Presented at the 2nd International Workshop on AI and Statistics,
Fort Lauderdale, Florida, January 1989.

ON RULE-BASED PROBABILISTIC INFERENCE: THEORETICAL PRINCIPLES AND PRACTICAL TECHNIQUES

Padhraic Smyth and Rodney M. Goodman
Department of Electrical Engineering 116-81
California Institute of Technology
Pasadena, CA 91125, USA

Introduction

In this paper we address the problem of probabilistic inference in a *rule*-based inference net. Rule-based inference is based on a set of probabilistic rules of the form

$$\text{If } \mathbf{Y} = y \text{ then } \mathbf{X} = x \text{ with probability } p \ ,$$

where $\mathbf{Y}$ and $\mathbf{X}$ are discrete random variables, with y and x being elements of their respective alphabets. $\mathbf{Y}$ may be a conjunction of basic propositions, $\mathbf{Y_1} = y_1, \ldots, \mathbf{Y_k} = y_k$, but $\mathbf{X} = x$ is restricted to being a simple proposition. Consider that we have N such random variables, including $\mathbf{Y_1}, \ldots, \mathbf{Y_k}$ and $\mathbf{X}$. In this sense any given rule set characterises to some degree of accuracy the joint probability distribution of the N variables. We have recently introduced [1] a measure for the information content of a rule, called the J-measure, i.e.,

$$J(\mathbf{X}; \mathbf{Y} = y) \ = \ p(y) \sum_x p(x|y) . \log(\frac{p(x|y)}{p(y)}) \tag{1}$$

which can be interpreted as the discrimination between the hypotheses that the variable $\mathbf{X}$ is dependent, or independent, of the event $\mathbf{Y}$=y. The ITRULE algorithm [2,3] uses the J-measure to induce the most informative set of probabilistic rules, based on sample data.

The primary goal of this paper is to investigate the appropriate *theoretical* equations for Bayesian updating using probabilistic rules, and then to analyse carefuly what assumptions and techniques can be used to render the updating *practical*. Our results may be interpreted as a quantitative demonstration of the recent assertion that probabilistic rule-based systems lack semantic modularity [4], i.e., we show clearly that rule transition probabilities in any knowledge base are, in general, *dynamic* rather than *static*. We propose a maximum-likelihood type of updating algorithm to minimise the effect on inference of such probability changes.

Basic concepts of rule-based updating

We begin by considering *singly* connected inference structures, where rules correspond to links, and nodes correspond to propositions of the form $\mathbf{X}$= x. We define the notion of a source, $\mathbf{S} = s$, which is a proposition whose probability has been determined via the external environment, i.e., a reference point from which to propagate *a posteriori* probabilities. A posteriori probabilities with respect to the source proposition s are denoted by the subscript s, e.g., $p_s(x)$. Note that $p_s(x) \neq p(x|s)$ unless $p_s(s) = 1$. In this sense we are dealing with the general case of *uncertain* evidence. In [3] we argue that

$$p_s(h|s) = p(h|s) \tag{2}$$

or that transition probabilities emanating directly from the source are invariant to changes at the source. More generally we find that

$$p_s(h_1, \ldots, h_n|e_1, \ldots, e_k, s) = p(h_1, \ldots, h_n|e_1, \ldots, e_k, s) \tag{3}$$

so that any transition probability conditioned on s is invariant to changes in s. The corollary to this statement is more interesting: in general we have that

$$p_s(h_1, \ldots, h_n, s|e_1, \ldots, e_k) \neq p(h_1, \ldots, h_n, s|e_1, \ldots, e_k) \tag{4}$$

and in particular,

$$p_s(h|e) \neq p(h|e) \quad . \tag{5}$$

This result states that rule transition probabilities are liable to change depending on changes in probabilities of propositions which are not directly linked to either the consequent or antecedent propositions. This means in turn that the use of *static* rule transition probabilities in a rule-based expert system represent an assumption on the part of the system designer unless those probabilities are 0 or 1, i.e., they are facts.

The general updating equation (for updating a *conditional* probability) is of the form

$$p_s(h|e) = \delta.\frac{p_s(s)}{p_s(e)} + \bar{\delta}.\frac{1 - p_s(s)}{p_s(e)} \tag{6}$$

where

$$\delta = p(h, e|s) = p_s(h, e|s) \quad \text{and} \quad \bar{\delta} = p(h, e|\bar{s}) = p_s(h, e|\bar{s}) \ , \tag{7}$$

i.e., δ and $\bar{\delta}$ do not depend on $p_s(s)$, they only depend on *a priori* probabilities. However the presence of higher-order terms on the right-hand side leads to the fact that to update a transition probability based on k prior pieces of evidence requires probability information up to order $k + 2$, a requirement which is generally impractical.

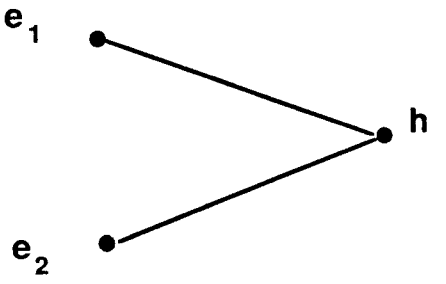

Figure 1: singly-connected inference net

Practical alternatives

Among the possible alternatives for reducing the complexity of equation (7) we conclude [3] that the widely-used conditional independence, or partial Markov, assumption of the form

$$p(e|h, s) = p(e|h) \tag{8}$$

is the most appropriate. Note that equation (8) is the so-called "weak" form of conditional independence and, hence, does not suffer from the inadequacies of the stronger "PROSPECTOR-type" assumptions (involving $\bar{h}$) as analysed by Johnson [5] among others. The conditional independence assumption effectively renders the transition probabilities static provided they are in the same "direction" as the source probability propagation (forward propagation). For probabilities on backward propagation paths the rule transition probabilities are liable to change. Hence in Figure 1 the probability $p_s(h|e_1) = p(h|e_1)$

if (8) holds, while $p_s(h|e_2) \neq p(h|e_2)$ in general whether (8) holds or not. Failure to keep track of the latter case implies that the system is not modelling *context*, with a consequent degradation in inference capabilities. For example a system based on backward chaining may over-estimate the relevance of e_2 in a given context, e.g., after s has been determined.

Problems and solutions due to multiply connected networks

It may be argued that, for uncertain reasoning, multiple-paths from evidence to hypotheses are more realistic than single-paths and, in addition, are more powerful if used correctly. Considering *multiple* paths from evidence to hypotheses, we find that the requirement for *point-valued a posteriori* probability estimates makes the problem particularly difficult. Relaxing this requirement to that of having lower bounds on the probability of each proposition and its negation makes it much more straightforward to handle multiple paths. In [3] we describe an interval-valued inference system based on bounding. For updating the *a posteriori* probability of a single hypothesis based on multiple pieces of evidence which are directly linked to the hypothesis, we use a *maximum-likelihood* type of updating rule, defined as follows:

$$t(h) = \max_{1 \leq i \leq n} \{p(h|e_i).p(e_i)\} \tag{9}$$

$$f(h) = \max_{1 \leq i \leq n} \{p(\bar{h}|e_i).p(e_i)\} \ , \tag{10}$$

where the e_i, $1 \leq i \leq n$, are the propositions (possibly conjunctive and by no means mutually exclusive) which are directly connected to h. $t(h)$ and $f(h)$ are lower bounds on $p(h)$ and $p(\bar{h})$ respectively. In a sense we are choosing the most likely event, of the events we know, in the probability space defined by the set $\{h, e_1, \ldots, e_n\}$. If this "most likely event" does not have a very high probability then the inference is cautious. On the other hand if the event has a high probability then both $p(e_j)$ and $p(h|e_j)$ must be near 1. In particular, if $p(e_j)$ and $p(h|e_j)$ are both 1, the inference is *certain* and facts are being propagated. We note in passing that Pearl has recently proposed (independently) a scheme quite similar in spirit, namely, MPE or "most probable explanation" [6].

Inference involves a trade-off. Stronger inference statements allow us to be more decisive, but these statements may not always be correct. The maximum likelihood equations seem an appropriate compromise. If $p(h|e_j) = 1$ then the updating rule is provably correct, since the transition probability corresponds to a fact and is invariant to any changes to other propositions. However when $p(h|e_j) < 1$ then the statement that

$$p(h) > p(h|e_j)p(e_j) \tag{11}$$

may not be true. Returning to our earlier notation, we see that the correct statement to make is that

$$p(h) > p_{\{e_1,\ldots,e_n,\ i \neq j\}}(h|e_j)p(e_j) \ . \tag{12}$$

Of course $p_{\{e_1,\ldots,e_n\}}(h|e_j)$ is our old friend, the unobtainable updated conditional probability. The transition probability from e_j to h may be affected by the probabilities of the other e_i.

We introduce the following notation. We have a proposition h, and n evidential propositions $e_1, \ldots, e_n$. For any proposition e_j, let $\mathbf{O_j}$ be a variable whose distribution is defined as the joint probability distribution of $\mathbf{E_1}, \ldots, \mathbf{E_i}, \ldots, \mathbf{E_n}$, $j \neq 1, i, n$, where each $\mathbf{E_i}$ is defined as a variable taking values in the alphabet $\{e_i, \bar{e_i}\}$. In words, $\mathbf{O_j}$ is defined as the joint variable of all the evidential propositions except for e_j. We denote $\sum_{e_j}$ or $\sum_{o_j}$ as the sum over all elements of the alphabet of $\mathbf{E_j}$ or $\mathbf{O_j}$ respectively. Using this notation, the correct updating equations are

$$p(h) = p_{o_j}(h|e_j)p(e_j) \tag{13}$$

$$= \left(\sum_{o_j} p(h|e_j, o_j) p_{o_j}(o_j|e_j) \right) p(e_j) \tag{14}$$

rather than the ML equation given by equations (9) and (10). The ML equation will be in error whenever

$$p(h|e_j, o_j) < p(h|e_j) \tag{15}$$

given that o_j and e_j are both true. This occurs on average with probability $p(o_j, e_j)$. Define the set W to be the set of component events of $\mathbf{O_j}$ such that equation (15) holds, i.e., $W = \{o_j : p(h|e_j, o_j) < p(h|e_j)\}$. Hence we define the average error T as the size of the error multiplied by the probability of its occurrence:

$$T = \Big(\sum_{o_j \in W} (p(h|e_j) - p(h|e_j, o_j)).p(o_j|e_j) \Big) p(e_j) \quad . \tag{16}$$

T is a measure of the average error in using the ML equations for updating. Obviously we would like T to be small if possible.

Theorem:

We have

$$T \leq \Big(p(h|e_j)\big(1 - p(h|e_j)\big) \Big).p(e_j) \tag{17}$$

$$\leq 0.25 p(e_j) \tag{18}$$

The proof of this result is given in [3]. From this theorem we see that as $p(h|e_j)$, the initial rule transition probability (not updated), approaches 1, the bound on the error goes to zero. This is a consequence of the fact that the closer $p(h|e_j)$ is to 1 initially, the less susceptible it is to changes in the probabilities of other propositions. As mentioned earlier, for the special case of $p(h|e_j) = 1$, it is invariant to change. The bound is attained if and only if all of the $p(h|e_j, o_j)$ (for $o_j \in W$) are equal to zero. Such zero terms, or even near-zero terms, have the interpretation that $p(\bar{h}|e_j, o_j) = 1 - p(h|e_j, o_j)$ must be near 1 and hence, either such terms will either be represented explicitly as rules, or if not, their overall probability of occurrence, $p(e_j, o_j, \bar{h})$ must be quite small. Using ITRULE to generate the rules in the network helps to keep T small, for if no rules exist for terms like $p(h|e_j, o_j)$ in the definition of T, then either $p(h|e_j, o_j) \approx p(h)$ or else $p(e_j, o_j)$ is very small. In either case the contribution to the error T should be small. Hence in the context of using rule-sets as generated by ITRULE, the bound on T is conjectured to be considerably larger than the actual value of T in practice.

Conclusions and future directions

The focus of our technique is rule-based or *event-based* reasoning rather than the more common variable-based reasoning. We have coupled the learning and inference phases in an implicit manner, such that the learning (via ITRULE) promotes the accuracy of the inference scheme (maximum likelihood) being used. Incorporating the ML equations into a *practical* inference algorithm will require further work on conflict resolution, non-monotonic reasoning, etc.

References

1. R. M. Goodman and P. Smyth, 'An information-theoretic model for rule-based expert systems,' presented at the 1988 International Symposium on Information Theory, Kobe, Japan.

4. R. M. Goodman and P. Smyth, 'Information-theoretic rule induction,' *Proceedings of the 1988 European Conference on Artificial Intelligence*, Pitman Publishing: London, August 1988.

3. P. Smyth, *The Application of Information Theory to problems in Decision Tree Design and Rule-Based Expert Systems*, Ph.D. Thesis, Department of Electrical Engineering, California Institute of Technology, May 1988.

4. D. Heckerman and E. J. Horvitz, 'The myth of modularity in rule-based systems for reasoning with uncertainty,' in *Uncertainty in Artificial Intelligence 2*, J. F. Lemmer and L. N. Kanal (Editors), Elsevier Science Publishers: Amsterdam, 1988, pp.23-34.

5. R. W. Johnson, 'Independence and Bayesian updating methods,' in *Uncertainty in Artificial Intelligence* , L. N. Kanal and J. F. Lemmer (Editors), Elsevier Science Publishers: Amsterdam, 1986, pp.197-201.

6. J. Pearl, 'Distributed revision of composite beliefs,' *Artificial Intelligence*, **33**, (1987) pp.173-215.