In AAAI-82: Proceedings of the National Conference on Artificial Intelligence, AAAI, 133-136, 1982.

REVEREND BAYES ON INFERENCE ENGINES: A DISTRIBUTED HIERARCHICAL APPROACH(*)(**)

Judea Pearl
Cognitive Systems Laboratory
School of Engineering and Applied Science
University of California, Los Angeles
90024

ABSTRACT

This paper presents generalizations of Bayes likelihood-ratio updating rule which facilitate an asynchronous propagation of the impacts of new beliefs and/or new evidence in hierarchically organized inference structures with multi-hypotheses variables. The computational scheme proposed specifies a set of belief parameters, communication messages and updating rules which guarantee that the diffusion of updated beliefs is accomplished in a single pass and complies with the tenets of Bayes calculus.

Introduction

This paper addresses the issue of efficiently propagating the impact of new evidence and beliefs through a complex network of hierarchically organized inference rules. Such networks find wide applications in expert-systems [1], [2], [3], speech recognition [4], situation assessment [5], the modelling of reading comprehension [6] and judicial reasoning [7].

Many AI researchers have accepted the myth that a respectable computational model of inexact reasoning must distort, modify or ignore at least some principles of probability calculus. Consequently, most AI systems currently employ ad-hoc belief propagation rules which may hinder both the inferential power of these systems and their acceptance by their intended users. The primary purpose of this paper is to examine what computational procedures are dictated by traditional probabilistic doctrines and whether modern requirements of local asynchronous processing render these doctrines obsolete.

We shall assume that beliefs are expressed in probabilistic terms and that the propagation of beliefs is governed by the traditional Bayes transformations on the relation $P(D|H)$, which stands for the judgmental probability of data D (e.g., a combination of symptoms) given the hypothesis H (e.g., the existence of a certain disease). The unique feature of hierarchical inference systems is that the relation $P(D|H)$ is computable as a cascade of local, more elementary probability relations involving intervening variables. Intervening variables, (e.g., organisms causing a disease) may or may not be directly observable. Their computational role, however, is to provide a conceptual summarization for loosely coupled subsets of observational data so that the computation of $P(H|D)$ can be performed by local processes, each employing a relatively small number of data sources.

The belief maintenance architecture proposed in this paper is based on a distributed asynchronous interaction between cooperating knowledge sources without central supervision similar to that used in the HEARSAY system [4]. We assume that each variable (i.e., a set of hypotheses) is represented by a separate processor which both maintains the parameters of belief for the host variable and manages the communication links to and from the set of neighboring, logically related variables. The communication lines are assumed to be open at all times, i.e., each processor may at any time interrogate its message-board for revisions made by its neighbors, update its own belief parameters and post new messages on its neighbors' boards. In this fashion the impact of new evidence may propagate up and down the network until equilibrium is reached.

The asynchronous nature of this model requires a solution to an instability problem. If a stronger belief in a given hypothesis means a greater expectation for the occurrence of a certain supporting evidence and if, in turn, a greater certainty in the occurrence of that evidence adds further credence to the hypothesis, how can one avoid an infinite updating loop when the two processors begin to communicate with one another? Thus, a second objective of this paper is to present an appropriate set of belief parameters, communication messages and updating rules which guarantee that the diffusion of updated beliefs is accomplished in a single pass and complies with the tenets of Bayes calculus.

A third objective is to demonstrate that proper Bayes inference can be accomplished among multi-valued variables and that, contrary to the claims made by Pednault, Zucker and Muresan [8], this does not render conditional independence incompatible with the assumption of mutual exclusivity and exhaustivity.

(*)The paper "An Essay Towards Solving a Problem in the Doctrine of Chances by the late Rev. Mr. Bayes", Phil. Trans. of Royal Soc., 1763, marks the beginning of the science of inductive reasoning.

(**) Supported in part by the National Science Foundation, Grant IST 80 19045.

Definitions and Nomenclature

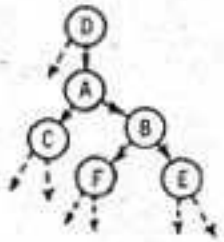
A node in an inference net represents a variable name. Each variable represents a finite partition of the world given by the variable values or states. It may be a name for a collection of hypotheses (e.g., identity of organism: ORG_1, ORG_2, ...) or for a collection of possible observations (e.g., patient's temperature: high, medium, low). Let a variable be labeled by a capital letter, e.g., A,B,C,..., and its various states subscripted, e.g., $A_1, A_2, \ldots$.

An inference net is a directed acyclical graph where each branch $(A) \rightarrow (B)$ represents a family of rules of the form: if A_i then B_j. The uncertainties in these rules are quantified by a conditional probability matrix, $\underline{M}(B|A)$, with entries: $M(B|A)_{ij} = P(B_j|A_i)$. The presence of a branch between A and B signifies the existence of a direct communication line between the two variables. The directionality of the arrow designates A as the set of hypotheses and B as the set of indicators or manifestations for these hypotheses. We shall say that B is a <u>son</u> of A and confine our attention to <u>trees</u>, where every node has only one multi-hypotheses father and where the leaf nodes represent observable variables.

In principle, the model can also be generalized to include some graphs (multiple parents), keeping in mind that the states of each variable in the tree may represent the power set of multi-parent groups in the corresponding graph.

Structural Assumptions

Consider the following segment of the tree:



The likelihood of the various states of B would, in general, depend on the entire data observed so far, i.e., data from the tree rooted at B, the tree rooted at C and the tree above A. However, the fact that B can communicate directly only with its father (A) and its sons (F and E) means that the influence of the entire network above B on B is completely summarized by the likelihood it induces on the states of A. More formally, let $D_d(B)$ stand for the data obtained from the tree rooted at B, and $D^u(B)$ for the data obtained from the network above B. The presence of only one link connecting $D^u(B)$ and (B) implies:

$$P(B_j|A_i, D^u(B)) = P(B_j|A_i) \qquad (1)$$

This structural assumption of local communication immediately dictates what is normally called "Conditional Independence": if C and B are siblings and A is their parent, then

$$P(B_j, C_k|A_i) = P(B_j|A_i) \cdot P(C_k|A_i) \qquad (2)$$

because the data $C=C_k$ is part of $D^u(B)$ and hence (1) implies $P(B_j|C_k, A_i) = P(B_j|A_i)$, from which (2) follows.

Note the difference between the weak form of conditional independence in (2) and the over-restrictive form adapted by Pednault et al. [8], who also asserted independence with respect to the complements $\bar{A}_i$.

Combining Top and Bottom Evidences

Our structural assumption (1) also dictates how evidences above and below some variable B should be combined. Assume we wish to find the likelihood of the states of B induced by some data D, part of which, $D^u(B)$, comes from above B and part, $D_d(B)$, from below. Bayes theorem, together with (1), yields the product rule:

$$P(B_i|D^u(B), D_d(B)) = \alpha P[D_d(B)|B_i] \cdot P[B_i|D^u(B)], \qquad (3)$$

where α is a normalization constant. This is a generalization of the celebrated Bayes formula for binary variables:

$$O(H|E) = \lambda(E)\, O(H) \qquad (4)$$

where $\lambda(E) = P(E|H)/P(E|\bar{H})$ is known as the likelihood ratio, and $O(H) = P(H)/P(\bar{H})$ as the prior odds [2].

Equation (3) generalizes (4) in two ways. First, it permits the treatment of non-binary variables where the mental task of estimating $P(E|\bar{H})$ is often unnatural, and where conditional independence with respect to the negations of the hypotheses is normally violated (i.e., $P(E_1, E_2|\bar{H}) \neq P(E_1|\bar{H})P(E_2|\bar{H})$). Second, it identifies a surrogate to the prior probability term for any intermediate node in the tree, even <u>after</u> obtaining some evidential data. According to (3), the multiplicative role of the prior probability in Equation (4) is taken over by the conditional probability of a variable based <u>only</u> on the evidence gathered by the network <u>above</u> it, excluding the data collected from below. Thus, the product rule (3) can be applied to any node in the network, without requiring prior probability assessments.

The root is the only node which requires a prior probability estimation. Since it has no network above, $D^u(B)$ should be interpreted as the available background knowledge which remains unexplicated by the network below. This interpretation renders $P(B_i|D^u(B))$ identical to the classical notion of subjective prior probability. The probabilities of all other nodes in the tree are uniquely determined by the arc-matrices, the data observed and the prior probability of the root.

Equation (3) suggests that the probability distribution of every variable in the network can be computed if the node corresponding to that variable contains the parameters

$$\lambda(B_i) \triangleq P(D_d(B)|B_i) \qquad (5)$$

and

$$q(B_i) \triangleq P(B_i|D^u(B)). \qquad (6)$$

$q(B_i)$ represents the anticipatory support attributed to B_i by its ancestors and $\lambda(B_i)$ represents the evidential support received by B_i from its diagnostic descendants. The total strength of belief in B_i

would be given by the product

$$P(B_i) = \alpha\lambda(B_i)\, q(B_i). \qquad (7)$$

Whereas only two parameters, $\lambda(E)$ and $\Omega(H)$, were sufficient for binary variables, an n-state variable needs to be characterized by two n-tuples:

$$\underline{\lambda}(B) = \lambda(B_1), \lambda(B_2), \ldots, \lambda(B_n)$$

$$\underline{q}(B) = q(B_1), q(B_2), \ldots, q(B_n).$$

Propagation of Information Through the Network

Assuming that the vectors $\underline{\lambda}$ and $\underline{q}$ are stored with each node of the network, our task is now to prescribe how the influence of new information spreads through the network. Traditional probability theory, together with some efficiency considerations [9], dictate the following propagation scheme which we first report without proofs.

1. Each processor computes two message vectors: $\underline{P}$ and $\underline{r}$. $\underline{P}$ is sent to every son while $\underline{r}$ is delivered to the father. The message $\underline{P}$ is identical to the probability distribution of the sender and is computed from $\underline{\lambda}$ and $\underline{q}$ using Equation (7). $\underline{r}$ is computed from $\underline{\lambda}$ using the matrix multiplication:

$$\underline{r} = \underline{\underline{M}} \cdot \underline{\lambda} \qquad (8)$$

where $\underline{\underline{M}}$ is the matrix quantifying the link to the father. Thus, the dimensionality of $\underline{r}$ is equal to the number of hypotheses managed by the father. Each component of $\underline{r}$ represents the diagnostic contribution of the data below the host processor to the belief in one of the father's hypotheses.

2. When processor B is called to update its parameters, it simultaneously inspects the $\underline{P}(A)$ message communicated by the father A and the messages $\underline{r}_1, \underline{r}_2, \ldots$, communicated by each of its sons and acknowledges receiving the latter. Using these inputs, it then updates $\underline{\lambda}$ and $\underline{q}$ as follows:

3. Bottom-up propagation: $\underline{\lambda}$ is computed using a term-by-term multiplication of the vectors $\underline{r}_1$, $\underline{r}_2, \ldots$:

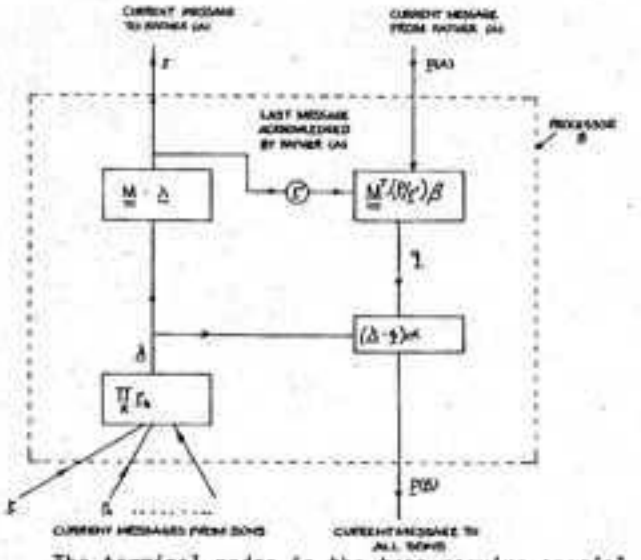
$$\lambda(B_i) = (\underline{r}_1)_i \times (\underline{r}_2)_i \times \ldots = \prod_k (\underline{r}_k)_i \qquad (9)$$

4. Top-down propagation: $\underline{q}$ is computed using:

$$q(B_i) = \beta \sum_j P(B_i|A_j) P(A_j)/(\underline{r}')_j \qquad (10)$$

where β is a normalization constant and $\underline{r}'$ is the last message from B to A acknowledged by the father A. (The division by $\underline{r}'$ amounts to removing from $\underline{P}(A)$ the contribution due to $D_B(B)$ as dictated by the definition of q in Equation (6)).

5. Using the updated values of $\underline{\lambda}$ and $\underline{q}$, the messages $\underline{P}$ and $\underline{r}$ are then recomputed as in step 1 and are posted on the message-boards dedicated for the sons and the father, respectively. This updating scheme is shown schematically in the diagram below, where multiplications and divisions of any two vectors stand for term-by-term operations.



The terminal nodes in the tree require special boundary conditions. Here we have to distinguish between the two cases:

1. Anticipatory node: an observable variable whose state is still unknown. For such variables, $\underline{P}$ should be equal to $\underline{q}$ and, therefore, we should set $\underline{\lambda}=(1,1,\ldots,1)$ (also implying $\underline{r}=(1,1,\ldots,1)$).

2. Data-node: an observable variable with a known state. Following Equation (5), if the j^{th} state of B was observed to be true, set $\underline{\lambda}$ = (0,0...0,1,0...) with 1 at the j^{th} position.

Similarly, the boundary conditions for the root node is obtained by substituting the prior probability instead of the message $\underline{P}(A)$ expected from the father.

A Token Game Illustration

Figure 2 shows six successive stages of belief propagation through a simple binary tree, assuming that updating is activated by changes in the belief parameters of neighboring processes. Initially (Figure 2a), the tree is in equilibrium and all terminal nodes are anticipatory. As soon as two data nodes are activated (Figure 2b), white tokens are placed on their links, directed towards their fathers. In the next phase, the fathers, activated by these tokens, absorb the latter and manufacture the appropriate number of tokens for their neighbors (Figure 2c), white tokens for their fathers and black ones for the children (the links through which the absorbed tokens have entered do not receive new tokens, thus reflecting the division of $\underline{P}$ by $\underline{r}'$). The root node now receives two white tokens, one from each of its descendants. That triggers the production of two black tokens for top-down delivery (Figure 2d). The process continues in this fashion until, after six cycles, all tokens are absorbed and the network reaches a new equilibrium.

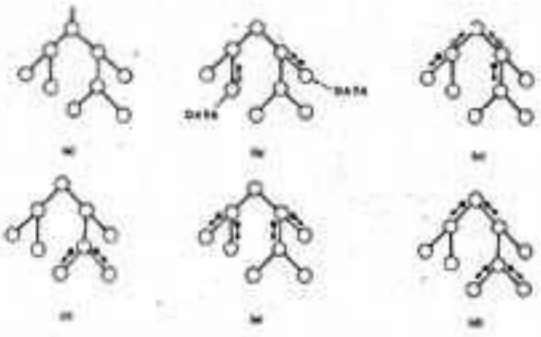


Figure 2

Properties of the Updating Scheme

1. The local computations required by the proposed scheme are efficient in both storage and time. For an m-ary tree with n states per node, each processor should store $n^2+mn+2n$ real numbers, and perform $2n^2+mn+2n$ multiplications per update. These expressions are on the order of the number of rules which each variable invokes.

2. The local computations are entirely independent of the control mechanism which activates the updating sequence. They can be activated by either data-driven or goal driven (e.g., requests for evidence) control strategies, by a clock or at random.

3. New information diffuses through the network in a single pass. Infinite relaxations have been eliminated by maintaining a two-parameter system ($\underline{\pi}$ and $\underline{\lambda}$) to decouple top and bottom evidences. The time required for completing the diffusion (in parallel) is equal to the diameter of the network.

A Summary of Proofs

From the fact that $\underline{\lambda}$ is only influenced by changes propagating from the bottom and $\underline{\pi}$ only by changes from the top, it is clear that the tree will reach equilibrium after a finite number of updating steps. It remains to show that, at equilibrium, the updated parameters $P(V_i)$, in every node V, correspond to the correct probabilities $P(V_i|D^u(V),D_d(V))$ or (see Equation (3)), that the equilibrium values of $\lambda(V_i)$ and $\pi(V_i)$ actually equal the probabilities $P(D_d(V)|V_i)$ and $P(V_i|D^u(V))$. This can be shown by induction bottom-up for $\underline{\lambda}$ and then top-down for $\underline{\pi}$.

Validity of $\underline{\lambda}$: $\underline{\lambda}$ is certainly valid for leaf nodes, as was explained above in setting the boundary conditions. Assuming that the λ's are valid at all children of node B, the validity of $\lambda(B)$ computed through steps (8) and (9) follows directly from the conditional independence of the data beneath B's children (Equation (2)).

Validity of $\underline{\pi}$: if all the λ's are valid, then π is valid for the root node. Assuming now that $\pi(A)$ is valid, let us examine the validity of $\underline{\pi}(B)$, where B is any child of A. By definition (equation (6)), $\underline{\pi}(B)$ should satisfy:

$$\pi(B_i)=P(B_i|D^u(B))=\sum_j P(B_i|A_j)P(A_j|D^u(A),D_d(S))$$

where S denotes the set of B's siblings. The second factor in the summation differs from $P(A_j)*P(A_j|D^u(A),D_d(A))$ in that the latter has also incorporated B's message $(\underline{\pi}')_j$ in the formation of $\lambda(A_j)$ (equation (9)). When we divide $P(A_j)$ by $(\underline{\pi}')_j$, as prescribed in (10), the correct probability ensues.

Conclusions

The paper demonstrates that the centuries-old Bayes formula still retains its potency for serving as the basic belief revising rule in large, multi-hypotheses, inference systems. It is proposed, therefore, as a standard point of departure for more sophisticated models of belief maintenance and inexact reasoning.

REFERENCES

[1] Shortliffe, E.H., and Buchanan, B.G., "A Model of Inexact Reasoning in Medicine". Math.Biosci., 23 (1975), 351-379.

[2] Duda, R.O., Hart, P.E. and Nilsson, N.J., "Subjective Bayesian Methods for Rule-Based Inference Systems". Tech. Note 124, AI Center, SRI International, Menlo Park, CA; also Proc. 1976 NCC (AFIPS Press).

[3] Duda, R., Hart, P., Barrett, P., Gashnig, J., Konolige, K., Reboh, R. and Slocum J., "Development of the Prospector Consultation System for Mineral Exploration". AI Center, SRI International, Menlo Park, CA, Sept. 1976.

[4] Lesser, V.R. and Erman, L.D., "A Retrospective View of HEARSAY II Architecture". Proc. 5th Int. Joint Conf. AI, Cambridge, MA, 1977, 790-800.

[5] DDI Handbook for Decision Analysis, Decision and Design Inc., McLean, VA, 1973.

[6] Rumelhart, D.E., "Toward an Interactive Model of Reading". Center for Human Info. Proc. CHIP-56, UC La Jolla, March 1976.

[7] Schum, D. and Martin, A., "Empirical Studies of Cascaded Inference in Jurisprudence: Methodological Consideration". Rice Univ., Psychology Research Report, #80-01, May 1980.

[8] Pednault, E.P.D., Zucker, S.W. and Muresan, L.V., "On the Independence Assumption Underlying Subjective Bayesian Updating". Art. Intel., Vol. 16, No. 2, May 1981, 213-222.

[9] Pearl, J., "Belief Propagation in Hierarchical Inference Structures". UCLA-ENG-CSL-8211, UC Los Angeles, January 1982.