Approximating the Noisy-Or model by Naïve Bayes

From: AAAI Technical Report SS-98-03. Compilation copyright © 1998, AAAI (www.aaai.org). All rights reserved.

John M. Agosta
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
johnmark@sri.com

Abstract

This paper gives an algebraic derivation of the posterior for both the noisy-or and naïve Bayes models, as a function of both input messages and probability table parameters. By examining these functions we show a technique where the naïve Bayes model may be used to approximate a logical-OR, rather than its typical interpretation as a logical-AND. The technique is to avoid the use of disconfirming evidence in the naïve Bayes model. A comparison with the posterior function for the noisy-or shows the quality of the logical-OR naïve Bayes approximation.

This approximation is key to the assumption of the underlying tree structure in a certain class of diagnostic Bayes' networks, where the tree structure mimics an *is-a* hierarchy. We argue that this is the correct causal structure. An example of applying the Bayes' network model to network intrusion detection illustrates this. This assumption that a tree-structured diagnostic Bayes' network can be formulated from an *is-a* hierarchy is a useful elicitation tool. We have addressed the quality of the numeric approximation in this assumption; the exact nature of cause in an *is-a* hierarchy remains an open question.

1 Introduction

In the construction and elicitation of a diagnostic Bayes' network, an important question arises about the proper direction of the arcs in the network. By definition the inverse of a diagnostic relationship is a causal one: The node at the origin of the arc is the *cause* of the node at the end of the arc. Correspondingly the node at the arc's end is an *effect* or the *evidence* for the predecessor node. This makes for a *causal network.*

In this paper we consider causal networks that tend to have a tree-like structure, with one–or a few–top level, or root nodes, fanning out to a large number of branch nodes. We look at the computational consequences of making the root the cause and the branches the effects, and vice versa.

The pattern where the root node is the cause of several conditionally independent evidentiary nodes is called a naïve Bayes model. An example is shown in Figure 1. A tree-like structured network is a hierarchical composition of naïve Bayes models. Imposed on this structure could be also arcs indicating causal dependencies that embellish the strict hierarchical structure; either indicating multiple parent causes for one node, or indicating dependencies among nodes at the same level.

The basic tree-like structure is a simplifying assumption that gives the domain expert a framework by which to organize diagnostic variables. The assumption of the tree-like structure is that a set of specific variables depend causally on (i.e., are effects of) one variable that generalizes the specific variables. Or, stated in another way, the assumption is that the naïve Bayes model can be applied to capture an *is-a* relationship between parents and children. The exact causal nature of this relationship is at the heart of the assumption. To suggest its plausibility: It may be for example, if the parent node is *He went home* and the children are *He went home by car, He went home by train, He went home on foot,* etc., then the relationship expressed is that the act of going home causes the person to employ some combination of means to get there.

Thus employing the tree-like structure is an elicitation strategy to exploit the expert's ability to organize the

domain by *is-a* relationships. It begs the question of whether this is appropriate use of causality or of the probability calculus of Bayes' networks. To shed some light on the question, this paper examines what different Bayes' networks models compute.

1.1 Probabilistic versions of the logical-AND and logical-OR

In the case that the assertion of all the effects of a cause are necessary to support the cause, the relationship between cause and effects is a probabilistic analog of a logical-AND over the effects. As will be explained in this paper, this is also the relationship that the naïve Bayes model calculates. A difficulty occurs in the case that the cause can be established by any one (or any subset) of the effects of the cause. One common method to treat this case is to model the relationship using a noisy-or model of the effects. However this makes the effects nodes predecessors to the cause node, as if cause and effect had traded places. See again Figure 1. Although one might argue that this is correct from a computational point of view, it destroys the causal semantics of the network. Fortunately, as will be shown, it is possible to use the naïve Bayes model, together with certain assumptions to approximate a logical-OR among effects. By comparing naïve Bayes and noisy-or models this paper examines the logical-OR approximation, and the quality of this approximation. To demonstrate this with a specific example, the paper uses the Bayes' network for a computer network intrusion detection system for which this approximation was derived.

2 Design of a Diagnostic Network for Computer Network Intrusion Detection

The approach that we have taken for detection of network intruders follows closely the model-based diagnostic Bayes' network methods that have been applied in traditional Bayes' network domains. Approaching intrusion detection as a diagnostic problem instead of by rule-based reasoning is novel in the network intrusion domain. To carry this out we modeled the interaction between the security features of the computer operating system and the methods by which an intruder can try to defeat them. We[1] began elicitation with a bottom-up approach. We collected a wide range of monitoring tests, then assembled clusters of tests under the possible causes that explain them. We

[1]Harold Javitz, Norman Nielsen, and Ric Steinberger are the Bayes' network's other authors.

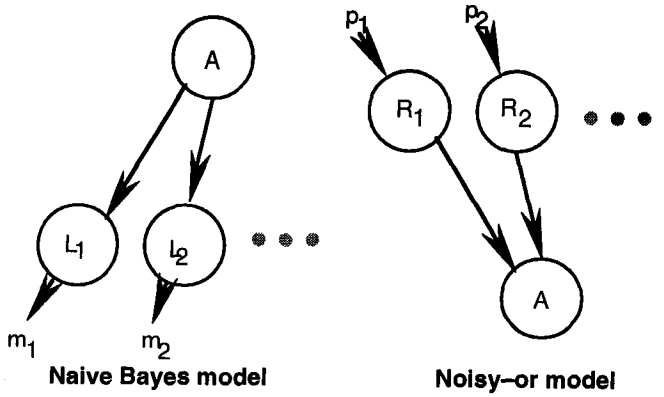

Figure 1: The naïve Bayes compared to the noisy-or model. Both reach conclusions about the event A based on evidence from two monitoring tests. In the naïve Bayes model, the two likelihood messages m are modified by the likelihood matrices L. In the noisy-or model, the two prior messages p are modified by the reliability matrices R.

represented the clustering of tests under causes by use of a naïve Bayes structure.

For example these tests make up one cluster:

Ping Test Evidence of repeated "pings" to establish the existence of a target computer at an IP address.

Strobe Test Evidence of scanning of successive port numbers for target TCP/IP services.

Probing Test Evidence of repeated attempts to connect to standard TCP/IP services (e.g. telnet) from remote locations.

These tests are each evidence that an intruder is searching the network to discover vulnerabilities in machines. In our representation *Discovery attempt* is a common cause for each of these activities. We capture this by making the three tests successor nodes to the *Discovery attempt* node, in a naïve Bayes model. Arguably there are also dependencies among the tests that we have not considered. An intruder may generate pings and attempt strobes to discover where next to attempt probing. Thus the suspiciousness of probing evidence would depend upon the state of test evidence for other tests. The arcs that capture these dependencies are largely ignored to keep the model simple. The fundamental relationship that we want the model to express is that strong evidence from *any* of these tests should confirm a *Discovery attempt* threat.

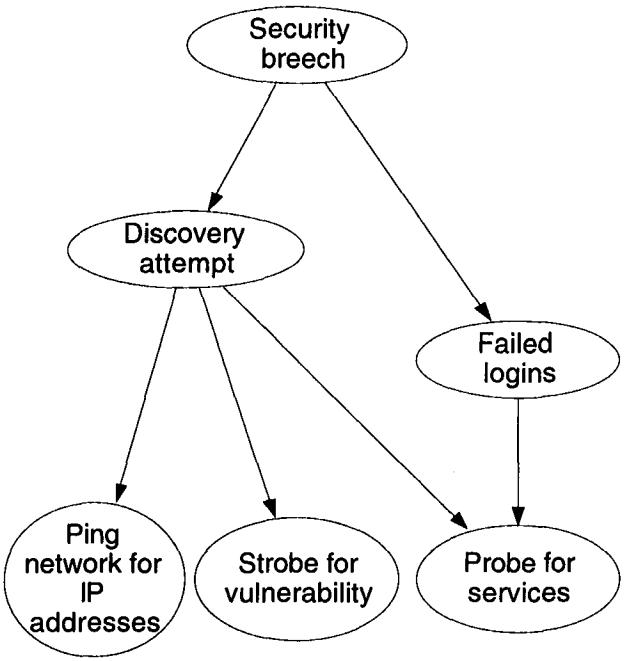

Figure 2: A fragment of the intrusion detection network consisting of two naïve Bayes levels. The leaves of the network are different monitoring tests that collect evidence used to detect intrusion attempts. Parents generalize over sets of monitoring tests that represent a common kind of threat. This model takes the form of a hierarchical naïve Bayes model. The additional arc in this network from *Failed logins* to *Probe for services* embellishes the hierarchical naïve Bayes model.

The network is built by stacking naïve Bayes models so that causes at one level become evidence at the next higher level. Thus if we look above the node *Discovery attempt*, we see that it is evidence for the cause *Security breech*, as is also *Failed logins*. See Figure 2.

Common to all of the network's naïve Bayes units is the relationship of a set of specific intrusion events to a more general intrusion event, where strong enough evidence of any of the specific events should confirm the general event. The "any of" nature of this relationship suggests a noisy-or relationship between the general event and set of specific events. This is inelegant for both representational and computational reasons. The problem is aggravated when we add an arc between *Failed logins* and *Probe for services*, indicating that the test of probing for services also detects unsuccessful attempts to login.

Consider applying noisy-or models to this network, by formulating the network with the same set of arcs, but that run in the opposite direction. This destroys the desirable multiple parent relationship formed by the *Probe for services* node. Gone is the desirable property of generating an "explaining away" relationship among its parents. Even worse is the spurious "explaining away" relationship now governed by the node *Security breech*. The explaining away property is the signature property of the noisy-or. It would not be exploited in the network where it is desired if the direction of the arcs were changed, and it would appear between test nodes, where there is no need for it.

Furthermore, the general tendency of the reversed-arc network is to increase the in-degree of nodes. This increases the cluster sizes of the join tree used to solve the network, which complicates the computation to solve the network.

To preserve the use of the naïve Bayes model instead of the noisy-or model for "any of" relationships, the general approach is to use only confirming evidence, and not to apply disconfirming evidence. The rest of this paper examines the quality of this approximation.

3 Comparisons of the models

As Figure 1 shows, the two models have a similar structure. The notation is a bit sloppy, but should be clear from context: Capital letters refer to both the node and its probability table, and sometimes the random variable for that node. Small letters refer to an instance of the random variable, or to a parameter in the probability table.

Common to both noisy-or and Bayes' network models

are a set of tests, or evidence that seek to confirm a threat, A. The models differ in the direction of their arcs, so they are not strictly equivalent. Yet we can make a correspondence. The prior messages p in the noisy-or model correspond to the likelihood messages m in the naïve Bayes model. In both models messages are modified by nodes—the reliability nodes R in the noisy-or model, and the likelihood nodes L in the naïve Bayes. In both the result of any message is to affect the posterior probability on the threat node, either $\mathrm{pr}\{A|m_i\}$ or $\mathrm{pr}\{A|p_i\}$. To compare the effect of messages on A's posterior, we will derive the algebraic relationship between messages and the threat node posterior for both models.

3.1 The noisy-or posterior as a function of prior messages

We can derive the probability of A as a function of the prior messages p_i by combining the probability tables of the nodes in the model. (The messages we consider are identical to the messages in Pearl's propagation scheme. We are looking only at the effects of prior messages, by considering only the predictive aspects of the noisy-or model. The inclusion also of likelihood messages would complicate the function.)

The probability table—of nodes R_i—for the individual reliabilities of the causes in the noisy-or are:

$$\mathrm{pr}\{R_i|P_i\} = \begin{pmatrix} r_i & 1-r_i \\ 0 & 1 \end{pmatrix}$$

The deterministic OR node probability table—for node A—is:

$$\mathrm{pr}\{A=a|R_1R_2\} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

The fact that causes work independently is enforced by the reliability of one cause r_1 being unaffected by the reliability of the second, r_2. By combining the nodes R_1 and R_2 with the deterministic OR node, A, we create the conventional noisy-or probability table:

$$\mathrm{pr}\{A=a|p_1p_2\} = \begin{pmatrix} r_1+r_2-r_1r_2 & r_1 \\ r_2 & 0 \end{pmatrix}$$

Taking expectation of the prior messages p_1 and p_2 over this matrix results in the expression we are looking for:

$$\begin{aligned} a_{no} &= (r_1+r_2-r_1r_2)p_1p_2 + r_1p_1(1-p_2) + r_2(1-p_1)p_2 \\ &= r_1p_1 + r_2p_2 - r_1p_1r_2p_2 \end{aligned}$$

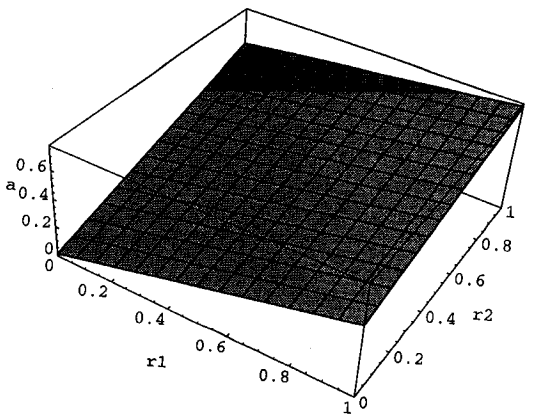

Figure 3: The noisy-or posterior as a function of input reliabilities.

We note that surprisingly the prior message parameter p and the reliability parameter r always appear as paired factors. We can thus conclude that:

Proposition 1 *The reliability, r and the prior message p are interchangeable in the predictive noisy-or model.*

For simplicity we will use $r/2$ in place of pr. Thus without loss of generality we have:

$$a_{no}(r_1, r_2) = r_1/2 + r_2/2 - r_1r_2/4$$

The graph of this function, shown in Figure 3 illustrates its properties. As is expected of a predictive relationship, the posterior of A is linear in both its prior messages. Increasing either message has the effect of increasing belief in A, and the rate of increase does not depend on the value of the other message. These properties of the noisy-or are the desired properties for the relationship between specific tests and the threat probabilities.

3.2 The naïve Bayes' posterior as a function of likelihood messages

The naïve Bayes model is diagnostic, so we solve for the probability of A as a function of the likelihood messages generated by the nodes L_i. Any predictive effects are captured in the prior parameter h, which has an overall scaling effect that is uninteresting for purposes of this comparison.

The likelihood matrices each have the form:

$$L_i = \begin{pmatrix} l_i & (1-l_i) \\ \overline{l_i} & (1-\overline{l_i}) \end{pmatrix} \text{ such that } l_i >> \overline{l_i}$$

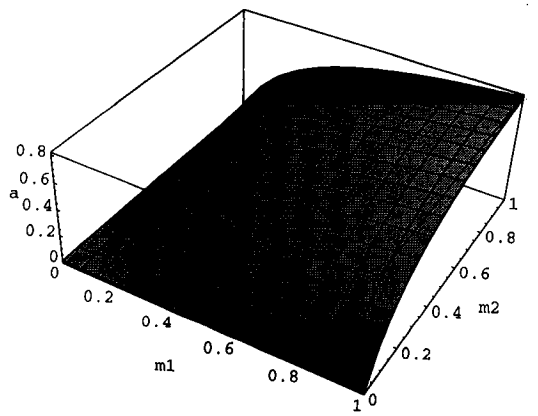

Figure 4: The naïve Bayes posterior as a function of input likelihoods.

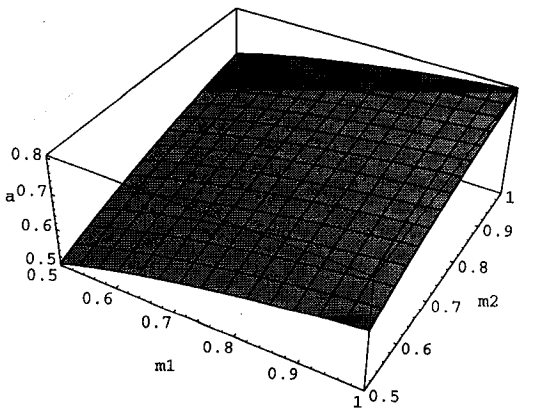

Figure 5: The naïve Bayes posterior as a function, restricted to confirming likelihoods.

Assuming that the likelihoods m entering these matrices multiply the second column in L_i by 0, and applying Bayes' rule to both likelihood matrices results in:

$$a_{nb} = \frac{l_1 l_2 h}{l_1 l_2 h + \overline{l_1}\,\overline{l_2}(1-h)}$$

Similarly to the treatment of the noisy-or function, we will concentrate on the interesting parameters l_1 and l_2, which play a role analogous to the r_i, and assume the others are held constant. In that case we can combine the rest of the parameters into a constant:

$$k = \frac{\overline{l_1}\,\overline{l_2}(1-h)}{h} << l_1, l_2 :$$

$$a_{nb}(l_1, l_2) = \frac{l_1 l_2}{l_1 l_2 + k}$$

As seen in the graph of this function, shown in Figure 4, the function has a definite curvature. It is zero whenever l_1 or $l_2 = 0$. This makes it function as a logical-AND when its arguments take on the extreme values of zero and one. By virtue of the curvature, partial support tends to result in greater belief than that for corresponding values of the arguments of the noisy-or.

3.3 Remedying the naïve Bayes model

Figure 4 suggests that if we restrict the domain of parameters to the upper quadrant of both functions then both functions are qualitatively the same. For purposes of comparison assume that the priors in all cases $= 1/2$. Then consider both functions restricted to the domain $[1/2, 1] \times [1/2, 1]$. The noisy-or function looks qualitatively the same as before. The naïve Bayes function has noticeably less curvature, as shown in Figure 5.

m_1	m_2	$m_1 \vee m_2$	naïve Bayes	noisy-or
$\perp$	$\perp$	$\perp$	1/2	7/16
$\top$	$\perp$	$\top$	2/3	5/8
$\perp$	$\top$	$\top$	2/3	5/8
$\top$	$\top$	$\top$	4/5	3/4

Table 1: Truth table for the probabilistic logical-OR. The qualitative relationship between the unknown ($\perp$) and confirming ($\top$) evidence for both models. By choice of the right threshold, both models express a probabilistic logical-OR.

The values of 1/2 indicate ignorance: the test parameters neither confirm or refute the threat to any degree. Ignorance is denoted by the symbol $\perp$. By choosing any threshold between 1/2 and 5/8 for the boundary between the values of confirmation $\top$ and ignorance $\perp$, both models calculate logical-ORs of the values. The resulting truth table is shown in Table 1, for both functions. The actual probabilities calculated by both functions are also shown.

As the title of the paper suggests, the equivalence is only approximate. By restriction of ranges of the evidence to discount disconfirming evidence we can set thresholds on output nodes so that strong evidence for any test will confirm the threat that it is associated with. This property of the naïve Bayes can be constructed by adjusting the values in the L_i matrices. Confirming evidence is enforced by keeping the ratio $l/\bar{l}$ large. Disconfirming evidence is discounted by keeping the ratio $(1-l)/(1-\bar{l})$ close to one. This

is possible if

$$\bar{l} << l << 1.$$

4 Discussion

This paper offers a practical but not entirely satisfactory solution to the problem it posed. The more fundamental problem is to better understand how causality applies to a set of specific conditions and a general condition. That should give a clue about how to construct a Bayes' network that captures best this causal relationship.

The kind of causal relationships considered in this paper are approximated by an *is-a* relationship between specific and general variables, e.g. between specific tests, and a general threat condition that they all test for. The concept of cause that applies is a more general one than the mechanistic concept of cause. It is more along the lines of Aristotle's use of the term, where cause is the basis for explanation.

In an *is-a* relationship, the probabilities used to in a Bayes' network do not express a stochastic relationship among events. The frequency interpretation of probability is of no use here. Rather probabilities capture the coarseness of the relationship between the concepts. The relationships in question are, for instance, if I call something a *Discovery attempt* what is the likelihood I would classify it also under the category of a *Breech of security*? Thus the model can focus on variables that are relevant, and avoid enumeration of variables containing irrelevant, exceptional conditions. The exceptional conditions become part of the "error" in the model.

There are probably better ways to express *is-a* relationships than either the noisy-or or the naïve Bayes models. This question has been addressed in a probabilistic context by Pearl, but just in the case of mutually exclusive conditions.[2]

5 Summary

We have shown that by restricting disconfirming evidence in a naïve Bayes model, the model can approximate a noisy-or for practical purposes. Furthermore this can be done by proper choice of probability table values that generate likelihood messages. The advantages of this are, first, that the Bayes' network keeps the causal representation with which it was constructed, so that multiple causes are not confused with multiple sources of support. Secondly, the connections in the network are fewer, easing the computational burdens. Unfortunately it is not evident whether this approximation can be made tighter. This approximation raises the larger question about how to correctly model *is-a* relationships.

[2] p. 335, J. Pearl, *Probabilistic Reasoning in Intelligent Systems,* (San Mateo, CA: Morgan Kaufmann, 1988)