Handling Uncertainty in Artificial Intelligence, and the Bayesian Controversy

by Donald Gillies

Contents

1. Introduction

This paper is divided into two parts. In the first part (sections 2 and 3), I will describe briefly how advances in artificial intelligence (AI) in the 1970s led to the crucial problem of handling uncertainty, and how attempts to solve this problem led in turn to the emergence of the new theory of Bayesian networks. I will try to focus in this historical account on the key ideas and will not give a full account of the technical details. Then, in the second part (section 4) I will consider the implications of these new results for the long-standing controversy between Bayesians and non-Bayesians.

2. The breakthrough with expert systems in the 1970s

Research in AI began in the 1950s and many important ideas were developed by the pioneers. Then in the 1970s a breakthrough was produced by the creation of expert systems. The lead here was taken by the Stanford heuristic programming group, particularly Buchanan, Feigenbaum, and Shortliffe. What they discovered was that the key to success was to extract from an expert the knowledge he or she used to carry out a specialised task, and then code this knowledge into the computer. In this way they were able to produce 'expert systems' which performed specific tasks at the level of human experts. One of the most important of these early expert systems (MYCIN) was concerned with the diagnosis of blood infections. This system will now be briefly described, and it will then be shown that its implementation led to the problem of how to handle uncertainty in AI.

MYCIN was developed in the 1970s by Edward Shortliffe and his colleagues in collaboration with the infectious diseases group at the Stanford medical school. The medical knowledge in the area was codified into rules of the form: IF such and such symptoms are observed, THEN likely conclusion is such and such. MYCIN's knowledge base comprised over 400 such rules which were obtained from medical experts. An example of such a rule will be given in a moment, but first it would be as well to present some evidence of MYCIN's success.

To test MYCIN's effectiveness a comparison was made in 1979 of its performance with that of nine human doctors. The program's final conclusions on ten real cases were compared with those of the human doctors, including the actual therapy administered. Eight other experts were then asked to rate the ten therapy recommendations and award a mark, without knowing which, if any, came from a computer. They were requested to give 1 for a therapy which they regarded as acceptable and 0 for an unacceptable therapy. Since there were eight experts and ten cases, the maximum possible mark was 80. The results were as follows:[1]

MYCIN	52	Actual therapy	46
Faculty-1	50	Faculty-4	44
Faculty-2	48	Resident	36
Inf dis fellow	48	Faculty-5	34
Faculty-3	46	Student	24

So MYCIN came first in the exam, though the difference between it and the top human experts was not significant.

Let us now examine one of MYCIN's rules. The following rule is given by Shortliffe & Buchanan:[2]

"If: (1) the stain of the organism is gram positive (S_1), and
(2) the morphology of the organism is coccus (S_2), and
(3) the growth conformation of the organism is chains (S_3)
Then: there is suggestive evidence (0.7) that the identity of the organism is streptococcus (H_1)"

In symbols this could be written: If S_1 & S_2 & S_3 , then there is suggestive evidence p that H_1, where p = 0.7. Here S_1, S_2, S_3 are the observations/symptoms, which support hypothesis H_1 to a particular degree. These rules were obtained from the medical experts. The numbers they contain such as 0.7 look like probabilities, and they too were obtained from the experts. The expert was in fact asked: "On a scale of 1 to 10, how much certainty do you affix to this conclusion?" The answer was then divided by 10.

It looks as if Shortliffe & Buchanan are using probability in the subjective sense to measure the degree of personal belief held by an expert. This at once raises the question of why subjective probabilities obtained from experts are preferred to objective probabilities obtained from data. Shortliffe & Buchanan do consider this question, and they answer[3] that in typical medical applications there is not enough data to obtain the requisite objective probabilities. This in turn is because of the inadequacy of hospital records, and the changes which are continually occurring in disease categories. It is interesting to note that only three years previously, another group working on computer diagnosis had reached exactly the opposite conclusion. This research group, working in Leeds, was headed by de Dombal. Their results are contained in de Dombal *et al.* (1972)[4], and Leaper *et al.* (1972)[5]. I will consider them in the final section of the paper.

De Dombal's approach was, however, largely ignored for the next twenty years, and nearly every researcher in the field made use of subjective probabilities. There are two possible reasons for this. First of all it may, in many cases, have been difficult to obtain objective probabilities from data. Secondly the general methodology of expert systems research, since it involved obtaining knowledge from the experts, may have encouraged the idea of obtaining probabilities as the degrees of belief of these experts. I will discuss further the question of objective versus subjective probabilities in the final section of the paper. Let us now return to a consideration of MYCIN, and the sample rule given earlier.

So far we have rather assumed that the figure 0.7 in the rule from MYCIN is an ordinary probability, but this is not the case, as Shortliffe & Buchanan make clear in the following passage:[6]

"… this rule at first seems to say $P(H_1 \mid S_1 \,\&\, S_2 \,\&\, S_3) = 0.7$, … . Questioning of the expert gradually reveals, however, that despite the apparent similarity to a statement regarding a conditional probability, the number 0.7 differs significantly from a probability. The expert may well agree that $P(H_1 \mid S_1 \,\&\, S_2 \,\&\, S_3) = 0.7$, but he becomes uneasy when he attempts to follow the logical conclusion that therefore $P(\text{not}.H_1 \mid S_1 \,\&\, S_2 \,\&\, S_3) = 0.3$. The three observations are evidence (to degree 0.7) *in favor* of the conclusion that the organism is a streptococcus and should not be construed as evidence (to degree 0.3) *against* streptococcus."

Shortliffe & Buchanan used this observation to motivate the introduction of a *non-probabilistic* model of evidential strength. Their measure of evidential strength was called a *certainty factor*, and certainty factors neither obeyed the standard axioms of probability theory, the Kolmogorov axioms, nor combined like probabilities.

Certainty factors were criticized by those who favoured a probabilistic approach, cf. Adams (1976)[7] and Heckerman (1986)[8], and in fact the next expert system we will consider (PROSPECTOR) did move more in the direction of standard probability.

PROSPECTOR, an expert system for mineral exploration, was developed in the second half of the 1970s at the Stanford Research Institute. A good general account of the system is given by Gaschnig in his 1982[9]. PROSPECTOR's most important innovation was to represent knowledge by an *inference network* (or *net*). This is motivated by Duda *et al.* in their 1976 as follows:[10]

"A collection of rules about some specific subject area invariably uses the same pieces of evidence to imply several different hypotheses. It also frequently happens that several alternative pieces of evidence imply the same hypothesis. Furthermore, there are often chains of evidences and hypotheses. For these reasons it is natural to represent a collection of rules as a graph structure or *inference net*."

A part of PROSPECTOR's inference network is shown in figure 1.

Figure 1

H_1 = There are massive sulfide deposits.
H_2 = There are clay minerals.
H_3 = There is a reduction process.
E_1 = Barite is overlying sulfide.
E_2 = Galena, sphalerite, or chalcopyrite fill cracks in rhyolite or dacite.
E_3 = There are bleached rocks.

Evidence E_1 is taken as supporting hypothesis H_1, and this is indicated by the arrow joining them in the inference network. Similarly E_2 supports hypothesis H_1, while E_3 supports H_3 which supports H_2 which supports H_1. Note how these rather complicated relations are simply and elegantly represented by the arrows of the network. Each inference arrow has a strength associated with it, and this obtained from the expert as in the case of MYCIN.

PROSPECTOR, however, differs from MYCIN in using subjective Bayesianism rather than certainty factors. This subjective Bayesianism is not entirely pure, since, as in the case of MYCIN, it is combined with the use of fuzzy logic formulae. This use of fuzzy logic tended to disappear in further developments.

In PROSPECTOR, Bayesianism is formulated using odds rather than probabilities. The odds on a hypothesis H [O(H)] are defined as follows:

$$O(H) = P(H)/P(\neg H)$$

Writing down Bayes theorem first for H and then for ¬H, we get

$$P(H \mid E) = P(E \mid H) P(H) / P(E)$$
$$P(\neg H \mid E) = P(E \mid \neg H) P(\neg H) / P(E)$$

So dividing gives

$$O(H \mid E) = \lambda(E) O(H) \qquad (1)$$

where $\lambda(E)$ is the likelihood ratio $P(E \mid H)/P(E \mid \neg H)$. (1) is the odds and likelihood form of Bayes theorem, and it is used in PROSPECTOR to change the prior odds on H to the posterior odds given evidence E.

Let us now consider the problems which arise if we have several different pieces of evidence $E_1, E_2, \dots , E_n$ say. We might in practice have to update using any subset of these pieces of evidence $E_i, E_j, \dots , E_k$ say, where (i, j, …k) is any subset of (1, 2, … , n). If we use (1), this would involve having values of $\lambda (E_i \& E_j \& \dots \& E_k)$ for all subsets of (1, 2, …n). When we remember that, on this approach the values of λ are obtained from the domain experts, we can see that obtaining the requisite values of λ is scarcely possible. Clearly some simplifying assumptions are necessary to produce a workable system, and the designers of PROSPECTOR therefore made the following two conditional independence assumptions:

$$P(E_1, \dots , E_n \mid H) = P(E_1 \mid H) \dots P(E_n \mid H) \qquad (2)$$

$$P(E_1, \dots , E_n \mid \neg H) = P(E_1 \mid \neg H) \dots P(E_n \mid \neg H) \qquad (3)$$

Given these assumptions, the whole problem of updating with many pieces of evidence becomes simple, and, in fact,

$$O(H \mid E_1 \& \dots \& E_n) = \lambda_1 \lambda_2 \dots \lambda_n O(H) \text{ where } \lambda_i = \lambda (E_i)$$

The only remaining problem was whether the conditional independence assumptions (2) and (3) are plausible. The search for a justification of these assumptions led, as we shall see in the next section, to the modification of the concept of inference network, and the emergence of the concept of *Bayesian network*.

3. The emergence of Bayesian networks in the 1980s

The concept of Bayesian network was introduced and developed by Pearl in a series of papers: Pearl (1982[15], 1985a[17], 1985b[11], 1986[12]), Kim & Pearl (1983)[16], and a book: Pearl (1988)[18]. An important extension of the theory was carried out by Lauritzen & Spiegelhalter (1988)[19], while Neapolitan's 1990 book[31] gave a clear account of these new ideas and helped to promote the use of Bayesian networks in the AI community. In what follows, I will comment on a few salient features of Bayesian networks which will be important when we consider their implications for the Bayesian controversy.

The actual term *Bayesian (or Bayes) network* was introduced in Pearl's 1985b where it is defined as follows:[11]

"Bayes Networks are directed acyclic graphs in which the nodes represent propositions (or variables), the arcs signify the existence of direct causal influences between the linked propositions, and the strengths of these influences are quantified by conditional probabilities."

This verbal account is illustrated by a diagram which is reproduced, with different lettering, in Figure 2.

Figure 2

If we compare the network of figure 2 with that of figure 1, two differences should be noted immediately. First of all the arrows in the inference network of figure 1 represent a relation of support holding between e.g. E_3 and H_3, while the arrows in the Bayesian network of figure 2 represent causal influences, so that, e.g. the arrow joining A to B means that A causes B. Secondly, corresponding to the first difference, we can say that, in a certain sense, the arrows of a Bayesian network run in the opposite direction to those of an inference network. Pearl puts this point as follows:[12]

"… in many expert systems (e.g. MYCIN), … rules point from evidence to hypothesis (e.g. if symptom, then disease), thus denoting a flow of mental inference. By contrast, the arrows in Bayes' networks point from causes to effects or from conditions to consequence, thus denoting a flow of constraints in the physical world."

This reversal of arrows from inference networks to Bayesian networks is illustrated in Figure 3, which shows one pair of nodes taken from the portion of PROSPECTOR's inference network shown in Figure 1.

Figure 3

Here E_3 = There are bleached rocks, while H_3 = There is a reduction process. From the point of view of an inference network (a), we regard the evidence of bleached rocks as supporting the hypothesis that there is a reduction process, while, from the point of view of a Bayesian network (b), we regard there being a reduction process as a cause of there being bleached rocks. In his 1993[13], Pearl gives an account of his discovery of Bayesian networks, and says that one factor that led him to the idea was his consideration of the concept of influence diagrams introduced by Howard and Matheson (1984).[14] Pearl decided to limit the influences specifically to causal influences.

Let us now make a few further points about Bayesian networks. If, in such a network, an arrow runs from node A to node B, then A is said to be a parent of B, and B a child of A. If a node has no parents, it is called a root, so that in figure 2, A is a root. In a Bayesian network, it is possible for a child to have several parents. Thus in figure 2, E has parents B and C. If, however, every child has at most one parent, the network is called a tree. As in the earlier case of PROSPECTOR's inference networks, in order to make computation feasible, some conditional independence assumptions have to be made. For a Bayesian network, these are that a node is conditionally independent given its parents of the rest of the network except its descendants.

From his definition of Bayesian networks, Pearl developed algorithms which allow Bayesian updating to take place in such networks. If one of the variables which represents an observation is set to a particular value, the changes brought about by this new information in all the probabilities throughout the tree can be computed in an efficient manner. Pearl began in his 1982[15] by developing an updating algorithm for a simple form of network, namely a tree. He then extended his algorithm to more complicated networks. Kim and Pearl (1983)[16] generalised from trees to Bayesian networks which are singly connected, i.e. there exists only one (undirected) path between any pair of nodes. Pearl in his 1986[12] tackled the further extension to Bayesian networks which are multiply connected. This problem was also investigated by Lauritzen & Spiegelhalter who in their 1988[19] solved it using the idea of reducing a multiply connected network to a tree of cliques. Their algorithm has been generally adopted by the AI community.

Let us now turn from these powerful mathematical developments to the consideration of a conceptual point. It will have been noted that two rather different definitions of Bayesian network have been given. The first definition is in terms of causes. Thus in figure 2 the arrows are taken as denoting a causal link between the two nodes which they join. The second definition is by contrast purely probabilistic. In figure 2 the variables A, B, C, D, E, F are taken to be random variables with a joint probability distribution, and the network becomes a Bayesian network if the relevant conditional independence assumptions are satisfied. I will henceforth use the term 'Bayesian network' for networks defined purely probabilistically in the manner just explained, and call the networks defined in terms of causes: 'causal networks'. Pearl tends, however, to use the terms 'Bayesian network' and 'causal network' interchangeably, because he believes the two notions to be closely connected. More specifically, his idea is that if in a network the parents of every node represent the direct causes of that node, then the relevant conditional independence assumptions will automatically be satisfied. As he says:[20]

"Causal utterances such as 'X is a direct cause of Y' were given a probabilistic interpretation as distinctive patterns of conditional independence relationships that can be verified empirically."

A suggested link between causality and conditional independence in fact goes back to Reichenbach (1956).[21] Reichenbach considers two events B and C say which are correlated. For example, in a travelling troupe of actors, B = the leading lady has a stomach upset, and C = the leading man has a stomach upset. We can explain such correlations, according to Reichenbach, by finding a common cause, namely that the leading lady and the leading man always have dinner together. The common stomach upsets occur when the food in the local restaurant has gone off. Denote ‘dining together’ by A. We then have the causal graph shown in figure 4.

Figure 4

Reichenbach then claimed that, conditional on A, B and C were no longer correlated but independent, i.e. P(B&C | A) = P(B | A) P(C | A). He also expressed this idea by saying that a common cause A screens one of its effects B off from the other C. Reichenbach’s causal fork is just a simple case of a Bayesian network. We can indeed apply his term ‘screening off’ to Bayesian networks by saying that in such networks, the parents of a node screen it off from all the other nodes in the network except its descendants.

We are now in a position to summarise the ingenious way in which Bayesian networks solved the problem of handling uncertainty in expert systems. In most of the domains considered, e.g medical diagnosis, a domain expert is very familiar with the various causal factors operating. It should therefore be an easy matter to get him or her to provide a causal network. By the addition of probabilities this can be turned into a Bayesian network. In earlier systems such as MYCIN or PROSPECTOR, conditional independence assumptions were made for the purely *ad hoc* and pragmatic reason of allowing the updating to become possible. For Bayesian networks, however, the causal information obtained from the expert provides a justification for making a set of conditional independence assumptions in the manner first suggested by Reichenbach. Moreover as Pearl, Lauritzen and Spiegelhalter have shown, this set of conditional independence assumptions is sufficient to allow Bayesian updating to become computationally feasible. Everything fits together in a most satisfying manner. There is only one weak link in the chain. It turns out, as we shall see in the next section, that it is possible to have a *bona fide* causal network in which the corresponding conditional independence assumptions are not satisfied.

4. Implications for the Bayesian controversy

The preceding sections have outlined some remarkable developments in AI. Let us now turn to a consideration of the implications of these developments for the Bayesian controversy. The Bayesianism versus non-Bayesianism debate has continued among philosophers of science for the last fifty years with no signs of abating. In the 1950s the major contenders were Carnap (in favour of Bayesianism) versus Popper (against Bayesianism). In the late 1980s and 1990s, we have had Howson & Urbach (1989)[22] in favour of Bayesianism, and Miller (1994)[23] against, while the most recent developments in the debate as seen by leading experts in the area are to be found in Corfield and Williamson, 2001.[24] The new results in AI are clearly relevant to this controversy, and indeed would seem to favour the Bayesian camp, though, as we shall see, this support is more qualified than might at first appear.

The Bayesian controversy, so I believe, involves two rather different issues. The first of these issues is the question of whether we should use the standard mathematical calculus of probability in handling uncertainty, or whether some other calculus might be appropriate. Here, of course, the Bayesians favour the use of the standard calculus. As an example of a non-Bayesian position we can take the view of Popper [see his (1934)[25], and, for a discussion, Gillies (1998)[26]] that the corroboration of universal laws of science

[C(H, E)] is not a probability function, i.e. does not satisfy the standard axioms of probability. In symbols the claim is that $C(H, E) \neq P(H \mid E)$.

As we have seen, this debate occurred also in the AI context. MYCIN used a non-probabilistic measure of evidential strength, and several other non-probabilistic approaches were proposed and developed by AI workers. [For some details, see Ng & Abramson (1990).[27]] However the development of AI has given a relatively unequivocal verdict. Probabilistic measures have proved much more successful in practice than non-probabilistic measures, and the latter have tended to disappear. AI has thus supported Bayesianism in this first sense. It should be added, however, that this does not give a decisive verdict against Popper's ideas on corroboration. Popper was considering the corroboration of hypotheses which were universal scientific laws. Most AI systems, however, have as hypotheses singular statements, such as 'this patient's infection is caused by streptococci' or 'that mountain range contains massive sulfide deposits'. It is possible that Bayesianism is appropriate for singular statements, while a non-Bayesian approach is appropriate for universal hypotheses as I have argued in an earlier work.[26]

Let us now turn to the second and rather different issue involved in the Baysian controversy. It can be most easily approached by considering the form that the debate has taken within statistics. Classical statisticians such as Neyman were strongly opposed to Bayesianism. Yet Neyman never used any formal system other than the standard mathematical theory of probability. Neyman was clearly not an anti-Bayesian in the sense we have just considered. In what sense, then, was he against Bayesianism? The answer is not immediately clear, since, because he accepted standard probability theory, Neyman *a fortiori* accepted Bayes theorem. The answer to this conundrum is that the second issue in the Bayesian controversy is really about the interpretation of probability. Neyman, following von Mises, regarded the objective interpretation of probability as the only valid one.[28] This meant that some applications of Bayes theorem were illegitimate in his eyes because they necessitated giving a degree of belief interpretation to some of the probabilities used. This applied particularly to the case of giving an *a priori* distribution to a fixed, but unknown, parameter θ. Since θ is fixed and does not vary randomly, it does not make sense to assign it an objective probability distribution, but, since it makes perfect sense for someone to have different degrees of belief in different possible values of θ, θ can easily be given a subjective probability distribution. Many Bayesian analyses involve giving *a priori* distributions to parameters such as $\theta_.$ and so become illegitimate to a strict objectivist such as Neyman. To sum up: the second issue involved in the Bayesian controversy is really about the relative merits of subjective versus objective interpretations of probability.

What have the AI developments given above shown as regards this controversy? It is immediately clear that they have lent support to the subjective interpretation of probability. Pearl has always argued for a subjective degree of belief interpretation of the probabilities in Bayesian networks, and this remains true of his latest, highly interesting, paper on the foundations of the subject (Pearl, 2001).[29] In this paper he describes himself as 'only half-Bayesian'. However his departure from standard Bayesianism arises because he thinks that prior probability distributions are inadequate to express background knowledge, and that one needs also to use causal judgements which cannot be expressed in probabilistic terms. As far as the interpretation of probability is concerned he remains faithful to the subjective, degree of belief, view which he says he adopted in 1971 after reading Savage.

Lauritzen and Spiegelhalter were also working in the tradition of subjective Bayesianism, but they seem less definitely committed to this view than Pearl. This is what they say:[30]

> "Our interpretation of probabilities is that of a subjectivist Bayesian …. This seems a convenient and appropriate view in an area concerned with the rational structuring and manipulation of opinion, and the subjectivist objectives of a coherent system of probabilities representing belief in verifiable propositions, successively updated on the basis of available evidence, appears to fit remarkably the objectives of expert systems research. However, many of the techniques presented here are appropriate in disciplines where graphical structures are used and a frequentist interpretation is more appropriate, such as complex pedigree analysis in genetics."

So Lauritzen and Spiegelhalter think that, in some cases at least, the probabilities in Bayesian networks might be given an objective interpretation. Neapolitan (1990)[31] is also favourable to objective probabilities in Bayesian networks. So although Bayesian networks were created within the tradition of subjective Bayesianism, it might nonetheless be possible to interpret the probabilities they contain objectively. Arguably this is likely to be a good strategy in many cases.

A first argument in favour of an objective interpretation is an appeal to the results, mentioned earlier, of de Dombal's group at Leeds. De Dombal and his group devised a computer-based diagnostic system using a straightforward statistical approach. The probability of a patient's having a particular condition given a set of symptoms was calculated using Bayes theorem, where the probabilities employed had been estimated from a large sample of previous patients. This approach did not attempt to encode any medical knowledge beyond what was contained in observed frequencies, and so was not an expert system. As we shall see, however, the system worked very well, and this was because it was designed for use in the following relatively simple situation.

Patients were admitted to the department of surgery of a Leeds hospital because of the onset of acute abdominal pain. The problem was to diagnose the cause of their pain, and, in particular, decide whether an operation was necessary. Leaving aside a small 'other' or 'miscellaneous' category comprising less than 4% of cases, the patients were diagnosed as having just one of the following seven conditions: appendicitis (just over a quarter of the cases), cholecystitis, small bowel obstruction, perforated duodenal ulcer, pancreatitis, diverticular disease, and, last but not least, non-specific abdominal pain, which accounted for about half the cases. The last condition is of course not a disease, but covers those cases where no cause for the pain could be found. This situation is obviously suitable for statistical treatment because of the limited number of mutually exclusive possibilities, and also because during the course of the treatment, often an operation, the cause of the pain could in most cases be definitely established, thereby establishing the correctness or incorrectness of the initial diagnosis.

The efficiency of de Dombal *et al.*'s computer system was compared with that of the hospital clinicians in a sample of 304 patients admitted between 1 January 1971 and 1 December 1971. The overall result[32] was that the computer system was correct in 91.8% of the cases, and the senior member of the clinical team handling the patient in 79.6% of the cases. Some of the detailed differences are also interesting. One of the most difficult, but at the same time crucial, problems in this area is to distinguish between appendicitis and non-specific abdominal pain. If a case of appendicitis is wrongly classified as non-specific abdominal pain, the result could be a delay in operating which results in the appendix perforating or forming an abscess. Conversely, however, if non-specific abdominal pain is wrongly diagnosed as appendicitis, this could result in the patient going through an entirely unnecessary operation – known in the business as a ‘negative laparotomy’.

De Dombal *et al.* describe the relative performance in this area of the computer and the humans as follows:[33]

> “… the computer system accurately classified 84 out of a possible 85 patients with acute appendicitis, … . This contrasts with the clinicians’ performance, where only 75 diagnoses of appendicitis were made, and six patients were originally classified as non-specific abdominal pain. … Moreover, although the computer erroneously classified six non-specific abdominal pain patients with the ‘appendicitis’ category, the corresponding figure for the clinical team was no fewer than 27 patients. … Had we slavishly followed the computer’s predictions, six negative laparotomies would have been performed, but in no case of appendicitis would surgery have been delayed. What actually happened was rather different. Twenty-odd negative laparotomies were performed, and six cases of appendicitis were ‘observed’ for over eight hours before the decision to operate was taken.”

These results are very interesting, but the next experiment of the de Dombal group was perhaps even more interesting (cf. Leaper et al, 1972)[5]. Once again the trial involved patients admitted to the professional surgical unit of the Leeds General Infirmary with abdominal pain of acute onset. The period covered again began on 1 January 1971, but this time continued longer until 31 May 1972 producing a larger number (472) of patients. Once again a comparison was made between the diagnoses of the computer system whose probabilities had been calculated from data obtained from a large number of previous patients, and the diagnoses of the senior clinician in charge of the patient. This time, however, a new comparison was introduced. As well as obtaining the probabilities from data, a variant of the computer system was produced in which the probabilities were obtained from estimates provided by the clinicians. In fact estimates were obtained from six clinicians, and the average taken. The general results of this trial were as follows:[34]

> “In the total series of 472 cases the overall accuracy of diagnoses made ‘on the spot’ by the clinical team was 79.7%, whereas the accuracy of the computer-aided system using values based on 600 surveyed cases was considerably higher (91.1%). … the overall accuracy of the computer-aided system using the clinicians’ estimates was a relatively unimpressive 82.2%.”

Once again the results in specific disease categories are also interesting:[35]

> “… in some instances (appendicitis, non-specific abdominal pain) the computer using estimates was more effective than the unaided clinician, but in others (diverticulitis, pancreatitis) it was much less effective. … the effectiveness of the computer using estimates *seemed to be related to the incidence of the diseases under study*. In respect of acute appendicitis (121 cases in 15 months) and non-specific abdominal pain (230 cases) the computer using estimates was relatively effective when compared with unaided clinician. But for other diseases such as diverticulitis (10 cases) and pancreatitis (14 cases) the computer using estimates proved to be less reliable.”

In effect, the computer system using objective probabilities obtained from data outperformed the human clinicians by a considerable margin, but if subjective probabilities obtained from these clinicians were used instead of the objective probabilities, this gain was wiped out, and in the case of relatively rare diseases the computer system performed worse than the human clinicians. Human doctors appear to be bad at estimating probabilities of diseases, and especially bad in the case of diseases which occur infrequently. It seems hard to avoid the conclusion of the de Dombal group which was the following:[35]

> "We suspect that the use of clinicians' estimates of probability may have been a cause of failure in some previous computer-aided diagnostic systems, and we conclude that in future computer-aided diagnostic systems there is no alternative to using carefully collated data from large-scale, real-life surveys rather than clinicians' estimates."

These results are very striking, but the issue is not simply one of a choice between different ways of interpreting probabilities. It should now be pointed out that this choice carries with it methodological implications. If we are interpreting the probabilities as objective, then any proposed value of a probability must be seen as a conjecture which could be right or wrong, and may therefore be in need of testing. Thus objective probabilities lead to a Popperian methodology of conjectures and refutations in which testing plays a central role. This is indeed the methodology of classical statistics.

Let us next contrast this with the use of subjective probabilities. Any such probability expresses the degree of belief of an individual at a particular moment. Further evidence does not refute the claim that that individual held that degree of belief at that time. It may however lead the individual to change his or her degree of belief in the light of the new evidence. In the Bayesian approach, the belief change takes place through Bayesian conditionalisation or updating, i.e. through the change from a prior probability $P(H)$ to a posterior probability $P(H \mid E)$. To sum up then. The use of objective probabilities goes with the Popperian methodology of statistical testing; while the use of subjective probabilities goes with the methodology of Bayesian conditionalisation. There have been examples which show that the use of a testing methodology can be advantageous in the construction of Bayesian networks.

One such example [see Sucar (1991)[36], and Sucar, Gillies, and Gillies (1993)[37]] concerned a medical instrument called an 'endoscope'. This allowed a doctor to put into the colon of a patient a small camera which transmitted an image of the interior of the colon to a television screen. In this image an expert could recognise various things in the interior of the colon. Let us take two such things as examples. One is called the 'lumen' which is the opening of the colon. Despite its name, it generally appeared as a large dark region; but sometimes it was smaller and surrounded by concentric rings. Another is called a 'diverticulum' and is a small malformation in the wall of the colon, which can cause some illnesses. A diverticulum generally appeared as a dark region, smaller than the lumen, and often circular. It was a problem then to program a computer to recognise from the image the lumen or a diverticulum. This is a typical problem of computer vision. To solve it, an attempt was made to construct a Bayesian network with the help of an expert in medical endoscopy.

Figure 5 shows only a small part of this network, but it is sufficient to illustrate the points which are to be made. L stands for the lumen which causes a large, dark, region region (LDR) to appear on the screen. This in turn produces values for the variables S (size of the region), and M and V (mean and variance of the light intensity of the region).

Figure 5

In every Bayesian network certain assumptions of independence or conditional independence are made. In this case, S, M, V must be conditionally independent of L and mutually conditionally independent, given LDR. Using a Popperian testing methodology, these assumptions were considered as conjectures which needed to be checked by statistical tests. These tests showed, however, that the conditional independence assumptions were not satisfied. In fact it turned out that, given LDR, M and V were strongly correlated rather than independent.

The response to this situation was to eliminate one of the two parameters M and V on the grounds that, since they were correlated, only one could give almost as much information as both. The results of this elimination were tested using a random sample of more than 130 images of the colon. It turned out that the elimination of one of the parameters gave better results than those obtained using all three parameters. For example using all three parameters (S, M, V), the system recognised the lumen correctly in 89% of cases, while, if it was modified by eliminating M, and using only (S, V) it recognised the lumen correctly in 97% of cases.[38] At first sight this seems a paradox, because these better results were obtained using less information. The explanation is simple however. Undoubtedly there is more information in all three parameters (S, M, V) than in only two (S, V). But the greater amount of information in the three parameters was used with mathematical assumptions of conditional independence which were not correct. The lesser amount of information in the two parameters S, V was, by contrast, used with true mathematical assumptions. So less information in a correct model worked better than more information in a mistaken model. Moreover, since the modified Bayesian network was simpler, the calculations using it were carried out more quickly. So, to conclude, the modified Bayesian network was more efficient, and gave better results. This shows the value of using objective probabilities, and a Popperian methodology of statistical testing.

The example also shows that it is possible to obtain a causal network from an expert for which the assumptions of conditional independence are not satisfied. Thus, although causal networks are useful heuristic guides for the construction of Bayesian networks, they are not infallible guides.

The conclusion I want to draw is that we might move towards a kind of synthesis between the Bayesian and non-Bayesian positions. For subjective Bayesians such as De Finetti, Lindley and Savage, the use of Bayes theorem to update beliefs in the light of evidence was a natural, indeed central, procedure. In the context of expert systems research, the adoption of this procedure led through PROSPECTOR and Pearl’s work to the concept of Bayesian network. It is very unlikely that any of the classical statisticians who emphasized objective probabilities and statistical testing would have taken this path. Although such classical statisticians would have accepted Bayes theorem as a consequence of the Kolmogorov axioms, it would not have been natural for them to think of using it, even with objective probabilities, for the purpose of updating in the light of evidence.

Nonetheless once Bayesian networks had emerged from the programme of subjective Bayesianism, it became clear that such networks could be improved in many cases by incorporating some ideas from classical statistics. These were (1) the use of objective rather than subjective probabilities, and (2) the use of statistical tests to check the assumptions underlying Bayesian networks. In particular, though causal relations in a network do suggest that the corresponding conditional independence relations are satisfied, this may not hold in some cases. Thus it could be worth using statistical tests to check whether the conditional independence assumptions, considered as conjectures, really hold. If it turns out that they do not hold, this could lead to modifications and improvements in the Bayesian network.

Notes

1. Jackson, P., *Introduction to expert systems*. Wokingham/ England: Addison-Wesley 1986.

2. E.H.Shortliffe & B.G.Buchanan, "A model of inexact reasoning in medicine", in: *Mathematical Biosciences*, 23, 1975, pp.351-79. The quotation is from p. 357.

3. *Ibid.,* pp. 352-5.

4. F.T.de Dombal, D.J.Leaper, J.R.Staniland, A.P.McCann, & J.C.Horrocks, "Computer-aided diagnosis of acute abdominal pain", in: *British Medical Journal,* 2, 1972, pp. 9-13.

5. D.J.Leaper, J.C.Horrocks, J.R.Staniland, & F.T.de Dombal, "Computer-assisted diagnosis of abdominal pain using 'estimates' provided by clinicians", in: *British Medical Journal*, 4, 1972, pp. 350-4.

6. *Ibid.,* p. 358.

7. J.B.Adams, "A probability model of medical reasoning and the MYCIN model", in: *Mathematical Biosciences*, 32, 1976, pp. 177-86.

8. D.Heckerman, "Probabilistic interpretations for MYCIN's certainty factors", in: L.N.Kanal / J.F.Lemmer (Eds.), *Uncertainty in Artificial Intelligence,* Amsterdam: North-Holland 1986, pp. 167-96.

9. J.Glaschnig, "Prospector: an expert system for mineral exploration", in: D.Michie (Ed.), *Introductory readings in expert systems*, New York: Gordon and Breach 1982, pp.47-64.

10. R.O.Duda, P.E.Hart, & N.J.Nilsson, "Subjective Bayesian methods for rule-based inference systems", in: *Proceedings of the National Computer Conference (AFIPS)*, 45, 1976, pp. 1075-82. The quotation is from p. 1076.

11. J.Pearl, "Bayesian networks: a model of self-activated memory for evidential reasoning", in: *Proceedings of the Cognitive Science Society*, Ablex, 1985b, pp. 329-34. The quotation is from p. 330.

12. J.Pearl, "Fusion, propagation and structuring in belief networks", in: *Artificial Intelligence*, 29, 1986, pp. 241-88. The quotation is from pp. 253-4.

13. J.Pearl, "Belief networks revisited", in: *Artificial Intelligence*, 59, 1993, pp. 49-56.

14. R.A.Howard & J.E.Matheson, "Influence diagrams", in: R.A.Howard / J.E.Matheson (Eds.), *The principles and applications of decision analysis,* Menlo Park/CA: Strategic Decisions Group 1984, Vol.2, pp. 721-62.

15. J.Pearl, "Reverend Bayes on inference engines: a distributed hierarchical approach", in: *Proceedings of the national conference on AI (ASSSI-82)*, 1982, pp. 133-6.

16. J.H.Kim & J.Pearl, "A computational model for combined causal and diagnostic reasoning in inference systems", in: *Proceedings of the 8th international joint conference on AI (IJCAI-85),* 1983, p. 190-3.

17. J.Pearl, "How to do with probabilities what people say you can't", in: *Proceedings of the second IEEE conference on AI applications,* Miami/Fl., 1985a, pp. 6-12.

18. J.Pearl, *Probabilistic reasoning in intelligent systems. Networks of plausible inference*. San Mateo/California: Morgan Kaufmann 1988.

19. S.L.Lauritzen & D.J.Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems (with discussion)", in: *Journal of the royal statistical society B*, 50, pp. 157-224.

20. Pearl, "Belief networks revisited", *loc. cit.* (note 13), p. 52.

21. H.Reichenbach, *The direction of time*. Berkeley: University of California press 1956.

22. C.Howson & P. Urbach, *Scientific reasoning; the Bayesian approach*. La Salle/Illinois: Open court 1989.

23. D.Miller, *Critical rationalism*. Chicago and La Salle/Illinois: Open court 1994.

24. D.Corfield /J. Williamson (Eds.), *Foundations of Bayesianism*. Dordrecht/Boston/London: Kluwer 2001.

25. K.R.Popper, *The logic of scientific discovery*. 6th impression of English translation of the original German 1934 edition. London: Hutchinson. 1972.

26. D.A.Gillies, "Confirmation theory", in: D.M.Gabbay/P.Smets (Eds.), *Handbook of defeasible reasoning and uncertainty management systems*, Dordrecht/Boston/London: Kluwer 1998, Volume 1, pp. 135-67. The arguments for Bayesianism for singular statements and non-Bayesianism for universal statements are on pp. 154-5.

27. K.Ng & B.Abramson, "Uncertainty management in expert systems", in: *IEEE Expert*, 5, 1990, pp. 29-48.

28. D.A.Gillies, "Debates on Bayesianism and the theory of Bayesian networks", in: *Theoria*, 64, 1998, pp. 1-22.

29. J.Pearl, “Bayesianism and causality, or, why I am only half-Bayesian”, in: D.Corfield/J.Williamson (Eds.), *Foundations of Bayesianism*. Dordrecht/Boston/London: Kluwer 2001, pp. 19-36.

30. Lauritzen & Spiegelhalter, “Local computations with probabilities on graphical structures and their application to expert systems”, *loc. cit.* (note 19), p. 159.

31. R.E.Neapolitan, *Probabilistic reasoning in expert systems. Theory and algorithms*. New York: John Wiley 1990.

32. de Dombal *et al.*, “Computer-aided diagnosis of acute abdominal pain”, *loc. cit.* (note 4), p. 9.

33. de Dombal *et al.*, “Computer-aided diagnosis of acute abdominal pain”, *loc. cit.* (note 4), p. 2.

34. Leaper *et al.*, “Computer-assisted diagnosis of abdominal pain using ‘estimates’ provided by clinicians”, *loc. cit.* (note 5), pp. 351-2.

35. Leaper *et al.*, “Computer-assisted diagnosis of abdominal pain using ‘estimates’ provided by clinicians”, *loc. cit.* (note 5), p. 353.

36. L.E.Sucar, *Probabilistic reasoning in knowledge-based vision systems*. PhD thesis, Imperial College, University of London 1991.

37. L.E.Sucar, D.F.Gillies, & D.A.Gillies, “Objective probabilities in expert systems”, in: *Artificial Intelligence*, 61, 1993, pp. 187-208.

38.*Ibid.*, p. 206 for further details.