The Bayesian Basis of Common Sense Medical Diagnosis[*]

Eugene Charniak

Department of Computer Science
Brown University
Providence, Rhode Island 02912

ABSTRACT

In this paper, we show that the objections most frequently raised against the use of Bayesian statistics within the AI-in-medicine community do not seem to hold. In particular, we will show that the independence assumptions required to make Bayesian statistics computationally feasible are not nearly as damaging as has been claimed. We will also argue that Bayesian statistics is perfectly compatible with heuristic solutions to the multiple disease problem.

I Introduction

While the mathematics of conditional probabilities in general, and Bayesian statistics in particular, would seem to offer a foundation for medical diagnosis (and other cases of decision making under uncertainty), such approaches have been rejected by most "artificial intelligence in medicine" researchers. Typically, Bayesian statistics have been rejected for the following reasons.

1) In its pure form it would require an impossible number of statistical parameters.

2) The only way to escape from (1) is to impose absurd statistical independence assumptions [7, 9].

3) And at any rate, Bayesian statistics only works for the single disease situation [3, 6]

In this paper we will argue that while objection (1) is correct, (2) is not, making (1) moot. Furthermore, while (3) seems to be valid, even there, Bayesian statistics is perfectly compatible with various heuristic solutions to the multiple-disease problem. To reject Bayesian statistics on the basis of (3) would be like rejecting closed-form solutions to differential equations because the toughest ones must be solved numerically. Thus we will claim that Bayesian statistics offers a realistic basis for practical medical diagnosis.

On the other hand, we will not argue that adoption of Bayesian statistics will lead to better diagnosis programs! Instead we will suggest that the common sense application of Bayesian statistics will lead to programs that will be hard to distinguish from current "non-Bayesian" programs. Or, to put it slightly differently, while programs such as MYCIN [8] and Internist [6] profess to be non-Bayesian, when one ignores what their authors say, and concentrate on what the programs do, they turn out to be remarkably Bayesian after all. Thus the goal here is one of clarification.

However, this is not to say that absolutely everyone is against Bayesian statistics. Duda, Hart and Nilsson [1] propose a scheme for combining a general inference system with Bayesian statistics. Indeed, the scheme presented here can be thought of as a continuation of their line of research, since we will assume, as they do, that the initial probabilities will be obtained from intuitive "guesses" by experts. They also have a nice discussion of how to handle the inevitable contradictory numbers that will arise from such a process. In related work, Pearl [4] shows how distributed systems can quickly update such probabilities.

However, to date there is no evidence that such work has had any influence on AI-in-medicine researchers, and it is not to hard to see why. All such systems require very strong assumptions, regarding both independence of symptoms and exclusivity of diseases — assumptions which are widely believed not to hold in the medical domain. These reservations have not been addressed. This we intend to do.

II Bayes's Theorem

Bayesian statistics is based upon the idea of conditional probabilities. If we have observed symptoms $s_1 \cdots s_n$ then the best disease to postulate would be that disease d_i that maximized $P(d_i | s_1 \cdots s_n)$, the conditional probability of d_i given $s_1 \cdots s_n$[1]. Unfortunately no one knows the necessary numbers to compare. Indeed, given that we would need such numbers for all possible subsets of symptoms, the number of such parameters is astronomical. When we are dealing with fifty or sixty findings, in all likelihood the particular combination is unique in the history of the universe. In such cases it is difficult to collect reliable statistics.

Now Bayes's formula offers a way to compute such numbers. The form we will use in this paper is this:

$$P(d_i | s_1 \cdots s_n) = \frac{P(d_i) * P(s_1 \cdots s_n | d_i)}{P(s_1 \cdots s_n)}$$

The standard form of Bayes's theorem has as its denominator the following term:

$$\sum_{j=1}^{m} P(s_1 \cdots s_n | d_j) * P(d_j)$$

But, this form requires the d_j's to be exhaustive and exclusive — things the first version does not require.

Unfortunately we do not have the numbers to plug into Bayes's theorem either. The standard answer to

[*] This paper came into being in an attempt to explain the limitations of Bayesian statistics. To my initial horror, the more I pressed, the less important the limitations seemed to be. However, in coming to this conclusion I had to have a lot of my confusions cleared up, and for this I would like to thank Stuart Geman, Drew McDermott, Jeffery Vitter, and the students of CS201 (particularly Jim Hendler and David Wittenberg) for their help. My thanks also to Graeme Hirst, who made helpful comments on an earlier draft of this paper. The confusions which remain are, of course, my own.

[1] Our use of the terms "symptom" and "disease" are somewhat loose. By "symptom" we simply mean anything which could suggest the presence of a disease, including the presence of a second disease. Nothing however hangs on this.

this problem is to make *statistical independence* assumptions. We shall return shortly to discuss the reasonableness of these assumptions. For the moment we will consider how they help the situation. We can simplify this last equation by making two such independence assumptions. The first is the independence of two symptoms. Formally, this assumption is this:

$$P(s_i \mid s_j) = P(s_i)$$

From this we can infer the following:

$$P(s_i \& s_j) = P(s_i) * P(s_j)$$

The other assumption we need is that two symptoms are not simply independent among people at large, but also in the subset of people suffering from the disease d.

$$P(s_i \mid s_j \& d) = P(s_i \mid d)$$

This is equivalent to:

$$P(s_i \& s_j \mid d) = P(s_i \mid d) * P(s_j \mid d)$$

By making these two independence assumptions for all subsets of symptoms we can reduce Bayes's theorem to:

$$P(d_i \mid s_1 \cdots s_n) = \frac{P(d_i) * P(s_1 \mid d_i) * \cdots P(s_n \mid d_i)}{P(s_1) * \cdots * P(s_n)}$$

However, even this reduced load of statistical parameters are not known, but rather have to be obtained from physician's subjective estimates. It seems plausible that physicians have some ideas about the conditional probabilities of symptoms given diseases. This is what one learns in medical school. However, the *prior* probabilities $P(d_i)$ and $P(s_j)$ are at best known only within a few orders of magnitude.[2] Nevertheless, to ignore them completely, as is done by several programs, really amounts to saying that they are all the same, an obviously bad approximation.

It is instructive to rewrite the last equation as:

$$P(d_i \mid s_1 \cdots s_n) = P(d) * \left[\frac{P(s_1 \mid d_i)}{P(s_1)} \right] * \cdots * \left[\frac{P(s_n \mid d_i)}{P(s_n)} \right]$$

$$= P(d_i) * I(d_i \mid s_1) \cdots * I(d_i \mid s_n)$$

where we have defined $I(d \mid s) = \frac{P(s \mid d)}{P(s)}$. This reformulation is useful because it suggests how to modify previous probability estimates when we get a new symptom. Initially we give every disease its prior probability $P(d_i)$. To take a new symptom s_j into account we multiply the previous probability of disease d_i by $I(d_i \mid s_j)$ to get the probability in light of the newest information.

Actually, we can make this even simpler. We virtually always modify probabilities by multiplying them by some factor. Because of this, and because probabilities vary over such a wide range (the prior probability of a cough is $\approx 10^{-1}$, but that of some rare disease might be $\approx 10^{-10}$) it makes sense to use the logarithm of probabilities rather than the probabilities themselves. Taking the log of both sides of the last equation gives us this:

$$\log(P(d_i \mid s_1 \cdots s_n)) = \log(P(d_i)) + \log(I(d_i \mid s_1)) + \cdots + \log(I(d_i \mid s_n))$$

Henceforth we will write this as follows:

$$LP(d_i \mid s_1 \cdots s_n) = LP(d_i) + LI(d_i \mid s_1) + \cdots + LI(d_i \mid s_n)$$

So instead of keeping around the various conditional probabilities, we really need to know the logarithm of the I factors, and we modify our belief in the presence of a disease by adding in a number. The result will look much like the Internist[6] updating system (except that Internist does not use the prior probabilities).

III Independence

Now let us turn to the independence assumptions we needed in order to get this far. As we have seen, the need for such assumptions has been a big argument against a Bayesian approach. However, a bit of reflection should convince you that wild non-independence of symptoms will kill any scheme whatsoever. When we take a new symptom into account we have to know how it modifies our belief in different diseases. Independence tells us that the modifications it makes are independent of the other symptoms we have seen. If such simplifying assumption cannot be used, then the modifications change drastically depending on the exact group of symptoms we have seen, and, as we have already noted, such numbers are unavailable on a priori grounds.

But are these independence assumptions the *right* assumptions to make? In this section we will show that they are not as bad as they have been made out.

A. Independence of Symptoms

We made two assumptions, the independence of symptoms, and the independence of symptoms given a disease.

$$P(s_i \mid s_j) = P(s_i)$$
$$P(s_i \mid s_j \& d) = P(s_i \mid d)$$

Of these, the second is more reasonable than the first, if only because the first is completely unreasonable. To see why the independence of symptoms is such a bad assumption, note that if we have two symptoms that are typically the results of the same diseases, then we will tend to see them together a lot. As such, they will not be independent. For example, vomiting and diarrhea go together so commonly that

$$P(\mathit{diarrhea} \& \mathit{vomiting}) \gg P(\mathit{diarrhea}) * P(\mathit{vomiting})$$

This example is only the most obvious. A little thought should convince you that if two or more symptoms tend to suggest the same disease, that virtually assures that they will not be independent.[3]

Indeed, the independence-of-symptoms assumption is *so* bad that it is fortunate that it doesn't matter. Look again at Bayes's formula with the two independence assumptions factored in.

[2] To make this even worse, these cannot be estimated relative to some observable universe, like the patients a doctor sees, but rather must be with respect to people as a whole. This is another undesirable side effect of the independence assumptions.

[3] There is a misunderstanding in the literature related to this point. Pednault et. al[5] claim to show that for a system that obeys both of the independence assumptions, plus exhaustiveness and mutual exclusivity of diseases, "no updating can take place". Subsequently Pearl[4] argues that Pednault must be wrong because he has a provably correct updating scheme. Pednault's result is correct, but it has nothing to do with the process of updating. As suggested from the above discussion, it is rather the case that any such system must also have the unfortunate property that none of the symptoms shed any light on the diseases. Thus, updating works correctly, but because the symptoms are completely uncorrelated with diseases, the posterior probabilities are the same as the prior probabilities. It is for this reason that "no updating can take place".

$$P(d_i \mid s_1 \cdots s_n) = \frac{P(d_i)*P(s_1 \mid d_i)* \cdots P(s_n \mid d_i)}{P(s_1)* \cdots *P(s_n)}$$

The independence of symptoms gives us the denominator. It is important to note that since the denominator does not mention the disease we are talking about, it will be the same for all diseases. Thus, the error we make by assuming independence of symptoms will be equally factored into our estimate of the probability of all diseases. So not only will this have no effect on the rank ordering of which diseases are most probable, but the ratio of the probabilities of two diseases will remain unaffected (or equivalently, the difference between two log's of probabilities will stay the same). This simply suggests that we should not base decisions on the absolute values of the probabilities, a conclusion that Pople [6], citing an article by Sherman argues for on empirical grounds[4]

B. Independence of Symptoms Given a Disease

The other assumption we made was the independence of symptoms *given a particular disease.*

$$P(s_1 \mid s_2 \& d) = P(s_1 \mid d)$$

This is a much better assumption than plain independence of symptoms, but, as opposed to the latter, if it fails for particular cases it *could* affect the ranking of disease possibilities.

Let us consider a typical case where it does fail. A person with arterial sclerosis (which we abbreviate *asc*) will sometimes show lung symptoms. Also, if the patient shows one such lung symptom then he is more likely to show a second. In other words:

$$P(\mathit{lung\text{-}sym}_2 \mid \mathit{asc} \& \mathit{lung\text{-}sym}_1) \gg P(\mathit{lung\text{-}sym}_2 \mid \mathit{asc})$$

Thus the two symptoms are not independent given asc.

Again, however, there is an easy solution to this problem. To see it, consider the "common sense" explanation of what is going on in this case. A doctor will explain these happenings by saying that asc can cause heart complications. In particular, it can cause one of two *pathological states* called *right heart syndrome* and *left heart syndrome*. In the left heart syndrome, blood backs up into the lungs, so we see various lung symptoms. Thus, seeing one lung symptom probably means that the patient has left heart syndrome, and is more likely to show other lung symptoms.

Such explanations form a typical argument for "causal" reasoning about diseases. Causal reasoning is indeed powerful, but the next step is usually to contrast it with a Bayesian approach. The two are not mutually exclusive. Indeed there is a Bayesian analog of this kind of reasoning. To see how it works, we first introduce the pathological state *left heart syndrome* into our Bayesian reasoning, by defining the following:

P(asc|left-heart-syndrome)
P(left-heart-syndrome|lung-symptom)

Then, by analogy with causal reasoning we have this:

$$P(d \mid s) = P(d \mid \mathit{path\text{-}state})*P(\mathit{path\text{-}state} \mid s)$$

Strictly, this equation only holds given two conditions. The first is this:

$$P(d \mid \mathit{not\ path\text{-}state} \& s)*P(\mathit{not\ path\text{-}state} \mid s) \ll P(d \mid \mathit{path\text{-}state} \& s)*P(\mathit{path\text{-}state} \mid s)$$

Informally this says that *s* is only related to *d* through *pathological state.* The second assumption is this:

$$P(d \mid \mathit{path\text{-}state} \& s) = P(d \mid \mathit{path\text{-}state})$$

That is, it requires that *d* and *s* are independent given the pathological state. In cases where causal reasoning is appropriate, both of these assumptions should hold.

Thus, to mimic causal reasoning, we would first calculate the new probability of the pathological state in light of the symptom observed, and then see how our belief in a disease is modified by our belief in the pathological state.

If we have two symptoms associated with the pathological state, when we observe the second the appropriate equation is this:

$$LP(d \mid s_1 \& s_2) = LP(d \mid s_1) + LI(\mathit{path\text{-}state} \mid s_2)$$

Given $LI(\mathit{path\text{-}state} \mid s_2) > LI(d \mid s_2)$, this equation correctly predicts the following relation:

$$LP(d \mid s_1 \& s_2) > LP(d \mid s_1) + LI(d \mid s_2)$$

This result is of interest because it shows how the introduction of the pathological state accounts for the fact that s_1 and s_2 are not independent given *d*.

The point is that by assuming the existence of this pathological state we remove this common objection to our independence assumptions. On the other hand, note that we had to introduce another statistical parameter to get rid of the independence assumption, namely $P(d \mid \mathit{path\text{-}state})$. But this is not a case of merely substituting one parameter for another. If, as seems reasonable, all of the symptoms s_i associated with *path-state* are independent given *path-state*, then we have introduced only one new parameter while accounting for many previous cases of symptom dependence.

C. What To Do When All Else Fails

In the last section we saw that we could ignore the non-independence of symptoms, and fix the non-independence of symptoms given a disease by the inclusion of commonly recognized pathological states. However, this latter technique will not always work.

One example is given by Szolovits and Pauker [9]. The probability of finding a second heart murmur given a diagnosis of heart disease will not be independent of the first murmer. While it would be possible to introduce a "pathological state" to account for this, my medical informant tells me that this is not natural (as opposed to the "right heart syndrome" situation), and instead the doctor must simply learn the relations.

But if a doctor must handle this as a special case, then it does not seem unreasonable for our program to do the same thing. Furthermore, we already have at hand a natural way to do this. The technique is provided by MYCIN [8]. In MYCIN, rather than restrict the program to inference rules of the form

symptom ==> disease (with likelihood = X)

(which would be the MYCIN equivalent of a Bayesian conditional probability), the program also allows

(and symptom1
symptom2 ...
symptom-*n*) ==> disease (with likelihood = X)

This latter situation will handle the case when the symptoms are not independent. Admittedly, we need to worry about cases where, say, all but one of the symptoms are present, and the last is unknown (a case MYCIN does not really handle well), but the extension is not all that

[4] Unfortunately, my (possibly early) copy of the Pople article has no reference information about the Sherman article.

complicated.

The point is that if there are cases where independence is violated, then reasonably obvious extensions to the "normal" independent case seem to do the job.

IV Multiple Diseases

The last of the major objections is that Bayesian statistics can only deal with situations where the alternatives are mutually exclusive. Thus there is a real problem in interpreting Bayesian results in cases where more than one disease may be present. If we know there is a single disease then we need only pick the one with the highest posterior probability. If there may be more than one, then what do we do?

There is a standard way to accommodate Bayesian statistics to the multiple-disease situation, and that is by creating new "diseases" which are really all pairs, triples, etc. of diseases. Of course, we will have to stop somewhere, or else we would have an infinite number of these new "diseases", but this number could be made large. As is commonly recognized, however, this "solution" is computationally untractable.

Thus Bayesian statistics probably says nothing useful about the multiple-disease situation. However, it is false that Bayesian statistics precludes handling multiple diseases. Firstly, the version of Bayes's theorem we used does not require the diseases to be mutually exclusive (the more typical version does, which is why we did not use the typical version). Next, suppose we use Bayes's theorem as in the single-disease case, and get a list of diseases, ordered by probability. Now, of course, more than one may have high probability. While Bayesian statistics says nothing about how to make use of such a list, there are well known heuristic techniques which take off from this point. In particular we refer here to the method used in Internist. There the highest ranking disease is used to define a subset of the previously postulated diseases which all explain roughly the same symptoms. Internist assumes, (in accordance with Occams Razor) that only one of these diseases will be present, and tries to distinguish between them by calling for further tests. If these support the previous highest ranking disease, then it is concluded to be present. At this point the "winning" disease, and its competitors, are all removed from the list of postulated diseases along with the symptoms the winner accounted for. If there are still remaining diseases on the list, the process is repeated, with the new probabilities computed on only the remaining symptoms. This goes on until no more diseases are left, or until the assumption of no further diseases is more plausible than any of the remaining diseases.

Now admittedly, this procedure owes nothing to Bayesian statistics. However, there is nothing in it incompatible with Bayesian statistics. We can start by collecting Bayesian statistics on the probabilities of each of the diseases. Indeed, this is virtually exactly what Internist does, although it is not labeled as such.

V Conclusion

Our conclusions can be summarized as follows:

1) While nobody has detailed probabilities for diseases, symptoms, etc, doctors do have order of magnitude estimates.

2) The independence assumptions are not nearly as bad as commonly portrayed. The lack of independence of symptoms is no problem, and the lack of independence of symptoms given diseases often (most of the time?) can be removed by introducing pathological states and causal reasoning – something most AI programs do anyway.

3) Where independence does break down completely, we can use techniques such as those in MYCIN. (Indeed, the lack of independence should be the only reason for introducing such techniques.)

4) As far as we can tell, multiple diseases cannot be handled by Bayesian methods. However, using Bayesian methods as if there were no multiple diseases will produce valid results up until the end, when it is no longer possible to pick the top disease as the "winner". Then we must use heuristic methods.

Overall our position in this paper has been conservative in that we have only endeavored to reconstruct already existing programs from a mathematically secure basis. It is possible to hope, however, that more might result from the recognition that mathematical rigor, and common sense practicality, do not necessarily conflict in the domain of medical diagnosis. The recognition that formal logic need not be divorced from knowledge representation theory [2] has lead to renewed interest in the insights of mathematicians and philosophers. Any statistician will recognize that the mathematics in this paper is a couple of hundred years old. Statistics has not stood still in that time.

VI References

1. R. O. Duda, P. E. Hart, and N. J. Nilsson, "Subjective bayesian methods for rule-based inference systems," in *Proceedings of the 1976 National Computer Conference*, AFIPS Press (1976).

2. Patrick J. Hayes, "In defense of logic," pp. 559-565 in *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, (1977).

3. Stephen G. Pauker and Peter Szolovits, "Analyzing and simulating taking the history of the present illness: context formation," pp. 109-118 in *Computational Linguistics in Medicine*, ed. W. Schneider and A.L. Sagvall Hein, (1977).

4. Judea Pearl, "Reverend Bayes on inference engines: a distributed hierarchical approach," pp. 133-136 in *Proceedings of the National Conference on Artificial Intelligence, 1982*, (1982).

5. E. P. D Pednault, S. W Zucker, and L. V Muresan, "On the independence assumption underlying subjective bayesian updating," *Artificial Intelligence* **16**(2) pp. 213-222 (1981).

6. Harry Pople, "Heuristic methods for imposing structure on ill structured problems: the structuring of medical diagnostics," pp. 119-185 in *Unknown*, (forthcoming).

7. Edward H Shortliffe and Bruce G Buchanan, "A model of inexact reasoning in medicine," *Mathematical Biosciences* **23** pp. 355-356 (1975).

8. Edward H. Shortliffe, *Computer-Based Medical Consultations: MYCIN*, American Elsevier Publishing Company, New York (1976).

9. Peter Szolovitz and S. G. Pauker, "Categorical and probabilistic reasoning in medical diagnosis," *Artificial Intelligence* **11** pp. 115-144 (1978).