# A Method of Computing Generalized Bayesian Probability Values For Expert Systems

Peter Cheese man

SRI International, 333 Ravenswood Road
Menlo Park, California 94025

Abstract

This paper presents a new method for calculating the conditional probability of any multi-valued predicate given particular information about the individual case. This calculation is based on the principle of Maximum Entropy (ME), sometimes called the principle of least information, and gives the most unbiased probability estimate given the available evidence. Previous methods for computing maximum entropy values shows that they are either very restrictive in the probabilistic information (constraints) they can use or combinatorially explosive. The computational complexity of the new procedure depends on the inter-connectedness of the constraints, but in practical cases it is small. In addition, the maximum entropy method can give a measure of how accurately a calculated conditional probability is known.

## §1 Introduction

Recently computer-based expert systems have been developed that store probabilistic knowledge obtained from experts and use this knowledge to make probabilistic predictions in specific cases. Similarly, analysis of data, such as questionnaire results, can reveal dependencies between the variables that can also be used to make probabilistic predictions. The essential problem in such systems is how to represent all known dependencies and relationships, and how to use such information to make specific predictions. For example, knowledge of interrelationships between factors such as age, sex, diet, cancer risk etc. should allow the prediction of say an individual's cancer risk, given information on the individual. However because of possible interactions between the factors, it is not sufficient to combine the effects of each factor separately.

The major problem faced by all such probabilistic inference systems is that the known constraints usually underconstrain the probability space of the domain. For example, if the space consists of 20 predicates, then $2^{20} - 1$ joint probability constraints are needed to fully specify all the probabilities. When a space is under- constrained, any desired probability usually has a range of possible values. The problem is to find a unique probability value within the allowed range that is the best estimate of the true probability, given the available information, and to determine how reliable this estimate is. Such an estimate is given by the method of maximum entropy (ME), sometimes called the method of least information. This method gives a probability value that is the least commitment value, subject to the constraints. To choose any other value has been shown by Shore and Johnson (1080) to be inconsistent, because any other choice would imply more information than was given in the problem.

This paper focuses on a type of expert system in which all the probability constraints are in the form of conditional probabilities or joint probabilities (sometimes called marginal probabilities because they occur in the margins of contingency tables). Such probability constraints may have come from an expert or from an analysis of data that has shown that particular subsets of factors are significantly correlated. The problem of making probabilistic predictions in under-constrained probability spaces is of sufficient importance that many solutions have been tried. One method is to acknowledge that the desired probabilities are under constrained and return the range of possible values consistent with the known constraints (rather than a point value). Such an approach is implicit in the method proposed by Shafer (1979).

Another method is to make the strong assumption of conditional independence when combining different evidence. This is the assumption behind PROSPECTOR (Duda *et al,* 197(5) and Dependence Trees (Chow and Liu 1968) and used most recently by Pearl (1982). Use of the conditional independence assumption with given conditional probabilities is usually sufficient to constrain the desired probabilities to a unique value. However, this assumption is not always satisfied by actual data and can lead to inconsistent and over-constrained probability values, as pointed out by Konolige (1979).

The main purpose of this paper is to introduce a new method for computing the maximum entropy probability of a predicate of interest, given specific evidence about related predicates, and subject to any linear probability constraints. This method avoids the combinatorial explosion inherent in previous methods without imposing strong limitations on the constraints that can be used, and it is therefore useful for computer-based expert systems.

## §2 The Maximum Entropy Method

The method of maximum entropy was first applied by Jaynes to the statistical mechanics problem of predicting the most likely state of a system given the physical constraints (e.g. conservation of energy). In Jaynes (1908), the maximum entropy method was used to provide prior probabilities for a Bayesian analysis. Lewis (1959) applied the method of least information (an equivalent method) to the problem of finding the best approximation to a given probability distribution baaed on knowledge of some of the joint probabilities (i.e., constraints on the possible distributions).

Ireland and Kullback (1968) applied the minimum discrimination information measure (yet another method equivalent to maximum entropy) to find the closest approximating probability distribution consistent with the known marginals in contingency table analysis. Konolige (1979) applied the least information method to expert systems, and this analysis has been extended by Lemmar and Barth (1982).

The mathematical framework used in this paper is defined below. Although the definitions are given for a space of four parameterized predicates, the framework applies to any number of predicates. The predicates are $A$, $B$,$C$, and $D$ where:

$A$ has possible values $A_i$ $\quad i=1$ *to* $I$ ;
$B$ has possible values $B_j$ $\quad j=1$ *to* $J$ ;
$C$ has possible values $C_k$ $\quad k=1$ *to* $K$ ;
$D$ has possible values $D_l$ $\quad l=1$ *to* $L$ ,

and $P_{ijkl}$ is the probability that $A$ has value $i$, $B$ has value $j$, $C$ has value $k$, and $D$ has value $l$.

For example, $A$ might be the predicate "soil-type" where $A_1$ has the value "clay," $A_2$ is "silt" and so on. Each value (category) of a predicate is assumed to be mutually exclusive and exhaustive of the other categories. Any predicate that is not currently exhaustive can be made so by adding an extra category "other" for anything that does not fit the existing categories. In terms of these predicates, the entropy function ($H$) is defined below:

$$H = -\sum_{ijkl} P_{ijkl}\, Log(P_{ijkl}) \tag{1}$$

$H$ is a measure of the uncertainty inherent in the component probabilities. For example, if one of the $P_{ijkl}$ values is 1 (and so all the rest are zero), then H is zero—i.e., there is no uncertainty, as expected. Conversely, $H$ can be shown to be a maximum when all the $P_{ijkl}$ values are equal (if there are no constraints)—this represents a state of maximum uncertainty.

Because the $P_{ijkl}$ values are probabilities, they must obey the following constraint:

$$\sum_{ijkl} P_{ijkl} = 1$$

In addition, any subset of the following constraints may be asserted:

joints; such as

$$\sum_{jkl} P_{ijkl} = P_i^A \tag{2}$$

$$\sum_{ij} P_{ijkl} = P_{kl}^{CD} \tag{3}$$

$$\sum_{i} P_{ijkl} = P_{jkl}^{BCD} \tag{4}$$

$P_{ijkl} = specific \quad value;$
$etc.,$

conditional probabilities, such as:

$$P(A_2|B_3) = \frac{P_{23}^{AB}}{P_3^B} = \frac{\sum_{kl} P_{kl}}{\sum_{ikl} P_{ikl}}$$

and probability assignment to arbitrary logical functions, e.g.

$$P(A_2 \Rightarrow B_3) = x \quad \text{(logical implication)}$$

implying $\quad \sum_{j} P_{2j}^{AB} = 1 - x \quad (j \neq 3)$

implying $\quad \sum_{jkl}^{j\neq 3} P_{2jkl} = 1 - x$

These constraints are given explicitly because their values differ significantly from their expected ME value. Such significant constraints could be specified by experts or found by a program that examines data looking for significant combinations. The main reason for calculating ME probability values on the basis of the known constraints is to be able to find any probability value without having to store the entire probability space. Only linear constraints involving equalities have been given above, but the ME method can be extended to include non-linear constraints as well. Note that (2), for example, is itself a set of constraints one for each value of i given. Also it is assumed here that if either the numerator or the denominator of a conditional probability constraint is given separately (as a joint probability constraint), then the conditional probability constraint is replaced by the equivalent joint probability (marginal) constraints. The last constraint indicates that a probability assignment to any logical formula is equivalent to a probability assignment to a subset of the total probability space, and so forms a simple linear constraint.

The principle of maximum entropy requires that a unique set of values for $P_{ijkl}$ be found that satisfies the given constraints and at the same time maximizes the value of // given by (1). A method for calculating this ME distribution is discussed in the following section. The reasons for accepting ME probability values as the best estimate of the true probability ane discussed in Javnes (1979) and Lewis (1959), and may be summarized as follows. In expert system applications, when all the significant constraints (e.g., marginals and conditionals) have been found, all the information about the domain is contained in these constraints. Any ME probability value calculated with these constraints has distributed our uncertainty *(H)* as evenly as possible over the underlying probability space in a way consistent with the constraints. Returning any non-ME value implies that extra information is being assumed because H is no longer a maximum.

The shape of a particular H distribution around the ME value indicates how well the particular calculated probability is constrained by the available information. A strongly peaked curve indicates that the value is highly localized around the ME value, whereas a relatively flat curve indicates that very little information about the calculated probability is available—i.e., it is essentially unknown. The difference between $H$ for an assumed probability and $H$ maximum (which occurs for the ME probability value) gives the amount of information assumed by choosing a non-ME probability value.

## §3 A New Method of Calculating Maximum Entropy Distributions

The first use of maximum entropy (least information) for estimating probability distributions in computer science is due

to Lewis (1959). He showed that if the given probabilities are conditionally independent then the underlying probability space can be represented by simple product formulas and that this is the maximum entropy distribution. This product form is the basis of Dependence Trees (Chow and Liu 1968) and the tree based Bayesian update method of Pearl (1982). An iterative techinique for computing the ME distribution given some of the joint probabilities without requiring conditional independence was developed by Brown (1959). This method was extended by Ku and Kullback (1969), but both authors put strong restrictions on the constrints that must be given, and their method combinatorially explodes if the space of predicates is large. The new method of computing ME distributions presented in this section avoids these difficulties.

The problem of optimizing a continuous function subject to constraints is a well-known one in applied mathematics and a general solution is the method of Lagrange multipliers. The specific problem of maximizing the entropy function (1) subject to constraints was first applied to the domain of statistical mechanics, and specifically to joint (marginal) constraints by Gokhale and Kullback (1978). This section derives the necessary formulae in a form suitable for efficient computation.

The first step is to form a new entropy function as defined below:

$$H' = -\sum_{ijkl} P_{ijkl}\, Log P_{ijkl} + \lambda\left(1 - \sum_{ijkl} P_{ijkl}\right) + \lambda_i\left(P_i^A - \sum_{jkl} P_{ijkl}\right) + \lambda_{23}^{AB}\left(P(A_2|B_3)\sum_{ikl} P_{i3kl} - \sum_{kl} P_{23kl}\right) + \cdots \quad (5)$$

The next step is to equate the derivative of (5) (with respect to each variable) to zero, giving:

$$\frac{\partial H'}{\partial P_{ijkl}} = -Log P_{ijkl} - 1 - \lambda - \lambda_i - \cdots - \lambda_{ij} - \cdots - \lambda_{23}^{AB} \cdots = 0$$

$$\text{implying} \quad Log P_{ijkl} = -[\lambda_0 + \lambda_i + \cdots + \lambda_{ij} + \cdots + \lambda_{23}^{AB} + \cdots] \quad \text{where} \quad (\lambda_0 = \lambda + 1) \quad (6)$$

$$\text{or} \quad P_{ijkl} = exp - (\lambda_0 + \lambda_i + \cdots + \lambda_{ij} + \cdots + \lambda_{23}^{AB} + \cdots) \quad (7)$$

$$\text{and} \quad \frac{\partial H'}{\partial \lambda_0} = 0 \;\Rightarrow\; \sum_{ijkl} P_{ijkl} = 1 \quad (8)$$

$$\frac{\partial H'}{\partial \lambda_i} = 0 \;\Rightarrow\; \sum_{jkl} P_{ijkl} = P_i^A \quad (9)$$

$$\cdots$$

$$\frac{\partial H'}{\partial \lambda_{23}^{AB}} = 0 \;\Rightarrow\; \sum_{kl} P_{kl} = P(A_2|B_3)\sum_{ikl} P_{ikl} \quad (9)$$

$$\text{etc.}$$

Equation (7) gives the ME distribution in terms of the Xs, so if the values of all Xs can be found, the ME space is known implicitly. Note that equation (6) is the so-called loglinear form, but here this form is a direct consequence of the maximization of // rather than an ad hoc assumption. From (5) it is clear that there is only one X per constraint and that these are the only unknowns. If equation (7) is substituted into (8-9) etc., (i.e., into each of the given constraints), then the resulting a set of simultaneous equations can be solved for the Xs. It is more convenient first to apply the following transformations:

$$\alpha_0 = e^{-\lambda_0} \quad \alpha_i = e^{-\lambda_i} \quad \alpha_{ij} = e^{-\lambda_{ij}} \quad etc.$$

$$\Rightarrow \quad P_{ijkl} = \alpha_0\alpha_i \cdots \alpha_{ij} \cdots \alpha_{23}^{AB} \cdots \quad (10)$$

i.e., the basic distribution $P_{ijkl}$ is given implicitly as a product of as. Equation (10) is the key to the new ME calculation method, as it implicitly gives the underlying probability space in terms of a product of parameters (the as), and there are only as many aS as there are constraints. Note that for any particular $P_{ijkl}$, only those aS with the corresponding indices appear in the product. With these substitutions, equation (8) becomes:

$$\alpha_0 \sum_{ijkl} \alpha_i\alpha_j \cdots \alpha_{ij}\alpha_{ik} \cdots \alpha_{ijk} \cdots = 1$$

and (9) becomes:

$$\alpha_2^A \alpha_{23}^{AB} \sum_{kl} \alpha_k\alpha_l\alpha_{kl}\alpha_{2k}^{AC} \cdots = P(A_2|B_3)\sum_{ikl} \alpha_i\alpha_k\alpha_l\alpha_{ik}\alpha_{3k}^{BC} \cdots$$

and so on (one equation for each constraint).

This set of simultaneous equations can be solved by any standard numerical techniques. However, in practice it is more common to need to update an existing solution by adding a new constraint. Such an update introduces a new corresponding (nonunity) a, and causes adjustments to some of the existing as. Even when a set of constraints is to be added, they can be introduced sequentially, thus an update method is always sufficient to compute the as. A suitable update method is to assume initially that all the as have their old value, then calculate a value for the new a from the new constraint equation. This new value is inserted into each of existing constraint equations in turn, and revised a values are calculated for the *a* corresponding to each constraint. This process is repeated until all the a values have converged on their new values. Current investigations are trying to determine which as change during update, the convergence of the method, and its computational cost.

## §4 Probabilistic Inference

The previous section describes a method for representing (implicitly) the underlying ME probability distribution. This section describes how to use such a representation to calculate the conditional probability of any desired predicate, given information about a specific case. Such a computation requires

summing over the probability space without creating a combinatorial explosion, as shown below.

Consider a hypothetical example with predicates $A_i, B_j, C_k$, $D_l$ and $E_m$, where each $A_i$ for example, could represent different age categories. If prior probabilities (constraints) are given for some of the predicate values and prior joint and conditional probabilities for combinations of values of different predicates, then the corresponding $\alpha$ values can be computed as explained in the previous section. The resulting probability space for a particular combination of given prior probabilities might be:

$$P_{ijklm} = \alpha_0\alpha_i\alpha_j\alpha_k\alpha_l\alpha_m\alpha_{ij}\alpha_{ik}\alpha_{jl}\alpha_{jm}\alpha_{km}\alpha_{lm}\alpha_{jlm}$$

If the prior probability of say a particular $A_i$ is required (i.e., it is not one of the given priors) then the value is given by:

$$P(A_i) = \alpha_0\alpha_i \sum_{jklm} \alpha_j\alpha_k\alpha_l\alpha_m\alpha_{ij}\alpha_{ik}\alpha_{jl}\alpha_{jm}\alpha_{km}\alpha_{lm}\alpha_{jlm}$$

$$= \alpha_0\alpha_i \sum_k \alpha_k\alpha_{ik}\left(\sum_m \alpha_m\alpha_{km}\left(\sum_j \alpha_j\alpha_{jm}\left(\sum_l \alpha_l\alpha_{lm}\alpha_{jl}\alpha_{jlm}\right)\right)\right)$$

Here, the $\sum$ has been recursively decomposed into its component partial sums allowing each partial sum to be computed as soon as possible and the resulting matrix then becoming a term in the next outer-most $\sum$. In the above example, this summation method reduces the cost of evaluating $\sum_{jklm}$ from O(J*K*L*M) (where J,..,M are the ranges of $j,..,m$ respectively) to O(J*L*M) i.e., the cost of evaluating the innermost $\sum$. Note that a different order of decomposition can produce higher costs i.e., the cost of $\sum$ evaluation is dependent on the evaluation order, and is a minimum when sum of the sizes of the intermediate matrices is a minimum. When there are a large number of predicates, the total computational cost of evaluating a $\sum$ is usually dominated by the largest intermediate matrix, whose size is partly dependent on the degree of interconnectedness of the predicate being summed over and the order of evaluation. The above summation procedure is also useful in updating the previous $\alpha$s when given new prior probabilities. In the above, $\alpha_0$ is a normalization constant that can be determined, once all values of $P(A_i)$ have been evaluated, from the requirement that $\sum_i P(A_i) = 1$. This normalization makes prior evaluation of $\alpha_0$ unnecessary.

To find the conditional probability of a predicate (or joint conditional probability of a set of predicates), all that is needed is to drop those predicates whose values are given from the total $\sum$. For example, to find the conditional probability of $A_i$ given that $D_2$ and $E_3$ are true, the correct formula is:

$$P(A_i/D_2,E_3) = \beta\alpha_i \sum_{jk} \alpha_j\alpha_k\alpha_{ik}\alpha_{j2}^{BD}\alpha_{j3}^{BE}\alpha_{k3}^{CE}\alpha_{j23}^{BDE}$$

$$= \beta\alpha_i\left(\sum_j \alpha_j\alpha_{j2}^{BD}\alpha_{j3}^{BE}\alpha_{j23}^{BDE}\right)\left(\sum_k \alpha_k\alpha_{ik}\alpha_{k2}^{CE}\right)$$

where $\beta$ is a normalization constant. Note that the more evidence there is concerning a particular case, the smaller the resulting . Abo, the conditional probability evaluation procedure is nondirectional because, unlike other expert systems, this procedure allows the conditional probability of any predicate to be found for any combination of evidence. That is, it has no specially designated evidence and hypothesis predicates.

The above probability evaluation method can be extented to include the case where the evidence in a particular case is in the form of a probability distribution over the values of a predicate that is different from the prior distribution, rather than being informed that a particular value is true. In this case, it is necessary to compute new as that correspond to the given distribution and use these new as in place of the prior corresponding as in probability evaluations such as those above. For instance, if a new distribution is given for $P(A_i)$, then the new as are given by:

$$\alpha_i^A(new) = \alpha_i^A(old)\frac{P_i^A(new)}{P_i^A(old)}$$

Note that the revised a values used in the above method are just multiplicative factors whose value is identical to the correction factors of Lemmar and Barth (1982), and so the methods are equivalent in this case. The major difference is that here the probability space is represented implicitly by the as, and the corresponding summation procedure will work even when the space cannot be partitioned.

The above conditional probability evaluation procedure (a type of expert system inference engine) has been implemented in LISP and has been tested on many well known ME examples. In ME conditional probability calculations when given specific evidence, it has been found that only short strong chains of prior joint or conditional probabilities can significantly change the probability of a predicate of interest from its prior value.

When a point probability value is computed by the proposed method, it is useful to also estimate its accuracy as well. There are two sources of uncertainty in a computed ME value. One is the possibility that the known constraints used are not the only ones operating in the domain. This type of uncertainty is hard to quantify and depends on the methods used to Dud the known constraints. If a constraint search is systematic (over the known data), then we can be confident that we know all the dependencies that can contribute to a specific ME value. If a constraint search is ad hoc, it is always possible that a major contributing factor has been overlooked. If any important factors are missing, the calculated ME probability values will differ significantly from the observed values. If such deviations are found, it indicates that factors are missing, and an analysis of the deviations often gives a clue to these missing factors.

The other source of uncertainty is the accuracy with which the constraints are known. This accuracy depends on the size of the sample from which the constraints were extracted or the accuracy of the expert's estimates. This uncertainty is also hard to quantify, but it provides a lower limit on the accuracy of any calculated value. In the analysis given here, the constraints were assumed to be known with complete accuracy.

## §5 Summary

This paper presents a new method of computing maximum entropy distributions and shows how to use these distributions and some specific evidence to calculate the conditional probability of a predicate of interest. Previous methods of computing maximum entropy distributions are either too restrictive IN the constraints allowed, or too computationally costly in non-trivial cases. The new method avoids both these difficulties. Justifications for preferring maximum entropy values are given, as are ways of estimating their certainty.

Further research is necessary to further improve the efficiency of this method, particularly by automatically finding the optimal $\sum$ evaluation order and discovery of approximations that would allow the exclusion from the summation of any predicates that could not significantly effect the final result. Such improvements should increase the usefulness of this ME computation technique as an expert system inference engine.

## References

Broun, D. T., "A Note on Approximations to Discrete Probability Distributions", Information and Control, 2, pp. 380 392 (1959)

Chow, C. K., Liu, C. N., "Approximating Discrete Probability Distributions with Dependence Trees", IEEE Trans, on Info. Theory, Vol IT-14, No. 3, May (1968)

Duda, R. O., Hart, P. E., and Nilsson, N. J., "Subjective Etayrsian Methods for Rule-Based Inference Systems.", Proc. AFIPS Nat. Compt. Conf., Vol 47, pp. 1075-1082, (1976)

Gokhale, D. V. and Kullback, S. "The Information in Contingency Tables", New York: Marcel Dekker, (1978)

Ireland, C. T. and Kullback, S. "Contingency Tables with Given Marginals", Biometrika, Vol. 55, pp. 179-188, March (1968)

Jaynes, E. T., "Prior Probabilities", IEEE TVans. on Systems Science and Cybernetics, Vol. SSC-4, No. 3, Sept. (1968)

Jaynes, E. T. "Where Do We Stand on Maximum Entropy", in "The Maximum Entropy Formalism", Levine and TVibus Eds. MIT. Press (1979)

Konolige, K. Appendix D in "A Computer-Based Consultant for Mineral Exploration" SRI Artificial Intelligence report Sept. (1979)

Ku, H. H. and Kullback S., "Approximating Discrete Probability Distributions" IEEE TVans. On Information Theory, Vol. IT-15, No. 4, July (1969)

Lemmer, J. F. and Barth, S. W., "Efficient Minimum Information Updating for Bayesian Inferencing in Expert Systems", pp. 424-427, Proc. National Conf. on Art. Intelligence, Pittsburgh, 1982

Lewis, P. M., "Approximating Probability Distributions to Reduce Storage Requirements." Infomation and Control, 2, pp. 214-225 (1959)

Pearl, J. "Distributed Bayesian Processing For Belief Maintenance in Hierarchical Inference Systems". Report UCLA-ENG-CSIv-82-11 Univ. of California, Los Angeles (1982)

Shafer, G. "A Mathematical Theory of Evidence", Princeton Univ. Press (1979)

Shore, J. E. and Johnson, R. W., "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy", IEEE Trans, on Info. Theory, Vol. IT-26, No. 1, pp. 26-37, Jan. (1980)