On Test Selection Strategies for Belief Networks

David Madigan, Department of Statistics, University of Washington*
and Fred Hutchinson Cancer Research Center

Russell G. Almond, StatSci division, MathSoft

November 1, 1994

Abstract

Decision making under uncertainty typically requires an iterative process of information acquisition. At each stage, the decision maker chooses the next best test (or tests) to perform, and re-evaluates the possible decisions. Value-of-information analyses provide a formal strategy for selecting the next test(s). However, the complete decision-theoretic approach is impractical and researchers have sought approximations.

In this paper, we present strategies for both myopic and limited non-myopic (working with known test groups) test selection in the context of belief networks. We focus primarily on utility-free test selection strategies. However, the methods have immediate application to the decision-theoretic framework.

1 Introduction

Graphical belief network researchers have developed powerful algorithms to propagate the effects of any piece of information to all the variables in the model in a manner analogous to forward chaining in rule-based expert systems (see, for example, Dawid, 1992). Comparatively little work has been done on the "backward chaining" problem: finding the most cost-effective information sources which will best increase information about a target variable. This is the *Test Selection* problem. There are two parts to the problem: choosing a metric for the value of information, and searching for potential information sources which maximize the metric. We consider both these issues and present workable approaches.

Many authors have recognized the importance of test selection in larger applications. Recent proposed approaches for probabilistic belief networks include: Jensen and Liang (1994), Heckerman, *et al.* (1993), and Almond (1993). In this paper, we propose a semi-automated myopic strategy which synthesizes the critiquing approach of Miller (1983) and Good's idea of a quasi-utility (Good and Card, 1971). We develop search strategies based on these ideas and demonstrate them in the context of a simple imaging application. We also address a limited form of the nonmyopic test selection problem, and propose and demonstrate a Markov chain Monte Carlo solution that generalizes the recent work of Heckerman, *et al.* (1993).

Section 2 reviews basic test selection techniques without particular reference to belief networks. In particular, it introduces *weights of evidence* and some of their properties, and the concept of "critiquing". Section 3 discusses weights of evidence and test selection in the context of belief

**Address for correspondence*: Department of Statistics, GN-22, University of Washington, Seattle, Washington 98195 (madigan@stat.washington.edu).

networks. Section 4 addresses the nonmyopic test selection problem and provides a connection with the decision theoretic approach to test seleciton.

2 Basic Test Selection

Two basic issues in test selection are independent of the graphical representation of our probability model: the choice of metric for the value of a test, and the search strategy for selecting a test according to that metric. Section 2.1 discusses our suggested metric, the expected weight of evidence, and Section 2.2 discusses model-independent search strategy issues.

2.1 Expected Weight of Evidence

Before performing a test, we should assess the amount of information that the test will provide. Generally, we should only pursue a test if the information gained from the test is worth the cost of testing. If we have a full decision model (e.g., an influence diagram), then the preferred metric is the *value of information* (Matheson, 1990, provides a review). For a purely probabilistic model, Good and Card (1971) suggest using a *quasi-utility*—a measure of information content which plays the role of a utility.

When discriminating between a single hypothesis H and its negation, Good and Card (1971) recommend the *expected weight of evidence* as a quasi-utility measure of the usefulness of a test, T:

$$EW(H:T) = \sum_{j=1}^{n} W(H:t_j)\Pr(t_j \mid H)$$

where $\{t_j, j = 1, ..., n\}$, represent the possible outcomes of the test, T. $W(H:t_j)$ is the weight of evidence for H provided by the evidence $T = t_j$. Specifically:

$$W(H:T_j) = \log \frac{\Pr(T = t_j \mid H)}{\Pr(T = t_j \mid \overline{H})}.$$

Informally, $EW(H:T)$ is the weight of evidence that will be obtained from T "on the average", when the H is true. Glasziou and Hilden (1989) provide a thorough review of the application of quasi-utilities to test selection and justify the choice of the weight of evidence. Good (1985) notes that expected weight of evidence is a generalization of entropy, and regards it as occupying "a central position in human inference." The weight of evidence is closely related to the decision-theoretic "value of perfect information" and Heckerman, Horvitz and Middleton (1993) (hereafter HHM) show how to transform value of information questions into questions involving weights of evidence (see Section 4). See Ben-Basset (1978), Good and Card (1971), and Pearl (1988) for discussion of alternative entropy-based quasi-utilities and multi-valued hypotheses.

It is clear that test selection strategies must account for test costs (and distinguish between "asking someone's age and doing an exploratory laparotomy", Good and Card, 1971). Our test selection strategy selects the test(s) that maximizes the expected weight of evidence for the current hypothesis per unit cost. The cost of a test is a summary quantity, expressed in quasi-utility units, combining all the undesirable aspects of testing. In the medical context, for instance, test costs include the risk of morbidity and mortality, the discomfort of the procedure, the delay in definitive treatment caused by waiting for the results, and the financial cost. For myopic or "greedy" test selection (i.e., looking only one test ahead), simply maximizing the expected weight of evidence per unit cost is effectively a miserly approach. This strategy continually shops around for bargains, when spending a little more freely may ultimately prove more cost effective (Glasziou and Hilden, 1989). We return to this issue in Section 4.2.

Not all mistakes are equal, and a sensible test selection strategy should also account for the cost of misclassification. Breiman, *et al.* (1984) and Glasziou and Hilden (1989) suggest a simple weighting scheme. Each disease is assigned a *degree of importance* which is then used to weight the probabilities in the expected weight of evidence. Suppose that to incorrectly classify a "H" subject as a "$\overline{H}$" is w times as regrettable as classifying a "$\overline{H}$" subject as a "H". In a sense, H is w times as important as $\overline{H}$. One way to think of this is to consider each case of H as w cases of H', where H' and $\overline{H}'$ have equal standing. In the transformed problem, the probability of H' becomes:

$$\Pr(H') = \frac{w\Pr(H)}{w\Pr(H) + \Pr(\overline{H})}.$$

The quasi-utility based on expected weight of evidence now becomes:

$$\sum_{j=1}^{n} W(H' : T_j)\Pr(T_j|H')$$

where

$$\Pr(T_j|H') = \Pr(T_j|H)\frac{\Pr(H)(w-1)+1}{\Pr(H|T_j)(w-1)+1}$$

and

$$W(H' : T_j) = W(H : T_j).$$

This begs the question of how the diseases should be weighted. We refer the interested reader to Glasziou and Hilden (1989), who discuss this issue at some length.

2.2 Critiquing

In this paper, we focus primarily on test selection for binary-valued hypotheses. Other authors have presented test selection metrics for multi-valued hypotheses (see, for example, Jensen and Liang, 1994), and our methods can use such metrics. However, in practice we have found that automated test selection strategies for multi-valued hypotheses (such as medical diagnoses) can exhibit disquieting behavior. Specifically, their line of reasoning may bear no resemblance to typical expert reasoning, moving from indicants relevant to one hypothesis, to indicants relevant to a different hypothesis. The artificial intelligence literature has discussed this problem at length–see, for example, Barr and Feigenbaum (1982, p.82).

To address this problem, we adopt a "critiquing" approach (Miller, 1983). The essential idea of critiquing is to elicit a suggested hypothesis from the user, say $H = h_0$. The system then elicits indicants which are chosen to maximize the probability of quickly accepting or rejecting h_0. Thus, once the user has suggested a hypothesis, the system only elicits indicants that are of direct and usually apparent relevance to that hypothesis. If the hypothesis is rejected, i.e., its probability falls below some defined threshold (Spiegelhalter and Knill-Jones, 1984), the user is prompted for an alternative suggestion, and so on (Miller, 1983, McSherry, 1986). Our experience with this approach in the medical context suggests that clinicians welcome the idea of interacting with the system in this way and that critiquing enhances confidence in the system.

3 Myopic Search

The previous section discussed test selection and weights of evidence without regard to the belief network structure. The application to belief networks, while computationally challenging, is straightforward in principle. First the user selects a hypothesis (H) to critique. Next we calculate the expected weight of evidence provided for H by *all* uninstantiated tests. Fortunately, the definition of the expected weight of evidence ensures that these calculations can be carried out in an efficient manner. Exactly two propagations are required: first, H is temporarily instantiated to "true" and $\Pr(T^i_j \mid H)$ is calculated for all uninstantiated tests $T^i, i = 1, \ldots, m$, with states $j = 1, \ldots, n_i$; second, H is temporarily instantiated to "false" and $\Pr(T^i_j \mid \overline{H})$ is again calculated for all i and j. A similar observation is made in Jensen and Liang (1994).

Unfortunately, in very large networks such as Pathfinder (Heckerman *et al.*, 1992), even these calculations are prohibitively expensive. The problem can be framed as a classical search problem looking for the most cost effective test. In the case of belief networks, the structure of the network provides a convenient structure for the search space which can be exploited to find efficient search strategies. In particular, expected weight of evidence is *monotonic* over a class of graphs called Berge Networks[1]—the farther you get from the hypothesis being critiqued, the lower the expected weight of evidence (Section 3.1). This property forms the basis of some simple search strategies in the belief network's "junction tree" and related Markov Tree models (Section 3.2).

3.1 Weight of evidence in Berge Networks and Markov Trees

Myopic test selection strategies can take advantage of the special structure of Berge networks (Madigan *et al.*, 1994). In a Berge network, the expected weight of evidence decreases in a monotone fashion away from the hypothesis, H. Therefore, for these networks, the test selection strategy can confine its attention to the immediate neighbors of H. This convenient property derives from the following basic result:

Theorem 1 (Monotonicity): In a belief network with three nodes (A, B and H), if B separates A from H (Figure 1), $EW(H : A) \leq EW(H : B)$.

Proof: See the appendix.

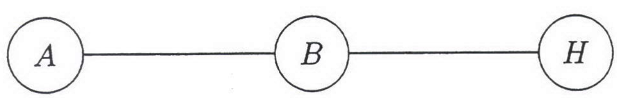


Figure 1: *A Simple Berge Network: The weight of evidence A provides for H can be no more than the weight of evidence B provides for H.*

Berge networks have the property that for any pair of connected nodes, the network is collapsible onto a unique "evidence chain" connecting the two nodes. A simple recursion argument extends the above monotonicity property to these chains–see Madigan *et al.* (1994) for details.

For non-Berge networks, no similar property exists. Consider the example of Figure 2. Although $EW(H : B, C) \geq EW(H : A)$, it could be true that $EW(H : A) > EW(H : B)$ and $EW(H : A) > EW(H : C)$. Simply searching all of the neighbors of the target hypothesis is not sufficient for non-Berge networks.

[1] A Berge network is a belief network with the property that clique intersections in the associated undirected graph (see Dawid, 1992) contain no more than one node.

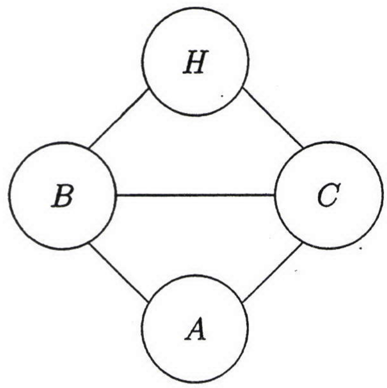


Figure 2: *A Simple Non-Berge Network. In this network, the weight of evidence provided by A for H could be greater than that provided by either B or C singly.*

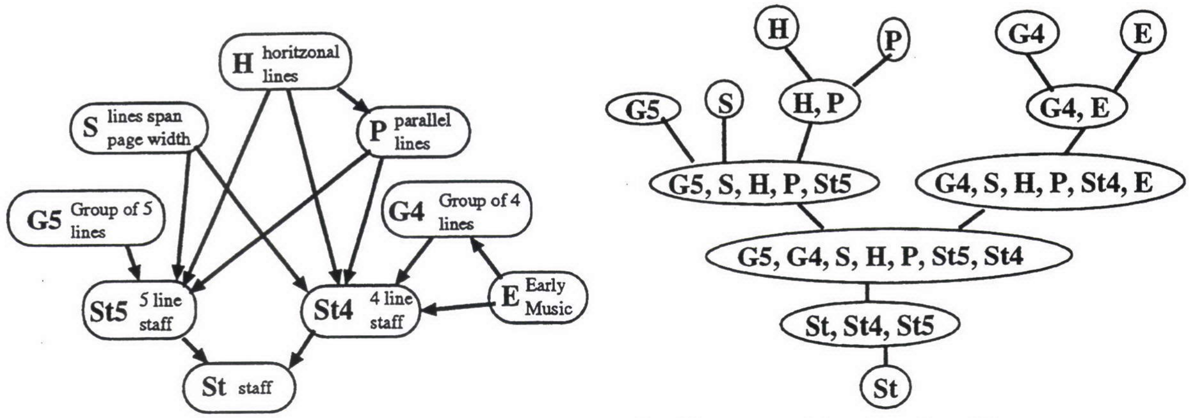
Figure 2 illustrates both the problem and a potential solution. If we cluster the variables B and C, the graph reduces to a simple Berge network (like Figure 1). In a general non-Berge network we can either cluster the variables to form a Berge network, or transform the network into a Markov tree.[2] Many popular algorithms for calculating probabilities in belief networks already build a Markov tree (Lauritzen and Spiegelhalter, 1988, Dawid, 1992, Almond, 1990). Since Markov trees are always Berge networks, monotonicity holds in the Markov tree model. The next section explores this approach.

3.2 Search Strategies in the Markov Tree

Ignoring costs, a simple branch and bound search using the Markov tree structure can find the best test. Almond (1990) recommends constructing the Markov tree by augmenting the tree of cliques or junction tree (Jensen, 1988) with nodes representing the individual variables (see Figure 3(b) for an example). Then, starting from the target hypothesis, the algorithm puts all the nodes connected to the hypothesis in the Markov tree on the "search boundary list". Next, it finds the node in the search boundary with the lowest expected weight of evidence for the hypothesis (i.e., the joint weight of evidence of all the variables in the node—see Section 4.1), and expands the node by placing its Markov tree-neighbors on the search boundary. When the node to be expanded represents a single variable corresponding to a test, it follows from the Monotonicity theorem this is the best test.

As an illustration, we show in Figure 3 a model for the task of classifying a group of lines as a musical staff (Almond *et al.*, 1994). The node $\{\mathbf{St}\}$ at the bottom of the Markov tree is the target hypothesis. In this example, lines which span the width of the page are rare, but staff lines are rarely less than a full page wide. Thus, $\{\mathbf{S}\}$ will have high expected weight of evidence. The search might start by looking at the $\{\mathbf{St}, \mathbf{St4}, \mathbf{St5}\}$ node which trivially has the highest weight of evidence since it contains $\{\mathbf{St}\}$. Next, the search considers $\{\mathbf{G5}, \mathbf{G4}, \mathbf{S}, \mathbf{H}, \mathbf{P}, \mathbf{St5}, \mathbf{St4}\}$ and then $\{\mathbf{G5}, \mathbf{S}, \mathbf{H}, \mathbf{P}, \mathbf{St5}\}$ if it has lower joint expected weight of evidence for $\{\mathbf{St}\}$ than $\{\mathbf{G4}, \mathbf{S}, \mathbf{H}, \mathbf{P}, \mathbf{St4}, \mathbf{E}\}$. Next, the search expands $\{\mathbf{G5}, \mathbf{S}, \mathbf{H}, \mathbf{P}, \mathbf{St5}\}$. If the single test variable, $\{\mathbf{S}\}$ has larger expected weight of evidence for $\{\mathbf{St}\}$ than either $\{\mathbf{G5}\}$ or $\{\mathbf{H}, \mathbf{P}\}$, we decide to do that test.

[2] A *Markov tree* is a tree whose nodes represent groups of variables and for which all the nodes containing any single variable forms a connected subtree. Both the tree of cliques and the junction tree are Markov trees. See Shenoy and Shafer (1990) or Almond (1990).



a. Network for Staff feature *b. Corresponding Markov Tree*

Figure 3: A model for determining if a group of lines is a musical staff and its corresponding Markov tree

Known background variables play a big role in eliminating unnecessary tests. For example, four line staffs are only relevant in early music. *A priori* knowledge that the music was not early (E false) would cause the node $\{\mathbf{G4}, \mathbf{S}, \mathbf{H}, \mathbf{P}, \mathbf{St4}, \mathbf{E}\}$ to have zero weight of evidence; that branch of the tree would never be searched.

Incorporating test costs requires a little extra care. Branch and bound search will still work if we have a overestimate of the value of the intermediate nodes or an underestimate of the cost. Therefore, we propagate minimum costs towards the target node. For example, the cost associated with the node $\{\mathbf{H}, \mathbf{P}\}$ is the lower of the costs of tests $\{\mathbf{H}\}$ and $\{\mathbf{P}\}$. Now when branch and bound search reaches a node corresponding to a single test variable, there will be no test which yield more expected evidence per unit cost.

4 Non-Myopic Search: accounting for feature groups

The test selection strategy described above is *myopic*—it only looks at the effect of extracting one feature at a time. HHM criticize myopic test selection strategies as unrealistic: tests often come bundled in groups. For example, a line detector will return a data structure containing the location, direction, length and thickness of the candidate line; an EKG will typically provide values for a slew of variables. Further, a single costly test (such as biopsy) can often be more cost effective than a myriad of cheaper tests, yet a myopic strategy will choose the cheaper test (HMM provide an example from Pathfinder).

There are two challenges in going from myopic to non-myopic search. First, we must be able to calculate the joint expected weight of evidence of the test group, which gets computationally expensive as the size of the group increases (Section 4.1). Second, in the case where the application does not define the test groups, we must *find* the sets of variables with the largest joint expected weight of evidence; Section 4.3 discusses this problem.

4.1 Joint Expected Weight of Evidence

If $T^1, T^2, \ldots, T^p$ form such a group, we need to calculate their joint weight of evidence:

$$EW(H : T^1, T^2, \ldots, T^p) = \sum_{\mathcal{T}} \Pr(T^1_{i_1}, T^2_{i_2}, \ldots, T^p_{i_p} \mid H) \log \frac{\Pr(T^1_{i_1}, T^2_{i_2}, \ldots, T^p_{i_p} \mid H)}{\Pr(T^1_{i_1}, T^2_{i_2}, \ldots, T^p_{i_p} \mid \overline{H})} \quad (1)$$

where the summation is over $\mathcal{T}$, all the possible values of the tests in the group. However, as HMM and Jensen and Liang indicate, the number of terms in the summation above can lead to computational difficulties with even a modest group size.

Simulation based methods provide a workable solution to this problem. The key point is that (1) above is precisely the expectation of:

$$\log \frac{\Pr(T^1, T^2, \ldots, T^p \mid H)}{\Pr(T^1, T^2, \ldots, T^p \mid \overline{H})} \quad (2)$$

with respect to $\Pr(T^1, T^2, \ldots, T^p \mid H)$. Both simple Monte Carlo in the Markov tree and Markov chain Monte Carlo (Hastings, 1970, Kong, 1991, Neal, 1993, Buntine, 1994) provide methods for approximating this expectation, although we only describe the latter. We construct an irreducible Markov chain $\{T(t) = T^1(t), T^2(t), \ldots, T^p(t)\}$, for $t = 1, 2, \ldots$ with state space $\mathcal{T}$ which has equilibrium distribution $\Pr(T^1, T^2, \ldots, T^p \mid H)$. Then for any well-behaved function $g(T(t))$ defined on $\mathcal{T}$, if we simulate this Markov chain for $t = 1, \ldots, N$, the average:

$$\hat{g} = \frac{1}{N} \sum_{t=1}^{N} g(T(t)) \quad (3)$$

converges with probability one to $E(g(T))$ as N goes to infinity. To compute (1) in this fashion we set $g(T(t)) = \log \frac{\Pr(T^1(t), T^2(t), \ldots, T^p(t) \mid H)}{\Pr(T^1(t), T^2(t), \ldots, T^p(t) \mid \overline{H})}$.

To implement the Markov chain we define a neighborhood $\text{nbd}(T)$ for each $T \in \mathcal{T}$ which is the set of elements of $\mathcal{T}$ which differ from T in just one of $T^1, T^2, \ldots, T^p$ (larger neighborhoods are also possible). Define a transition matrix q by setting $q(T \to T') = 0$ for all $T' \notin \text{nbd}(T)$ and $q(T \to T')$ constant for all $T' \in \text{nbd}(T)$. If the chain is currently in state T, we proceed by drawing T' from $q(T \to T')$. We accept it with probability:

$$\min \left\{ 1, \frac{\text{pr}(T' \mid H)}{\text{pr}(T \mid H)} \right\}. \quad (4)$$

Otherwise the chain stays in state T. Diagnostics exist for assessing how many cycles are needed and how many should be discarded (Raftery and Lewis, 1992), for assessing convergence (Geyer, 1993), and for overcoming difficulties with multimodal discrete distributions (Lin, 1992).

As an illustration, we consider the coronary artery disease study of Detrano, *et al.* (1989). The study attempted to find clinical variables to predict the presence of coronary artery disease without an intrusive angiograph, and measured a number of clinical and test variables for 303 patients referred for to the Cleveland Clinic for coronary angiography. The data are available through the Murphy and Aha (1992) repository. The variables are as follows (the number in parentheses is the number of states for each variable):

Clinical Data: *Age* (3), *Sex* (2), *Rest-Bp* (3; Systolic blood pressure at rest), *Chest-Pain* (4).

Routine Test Data: *Chol* (4; Serum cholestoral in mg/dl), *Fast-Bsug* (2; Fasting Blood Sugar in mg/dl), *Rest-Ecg*(3; Electrocardiographic results at rest).

Exercise Data: *Max-Heart-Rate* (3), *Exer-Angina* (2; Exercise induced angina), *Old-Peak* (3; ST depression induced by exercise relative to rest), *Slope-Peak* (3; The slope of the peak exercise ST segment), *Exer-Thal-Defects* (3; Exercise thallium scintigraphic defects).
Experimental Non-invasive Test: *Colored-Floro* (4; Number of blood vessels colored by fluoroscopy).
Outcome Variables: *Health-State* (5), *Healthy?* (2).

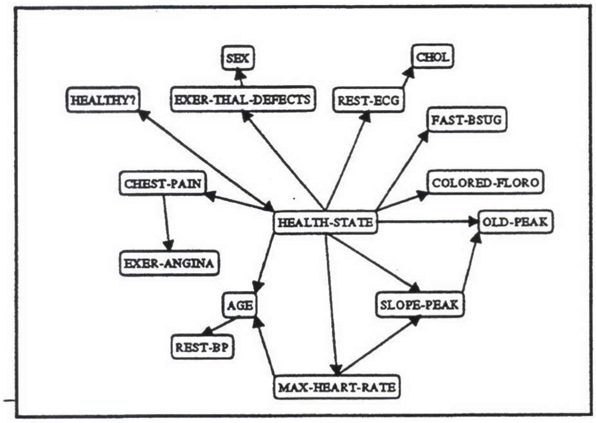


Figure 4: Belief network for the Heart model

Almond and Madigan (1993) selected a belief network model for these data and we show this model in Figure 4. Treating *Healthy*? as the target variable, we used the Markov chain Monte Carlo (MCMC) algorithm to compute the expected weight of evidence (EW) for a number of test groups, and we show the results in Table 1. This example is sufficiently small that we can also calculate the exact expected weight of evidence for each group. The results show that within 1,000 samples, the MCMC algorithm provides a reasonable approximation to the exact expected weight of evidence. Therefore, the MCMC approach will be useful in situations where the number of possible outcomes of the test group is larger than 1,000. Frequently the rank order of the expected weights of evidence for different test groups will be of primary importance. In that case, as few as 100 samples will often prove adequate.

4.2 Relationship to Decision-Theoretic Test Selection

The simulation approach outlined above provides a general solution to the problem addressed by HHM in the context of a specific class of graphs (graphs where the tests in the group are conditionally independent given the diseases or form a Markov chain structure in the graph). They show that the calculation of the value of perfect information for a test group, reduces to the assessment of a number of inequalities like:

$$\Pr\left(\log \frac{\Pr(T^1, T^2, \ldots, T^p \mid H)}{\Pr(T^1, T^2, \ldots, T^p \mid \overline{H})}\right) > W^* \tag{5}$$

where W^* depends on the actual utilities. To compute this using the Markov chain, set:

$$g(T) = \mathrm{I}\left(\log \frac{\Pr(T^1, T^2, \ldots, T^p \mid H)}{\Pr(T^1, T^2, \ldots, T^p \mid \overline{H})} > W^*\right), \tag{6}$$

where I is an indicator function. This approach does not require any constraints on the graph.

Table 1: Markov chain Monte Carlo computation of the expected weight of evidence for test groups in the coronary artery disease example

Indicant Group	*Exact EW*	*EW at 500 samples*	*EW at 1,000 samples*	*EW at 10,000 samples*
Serum Cholesterol: Chol fast-bug	0.15	0.57	0.14	0.18
Treadmill: exer-angina exer-thal-defects max-heart-rate slope-peak	45.87	52.81	45.89	42.86
Initial Observations: age sex chest-pain	24.96	26.98	26.67	24.92

4.3 Searching for Groups

A general solution to the nonmyopic test selection problem requires that we *find* the groups of variables which are most cost effective relative to some target hypothesis. Unfortunately, the individual test variables might be widely scattered through the Markov tree model and we do not have a fully satisfactory solution to the combinatorial problem that arises. HHM (Section VII) provide some suggestions.

Appendix: Some Properties of Expected Weights of Evidence

Proposition 1: The expected weight of evidence, $EW(H : T)$ is non-negative.

Proof: This proposition can be proven by use of the inequality (Gallager, 1968):

$$\ln(z) \leq z - 1; \qquad z > 0,$$

with equality when $z = 1$. We will show that $-EW(H : T) \leq 0$:

$$\begin{aligned} -EW(H:T) &= \sum_{j=1}^{n} \Pr(T_j \mid H) \log \frac{\Pr(T_j \mid \overline{H})}{\Pr(T_j \mid H)} \\ &\leq \sum_{j=1}^{n} \Pr(T_j \mid H) \left(\frac{\Pr(T_j \mid \overline{H})}{\Pr(T_j \mid H)} - 1 \right) \\ &= \sum_{j=1}^{n} \left(\Pr(T_j \mid \overline{H}) - \Pr(T_j \mid H) \right) \\ &\equiv 0. \end{aligned}$$

We need the following extension of Proposition 1 in the proof of Theorem 1.

Proposition 2: The conditional expected weight of evidence, $EW(H:T\mid S)$ is non-negative.

Proof: For notational simplicity we omit the subscripts in what follows.

$$\begin{aligned}
-EW(H:T\mid S) &= \sum_{S,T}\Pr(S,T\mid H)\log\frac{\Pr(T\mid \overline{H},S)}{\Pr(T\mid H,S)}\\
&\leq \sum_{S,T}\Pr(S,T\mid H)\left(\frac{\Pr(T\mid \overline{H},S)}{\Pr(T\mid H,S)}-1\right)\\
&= \sum_{S,T}\left(\Pr(S,T\mid \overline{H})\frac{\Pr(S\mid H)}{\Pr(S\mid \overline{H})}-\Pr(S,T\mid H)\right)\\
&= \left(\sum_{S}\Pr(S\mid H)\sum_{T}\Pr(T\mid S,\overline{H})\right)-1\\
&\equiv 0.
\end{aligned}$$

The joint expected weight of evidence can be decomposed into a marginal weight of evidence and a conditional weight of evidence:

Proposition 3: $EW(H:A,B)=EW(H:A)+EW(H:B\mid A)$

Proof: We have that:

$$\begin{aligned}
EW(H:A)+EW(H:B\mid A) &= \sum_{A}\Pr(A\mid H)\log\frac{\Pr(A\mid H)}{\Pr(A\mid \overline{H})}+\sum_{A,B}\Pr(A,B\mid H)\log\frac{\Pr(B\mid H,A)}{\Pr(B\mid \overline{H},A)}\\
&= \sum_{A}\Pr(A\mid H)\log\frac{\Pr(A\mid H)}{\Pr(A\mid \overline{H})}+\\
&\quad \sum_{A,B}\Pr(A,B\mid H)\log\frac{\Pr(A,B\mid H)\Pr(A\mid \overline{H})}{\Pr(A,B\mid \overline{H})\Pr(A\mid H)}\\
&= \sum_{A,B}\Pr(A,B\mid H)\log\frac{\Pr(A,B\mid H)}{\Pr(A,B\mid \overline{H})}\\
&= EW(H:A,B)
\end{aligned}$$

Proposition 4 (Independence): If T is conditionally independent of H given S, then $EW(H:T\mid S)=0$.

Proof: By independence, $\Pr(T\mid H,S)=\Pr(T\mid S)$. Therefore $W(H:T\mid S)=0$ and the result follows.

Theorem 1 (Monotonicity): For a Berge network with three nodes where B separates A from H (Figure 1), $EW(H:A)\leq EW(H:B)$.

Proof: From Proposition 3,

$$EW(H:A)+EW(H:B\mid A)=EW(H:A,B)=EW(H:B)+EW(H:A\mid B)$$

By Proposition 2 we have that $EW(H:B\mid A)\geq 0$. Because of the topology of the graph, A is independent of H given B and hence by Proposition 4 $EW(H:A\mid B)=0$. The result follows.

References

Almond, R.G. [1990]. *Fusion and Propagation in Graphical Belief Models: An Implementation and an Example.* Ph.D. dissertation and Harvard University, Department of Statistics Technical Report S-130. Revised version to be published as a monograph from Chapman and Hall.

Almond, R.G. [1993]. "Lack of Information Based Control in Expert Systems." In Hand, D.J (ed). *Artificial Intelligence Frontiers in Statistics: AI and Statistics III*, Chapman and Hall, pp 82–89.

Almond, R.G., M.Y. Jaisimha, E. Arbogast, and S.F. Elston[1994]. "Intelligent Image Browsing and Feature Extraction." *StatSci Research Report No. 25,* 1700 Westlake Ave, N. Suite 500, Seattle, WA 98117.

Almond, R.G., and D. Madigan[1993]. "Using GRAPHICAL-BELIEF to Predict Risk for Coronary Artery Disease." *StatSci Research Report No. 19,* 1700 Westlake Ave, N. Suite 500, Seattle, WA 98117.

Barr, A. and Feigenbaum, E. [1982]. *Handbook of Artificial Intelligence,* volume 2. Kaufmann, Los Altos.

Ben-Basset, M. [1978]. "Myopic policies in sequential classification." *IEEE Transactions on Computing,* (**27**, 170–174.

Breiman, L., J. Friedman, and C. Stone[1984]. *Classification and Regression Trees.* Wadsworth International.

Buntine, W.[1994]. "Learning with Graphical Models." *Technical Report FIA-94-03,*, NASA Ames Research Center.

Dawid, A. P. [1992]. "Applications of a general propagation algorithm for probabilistic expert systems." *Statistics and Computing,* (**2**), 25–36.

Detrano, R., A. Janosi, W. Steinbrunn, M. Pfisterer, J-J. Schmid, S. Sandhu, K.H. Guppy, S. Lee, and V. Froelicher [1989]. "International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease." *American Jouran of Cardiology,* (**64**) 304–310.

Gallager, R.G.[1968]. *Information theory and reliable communication.* John Wiley : New York.

Geyer, C.J. [1993]. "Practical Markov chain Monte Carlo." *Statistical Science*, **8**, .

Glasziou, P. and J. Hilden[1989]. "Test Selection Measures." *Medical Decision Making* (**9**), 133-141.

Good, I.J. [1985]. "Weight of Evidence : a brief survey." In : *Bayesian Statistics 2*, Bernardo, J.M., DeGroot, M.H., Lindley, D.V., and Smith, A.F.M., eds, North Holland : New York, 249–269.

Good, I. J. and Card, W. [1971]. "The diagnostic process with special reference to errors." *Method of Inferential Medicine.* (**10**), 176–188.

Hastings, W.K. [1970]. "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, **57**,97–109.

Heckerman, D., E. Horvitz, and B. Middleton[1993]. "An Approximate Nonmyopic Computation for Value of Information." *IEEE Transaction of Pattern Analysis and Machine Intelligence* (**15**), 292–298.

Heckerman, D., E. Horvitz, and B.N. Nathwani[1992]. "Toward normative expert systems: Part I. The Pathfinder Project." *Methods of Information in Medicine* (**31**), 90–105.

Jensen, F.V., [1988]. "Junction trees and decomposable hypergraphs." JUDEX Research Report, Aalborg, Denmark.

Jensen, F.V. and J. Liang[1994]. "drHugin: A system for hypothesis driven data request." In: *Bayesian Belief Networks and Probabilistic Reasoning*, Gammerman, A., ed., UNICOM : London, to appear.

In : *Bayesian Statistics 2*, Bernardo, J.M., DeGroot, M.H., Lindley, D.V., and Smith, A.F.M., eds, North Holland : New York, 249–269.

Kong, A.[1991]. "Efficient Methods for Computing Linkage Likelihoods of Recessive Diseases in Inbred Pedigrees." *Genetic Epidemiology*, **8**, 81–103.

Lauritzen, S.L. and Spiegelhalter, D.J. [1988]. "Local Computation with Probabilities on Graphical Structures and their Application to Expert Systems (with discussion)." *Journal of the Royal Statistical Society (Series B)*, **50**, 205–247.

Lin, S. [1992]. "On the performance of Markov chain Monte Carlo methods on pedigree data and a new algorithm." *Technical Report 231*, Department of Statistics, University of Washington.

Madigan, D., Mosurski, K., and Almond, R.G. [1994]. "Explanation in Belief Networks." Submitted for publication.

Madigan, D. and Almond, R.G.[1993]. "Test Selection Strategies for Belief Networks" StatSci Research Report 20.

McSherry, D.M.G.[1986]. "Intelligent dialogue based on statistical models of clinical decision making." *Statistics in Medicine*, **5**,497–502.

Matheson, J.E.[1990]. "Using Influence Diagrams to Value Information and Control." *Influence Diagrams, Belief Nets and Decision Analysis,* Oliver, Robert M. and Smith James Q. (ed.) John Wiley & Sons.

Miller, P. [1983]. "ATTENDING: Critiquing a physician's management plan." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* (**5**, 449–461.

Murphy, P.M. and Aha, D.W.[1992]. *UCI Repository of Machine Learning Databases.* Online database maintained at the Department of Information and Computer Science, University of California, Irvine, CA.

Neal, R.M. [1993]. "Probabilistic inference using Markov chain Monte Carlo methods." *Technical Report CRG-TR-93-1*, Department of Computer Science, University of Toronto.

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

Raftery, A.E. and Lewis, S.L. [1992]. "How many iterations in the Gibbs sampler?" In : *Bayesian Statistics 4*, Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M., eds, Oxford University Press : Oxford, 763–773.

Shenoy, P.P. and Shafer, G. [1990]. "Axioms for Probability and Belief-Function Propagation." in *Uncertainty in Artificial Intelligence,* 4, 169-198.

Spiegelhalter, D.J. and Knill-Jones, R.P. [1984]. "Statistical and knowledge based approaches to clinical decision support systems, with an application in gastroenterology (with discussion)." *Journal of the Royal Statistical Society (Series A)* (**147**), 35–77.