A Tractable Inference Algorithm for Diagnosing Multiple Diseases*

David Heckerman
Medical Computer Science Group
Knowledge Systems Laboratory
Departments of Computer Science and Medicine
Stanford, California 94305

Abstract

We examine a probabilistic model for the diagnosis of multiple diseases. In the model, diseases and findings are represented as binary variables. Also, diseases are marginally independent, features are conditionally independent given disease instances, and diseases interact to produce findings via a noisy OR-gate. An algorithm for computing the posterior probability of each disease, given a set of observed findings, called quickscore, is presented. The time complexity of the algorithm is $O(nm^{-2m^+})$, where $n$ is the number of diseases, $m^+$ is the number of positive findings and $m^-$ is the number of negative findings. Although the time complexity of quickscore is exponential in the number of positive findings, the algorithm is useful in practice because the number of observed positive findings is usually far less than the number of diseases under consideration. Performance results for quickscore applied to a probabilistic version of Quick Medical Reference (QMR) are provided.

1 Introduction

One of the most common criticisms of the use of probability theory in expert-system applications is that the theory is impractical to apply in realistic situations [1]. In attempts to answer this criticism, several researchers at Stanford, myself included, have undertaken the task of converting to a probabilistic framework Quick Medical Reference (QMR) [10], one of the largest medical expert systems in existence.

We have made a straightforward and tractable tranformation of the QMR knowledge base to a probabilisitic model [6,8]. Like the heuristic algorithms in QMR, the model allows for any combination of diseases to be present in a patient. Unfortunately, this feature leads to a time complexity for inferring the probability of each disease given a set of findings that is exponential in the number of diseases ($O(2^n)$) for all known algorithms. As there are over 600 diseases in QMR, these algorithms are intractable. Furthermore, this problem is known to be NP-hard [2].

In this paper, I present quickscore, an algorithm with time complexity that is exponential in the number of positive findings. More precisely, the algorithm has a time complexity of $O(nm^{-2m^+})$, where $m^+$ is the number of positive findings and $m^-$ is the number of negative findings. Although quickscore has an exponential time complexity, it is useful in practice because the number of observed positive findings is often far less than the number of diseases under consideration. For many realistic patient cases, quickscore implemented on a Macintosh II produces an answer in less than 1 minute of real time.

*This work was supported by the NSF under Grant IRI-8703710, and by the NLM under Grant R01-LM04529.
Diseases are marginally independent. Findings are conditionally independent given a disease instance.

2 The QMR model

The probabilistic version of QMR is called QMR-DT for Decision-Theoretic QMR. A belief network for QMR-DT, is shown in Figure 1. Each of the $n$ nodes in the upper layer of the network represents a disease that may be present or absent in a patient. Each of the $m$ nodes in the lower layer represents a finding that may be observed to be present or absent in the patient, or that may not be observed at all. The problem of interest is to compute the probability of each disease given a set of findings.

As indicated by the network in Figure 1, we assume diseases to be marginally independent. Also, we assume that findings are conditionally independent given any disease instance, where a disease instance is an assignment of true or false to each disease. A disease instance for $n = 3$, for example, is $d_1$ present, $d_2$ present, and $d_3$ absent.

We also assume that diseases act independently to cause any given finding to be present. A belief network that represents this independency for two diseases is shown in Figure 2. The node labeled $d_i$ causes $f$ represents the event that $d_i$ causes finding $f$ to be present. The node with the double boundary, labeled OR, is a deterministic node that says finding $f$ will be present if $d_1$ causes $f$ to be present, $d_2$ causes $f$ to be present, or both $d_1$ and $d_2$ cause $f$ to be present. The arc between the node labeled $d_1$ and the node labeled $d_1$ causes $f$ reflects our assumption that the presence or absence of $d_1$ influences the probability that $d_1$ causes $f$ to be present. In particular, we assume that, if $d_1$ is present, it may cause $f$ to be present with some probability. If $d_1$ is absent, we assume that the disease cannot act to cause $f$. The same set of assumptions holds for the disease $d_2$. Finally, the missing arc between the two oval nodes reflects our assumption that knowing whether or not $d_1$ causes $f$ to be present does not affect the probability that $d_2$ causes $f$ to be present. We call this form of conditional independence causal independence to distinguish it from the type of conditional independence that is commonly represented in belief networks.

Suppose that these assumptions are true, and that only the two diseases $d_1$ and $d_2$ can cause $f$ to be present. Under these assumptions, we can compute the probability of $f$ given any instance of the pair $\{d_1, d_2\}$ from the two assessments $p_1$ and $p_2$, where $p_i$ is the probability that $d_i$ causes $f$,

\[ p(f^+ | \text{only } d_i) = p_i \]

and

\[ p(f^- | \text{only } d_i) = 1 - p_i \]

where $f^+$ and $f^-$ denote the presence and absence of finding $f$, respectively. That is, the probability that $d_i$ causes $f$ is just the probability of $f$ given that only disease $d_i$ is present. If both $d_1$ and $d_2$
If disease $d_1$ is present, then with some probability it will act to cause finding $f$. This probability does not depend on whether $d_2$ acts to cause $f$. A reciprocal relationship holds for disease $d_2$. If either disease acts to cause $f$, then $f$ will be observed to be true.

are present, finding $f$ will be absent only if both diseases fail to cause $f$ to be present. Therefore,

$$p(f^-|d_1^+, d_2^+) = (1 - p_1)(1 - p_2) = p(f^-|\text{only } d_1)p(f^-|\text{only } d_2)$$

where $d_i^+$ denotes the presence of disease $d_i$.

More generally, suppose that $n$ diseases can potentially cause finding $f$, and that these $n$ diseases are the only causes of $f$. As in the simpler case,

$$p(f^+|\text{only } d_i) = p_i$$

where $p_i$ is the probability that $d_i$ causes $f$. Let $D_i$ denote a particular instance of the $n$ diseases and let $D_i^+$ denote the set of diseases that are present in the instance $D_i$. Given instance $D_i$, $f$ will be absent only if none of the diseases in $D_i^+$ cause $f$ to be present. It follows that

$$p(f^-|D_i) = \prod_{d \in D_i^+} p(f^-|\text{only } d)$$

Thus, using Equation 1, we can compute the probability of $f$ given any of the $2^n$ disease instances from only $n$ assessments.

The assumptions underlying Equation 1, including the assumption of causal independence, have been described previously [3], and several researchers have suggested that these assumptions be used to model medical domains [4,13]. Pearl has called this canonical model of cause and effect the noisy OR-gate [9,11]. Henrion [7] has extended the model to accommodate situations where a finding may appear in the absence of all explicitly represented diseases, a leaky OR-gate.

The leaky OR-gate model was assumed implicitly by the developers of the QMR expert system [10]. In conjunction with the independence assumptions shown in Figure 1, the model made
tractable the transformation of QMR into a probabilistic framework. As we see in the following section, these assumptions also accommodate an inference algorithm that is tractable for many patient cases. In Section 5, we examine the appropriateness of these assumptions, and discuss how they might be relaxed.

3 The Quickscore Algorithm

The goal of quickscore is to compute the probability of each disease $d_i$, $i = 1, 2, \ldots, n$, given a set of positive findings $f \in F^+$ and a set of negative findings $f \in F^-$, under the assumptions described in the previous section. To understand how quickscore works, we first compute the probability that a single finding $f$ will be absent. Using the expansion rule, we get

$$p(f^-) = \sum_{D_k \in D} p(f^-|D_k)p(D_k)$$

(2)

where $D$ is the set of all disease instances. Using the assumption that diseases are marginally independent and the assumptions underlying the noisy OR-gate, Equation 2 becomes

$$p(f^-) = \sum_{D_k \in D} \left[ \prod_{d \in D_k^+} p(f^-|\text{only } d) \prod_{d \in D_k^+} p(d^+) \prod_{d \in D_k^-} p(d^-) \right]$$

(3)

where $D_k^-$ denotes the set of diseases that are absent in the instance $D_k$.

Now consider the expression

$$\prod_{i=1}^n \left[ p(f^-|\text{only } d_i) p(d_i^+) + p(d_i^-) \right]$$

(4)

If we multiply out this expression, we see that it is just the right-hand side of Equation 3. Thus, we obtain

$$p(f^-) = \prod_{i=1}^n \left[ p(f^-|\text{only } d_i) p(d_i^+) + p(d_i^-) \right]$$

(5)

The difference in time complexity between the computations of Equations 3 and 5 is striking. The computation in Equation 3 is a sum over $2^n$ terms, whereas the computation in Equation 5 is a linear product over $n$ sums. Pearl [12] (pp. 187-188) was the first to note the equivalence between these two computations. As we shall see, quickscore derives its speed from this equivalence.

Under the assumption that findings are conditionally independent given any disease instance, we can employ the transformation described in the previous paragraph to compute the probability that the set of negative findings, $F^-$, are observed. We have

$$p(F^-) = \prod_{i=1}^n \left( \left[ \prod_{f \in F^-} p(f^-|\text{only } d_i) \right] p(d_i^+) + p(d_i^-) \right)$$

(6)

The situation is more complex for positive findings. Let us first examine the simple case where $F^+ = \{f_1, f_2\}$. Applying the expansion rule to $p(f_1^+, f_2^+)$ and using the assumption of conditional independence of findings, we get

$$p(f_1^+, f_2^+) = \sum_{D_k \in D} p(f_1^+|D_k)p(f_2^+|D_k)p(D_k)$$

(7)
Since \( p(f_j^+|D_k) = 1 - p(f_j^-|D_k) \), Equation 7 becomes
\[
p(f_1^+, f_2^+) = \sum_{D_k \in D} p(D_k) - \sum_{D_k \in D} p(f_1^-|D_k)p(D_k) - \sum_{D_k \in D} p(f_2^-|D_k)p(D_k) + \sum_{D_k \in D} p(f_1^-|D_k)p(f_2^-|D_k)p(D_k)
\] (8)

The first sum in Equation 8 is equal to 1. The remaining terms are in the same form as the right-hand side of Equation 2. Thus, using the algebraic transformations derived previously, we obtain
\[
p(f_1^+, f_2^+) = 1 - \prod_{i=1}^{n} \left[ p(f_i^-|\text{only } d_i)p(d_i^+) + p(d_i^-) \right] - \prod_{i=1}^{n} \left[ p(f_i^+|\text{only } d_i)p(d_i^-) + p(d_i^+) \right] + \prod_{i=1}^{n} \left[ p(f_i^-|\text{only } d_i)p(f_i^+|\text{only } d_i)p(d_i^+) + p(d_i^-) \right]
\] (9)

More generally,
\[
p(F^+) = \sum_{F' \in 2^{F^+}} (-1)^{|F'|} \prod_{i=1}^{n} \left( \prod_{f \in F'} p(f^-|\text{only } d_i) \right) p(d_i^+) + p(d_i^-)
\] (10)

where \( 2^{F^+} \) denotes the power set of \( F^+ \), and \( |F'| \) denotes the number of elements in set \( F' \).

In the most general case where some findings are present and some are absent, we can combine Equation 6 and Equation 9 to obtain
\[
p(F^+, F^-) = \sum_{F' \in 2^{F^+}} (-1)^{|F'|} \prod_{i=1}^{n} \left( \prod_{f \in F' \cup F^-} p(f^-|\text{only } d_i) \right) p(d_i^+) + p(d_i^-)
\] (11)

It is now a simple matter to compute \( p(d_i|F^+, F^-) \). First, we compute \( p(F^+, F^-) \). Then, we compute \( p(d_i|F^+, F^-) \) by setting \( p(d_i^+) = 1 \) in Equation 11. The sought-after probability is then
\[
p(d_i|F^+, F^-) = \frac{p(d_i|F^+, F^-)}{p(F^+, F^-)}
\] (12)

by the product rule of probability.

The quickscore algorithm can provide intermediate results. In particular, suppose that we order the findings in \( F^+ \)—say, \( f_1, f_2, \ldots f_m^+ \). In Equation 11, we can first compute the term in the power set of \( F^+ \) corresponding to \( f_1 \) alone. We can then compute the terms in the power set that correspond to combinations of only \( f_1 \) and \( f_2 \). Continuing in this way, the probability of each \( d_i \) given the first \( j \) findings can be recovered from the algorithm at any time, where time is an exponential function of \( j \). The fact that quickscore can provide intermediate results may prove useful. The QMR knowledge base contains a partial ordering of findings by their general clinical importance in determining a diagnosis. This ordering might be used to degrade gracefully the performance of quickscore with increasing time constraints.
Run times for quickscore implemented in LightSpeed Pascal on a Macintosh II are plotted against the number of positive findings in a case. Nine positive findings typically can be scored in less than 1 minute.

4 Run-Time Performance of Quickscore

Quickscore has been implemented in LightSpeed Pascal on the Macintosh II computer. Figure 3 shows the run time of the algorithm for cases of various size. The cases are taken from a library of classic cases used by the QMR research team to test periodically the diagnostic accuracy of the heuristic knowledge base. These cases contain only positive findings. As the graph indicates, nine findings can typically be scored in less than 1 minute. Overall, in 25 percent of the 400 cases in the library, quickscore requires 15 minutes or less to score each case.

5 Relaxing the Assumptions

The diagnostic model described in this paper contains several assumptions that may not be appropriate for many medical and nonmedical domains. For example, some diseases and findings can occur with different degrees of severity, and hence are not two-valued. In addition, certain diseases cause others to be present. Consequently, all diseases are not marginally independent. Furthermore, some findings are conditionally dependent, and certain diseases are caused by findings rather than vice-versa (e.g., a history of alcoholism tends to cause cirrhosis of the liver).

None of these observations pose a significant barrier to the translation of the QMR's knowledge base to a probabilistic framework. For example, several researchers have generalized the noisy OR-gate model, constructing other prototypical models that embody the assumption of causal independence [9,5]. We can use these prototypical models to accommodate multiple-valued diseases and findings, and to accommodate causal interactions among diseases. Also, by introducing hidden or unobservable pathophysiologic states, we can capture many of the dependencies among findings. We can use the noisy OR-gate and its extensions to model interactions between disease and
pathophysiologic state.

We cannot extend quickscore in a straightforward manner to treat some of the extensions to the current QMR-DT model. In particular, marginally dependent diseases and conditionally dependent findings pose a significant challenge. However, the fact that the noisy or-gate and similar prototypical models will be used to model these dependencies offers promise for a solution.

6 Conclusion

Despite its shortcomings, the current QMR-DT model may significantly outperform its heuristic counterpart. If evaluations of QMR-DT and QMR currently underway verify this conjecture, the quickscore algorithm will become an invaluable component of the expert system. In either case, quickscore is likely to be a useful tool for knowledge engineers who develop probabilistic or decision-theoretic expert systems in other domains.

7 Acknowledgments

I thank Gregory Copper for encouraging me to implement the quickscore algorithm, Michael Shwe for assisting with the implementation, the entire QMR-DT group at Stanford for useful discussions, and Lyn Dupre for reviewing several drafts of this manuscript.

References


