EFFICIENT MINIMUM INFORMATION UPDATING FOR BAYESIAN INFERENCING IN EXPERT SYSTEMS

John F. Lemmer and Stephen W. Barth

PAR Technology Corporation

ABSTRACT

This short paper presents a new algorithm for minimum information Bayesian Inferencing within Expert Systems. This algorithm is as efficient in both time and space as previously reported work [3] but always provides a minimum information result. In addition to describing the new algorithm, we will prove that it does indeed satisfy minimum information criteria. Since both algorithms are substantially different from the "Bayesian" approaches in well known expert systems such as the original Prospector [1], AL/X [8], and MYCIN [9], and from the approach of Kulikowski [5], background is provided to show the motivation for using the minimum information approach to Bayesian updating.

I. BACKGROUND

There is a naturally defined probability space underlying the rule base of any expert system that is generally diagnostic. MYCIN and Prospector are examples of such systems, whereas R1 [7], and KNOBS [2], are planning systems that cannot be characterized in this way. Rules in expert systems have the general form:

IF <antecedent> THEN <consequent>

An example of such a rule in a medical system might be

IF fever and rash THEN measles

In diagnostic systems inferential weights are associated with such rules. For example, if fever and rash always implied measles, then a weight of 1 might be assigned to this rule. If it were sometimes possible to have fever and rash but not have measles, then a weight less than 1 would be assigned.

Both the antecedent and the consequent in such rules can be modeled as events defined over a probability space. In such a model, the inferential weight might be interpreted as the conditional probability that the consequent event will occur, given that the antecedent event has occurred. If the antecedent event has occurred with probability 1 and the conditional probability has been properly assessed by the expert, all is well with this interpretation. However, difficulties arise in the cases where the occurrence of the antecedent event is not certain or when more than one rule contains the same consequent event.

The difficulties just described reside in determining how to combine various amounts of uncertain information. Suppose the rule base contains the rules:

IF a THEN z (w_1)
IF b THEN z (w_2)

where the w_i are now to be interpreted as conditional probabilities. If events a and b are both determined to have occurred with certainty but the w_i are both less than 1, what can be said of the (posterior) probability of z having also occurred? In general the answer to this question depends on the correlations among the events a, b, and z. For example, if a and b always occur together (are perfectly correlated), then the second rule contains no information not in the first (and a consistent expert would have provided $w_1 = w_2$). A related difficulty would occur in applying the measles rule if it were almost certain that a fever was present but the presence of the rash could not be reliably determined.

To deal with the problem of correlations among the events defined by the rule base, it is convenient to conceive of a joint probability distribution defined over all the events occurring in the rules [3], [6]. Then, if this joint distribution could be estimated and some subset of events defined in this distribution observed with certainty, Bayes' Rule could be applied to determine the posterior probability of the unobserved events. Two facts would seem to render this conception useless: the difficulty of estimating the joint distribution and the difficulty of observing sufficient interesting events with certainty. (This latter problem would be especially severe in hierarchical inference systems where events which are antecedents in one rule are actually inferred as consequents of some other rule.)

Considerable work has been done in developing practical techniques for estimating meaningful portions of the underlying distribution, and in developing methods for dealing with uncertain observations. Konolige [3] and Lemmer [6] have both suggested defining subsets of the complete

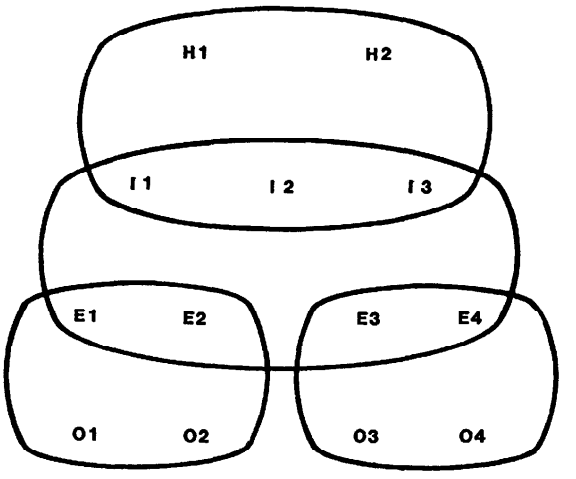
set of events where the subsets are chosen to contain only events whose interrelations are thought to be important. Lemmer has suggested methods for estimating marginal distributions over each of the subsets in such a way that all the marginals can be considered components of the same underlying complete distribution. Both have defined procedures for estimating these components from incomplete and possibly inconsistent information. Konolige has suggested a computationally expensive approach which results in an estimate satisfying a minimum information criterion. Lemmer has suggested a more efficient procedure which only approximates a minimum information criterion. The rest of this paper addresses the problem of dealing with uncertain observations.

II. EXTENDED BAYESIAN UPDATING

The component distributions which can be estimated by the techniques referenced above are the prior probabilities for the occurrence of the events defined by the rule base. As data enters the expert system, it becomes desirable to update this prior information in the light of the newly acquired data. In the situation described above, Bayes' rule will not, in general, be applicable. Why this is so will be suggested by the example to follow. The example will also be used to introduce an extended updating rule which will always be applicable.

The rule itself will be defined initially in terms of the example, but will later be generalized. We will then prove that the rule is a minimum information rule. (For the significance of minimum information [3] may be consulted.) We will also show that Konolige's rule is a special case of the rule presented here.

An expert system to be used for hierarchical inference will normally explicitly define events which may be categorized as either hypotheses of interest, intermediate hypotheses, or base events. The rule base may also implicitly define other events, observations as it were, whose occurrence implies the occurrence of some explicitly defined event of the system. The diagram of Figure 1 defines sets of local event groups (LEGs) consisting of sets of events of these types whose interactions are thought by the expert to be significant to the inferences to be performed by the system. This diagram is meant to suggest that the expert is comfortable deciding on the truth of the hypotheses, H1 and H2, while only considering the truth of the intermediate hypotheses, I1, I2, and I3. Likewise, he considers only observations O1 and O2 to be significant to the occurrence of base events E1 and E2. Using the procedures of Konolige or Lemmer, which were mentioned in the previous section, each of these LEGs would define a set of events over which a component marginal distribution (CMD) would be estimated.



SOME EVENTS IN A DIAGNOSTIC SYSTEM

Figure 1

There are a number of reasons why only CMDs are estimated rather than the entire underlying distribution. The foremost reason is that practical systems contain so many events that it would be infeasible to estimate or store the entire distribution. The second reason is inherent in the very concept of an LEG. The concept implies that the expert is not realistically able to aid in the estimation of the interaction between variables in different event groups because he does not consider their interactions important. However, the use of LEGs does rule out, in general, the possibility of the direct application of Bayes' rule.

Bayes' rule cannot be applied in general because the rule requires that the conditioning event have probability 1. That is, before applying

$$p'(a) = p(a|b) = \frac{p(b|a)\,p(a)}{p(b)}$$

it must be the case that the posterior probability of b, p'(b), is 1. This criterion will not normally be met when the effect of an observation is propagated through a set of LEGs. Consider the example of Figure 1. If observations O1 and O2 have been made, these events are a posteriori certain and the CMD defined over the events E1, E2, O1, and O2, CMD([E1, E2, O1, O2]), can be directly updated with Bayes' rule. Logically, it should be possible to to now update CMD([I1, I2, I3, E3, E4, E1, E2]) using the posterior information concerning E1 and E2. But, in general, the posterior probability of these two events will not be equal to 1, so that Bayes' rule cannot be applied to this second CMD.

A rule will now be presented which is always applicable, which will act exactly like Bayes' rule whenever updating is done relative to events with posterior probability equal to 1, and which will always satisfy a minimum information criterion. This rule will specify a simple computation to be performed on the prior probability value associated with each of the 2**n joint events normally used to define a binary distribution over n events. One of the joint events used to define CMD([I1, I2, I3, E3, E4, E1, E2]) is, for example,

$$I1*-I2*I3*-E3*E4*E1*-E2 \qquad (1)$$

where "*" means "and" and "-" means "not". Likewise the joint event

$$E1*-E2 \qquad (2)$$

would provide a part of the specification for the posterior distribution, CMD'([E1, E2]). (Note that CMD'([E1, E2]) is readily computable from CMD'([E1, E2, O1, O2]) which in this example has been assumed to be computable using Bayes' rule.

Call the event in (1), x, and the event in (2), y. Then the updating rule proposed here is

$$p'(x) = p(x)\,\frac{p'(y)}{p(y)} \qquad (3)$$

This rule can be applied to each joint event, x, used to specify the CMD being updated. The result will be a new updated CMD which can then be used to propagate the effect of the observations to other CMDs. Continuing the example, once CMD'([I1, I2, I3, E3, E4, E1, E2]) has been computed, CMD'([I1, I2, I3]) can be readily computed and used to produce the updated CMD, CMD'([H1, H2, I1, I2, I3).

We will now show that the procedure implied by (3) computes a posterior component distribution, CMD' which is as close as possible to the prior, CMD, subject to a simple, necessary constraint. The measure of closeness to be used is the measure of discriminant information described in [4]. To define the information measure and present the proof we must first introduce some precise notation. Let F be a set of binary events over which a joint distribution is defined, and let x be one such event. In the example of Figure 1, we could have F=[I1, I2, I3, E1, E2, E3, E4] and x could be (1). Let CMD (F) be a prior probability distribution over the joint events definable over the binary events in F. Thus, we can write

$$CMD(F) = [p(x) : x \text{ definable over } F],$$

meaning that CMD (F) is a set containing the probability for every possible joint event x. Similarly, we can have the posterior distribution

$$CMD'(F) = [p'(x) : x \text{ definable over } F].$$

With this notation the discriminant information between CMD(F) and CMD'(F) is defined as

$$D = \sum_x p'(x)\,\ln\frac{p'(x)}{p(x)} \qquad (4)$$

In order to relate (4) to the rule (3) we define G to be any subset of F such that a posterior distribution is available for G. Let y stand for any joint event definable over G. In the example we could have G=[E1, E2] and (2) would be an example of y. Note that with x referring to (1) and y referring to (2), event y will occur whenever event x has occurred, or, $x \subset y$. Note also that the choice of an x uniquely defines a y. Thus the algorithm: "for all x definable over F compute p'(x) by (3)" computes the updated distribution CMD'(F).

We now show that the algorithm minimizes D as defined by (4), subject to the constraint that for all y, p'(y) computed from CMD'(F) will equal the p'(y) computed from CMD'(G); that is:

$$p'(y) = \sum_{x \mid x \subset y} p'(x) \quad : \quad \frac{p'(y)CMD'(G)}{p'(x)CMD'(F)} \qquad (5)$$

This constraint simply means that p'(y) has the same value in any LEG.

The proof is by the method of La Grange multipliers. Define a(x) such that p'(x) = a(x) p(x). We will show that choosing

$$a(x) = \frac{p'(y)}{p(y)} \quad : \quad x \subset y \qquad (6)$$

as was done in (3), will minimize (4) subject to (5). According to the method of La Grange we want to choose the a(x) to minimize

$$D = \sum_x a(x)p(x)\ln\frac{a(x)p(x)}{p(x)} + \sum_y \lambda_y \left(p'(y) - \sum_{x \subset y} a(x)p(x)\right)$$

To minimize D we set all the partial derivatives equal to 0, yielding

$$\frac{\partial D}{\partial a(x)} = p(x)\,(1 - \ln a(x) - \lambda_y) = 0 \qquad (7)$$

and

$$\frac{\partial D}{\partial \lambda_y} = p'(y) - \sum_{x \subset y} a(x)p(x) = 0 \qquad (8)$$

From (7) we have

$$\ln a(x) = \lambda_y - 1, \quad x \subset y \qquad (9)$$

Thus, given any y, for every x y, a(x) will have the same value. Call this value a(y). Thus from (8)

$$p'(y) = a(y) \sum_{x \subset y} p(x) = a(y)p(y) \qquad (10)$$

so that

$$a(y) = \frac{p'(y)}{p(y)} \qquad (11)$$

which is exactly the formula (3).

The rule proposed by Konolige is a special case of (3). To see this, rewrite (3) as

$$p'(x) = p(x/y)\ p'(y)$$

In our case p'(y) is the probability of some joint event taken from CMD'(G). However, Konolige replaces p'(y) with an estimate formed by a product of the posterior probabilities of the individual events in the joint event y. In terms of the example, his rule replaces p'(E1*-E2) by p'(E1) (1-p'(E2)). Thus his rule is minimum information if the events in the updating set, G, are independent. Thus his rule for updating cannot make use of the correlation information in the prior distribution.

III. CONCLUSIONS

The rule presented here in (3) is equivalent to Konolige's rule in computational complexity, but the Konolige rule is a special case (in the minimum information sense) of (3). In particular, (3) makes better use of correlation information. Thus (3) would seem to be superior in all cases.

REFERENCES

[1] Duda, R.O. Hart, P.E., and Gaschnig, J. "Model Design in the PROSPECTOR Consultant System for Mineral Exploration", Expert Systems in the Micro-electronic Age (ed. D. Michie) Edinburgh: Edinburgh University Press, 1977 pp. 153-167

[2] Engelman, C. Scarl, E.A., and Berg, C. "Interactive Frame Instantiation," Proceedings of the First Annual National Conference on Artificial Intelligence, The American Association for Artificial Intelligence, Menlo Park, CA 1980 pp. 184-186

[3] Duda, .O. Hart, P.E. Konolige, K., Reboh, R. A Computer-Based Consultant for Mineral Exploration, Appendix D, Final report for SRI Project 6415. Menlo Park, CA; SRI International, 1979

[4] Ku, H.,H., and Kullback, S., "Approximating Discrete Probability Distribution," IEEE Transactions on Information Theory, Vol. IT-15, No. 4, July 1969.

[5] Kulikowski, and Weiss, S., "Strategies of Data Base Utilization in Sequential Pattern Recognition." Proceedings of the 1972/IEEE Conference on Decision and Control and 115th Symposium on Adaptive Processes, New York: IEEE, 1972, pp. 103-105

[6] Lemmer, J. "Algorithms for Incompletely Specified Distributions in a Generalized Graph Model for Medical Diagnosis," Ph.D. Thesis, University of Maryland, 1976

[7] McDermott, J. "R1: an Expert in the Computer Systems Domain", Proceedings of the First Annual National Conference on Artificial Intelligence, The American Association for Artificial Intelligence, Menlo Park, CA, 1980 pp. 269-271

[8] Michie, D., and Paterson, A., AL/X User Manual, Oxford: Intelligent Terminals Ltd., 1981

[9] Shortcliffe, E.H., Computer-Based Medical Consultation: MYCIN, New York: American Elsevier Publishing Co., 1976