Representing diagnostic knowledge for probabilistic Horn abduction

David Poole
Department of Computer Science,
University of British Columbia,
Vancouver, B.C., Canada V6T 1Z2
poole@cs.ubc.ca

Abstract

This paper presents a simple logical framework for abduction, with probabilities associated with hypotheses. The language is an extension to pure Prolog, and it has straight-forward implementations using branch and bound search with either logic-programming technology or ATMS technology. The main focus of this paper is arguing for a form of representational adequacy of this very simple system for diagnostic reasoning. It is shown how it can represent model-based knowledge, with and without faults, and with and without non-intermittency assumptions. It is also shown how this representation can represent any probabilistic knowledge representable in a Bayesian belief network.

1 Introduction

Determining what is in a system from observations (diagnosis and recognition) are an important part of AI. There have been many logic-based proposals of what a diagnosis is [Reiter, 1987; de Kleer and Williams, 1987; Poole, 1989; de Kleer *et al.*, 1990]. One problem with these proposals is that for any problem of a reasonable size there are far too many "logical possibilities" to handle (for a human or a computer). For example, when considering fault models [de Kleer and Williams, 1989; Poole, 1989], there is almost always an exponential number of logical possibilities (e.g., each component could be in its normal state or in the abnormal unknown state). For practical problems, we find that many of the logically possible diagnoses are so unlikely that it is not worth considering them. There is a problem, however, in removing the unlikely possibilities a priori: it may happen that the unlikely occurrence is the actual truth in the world.

Such analysis of the combinatorial explosions would tend to suggest that we need to take into account probabilities of the diagnoses [de Kleer and Williams, 1987; Peng and Reggia, 1990; Neufeld and Poole, 1987], and not generate the unlikely diagnoses. Similar experience has been found in natural language understanding [Hobbs *et al.*, 1988; Goldman and Charniak, 1988].

Probabilistic models of diagnostic reasoning [Pearl, 1988; Heckerman and Horvitz, 1990; Andreassen *et al.*, 1987], being purely propositional by nature, do not have the modelling power of the logic-based models. This paper points to one direction in which probabilistic diagnostic frameworks can be extended to a non-propositional form.

This paper presents a very simple form of abduction, where the background knowledge is Horn, and the assumptions are atomic. Associated with hypotheses are probabilities. The main features of the approach are:

- We are trying to carry out a empirical study of automated reasoning. In order to carry this out we try to determine where very simple frameworks work and fail. The best way to show that we need certain features is to try to do without them. It is in this spirit that we try to use the simplest framework that seems plausible, and only add features when they can be demonstrated to be needed.
- We are trying to get a good compromise between representational (epistemic) adequacy and procedural (heuristic) adequacy [McCarthy and Hayes, 1969].
- As a prima facie case for representational adequacy, we note that the language incorporates pure Prolog as a special case, and also an ATMS[1] [Reiter and de Kleer, 1987], and the language can represent any probabilistic information that can be represented in a Bayes net [Pearl, 1988] (see section 4). We also demonstrate representational adequacy by showing how some common diagnostic representational problems can be represented in this framework. The representational adequacy can only be verified empirically, and we are currently trying to test the framework on a variety of problems.
- It is straight forward to implemented using either logic programming [Poole, 1991] or ATMS [de Kleer,

[1] Note that we are using the assumption based framework as the object language and not as a book keeping mechanism for a problem solver.

1986] technology. In this paper we use a specification of what is to be implemented that is independent of the actual implementation strategy used. Once we have the specification of what it is we want to compute, we can then compare different implementation strategies to determine which is more efficient in space and/or time.

In all of the implementations, we do not generate the unlikely explanations unless we need to. Hopefully we can cut down on the combinatorial explosions that are inherent in considering the set of all logically possible explanations, but this is beyond the scope of this paper.

2 The System

2.1 Abductive Framework

The formulation of abduction used is in terms of Theorist [Poole *et al.*, 1987; Poole, 1988a].

Given a language L, and a consequence relation (written $\models$) on L, and an abductive scheme is a pair $\langle F, H\rangle$ where F and H are sets of sentences in L.

Definition 2.1 [Poole *et al.*, 1987; Poole, 1988a] If g is a ground formula, an **explanation** of g from $\langle F, H\rangle$ is a set $D \subseteq H$ such that

- $F \cup D \models obs$ and
- $F \cup D \not\models \bot$

where $\bot$ is an atom representing *false*. The first condition says that, D is a sufficient cause for *obs*, and the second says that D is possible (i.e., $F \cup D$ is consistent).

Definition 2.2 A **minimal explanation** of g is an explanation of g such that no strict subset is also an explanation.

Associated with each minimal explanation D, is a measure $\mu(D)$ [Neufeld and Poole, 1987]. This measure could be an assumption cost that is added [Hobbs *et al.*, 1988], but in this paper we investigate the use of probability as a measure over explanations.

2.2 Probabilistic Horn abduction

In probabilistic Horn abduction we restrict the language L to be $Horn$ clauses.

We use the normal Prolog definition of an atomic symbol [Lloyd, 1987]. A Horn clause is of the form:

$$\begin{array}{rl} a. & \\ a & \leftarrow a_1 \wedge ... \wedge a_n. \\ \mathit{false} & \leftarrow a_1 \wedge ... \wedge a_n. \end{array}$$

where a and each a_i are atomic symbols. *false* is a special atomic symbol that is not true in any interpretation[2]. All variables in F are assumed universally quantified.

We restrict the elements of H to be ground instances of atoms. If we are given a set of open atoms as possible hypotheses we mean the the set of ground instances of these atoms.

2.3 Probabilities

The measure we use is the probability of the explanation.

Associated with each possible hypothesis (i.e., with each ground instance of an open possible hypothesis) is a prior probability. The aim is to compute the posterior probability of the explanations given the observations. Abduction gives us what we want to compute the probability of and probability theory gives a measure over the explanations [Neufeld and Poole, 1987].

We use the declaration

$$\mathit{assumable}(h, p).$$

where h can contain free variables, to mean each ground instance of h is in H with prior probability p.

To compute the posterior probability of an explanation $H = \{h_1, ..., h_n\}$ given observation *obs*, we use Bayes rule and the fact that $P(obs|H) = 1$ as the explanation logically implies the observation:

$$\begin{array}{rcl} P(H|obs) & = & \dfrac{P(obs|H) \times P(H)}{P(obs)} \\ & = & \dfrac{P(H)}{P(obs)} \end{array}$$

The value, $P(obs)$ is the prior probability of the observation, and is a constant factor for all explanations. We compute the prior probability of the conjunction of the hypotheses using:

$$\begin{array}{rcl} P(h_1 \wedge ... \wedge h_{n-1} \wedge h_n) & = & P(h_n|h_1 \wedge ... \wedge h_{n-1}) \\ & & \times P(h_1 \wedge ... \wedge h_{n-1}) \end{array}$$

The value of $P(h_1 \wedge ... \wedge h_{n-1})$ forms a recursive call, with $P(true) = 1$. The only other thing that we need to compute is

$$P(h_n|h_1 \wedge ... \wedge h_{n-1})$$

The first thing to notice is that if h_n is inconsistent with the other hypotheses, then its probability is zero. These are exactly the cases that are removed by the inconsistency check. Similarly if h_n is implied be the other hypotheses, its probability is one. This will never be the case if the explanations are minimal. While any method can be used to compute this conditional probability, the

[2] Notice that we are using Horn clauses differently from how Prolog uses Horn clauses. In Prolog, the database consists of definite clauses, and the queries provide the negative clauses [Lloyd, 1987]. Here the database consists of definite and negative clauses, and we build a constructive proof of an observation.

assumption of conditional independence is often an appropriate assumption in many domains [de Kleer and Williams, 1987; Peng and Reggia, 1990]. We make this assumption here and in later sections we show how to allow arbitrary probabilistic interactions, without changing the underlying system. The system uses the following assumption:

Assumption 2.3 Logically independent instances of hypotheses are probabilistically independent.

Definition 2.4 A set H of hypotheses are **logically independent** (given F) if there is no $S \subset H$ and $h \in H \backslash S$ such that

$$F \cup S \models h \quad \text{or} \quad F \cup S \models \neg h$$

The assumptions in a minimal explanation are always logically independent. Minimality ensures that no hypothesis in an explanation can be implied by other hypotheses in the explanation. Consistency ensures the negation of a hypothesis cannot be implied by other hypotheses.

Under assumption 2.3, if $\{h_1, ..., h_n\}$ are part of a minimal explanation, then

$$P(h_n | h_1 \wedge ... \wedge h_{n-1}) = P(h_n)$$

thus

$$P(h_1 \wedge ... \wedge h_n) \quad = \quad \prod_{1=1}^{n} P(h_i)$$

To compute the prior of the explanation we multiply the priors of the hypotheses. The posterior probability of the explanation is proportional to this.

One problem that arises is in determining the value of $P(obs)$.

When using abduction we often assume that the diagnoses are covering. This can be a valid assumption if we have anticipated all eventualities, and the observations are within the domain of the expected observations (usually if this assumption is violated there are no explanations). This is also supported by recent attempts at a completion semantics for abduction [Poole, 1988b; Console *et al.*, 1989; Konolige, 1990]. The results show how abduction can be considered as deduction on the "closure" of the knowledge base that includes statements that the given causes are the only causes. The closure implies the observation are logically equivalent to the disjunct of its explanations. We make this assumption explicit here:

Assumption 2.5 *The diagnoses are covering.*

For the probabilistic calculation we make an additional assumption:

Assumption 2.6 *The diagnoses are disjoint (mutually exclusive).*

It turns out to be straightforward to ensure that these properties hold, for observations that we can anticipate[3]. We make sure that the rules for each possible subgoal are disjoint and covering (see section 3.1).

Under these assumptions, if $\{e_1, ..., e_n\}$ is the set of all explanations of obs:

$$\begin{aligned} P(obs) &= P(e_1 \vee e_2 \vee ... \vee e_n) \\ &= P(e_1) + P(e_2) + ... + P(e_n) \end{aligned}$$

2.4 Implementation

The very simple definition of the framework makes implementation straight forward (although some difficult problems do arise in trying to make it very efficient). We are currently experimenting with implementations based on Logic programming technology and based on ATMS technology (similar to [de Kleer and Williams, 1989]). Both implementations keep a priority queue of sets of hypotheses that could be extended into explanations ("partial explanations"). At any time the set of all the explanations is the set of already generated explanations, plus those explanations that can be generated from the partial explanations in the priority queue. It is possible to put a bound on the probability mass in the queue, and this allows us to estimate errors on the results before the computation is completed (forming an "anytime" algorithm). See [Poole, 1992] for details.

The difference between these two represents a difference between "interpreted" and "compiled" approaches [Reiter and de Kleer, 1987]. As far as the rest of the paper is concerned, it is irrelevant as to how the system is implemented. Given a specification of what it is we want to compute we can now experiment with trade-offs between various implementation strategies.

Note that the problem is NP-complete [Provan, 1988], thus we are never going to expect efficient polynomial worst-case algorithms. The best we can expect is good average-case behaviour; but this is, of course, what we are interested in.

3 Representational Methodology

Once we have a tool, it is important to know how to use it. The problem of a representational methodology [Poole, 1990] is an important and much overlooked part of automated reasoning research.

It may seem that the assumptions used in designing the system were so restrictive that the system would be useless for real problems. In this section, I argue that this is not the case.

3.1 Disjoint and Covering Explanations

For our probabilistic analysis (section 2.3), we assumed that the explanations were disjoint and covering. If we

[3] Like other systems (e.g., [Pearl, 1988]), we have to assume that unanticipated observations are irrelevant.

want our probabilities to be correct[4], we must ensure that the explanations are disjoint and covering.

If the rules for an atom a are not covering, we can invent another cause for the goal representing "all the other possible causes" of the atom [de Kleer and Williams, 1989; Poole, 1989], and add

$$a \leftarrow \text{some\_other\_reason\_for\_a}.$$
$$assumable(\text{some\_other\_reason\_for\_a}, p).$$

Where p is the prior probability that something else would have caused a.

We can locally ensure that any explanations generated are disjoint. The following proposition can be easily proved:

Proposition 3.1 If for any two rules with the same consequent ($a \leftarrow b_1$, and $a \leftarrow b_2$), the antecedents are inconsistent ($F \models b_1 \wedge b_2 \Rightarrow false$), then the minimal explanation are disjoint.

Although disjointedness of explanations places a restriction on the knowledge base, it does not place a restriction on the sorts of knowledge that we can represent. In general, if we have rules

$$\begin{array}{rcl} a & \leftarrow & b_1. \\ & \vdots & \\ a & \leftarrow & b_n. \end{array}$$

these can be made disjoint by adding hypotheses $h_1, ..., h_n$ to the rules

$$\begin{array}{rcl} a & \leftarrow & h_1 \wedge b_1. \\ & \vdots & \\ a & \leftarrow & h_n \wedge b_n. \end{array}$$

and making sure these rules are disjoint by having, for each different i and j, the fact

$$false \quad \leftarrow \quad h_i \wedge h_j.$$

We need to associate a probability with each hypothesis such that $\sum_i P(h_i) = 1$. This probability represents the probability that the particular body was "the cause" for a.

Sometimes we can make the rules naturally disjoint, by ordering the rules and making sure that the bodies of rules are false if the bodies of previous rules are true.

Example 3.2 Suppose we want to represent an "and-gate" that should have value 0 if either of the inputs are zero. Suppose we represent the proposition that port G has output V at time T as $val(G, V, T)$. We can ensure that the explanations are disjoint locally by ensuring that only one body can ever be true:

$$\begin{array}{rcl} val(out(G), 0, T) & \leftarrow & and\_gate(G) \wedge ok(G) \\ & & \wedge val(input(1, G), 0, T). \\ val(out(G), 0, T) & \leftarrow & and\_gate(G) \wedge ok(G) \\ & & \wedge val(input(1, G), 1, T) \\ & & \wedge val(input(2, G), 0, T). \\ val(out(G), 1, T) & \leftarrow & and\_gate(G) \wedge ok(G) \\ & & \wedge val(input(1, G), 1, T) \\ & & \wedge val(input(2, G), 1, T). \end{array}$$

This has repercussions in biasing the most likely explanation to the first rule which is more general than the others. To make it more fair the first rule could be split into two cases depending on the value of input 2. This problem of the most likely diagnosis depending on the representation seems endemic to approaches that try to find the diagnosis (either explanation or interpretation) that is "most likely" [Pearl, 1988; Poole and Provan, 1990].

3.2 Parametrizing Hypotheses

The next important part of the methodology for abduction concerns parametrizing possible hypotheses and the interaction with the independence assumption. I have argued elsewhere [Poole, 1989; Poole, 1990] that there is much power obtainable and subtlety involved in parametrizing hypotheses appropriately.In this section we expand on previous analysis [Poole, 1990], and show how probabilities affect parametrization considerations by considering some case studies on different proposals.

3.2.1 Hypotheses with indeterminate output

As an example, suppose we have a gate G that takes two values as input, and outputs a value that can be in the range 1 to n. Suppose we want to represent the gate being in an unknown state (this is applicable whether or not we have fault models [de Kleer and Williams, 1989; Poole, 1989]). Suppose we represent the proposition that gate G has output V at time T as $val(G, V, T)$.

We cannot representing the hypothesis that the gate is in the unknown state by using the hypothesis $u(G)$ and the fact

$$val(out(G), V, T) \leftarrow u(G).$$

The problem is that the above fact states that a gate in the unknown state produces *all* values of output, rather than saying that it produces some output. Knowing a gate is in an unknown state does not imply any value for the output.

When there are no probabilities involved [Poole, 1990; Poole, 1989] we parametrize the hypothesis by the values

[4] It may be the case that they are "good enough" for any decisions that we may want to make, even though they are not accurate.

on which it depends. This could be done by having the hypothesis $produces(G, V, T)$ and the rule

$$val(out(G), V, T) \leftarrow produces(G, V, T).$$

We would say that a port has only one value at a time by having the constraint

$$false \leftarrow val(P, V_1, T) \wedge val(P, V_2, T) \wedge V_1 \neq V_2$$

Suppose we know that gate g_1 has probability ϵ of being in the unknown state. If we assume that each possible output value has equal chance, and that there are n possible output values, then the prior probability that it produces output value V is ϵ/n.

$$assumable(u(g_1, V, T), \frac{\epsilon}{n})$$

When we have more than one observation, there is another problem. For the probabilities we assumed that the hypotheses were independent. We would not expect that

$$P(u(g_1, 1, t_2)|u(g_1, 1, t_1)) = P(u(g_1, 1, t_2))$$

Once we know that the gate is in an unknown state at time t_1 it should not be so unlikely that it is in an unknown state at time t_2. Put another way, once we have paid the price once for assuming that the gate is in an unknown state at time t_1 we should not pay the price again for assuming that it is in an unknown state at time t_2.

To work in general, we need a mixture of the above two ideas. Suppose a gate G has probability of ϵ of being in the unknown state, and that there are n possible output values, each of which has an equal prior chance of being produced by a gate in the unknown state. This can be represented as the hypotheses

$$assumable(u(G), \epsilon)$$
$$assumable(produces(G, V, T), \tfrac{1}{n})$$

and the rule

$$val(out(G), V, T) \leftarrow u(G) \wedge produces(G, V, T).$$

$u(G)$ means G is in the unknown state, and $produces(G, V, T)$ means that given gate G is broken, it produces value V at time T. We assume once that the gate is broken, and then make other assumptions of what values it is producing at different times.

It is interesting to note that this analysis of dividing by n can be done when building the knowledge base and does not need to be carried out dynamically (as [de Kleer and Williams, 1989] seem to need to do). This means that distributions other than the uniform distribution can be given if appropriate.

3.2.2 Intermittent versus non-intermittent faults

Because of the way we parametrized the hypotheses, the above representation of faults says that the output is only a function of the time. The hypothesis $prod(G, V, T)$ and the above rules places no constraints on the values of the outputs at different times. This is a way to represent the fact that the gate can have an intermittent fault (it depends only on the time of observation). There is no constraint that says the gate produces the same output when given the same inputs at different times.

We can give the non-intermittency assumption by saying that the fault only depends on the input and not on the time. This can be done instead by having the hypothesis $prod(G, V, I_1, I_2)$ (meaning gate G produces output V when given I_1 and I_2 as input) and a rule

$$\begin{aligned} val(out(G), V, T) \quad \leftarrow \quad & u(G) \wedge prod(G, V, I1, I2) \\ & \wedge val(input(1, G), I_1, T) \\ & \wedge val(input(2, G), I_2, T). \end{aligned}$$

With the same integrity constraint as before, it is inconsistent to assume that the gate has different outputs for the same input.

3.3 Causation events

When using abduction we run into the problem of a cause not actually implying a symptom. For example, having a cold does not imply sneezing, but could cause sneezing. To implement this idea we introduce another hypothesis that the cold caused the sneezing. This idea is analogous to the notion of a "causation event" of Peng and Reggia [1990].

To implement the causation events, we can use the relations $has\_disease(D)$ to mean that the patient has disease D; $actually\_causes(D, M)$ to mean that disease D "actually caused" manifestation M; and $has\_manifestation(M)$ to mean that the patient has manifestation M.

We can say that a manifestation is caused by the disease that actually causes it by:

$$\begin{aligned} has\_manifestation(M) \quad \leftarrow \quad & has\_disease(D) \\ & \wedge actually\_causes(D, M). \end{aligned}$$

We can use the rule to say that there is only one actual cause of a manifestation by:

$$\begin{aligned} false \quad \leftarrow \quad & actually\_causes(D_1, M) \\ & \wedge actually\_causes(D_2, M) \\ & \wedge different(D_1, D_2). \end{aligned}$$

This rule ensures that the explanations for having a manifestation are disjoint.

The conjunction

$$has\_disease(D) \wedge actually\_causes(D, M)$$

corresponds to Peng and Reggia [1990]'s causation event $M : D$. The completion semantics of abduction [Poole, 1988b; Console *et al.*, 1989; Konolige, 1990] show that, under the covering explanation assumption, we implicitly have the relationship

$$\begin{aligned} manifestation(M) \quad \equiv \bigvee_j (\quad & has\_disease(D_j) \\ & \wedge actually\_causes(D_j, M)) \end{aligned}$$

We have the possible hypothesis

$$assumable(actually\_causes(d_i, m_j), p_{ij})$$

where p_{ij} is the "conditional causal probability" ("causal strength") of [Peng and Reggia, 1990]. It can be seen as the fraction of the cases of d_i being true that m_j is actually caused by d_i.

We also have the possible hypotheses

$$assumable(has\_disease(d_i), p_i)$$

where p_i is the prior probability of the disease d_i.

4 Representing Bayesian networks

In this section we give the relationship between Bayesian networks and our probabilistic abduction. The analysis here is, in some sense, the dual of the analysis given by Charniak and Shimony [1990]. We show how any probabilistic knowledge that can be represented in a Bayesian network, can be represented in our formalism.

Suppose we have a Bayes net with random variables $a_1, ..., a_n$, such that random variable a_i can have values $v_{i,1}, ..., v_{i,n_i}$. We will represent random variable a_i having value $v_{i,j}$ as the proposition $a_i(v_{i,j})$.

The first thing we need to do is to state that the values of variables are mutually exclusive. For each i and for each j, k such that $j \neq k$, we have the rule

$$false \leftarrow a_i(v_{i,j}) \wedge a_i(v_{i,k})$$

A Bayes net [Pearl, 1988] is a directed acyclic network where the nodes represent random variables, and the arcs represent a directly influencing relation. Terminal nodes of a Bayes net are those variables that do not influence any other variables. A composite belief [Pearl, 1987] is an assignment of a value to every random variable.

Suppose variable a is directly influenced by variables $b_1, ..., b_m$ in a Bayes network. This can represented in our system by the rule:

$$a(V) \leftarrow b_1(V_1) \wedge ... \wedge b_m(V_m) \wedge caused\_a(V, V_1, ..., V_m)$$

Here the intended interpretation of

$$caused\_a(V, V_1, ..., V_m)$$

is that a has value V because b_1 has value V_1,..., and b_m has value V_m.

Associated with the Bayes net is a contingency table [Pearl, 1988] which gives the marginal probabilities of the values of a depending on the values of $b_1, ..., b_m$. This will consist of probabilities of the form

$$P(a = v | b_1 = v_1, ..., b_m = v_m) = p$$

This is translated into the assertion

$$assumable(caused\_a(v, v_1, v_2, ..., v_m), p).$$

The following propositions can be proved [Poole, 1992]:

Lemma 4.1 The minimal explanations of the terminal variables having particular values correspond to the composite beliefs in the Bayes net with the terminals having those values. The priors for the explanations and the composite beliefs are identical.

As the same procedure can be used to get from the priors of composite hypotheses and explanations to the posteriors given some observations, the following theorem is a direct corollary of lemma 4.1.

Theorem 4.2 If the observed variables include all terminal variables, the composite beliefs with the observed variables having particular values correspond exactly to the explanations of the observations, and with the same posterior probability.

If the observed variables do not include all terminal values, we need to decide what it is that we want the probability of [Poole and Provan, 1990]. If we want to commit to the value of all variables, as in the composite belief of Pearl [1988], then we consider the set of possible observations that include assigning values to terminal nodes. That is, if o was our observation that did not not include observing a value for variables a_i, then we need to consider the observations $o \wedge a_i(v_{i,1}), ..., o \wedge a_i(v_{i,n_i})$. To find the accurate probabilities we need to normalise over the sum of all of the explanations. Whether or not we want to do this is debatable.

It is not only the probability of a composite hypothesis that has a characterisation in terms of explanations.

Let $expl(a)$ be the set of minimal explanations of proposition a. Define

$$\mathcal{M}(a) = \sum_{E \in expl(a)} P(E)$$

Lemma 4.3 If H is a set of assignments to variables in a Bayesian Network, and H' is the analogous propositions to H in the corresponding probabilistic Horn abduction system, then

$$P(H) = \mathcal{M}(H')$$

A simple corollary of the above lemma can be used to determine the posterior probability of a hypothesis based on some observations:

Theorem 4.4

$$P(x_i(v_i)|obs) = \frac{\mathcal{M}(obs \wedge x_i(v_i))}{\mathcal{M}(obs)}$$

The denominator can be obtained by finding the explanations of the observations (or can be approximated by finding some of the explanations that cover some proposition of the probability mass). The numerators can be obtained by explaining $x_i(v_i)$ from these explanations (see [Poole, 1991]).

What is important about the comparison with the Bayes net is that any probability distribution that can be represented as a Bayes net can be represented using the probabilistic Horn abduction. The opposite is not the case, however, because our Horn abduction is not restricted to a propositional language.

5 Comparison with other diagnostic systems

The closest work to that presented here, namely the work of de Kleer and Williams [1987; 1989] and Peng and Reggia [1990], both incorporate probabilistic knowledge to find the most likely diagnoses.

5.1 de Kleer and Williams

de Kleer and Williams [1987; 1989] have explored the idea of using probabilistic information in consistency-based diagnosis (see [Poole, 1988b; Poole, 1989; Console *et al.*, 1989; Konolige, 1990] for comparisons between abductive and consistency-based diagnoses).

They differ from us in what they compute the probability of. de Kleer and Williams are finding the most likely interpretations (this is the same as the diagnoses of Peng and Reggia [1990] and the composite beliefs of Pearl [1987], but is different to the kernel or minimal diagnoses of de Kleer, Mackworth and Reiter [1990]). We are computing the most likely explanations; we want to remain agnostic about the value of the irrelevant hypotheses. de Kleer and Williams cannot distinguish between the remaining diagnoses that differ in substantial ways from the most likely interpretation, and those that differ only in varying values that are irrelevant to the diagnosis. In our system, hypotheses that are not part of an explanation are ignored, and play no part in the probability of a diagnosis.

We differ in the use of the assumption-based framework. We are using the assumption-based reasoning, with variables, as the object language. They use the ATMS as a book keeping mechanism for their diagnostic engine.

5.2 Peng and Reggia

Peng and Reggia [1990] also consider an abductive definition of diagnosis and incorporate probabilities, and best-first search. Like [de Kleer and Williams, 1989; Pearl, 1987] they are trying to find probabilities of interpretations. We also do not assume that the set of manifestations is complete. The main difference, however, is in the underlying language. They use the notion of "hyper-bipartite" graphs made up of causation relations on sets of manifestations (can be observed), disorders (can be hypothesised), and pathological states. We, however, allow the full power of Horn clauses. We can represent the probabilistic knowledge of Peng and Reggia (see section 3.3).

6 Conclusion

This paper presented a simple but powerful mechanism for diagnostic reasoning and showed how it can be used to solve diagnostic representation problems. One main advantage of the simple specification of what we want to compute is that we can investigate different implementation techniques to determine which works best in practice.

One question that needs to be asked is whether a set of most likely explanations is really what we want to compute [Poole and Provan, 1990]. We conjecture that for real problems, the probability mass of the most likely explanations will be so close to one to make the question moot. By ignoring the large number of very unlikely explanations, we will not make many mistakes. Whether this is true in practice remains to be seen.

We are also investigating the use of the abductive framework for differential diagnoses, and for making decisions, but that is beyond the scope of this paper.

Acknowledgements

Thanks to Alan Mackworth and Mike Horsch for valuable comments on a previous version of this paper, and Mike for work on the implementation. This research was supported under NSERC grant OGPOO44121, and under Project B5 of the Institute for Robotics and Intelligent Systems.

References

[Andreassen *et al.*, 1987] S. Andreassen, M. Woldbye, B. Falck, and S. K. Andersen. MUNIN - a causal probabilistic network for interpretation of electromyographic findings. In *IJCAI-87*, pages 366–372, Milan, Italy, August 1987.

[Charniak and Shimony, 1990] E. Charniak and S. E. Shimony. Probabilistic semantics for cost based abduction. In *AAAI-90*, pages 106–111, Boston, July 1990.

[Console *et al.*, 1989] L. Console, D. Theseider Dupré, and P. Torasso. Abductive reasoning through direct deduction from completed domain models. In W. R. Zbigniew, editor, *Methodologies for Intelligent Systems 4*, pages 175–182. Elsiever Science Publishing Co., 1989.

[de Kleer and Williams, 1987] J. de Kleer and B. C. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32(1):97–130, April 1987.

[de Kleer and Williams, 1989] J. de Kleer and B. C. Williams. Diagnosis with behavioral modes. In *IJCAI-89*, pages 1324–1330, Detroit, August 1989.

[de Kleer *et al.*, 1990] J. de Kleer, A. K. Mackworth, and R. Reiter. Characterizing diagnoses. In *AAAI-90*, pages 324–330, Boston, July 1990.

[de Kleer, 1986] J. de Kleer. An assumption-based TMS. *Artificial Intelligence*, 28(2):127–162, March 1986.

[Goldman and Charniak, 1988] R. P. Goldman and E. Charniak. A probabilistic ATMS for plan recognition. In *Proceedings of the Plan Recognition Workshop, 1988*, August 1988.

[Heckerman and Horvitz, 1990] D. E. Heckerman and E. J. Horvitz. Problem formulation as the reduction of a decision model. In *UAI-90*, pages 82–89, Cambridge, MA, July 1990.

[Hobbs *et al.*, 1988] J. R. Hobbs, M. E. Stickel, P. Martin, and D. Edwards. Interpretation as abduction. In *Proc. 26th Annual Meeting of the Association for Computational Linguistics*, pages 95–103, Buffalo, June 1988.

[Konolige, 1990] K. Konolige. Closure + minimization implies abduction. technical report ??, SRI International, Menlo Park, CA, 1990.

[Lloyd, 1987] J. W. Lloyd. *Foundations of Logic Programming.* Symbolic Computation Series. Springer-Verlag, Berlin, second edition, 1987.

[McCarthy and Hayes, 1969] J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In M. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, 1969.

[Neufeld and Poole, 1987] E. M. Neufeld and D. Poole. Towards solving the multiple extension problem: combining defaults and probabilities. In *Proc. Third Workshop on Reasoning with Uncertainty*, pages 305–312, Seattle, July 1987.

[Pearl, 1987] J. Pearl. Distributed revision of composite beliefs. *Artificial Intelligence*, 33(2):173–215, October 1987.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA, 1988.

[Peng and Reggia, 1990] Y. Peng and J. A. Reggia. *Abductive Inference Models for Diagnostic Problem-Solving.* Symbolic Computation – AI Series. Springer-Verlag, New York, 1990.

[Poole and Provan, 1990] D. Poole and G. Provan. What is an optimal diagnosis? In *Proc. Sixth Conference on Uncertainty in AI*, pages 46–53, Boston, July 1990.

[Poole *et al.*, 1987] D. Poole, R. Goebel, and R. Aleliunas. Theorist: A logical reasoning system for defaults and diagnosis. In N. Cercone and G. McCalla, editors, *The Knowledge Frontier: Essays in the Representation of Knowledge*, pages 331–352. Springer-Verlag, New York, NY, 1987.

[Poole, 1988a] D. Poole. A logical framework for default reasoning. *Artificial Intelligence*, 36(1):27–47, 1988.

[Poole, 1988b] D. Poole. Representing knowledge for logic-based diagnosis. In *International Conference on Fifth Generation Computing Systems*, pages 1282–1290, Tokyo, Japan, November 1988.

[Poole, 1989] D. Poole. Normality and faults in logic-based diagnosis. In *IJCAI-89*, pages 1304–1310, Detroit, August 1989.

[Poole, 1990] D. Poole. A methodology for using a default and abductive reasoning system. *International Journal of Intelligent Systems*, 5(5):521–548, December 1990.

[Poole, 1991] D. Poole. Search-based implementations of probabilistic Horn abduction. Technical report, Department of Computer Science, University of British Columbia, Vancouver, B.C., Canada, 1991.

[Poole, 1992] D. Poole. Probabilistic Horn abduction and Bayesian networks. Technical report, Department of Computer Science, University of British Columbia, Vancouver, B.C., Canada, 1992.

[Provan, 1988] G. Provan. A complexity analysis for assumption-based truth maintenance systems. In B. M. Smith and R. Kelleher, editors, *Reason Maintenance Systems and Their Applications*, pages 98–113. Ellis Howard, 1988.

[Reiter and de Kleer, 1987] R. Reiter and J. de Kleer. Foundations of assumption-based truth maintenance systems: preliminary report. In *AAAI-87*, pages 183–188, Seattle, July 1987.

[Reiter, 1987] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95, April 1987.