked by strength and length. Our foremost objective is to tell the user how the evidence influences the hypothesis. For the explanation to be accurate, BANTER needs to identify the strongest paths; for it to be concise, we choose the shortest paths among those that are equally strong. The method of identifying paths of influence is described in greater detail elsewhere [11].

RESULTS

We applied BANTER to a Bayesian network model of acute gallbladder disease (Figure 1) [5]. The two principal diagnoses are gallstones and cholecystitis; appendicitis, gastroenteritis, small bowel obstruction, and abdominal pain of unknown cause serve as alternative diagnoses. The remaining nodes represent a patient's history, physical findings, and test results. Using the story template file (Figure 2), BANTER generates a story problem (Figure 3).

Querying the Knowledge Base

The user queries the knowledge base by setting up a scenario. A scenario is created by specifying a set of known values for the history and physical findings, as well as a set of diseases of interest. This is done by clicking on nodes in windows displaying for history, physical findings, and diseases of interest. The user now can ask the system to compute the posterior probability of the selected diseases or to determine the best tests to rule in and rule out the selected

```
{SEX:Mr. Jones:Mrs. Jones:The patient}
{AGE:is % years old, and} presents with [HISTORY], and denies (HISTORY).
{SEX:His:Her:The patient's} {TEMPERATURE:temperature is %.%.}
{SEX:His:Her:The patient's} {WBC-COUNT:WBC count is %.%.}
Physical examination reveals [PHYSICAL-FINDINGS], and no evidence of (PHYSICAL-FINDINGS).
What is the best test to <BOOLEAN:rule in:rule out> [DISEASE]?
```

Figure 2. Story template for gallbladder-disease model.

```
Mrs. Jones is 41 years old, and presents with ANOREXIA, and denies VOMITING, DIARRHEA,
     OBSTIPATION, SIMILAR-SX-PREVIOUSLY.  Her WBC count is 12.6.  Physical examination reveals
     GUARDING, and no evidence of RIGIDITY, REBOUND-TENDERNESS, ABNORMAL-BOWEL-SOUNDS.

What is the best test to rule out GALLSTONES?
```

Figure 3. Story problem generated by BANTER.

diseases.

Requesting an Explanation

The user can obtain an explanation of the reasoning that lead the system to select these tests (Figure 4). The system starts by explaining how the known history and physical findings influence the probability of gallstones. Having explained how the pretest probability of gallstones was arrived at, the system continues by explaining each possible test further influences the probability of gallstones.

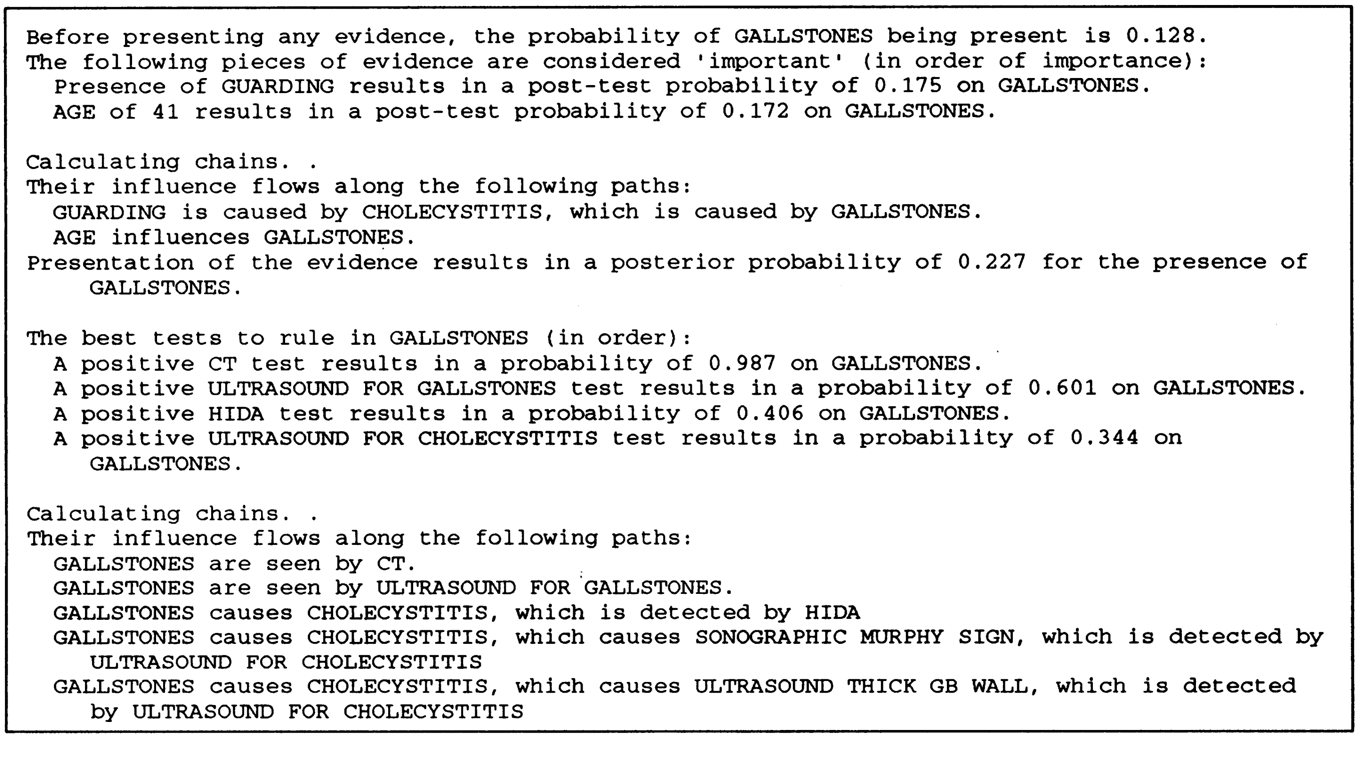
Quizzing the User

In addition to asking the system to perform computations, the user can ask to be quizzed in the selection of optimal diagnostic procedures. This can be done in two ways. The user can specify a scenario and choose the test he or she thinks best to rule in or rule out the selected disease. If an answer is incorrect, the system tells the user which tests are preferable to the one selected and can explain its reasoning. The second way the user can be quizzed is by selecting the "story" action. In this mode, the system randomly selects a patient history, a set of physical findings, and a disease of interest, and presents the scenario to the user in English (Figure 3). The user can select an answer from the quiz menu and continue as described above.

DISCUSSION

BANTER transforms the information contained in a Bayesian network into an easily intelligible form for medical education and clinical decision support. BANTER quizzes and tutors users on the evaluation of diagnoses and optimal selection of diagnostic procedures. Since almost all the system's reasoning is performed using the Bayesian network knowledge base, configuring the system to work with a given network requires little effort. On the other hand, since nothing in the system's functionality indicates that it is using a Bayesian network for its reasoning, the complex details of the representation are hidden from the user.

Future research will focus on (1) explanations in extremely large networks, (2) more informative explanations, and (3) rigorous evaluation. In the newly emerging models that contain thousands of nodes, inference will become too slow to provide acceptable interaction and the explanations produced by the current algorithm will become too lengthy. For a given problem, typically only a portion of a given network model will be relevant. We have developed a technique for specifying a Bayesian network as a collection of rules in probability logic and generating that portion of the network relevant to a given computation [12]. Integrating this technique



```
Before presenting any evidence, the probability of GALLSTONES being present is 0.128.
The following pieces of evidence are considered 'important' (in order of importance):
  Presence of GUARDING results in a post-test probability of 0.175 on GALLSTONES.
  AGE of 41 results in a post-test probability of 0.172 on GALLSTONES.

Calculating chains. .
Their influence flows along the following paths:
  GUARDING is caused by CHOLECYSTITIS, which is caused by GALLSTONES.
  AGE influences GALLSTONES.
Presentation of the evidence results in a posterior probability of 0.227 for the presence of
     GALLSTONES.

The best tests to rule in GALLSTONES (in order):
  A positive CT test results in a probability of 0.987 on GALLSTONES.
  A positive ULTRASOUND FOR GALLSTONES test results in a probability of 0.601 on GALLSTONES.
  A positive HIDA test results in a probability of 0.406 on GALLSTONES.
  A positive ULTRASOUND FOR CHOLECYSTITIS test results in a probability of 0.344 on
     GALLSTONES.

Calculating chains. .
Their influence flows along the following paths:
  GALLSTONES are seen by CT.
  GALLSTONES are seen by ULTRASOUND FOR GALLSTONES.
  GALLSTONES causes CHOLECYSTITIS, which is detected by HIDA
  GALLSTONES causes CHOLECYSTITIS, which causes SONOGRAPHIC MURPHY SIGN, which is detected by
     ULTRASOUND FOR CHOLECYSTITIS
  GALLSTONES causes CHOLECYSTITIS, which causes ULTRASOUND THICK GB WALL, which is detected
     by ULTRASOUND FOR CHOLECYSTITIS
```

Figure 4. Explanations generated by BANTER.

into BANTER will significantly reduce the complexity of inferences in very large networks.

BANTER provides more informative explanations by associating semantic information with Bayesian networks. Instead of displaying paths of influence with arrows, we indicate how each node influences its successor with terms like "causes" or "detects." Including abstraction information may make explanations more informative and more concise. Rather than explain only the current scenario, one could explain a more general scenario, of which the current one is an instance. For example, in the case of cholecystitis elevating temperature, one could additionally tell the user that any inflammatory disease, of which cholecystitis is an instance, has the tendency to elevate temperature.

We currently are evaluating BANTER's explanatory content and style in tests with physicians at various levels of training. In addition to the model of gallbladder disease described above, we are applying BANTER to belief-network models for diagnosis of liver lesions by magnetic resonance imaging [4] and echocardiographic diagnosis (Díez FJ, personal communication).

Because of its generality and ease of use, BANTER can be used in a wide variety of settings where belief networks models have been formulated. Its ability to quiz users and provide explanations – without explicit reference to a belief network model – makes the system useful for clinical decision support and medical education.

Acknowledgments

We thank Dr. Finn V. Jensen and the HUGIN group at Aalborg University for generously providing us the use of HUGIN for this work. This work was supported in part by NSF grant IRI-9207262 (PH) and the 1993 American Roentgen Ray Society Scholarship (CEK).

References

[1] Andreassen S, Woldbye M, Falck B, Andersen SK. MUNIN — a causal probabilistic network for interpretation of electromyographic findings. In: Proceedings of the International Joint Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 1987: 366-372.

[2] Shwe MA, Middleton B, Heckerman DE, et al. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I. The probabilistic model and inference algorithms. Methods Inf Med 1991; 30:241-255.

[3] Andreassen S, Jensen FV, Olesen KG. Medical expert systems based on causal probabilistic networks. Int J Biomed Comput 1991; 28:1-30.

[4] Tombropoulos R, Shiffman S, Davidson C. A decision aid for diagnosis of liver lesions on MRI. In: Safran C, ed. Proceedings of the Seventeenth Annual Symposium of Computer Applications in Medical Care. New York: McGraw-Hill, 1993: 439-443.

[5] Haddawy P, Kahn CE Jr, Butarbutar M. A Bayesian network model for radiological diagnosis and procedure selection: work-up of suspected gallbladder disease. Med Phys 1994; 21:1185-1192.

[6] Andersen SK, Olesen KG, Jensen FV, Jensen F. HUGIN — a shell for building Bayesian belief universes for expert sytems. In: Proceedings of the Eleventh International Joint Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 1989: 1080-1085.

[7] Srinivas S, Breese J. IDEAL: a software package for analysis of influence diagrams. In: Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence. San Mateo, CA: Morgan Kaufmann, 1990: 212-219.

[8] Suermondt HJ. Explanation in Bayesian Belief Networks [Ph.D. thesis]. Medical Information Sciences, Stanford University, 1992.

[9] Suermondt HJ, Cooper GF. An evaluation of explanations of probabilistic inference. Comput Biomed Res 1993; 26:242-254.

[10] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann Publishers, 1988.

[11] Haddawy P, Jacobson J, Kahn CE Jr. An educational tool for high-level interaction with Bayesian networks. In: Proceedings of the 6th IEEE International Conference on Tools with Artificial Intelligence. New Orleans, LA: IEEE Press, 1994: (in press).

[12] Haddawy P. Generating Bayesian networks from probability logic knowledge bases. In: Uncertainty in Artificial Intelligence: Proceedings of the Tenth Conference. San Mateo, CA: Morgan Kaufmann, 1994: 262-269.