RESEARCH NOTE

Using Crude Probability Estimates to Guide Diagnosis

Johan de Kleer
Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA

ABSTRACT

In order to identify the faulty components of a malfunctioning device in the fewest number of measurements, model-based diagnosis often uses a minimum entropy technique to select the next best measurement. This technique seems critically dependent on the availability of failure probabilities for components. Unfortunately, in many cases this information is unavailable or unknown. However, if we can assume that all components fail independently with equal probability and that components fail with very small probability, then it is possible to exploit the intuitions of the technique even when the exact probabilities are unknown. In addition, the computation required is much simpler. This approach can be generalized if the set of components can be partitioned such that each of the components of a partition fail with equal probability but are much more or less likely to fail than those of other partitions.

1. Introduction

Diagnosis is the process of identifying which components of a properly designed device are malfunctioning and thereby preventing the device from achieving its purpose. Model-based diagnosis seeks to develop a general theory of diagnosis based on first-principles reasoning about the behavior of the device's individual components. The model-based diagnosis process [5] consists of two distinct phases. The objective of the first phase is to identify sets of faulty components (called diagnoses) which explain the observed symptoms. The objective of the second phase is to gather additional evidence (e.g., by making measurements) which best differentiates among the diagnoses. This paper presents a simple process for identifying the best measurement to make next.

Early approaches to model-based diagnosis [2, 4, 6] assumed that the device had only one fault and incorporated ad hoc heuristics to identify the best measurement to make next. For example, a half-split method was often

Artificial Intelligence **45** (1990) 381–391

proposed. But these heuristics are neither generally applicable nor optimal. The general diagnostic engine (GDE) [8] makes no assumption about the number of faults in the device and uses a general minimum entropy technique [1] to select the best measurement to make next. Although general, this technique seems critically dependent on the availability of failure probabilities for components. In many cases this information may be only approximately known, unavailable, or unknown.

Nevertheless, it is still possible to exploit the intuitions of the minimum entropy technique even when the exact probabilities are unknown. In addition to the usual GDE assumptions (see Section 3) we assume that all components fail (a) with equal probability and (b) with extremely small probability. Given these two additional assumptions it is straightforward to construct a probing strategy rooted in the minimum entropy approach even when the exact failure probabilities are unavailable. The strategy has the additional advantage that it is far simpler computationally. Therefore, even in cases where these assumptions do not completely hold it may be expedient to use the approach.

As a consequence of these assumptions, we are only concerned with diagnoses of minimum cardinality. (A candidate is represented by the set of faulty components and a diagnosis is a candidate consistent with the evidence.) If the minimum cardinality of diagnoses is q, then we are only interested in candidates with exactly q faults. Every candidate with less than q faults has been eliminated and every diagnosis with more than q faults has negligible probability compared to any q-fault diagnosis. As the candidates are eliminated as diagnoses, the minimum cardinality of diagnoses may increase. Therefore, the probing strategy first proposes measurements which differentiate among single-fault diagnoses, then when all single-fault diagnoses are eliminated, it proposes probes which differentiate among double-fault diagnoses, etc.

The proposed approach focuses on the same class of diagnoses that the minimum cardinality principle of generalized set covering (GSC) does [14, 15, 18]. Reiter [19] argues that the minimum cardinality principle misses intuitively plausible diagnoses. Nevertheless, given our assumptions (which include that components fail with very small and independent probabilities) the minimal cardinality principle is perfectly valid (as argued in [14, 15, 18]).

The proposal of this paper lies conceptually between the simple probing strategies of the earlier model-based reasoning work and GDE and could be developed from first-principles considerations alone. Nevertheless, this paper takes the minimum entropy approach of GDE as given, and shows how it reduces to a very simple algorithm given the preceding two assumptions. However, it is unnecessary to presume the full ATMS-based [7] framework of GDE. In particular, any approach which tests candidates for consistency with evidence and which predicts measurement outcomes is adequate (e.g., [19]).

In many cases the minimum entropy technique reduces to the following simple case. Suppose that all candidates having less than q faults have been shown to be inconsistent with the evidence. Consider measuring some variable x which has values v_k. The score of measuring x is:

$$\$(x) = \sum c_k \ln c_k , \tag{1}$$

where c_k is the number of diagnoses of size q which predict $x = v_k$. The best measurement has the lowest score. Typically most of the candidates of size q will have been eliminated by this point so there are relatively few diagnoses to consider. The function is very simple to compute and produces the intuitive result. The following two examples illustrate this simple case.

2. Examples

Consider the sequence of four buffers shown in Fig. 1 where the input is one and the output zero. There are four single-fault ($q = 1$) diagnoses: $[A]$, $[B]$, $[C]$ and $[D]$. (Diagnosis $[A]$ indicates that A is faulted and B, C and D are normal.) The possible outcomes divide the diagnoses as follows (the second line is read: Diagnoses $[B]$, $[C]$ and $[D]$ predict that $X = 1$). Many approaches to model-based diagnosis can compute these sets—the exact method used is not relevant here.

$$\begin{aligned}
X &= 0, \quad S = \{[A]\}, \\
X &= 1, \quad S = \{[B][C][D]\}, \\
Y &= 0, \quad S = \{[A][B]\}, \\
Y &= 1, \quad S = \{[C][D]\}, \\
Z &= 0, \quad S = \{[A][B][C]\}, \\
Z &= 1, \quad S = \{[D]\}.
\end{aligned}$$

Therefore, the scores for the possible probe points are (as in [8] we use natural logarithms):

$$\$(X) = 3 \ln 3 + 1 \ln 1 = 3.3,$$

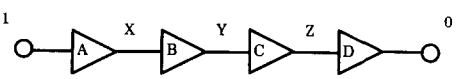

Fig. 1. A, B, C and D are buffers, the input is one and the symptomatic output is zero. X, Y and Z are possible points to measure next.

$$\$(Y) = 2 \ln 2 + 2 \ln 2 = 2.8,$$
$$\$(Z) = 1 \ln 1 + 3 \ln 3 = 3.3.$$

Therefore, measuring Y is best. A simple half-split heuristic identifies the same measurement point. However, the minimum entropy approach applies in the general case where the topology is arbitrary and not just a linear chain of components.

Consider the simple example of Fig. 2. Given the inputs, $A = 3$, $B = 2$, $C = 2$, $D = 3$, $E = 3$ and outputs $F = 10$ and $G = 12$, there are two single-fault diagnoses: $[M_1]$ and $[A_1]$. The possible outcomes divide the diagnoses as follows:

$$X = 4, \quad S = \{[M_1]\}$$
$$X = 6, \quad S = \{[A_1]\}$$
$$Y = 6, \quad S = \{[A_1][M_1]\}$$
$$Z = 6, \quad S = \{[M_1][A_1]\}$$

Therefore, the scores for the possible probe points are:

$$\$(X) = 1 \ln 1 + 1 \ln 1 = 0$$
$$\$(Y) = 2 \ln 2 = 1.4,$$
$$\$(Z) = 2 \ln 2 = 1.4.$$

Therefore measuring X is best. The zero score indicates that, no matter what, at least one of the two diagnoses is eliminated.

It would also have been easy to invent a simple rule of thumb for this example which circumvents any need for numerical scoring. However, when

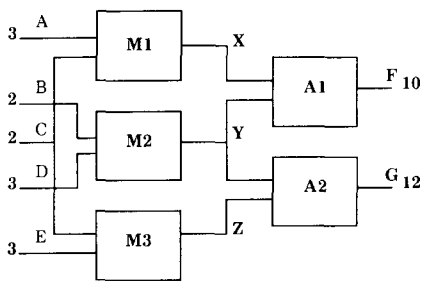

Fig. 2. A, B, C, D and E are input terminals, F and G are output terminals, X, Y and Z are internal probe points, M_1, M_2 and M_3 are multipliers and A_1 and A_2 are adders.

the number of diagnoses and possible outcomes grows (and some diagnoses fail to predict some measurements), it becomes harder and harder to invent simple rules of thumb to determine the best measurement to make next. The simple scoring function we develop is easy to evaluate and provides the optimal measurement(s) to make next (given our presuppositions).

3. Probability of a Diagnosis

As components fail independently,[1] the prior probability that a particular diagnosis C_l is the correct one is given by:

$$p_l = \prod_{c \in C_l} p(c \in C_a) \prod_{c \notin C_l} [1 - p(c \in C_a)] . \tag{2}$$

We assume that all components fail with equal (small) probability:

$$p(c \in C_a) = \varepsilon .$$

Therefore,

$$p_l = \varepsilon^{|C_l|}(1 - \varepsilon)^{n - |C_l|} , \tag{3}$$

where n is the number of components in the device and $|C_l|$ is the number of faulted components in C_l. If ε is small enough, then p_l can be approximated:

$$p_l = \varepsilon^{|C_l|} . \tag{4}$$

As candidates are eliminated as diagnoses, the probabilities of the other diagnoses increase. Bayes' rule allows us to calculate the conditional probability of the diagnoses given that point x_i is measured v_{ik}:

$$p(C_l | x_i = v_{ik}) = \frac{p(x_i = v_{ik} | C_l) p(C_l)}{p(x_i = v_{ik})} . \tag{5}$$

The denominator, $p(x_i = v_{ik})$, is just a normalization. $p(C_l)$ is computed in the preceding step (or is the prior). Finally, $p(x_i = v_{ik} | C_l)$ is determined as follows:

(1) If $x_i = v_{ik}$ is predicted by C_l given the preceding evidence, then $p(x_i = v_{ik} | C_l) = 1$.
(2) If $x_i = v_{ik}$ is inconsistent with C_l and the preceding evidence, then $p(x_i = v_{ik} | C_l) = 0$.

[1] This independence assumption which is widely made in model-based diagnosis is somewhat suspect. In some domains dependent failures are quite common.

(3) If $x_i = v_{ik}$ is neither predicted by nor inconsistent with C_l and the preceding evidence, then we make the presupposition[2] that every possible value for x_i is equally likely. Hence, $p(x_i = v_{ik}|C_l) = 1/m$, where m is the number of possible values x_i might have (in a conventional digital circuit $m = 2$). Intuitively, this provides a bias for diagnoses which predict a measurement over those that do not. This strong bias towards diagnoses that predict the symptoms is advocated by the abductive approach [16] to diagnosis but arises here as a direct consequence of applying Bayes' rule.

If m is constant[3] for all probe points then the probability of a relevant diagnosis (i.e., one not yet eliminated) after multiple probes (E) is:

$$p(C_l|E) = \frac{\varepsilon^q}{Nm^{f_l}}, \tag{6}$$

where f_l is the number of times diagnosis C_l failed to predict a measurement outcome in the sequential diagnosis and N is the normalization. Therefore, the current probability of any diagnosis can be characterized by two integers: (1) the number of faults it hypothesizes, and (2) the number of times it failed to predict a measurement outcome.

If $\varepsilon \ll 1/m^{f_l}$ for the minimal cardinality diagnoses, then we can discard all but the minimal cardinality diagnoses. In terms of Bayes' rule, this condition ensures that $p(x_i = v_{ik}|C_l)$ always has a significant value. In practice, this condition is more than met because $\varepsilon \ll 1$ and f_l is rarely nonzero for the minimal cardinality diagnoses. If for some reason the condition was not met, then some minimal cardinality diagnoses might have lower posterior probability than some higher cardinality diagnoses.

4. The General Case

This section shows how the reduced algorithm presented in the introduction is a correct instantiation of the minimum entropy technique given our assumptions. In addition, it also shows the correction to make if some diagnosis fails to predict a measurement outcome (although this occurs relatively rarely, the algorithm must account for it).

[2] This presumes that (1) the prediction engine is logically complete, (2) when components fail, every possible output is equally likely, and (3) for each instance we have to estimate $p(x_i = v_{ik}|C_l)$ by $1/m$ that the value of x_i is conditionally independent of all the other x_i given C_l. All three are invalid in general, however, without additional information (e.g., such as fault modes [9]), it is hard to come up with a better heuristic.

[3] This is easily generalized to the case where m_i varies with x_i. However, this formulation presumes that x_i is a discrete variable or a discretization of some other variable (in the case of analog diagnosis).

According to the minimum entropy technique the best probe x_i is the one which minimizes the expected posterior entropy of the diagnoses after measuring x_i. We need to minimize (over all probe points x_i) [8]:

$$\sum_{k=1}^{m} \left[p(S_{ik}) + \frac{p(U_i)}{m} \right] \ln \left[p(S_{ik}) + \frac{p(U_i)}{m} \right] - p(U_i) \ln \frac{1}{m} , \tag{7}$$

where S_{ik} is the set of diagnoses which predict that probing point x_i will obtain value v_{ik}, and U_i is the set of diagnoses which predict no value for x_i. This expression is easily evaluated using the two integer scores for each diagnosis (from equation (6)).

If we only need consider the minimum cardinality diagnoses of size q, then (7) is easily evaluated. As the probabilities of all diagnoses must sum to 1, from (6) we obtain:

$$N = \sum_{C_j} \frac{\varepsilon^q}{m^{f_j}} = \varepsilon^q \sum_{C_j} \frac{1}{m^{f_j}} , \tag{8}$$

where the C_j are the minimum cardinality diagnoses. Hence, equation (6) is independent of ε:

$$p(C_l|E) = \frac{1}{m^{f_l} \sum_{C_j} \frac{1}{m^{f_j}}} . \tag{9}$$

Hence, equation (7) is independent of ε and a simple function of the f_j.

In the common case (discussed in the introduction), where diagnoses always predict outcomes, $p(U_i) = 0$ and $f_j = 0$ always. In this case, (7) immediately simplifies to:

$$\sum_{k=1}^{m} \frac{c_{ik}}{N'} \ln \left[\frac{c_{ik}}{N'} \right] , \tag{10}$$

where c_{ik} is the number of diagnoses (of size q) predicting that $x_i = v_{ik}$ and N' is the number of diagnoses (of size q). As N' is constant over all probe points we need only minimize,

$$\sum_{k=1}^{m} c_{ik} \ln c_{ik} . \tag{11}$$

This is intuitively appealing because this is essentially the entropy of the counts of the number of diagnoses supporting each outcome.

5. Problems with This Strategy

There are a number of reasons why the minimum cardinality of the diagnoses may increase during the diagnostic session. The number of faults in the device may be initially unknown, and as evidence accumulates all diagnoses of size q may be eliminated, and the diagnostic process must then consider the $q+1$ faults. Even when a particular q-fault diagnosis has been identified, we may want to increase our confidence that it is the correct diagnosis by gathering evidence to eliminate as many $(q+1)$-fault diagnoses as possible. Unfortunately, the probes which best differentiate among the q-faults diagnoses are not necessarily those which best differentiate among both the q- and $(q+1)$-fault diagnoses. As a consequence, we make more measurements than needed. In essence our strategy trades off optimal measurements in some cases for computational simplicity in general.

Any component which does not appear in any q-fault diagnosis is effectively considered normal. As a consequence there may be a very large number of potential probes all with the same score. For example, measuring on either side of a normal component has equal score. However, the different measurement points might discriminate among the as yet unconsidered higher cardinality diagnoses. This issue is simply illustrated in Fig. 3.

Suppose the input is zero and the output one. Thus, there are only two $(q=1)$ diagnoses: $[A]$ and $[B]$. The possible outcomes divide the diagnoses as follows:

$$
\begin{aligned}
&T=0, && S=\{[B]\},\\
&T=1, && S=\{[A]\},\\
&OUT_2=0, && S=\{[B]\},\\
&OUT_2=1, && S=\{[A]\}.
\end{aligned}
$$

Thus T and OUT_2 have equal scores: $1 \ln 1 + 1 \ln 1 = 0$. But this is clearly suboptimal as can be seen by including the double faults:

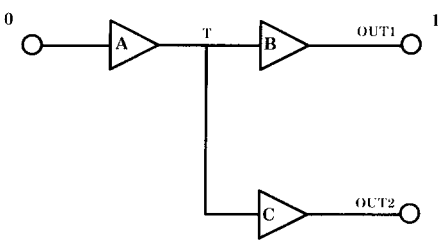

Fig. 3. A, B and C are buffers, the input is zero, and output OUT_1 is one. Does measuring T or OUT_2 provide the most information?

$$T = 0, \qquad S = \{[B][B, C]\},$$
$$T = 1, \qquad S = \{[A][A, C]\},$$
$$OUT_2 = 0, \qquad S = \{[B]\},$$
$$OUT_2 = 1, \qquad S = \{[A]\}.$$

(Note that most double faults do not predict measurement outcomes and are not included. For example, $[A, B]$ does not predict a value for T or OUT_2.) The table shows that measuring T will eliminate some double faults from consideration, while measuring OUT_2 will not. So T is the better probe point because it provides more information.

A possible way to avoid this difficulty is that if the strategy scores a set of probe points equally, to consider higher cardinality $(q + 1)$ faults in the scoring using some value for ε. Notice that in this example, any $\varepsilon > 0$ indicates T is a better measurement.

The example circuit of Fig. 3 is artificially simple to illustrate the point. Intuitively, suppose that there were an arbitrary sequence of additional buffers attached to OUT_2, then measuring any of their outputs would have equal score considering q faults. No matter how many buffers were in the sequence, T would always be the better measurement. More realistic examples are outside the scope of this paper.

The same problem arises in many guises. Suppose there is only one diagnosis. Although this is an indication that the diagnostic task is complete, it does not allow further discrimination: If there is only one diagnosis, then all measurement points have the same zero score. Thus the approach indicates no useful points to measure when presented with a failing device with no symptoms. GDE, which is not restricted to minimal cardinality diagnoses, can choose a good measurement even in this case.

The diagnosis process may reduce the possible diagnoses to an indistinguishable set. Additional measurements motivated by higher cardinality diagnoses can help shift the probabilities of individual component failure. By considering higher cardinality diagnoses, GDE handles this case as well.

Any probing strategy which focuses on the more probable diagnoses at the expense of the less probable diagnoses (another example is [9]) suffers from analogous difficulties. This is the price paid for the computational advantages provided by focusing.

6. A Generalization

The approach outlined in this paper can be generalized in the case where certain groups of components are far more likely to fail than others. For example, transistors may be far more likely to fail than resistors, and resistors far more likely to fail than wires. The set of components can be partitioned

into D_i where the failure probability of each component in D_i is ε_i. We assume that for $j > i$:

$$\frac{\varepsilon_i^{|D_i|}}{m^f} \gg \varepsilon_j \,, \tag{12}$$

where f is the maximum number of measurements that will be made. This condition ensures that any failures within components of partition D_i will always dominate failures in later partitions. This equation is obtained from (6) by assuming the lowest probability diagnosis involving components from partition D_i. The worst case occurs if all the components are faulted and none of the measurements are predicted. (Although many groups of components may not have different enough failure probabilities to satisfy this condition well, in many cases it is a useful technique for limiting the computation of the diagnostician.)

Instead of associating a single integer cardinality with each candidate, we associate a tuple $[t_1, \ldots, t_p]$ where t_k is the number of faulted components it selects from D_k. Thus, the prior probability of a diagnosis is:

$$p_i = \varepsilon_1^{t_1} \cdots \varepsilon_k^{t_k} \,. \tag{13}$$

We say that tuple $[a_1, \ldots, a_k]$ is greater than $[b_1, \ldots, b_k]$ if there is an integer c such that $a_c > b_c$ and for all $j < c$, $a_j = b_j$. If ε_i is sufficiently larger than $\varepsilon_{j>i}$, then at any point we need only consider diagnoses with the greatest tuple—all others can be ignored because their probabilities are vanishingly (relatively) small. Therefore, the only diagnoses of interest have the identical prior probabilities and all the equations developed in the paper go through directly. For example, in the simple case the best measurement to make next is the one which minimizes,

$$\sum c_{ik} \ln c_{ik} \,, \tag{14}$$

where c_{ik} is the number of diagnoses with the largest tuple which predict that x_i will be measured to be v_{ik}.

ACKNOWLEDGEMENT

John Lamping pointed out that minimum entropy calculation was, in fact, extremely simple to carry out and provided many useful insights. Daniel G. Bobrow, Bernardo Huberman, Olivier Raiman, Mark Stefik and Peter Struss provided useful comments on early drafts.

REFERENCES

1. M. Ben-Bassat, Myopic policies in sequential classification, *IEEE Trans. Comput.* **27** (1978) 170–178.

2. J.S. Brown, R.R. Burton and J. de Kleer, Pedagogical, natural language and knowledge engineering techniques in SOPHIE I, II and III, in: D. Sleeman and J.S. Brown, eds., *Intelligent Tutoring Systems* (Academic Press, New York, 1982) 227–282.
3. R. Davis, H. Shrobe, W. Hamscher, K. Wieckert, M. Shirley and S. Polit, Diagnosis based on description of structure and function, in: *Proceedings AAAI-82*, Pittsburgh, PA (1982) 137–142.
4. R. Davis, Diagnostic reasoning based on structure and behavior, *Artificial Intelligence* **24** (1984) 347–410.
5. R. Davis and W. Hamscher, Model-based reasoning: Troubleshooting, in: H.E. Shrobe and the American Association for Artificial Intelligence, eds., *Exploring Artificial Intelligence* (Morgan Kaufmann, San Mateo, CA, 1988) 297–346.
6. J. de Kleer, Local methods of localizing faults in electronic circuits, AIM-394, AI Lab., MIT, Cambridge, MA (1976).
7. J. de Kleer, An assumption-based TMS, *Artificial Intelligence* **28** (1986) 127–162; also in: M.L. Ginsberg, ed., *Readings in Nonmonotonic Reasoning* (Morgan Kaufmann, San Mateo, CA, 1987) 280–297.
8. J. de Kleer and B.C. Williams, Diagnosing multiple faults, *Artificial Intelligence* **32** (1987) 97–130; also in: M.L. Ginsberg, ed., *Readings in Nonmonotonic Reasoning* (Morgan Kaufmann, San Mateo, CA, 1987) 372–388.
9. J. de Kleer and B.C. Williams, Diagnosis with behavioral modes, in: *Proceedings IJCAI-89*, Detroit, MI (1989) 104–109.
10. M.R. Genesereth, The use of design descriptions in automated diagnosis, *Artificial Intelligence* **24** (1984) 411–436.
11. G.A. Gorry and G.O. Barnett, Experience with a model of sequential diagnosis, *Comput. and Biomed. Res.* **1** (1968) 490–507.
12. W.C. Hamscher, Model-based troubleshooting of digital systems, Artificial Intelligence Laboratory, TR-1074, MIT, Cambridge, MA (1988).
13. L.J. Holtzblatt, Diagnosing multiple failures using knowledge of component states, in: *IEEE Proceedings on AI Applications* (1988) 139–143.
14. Y.P. Peng and J.A. Reggia, Plausibility of diagnostic hypotheses: The nature of simplicity, in: *Proceedings AAAI-86*, Philadelphia, PA (1986) 140–145.
15. Y.P. Peng and J.A. Reggia, A probabilistic causal model for diagnostic problem-solving, *IEEE Trans. Syst. Man Cybern.* **17** (1987) 146–162, 395–406.
16. D. Poole, Representing knowledge for logic-based diagnosis, in: *Proceedings International Conference on Fifth Generation Computer Systems* (1988) 1282–1290.
17. O. Raiman, Diagnosis as a trial: The alibi principle, IBM Scientific Center (1989).
18. J.A. Reggia, D.S. Nau and P.Y. Wang, A formal model of diagnostic inference, *Inf. Sci.* **37** (1983) 227–285.
19. R. Reiter, A theory of diagnosis from first principles, *Artificial Intelligence* **32** (1987) 57–95; also in: M.L. Ginsberg, ed., *Readings in Nonmonotonic Reasoning* (Morgan Kaufmann, San Mateo, CA, 1987) 352–371.
20. P. Struss, Extensions to ATMS-based diagnosis, in: J.S. Gero, ed., *Artificial Intelligence in Engineering: Diagnosis and Learning* (Elsevier, Amsterdam, 1988) 3–28.
21. P. Struss and O. Dressler, "Physical negation": Integrating fault models into the general diagnostic engine, in: *Proceedings IJCAI-89*, Detroit, MI (1989) 1318–1323.

Received September 1989; revised version received May 1990