Diagnosis as Refutation

Sheila McIlraith
Department of Computer Science
University of Toronto
Toronto, M5S 1A4. Canada
email: mcilrait@ai.toronto.edu phone: 416-978-1929
fax: 416-978-1931

Abstract

Two popular approaches to model-based diagnosis are abductive diagnosis and consistency-based diagnosis. Given a description of a system together with an observation of the system's behavior, both approaches conjecture a space of candidate diagnoses. Rather than characterize the diagnoses, which are defeasible, we characterize the space of refuted, or eliminated diagnoses and show their applicability to various diagnostic reasoning tasks. In keeping with previous work, we exploit the notion of prime implicants/implicates to provide our characterization. Generating refuted diagnoses is at the computational core of both consistency-based and abductive diagnoses. We demonstrate this simple intuitive correspondence. As a by-product, we extend the general results on characterizing diagnosis [1] by providing a new characterization of abductive explanation and abductive diagnosis.

Introduction

Most characterizations of diagnosis legitimately presuppose that the objective of diagnostic reasoning is restricted to the generation of the space of candidate diagnoses. This is not the case. Diagnostic reasoning, whether it be consistency-based, abductive or otherwise, is used within the context of many different hypothetical reasoning tasks in domains such as vision, planning, and language understanding. In such applications, it is sometimes more relevant to know which diagnoses have been soundly eliminated, rather than to characterize the space of defeasible diagnoses which can currently be conjectured. As a simple example, take the task of administering first aid. There is a critical path to performing first aid which the attendant must follow in diagnosing and treating an injured person: eliminate the possibility that the airway is blocked (or treat); eliminate the possibility that the patient is not breathing (or treat); eliminate the possibility that the heart has stopped (or treat); ... eliminate the possibility of spinal injury ..., and so on. In this application, an objective of diagnostic reasoning is to explicitly characterize the refuted diagnoses in a goal-directed manner, rather than to characterize the space of candidate diagnoses. Similarly, in the case of differential diagnosis [3], [2] where we may be given a small set of candidate diagnoses a priori, the objective of diagnostic reasoning is not candidate generation, but rather candidate refutation.

This paper provides three separate contributions under the unifying theme of *diagnosis as refutation*. First, we define the notion of a refuted diagnosis, we characterize the space of refuted diagnoses, and propose their relevance to certain hypothetical reasoning tasks. Continuing the theme of generating refutations, we show the simple correspondence between the space of refuted diagnoses and the space of consistency-based and abductive diagnoses. Defining the computational core in this arguably more intuitive manner, not only helps us in understanding the nature of consistency-based and abductive diagnosis, but also assists a developer in addressing the computational complexity of a specific diagnosis application. As a by-product of demonstrating the relationship between abduction and refuted diagnoses, we have extended the results on the characterization of diagnosis [1] for abductive explanations and abductive diagnoses.

Preliminaries

In keeping with previous work on diagnosis, we assume a language of first-order logic with equality and follow the terminology provided in [1].

Definition 1 *A system is a triple (SD, COMPS, OBS) where:*

- *SD, the system description, is a set of first-order sentences.*
- *$COMPS$, the system components, is a finite set of constants.*
- *OBS, a set of observations, is a set of first-order sentences.*

SD, the system description serves as the relevant background knowledge for describing the system under analysis. For example, in the case of circuit diagnosis, SD might describe the individual circuit components, their normal input/output behavior, their fault models, the topology of their interconnections, and the legal combinations of circuit inputs (e.g. [2], [5]). The predicate AB denotes the abnormal ($AB(c)$) and normal

($\neg AB(c)$) behavior of components $c \in COMPS$. A diagnosis is a conjunction of literals denoting the normal or abnormal behavior of each component in $COMPS$. Although we maintain the use of the predicate AB for continuity with previous work, it is not necessary and is not always an intuitive way of encoding an application domain. For example, in the case of medical diagnosis, the set of components, $COMPS$ could represent various diseases, the set of observations, OBS could represent various symptoms. The diagnoses would simply be a conjunction of the literals c or $\neg c$ for every $c \in COMPS$, eliminating the use of the AB predicate altogether. For notational simplicity, we have assumed that any initial conditions such as inputs, sometimes designated separately as I (for input) or A (for achievable literal) are included in SD and are not explicitly identified. This is in keeping with the literature on abduction, but differs from [1].

Definition 2 (AB-literal) *An AB-literal is $AB(c)$ or $\neg AB(c)$ for some $c \in COMPS$.*

Definition 3 (AB-hypothesis) *Given two mutually exclusive sets of components $\triangle_1, \triangle_2 \subseteq COMPS$, define an AB-hypothesis $\mathcal{D}(\triangle_1, \triangle_2)$ of the system $(SD, COMPS, OBS)$ to be the conjunction:*
$[\bigwedge_{c \in \triangle_1} AB(c)] \wedge [\bigwedge_{c \in \triangle_2} \neg AB(c)]$.

In our notation, we distinguish between a *diagnosis* which labels all of the components $c \in COMPS$ as either normal or abnormal, and an *AB-hypothesis* which need not label every component in $COMPS$. In subsequent sections, we will distinguish the generic term diagnosis from consistency-based diagnosis and abductive diagnosis.

Definition 4 (AB-clause) *An AB-clause is a disjunction of AB-literals containing no complementary pair of AB-literals. A positive AB-clause is an AB-clause all of whose literals are positive.*

Definition 5 (Covers) *A conjunction C of literals covers a conjunction D of literals iff every literal of C occurs in D.*

Prime Implicates and Implicants Prime implicants/implicates have been employed in the analysis of logic circuits for some time [8]. More recently, they have been used to provide a propositional logic characterization of abduction [6] and consistency-based diagnosis [1]. As it happens, the principal task of an assumption-based truth maintenance system is the computation of certain prime implicates of a background theory Σ [6]. Despite the high complexity associated with the computation of prime implicates, ATMSs are very frequently used as implementation tools in abductive and diagnostic reasoning systems. The application of prime implicates and implicants to first-order theories with universal quantification is problematic. However, prime implicates and implicants may be used for first-order theories with finite domains, since they can be expressed propositionally. Further, with model-based diagnosis, our system $(SD, COMPS, OBS)$ is composed of a finite set of components, $COMPS$. The scope of the literals AB and $\neg AB$ is limited to the components $COMPS$. Consequently, even though our system description, SD is first order, we may employ the prime implicants of AB-clauses and the AB-clauses which are prime implicates of a set of sentences since they range over a finite domain.

Definition 6 (Prime Implicate) *Suppose Σ is a set of first-order sentences. A disjunction of ground literals C is an implicate of Σ iff Σ entails C. C is a prime implicate of Σ iff the only implicate of Σ covering C is C itself.*

Definition 7 (Prime Implicant) *Suppose Σ is a set of first-order sentences. A conjunction of ground literals π containing no pairs of complementary literals is an implicant of Σ iff π entails each sentence in Σ. π is a prime implicant of Σ iff the only implicant of Σ covering π is π itself.*

Refuted Diagnoses

To refute a diagnosis or an hypothesis is to prove its falsity. Returning to our first aid example, given the axiom $blocked - airway \supset \neg speak$ and the observation *speak*, we can infer $\neg blocked - airway$. Thus, $blocked - airway$ is a refuted hypothesis/diagnosis. In this section, we formally define the notion of a refuted hypothesis and a refuted diagnosis. We characterize the space of all refuted diagnoses in terms of minimal refuted AB-hypotheses and in terms of prime implicates. Finally, we propose the relevance of refuted diagnoses within the context of several different diagnostic reasoning tasks.

Definition 8 (Refuted AB-hypothesis)
$\mathcal{D}(\triangle_1, \triangle_2)$ is a refuted AB-hypothesis of $(SD, COMPS, OBS)$ iff

- $SD \cup OBS \models \neg\mathcal{D}(\triangle_1, \triangle_2)$
- *$SD \cup OBS$ is satisfiable*

Recall, we distinguish between an AB-hypothesis, which is a labeling of AB or $\neg AB$ for some subset of components in $COMPS$, and a diagnosis, which is a labeling for *all* components in $COMPS$.

Definition 9 (Refuted diagnosis)
$\mathcal{D}(\triangle_1, \triangle_2)$ is a refuted diagnosis of $(SD, COMPS, OBS)$ iff $\mathcal{D}(\triangle_1, \triangle_2)$ is a refuted AB-hypothesis and $\triangle_1 \cup \triangle_2 = COMPS$.

We observe that unlike consistency-based or abductive diagnoses which are defeasible upon the addition of subsequent observations [1], a refuted diagnosis remains refuted forever.

Remark 1 (Monotonicity)
If $\mathcal{D}(\triangle_1, \triangle_2)$ is a refuted diagnosis of $(SD, COMPS, OBS_1)$ and $SD \cup OBS_1 \cup OBS_2$ is satisfiable, then $\mathcal{D}(\triangle_1, \triangle_2)$ is a refuted diagnosis of $(SD, COMPS, OBS_1 \cup OBS_2)$.

Rather than individually enumerate the space of refuted diagnoses of $(SD, COMPS, OBS)$ (of which there could be up to $2^{|COMPS|}$), we seek a parsimonious characterization. To this end we define the notion of a minimal refuted AB-hypothesis and show its correspondence to the prime implicates of $SD \cup OBS$.

Definition 10 (Min refuted AB-hypothesis) *$\mathcal{D}(\triangle_1, \triangle_2)$ is a minimal refuted AB-hypothesis for $(SD, COMPS, OBS)$ iff for no proper subconjunct $\mathcal{D}(\triangle_1', \triangle_2')$ of $\mathcal{D}(\triangle_1, \triangle_2)$ is $\mathcal{D}(\triangle_1', \triangle_2')$ a refuted AB-hypothesis for $(SD, COMPS, OBS)$.*

Theorem 1 *If $\mathcal{D}(\triangle_1', \triangle_2')$ is a refuted AB-hypothesis of $(SD, COMPS, OBS)$, then $\mathcal{D}(\triangle_1, \triangle_2)$ is a refuted AB-hypothesis of $(SD, COMPS, OBS)$ for any $\triangle_1, \triangle_2$ such that $COMPS \supseteq \triangle_1 \supseteq \triangle_1'$, $COMPS \supseteq \triangle_2 \supseteq \triangle_2'$.*

Corollary 1 *If $\mathcal{D}(\triangle_1', \triangle_2')$ is a minimal refuted AB-hypothesis of $(SD, COMPS, OBS)$ then $\mathcal{D}(\triangle_1, \triangle_2)$ is a refuted diagnosis of $(SD, COMPS, OBS)$ for any $\triangle_1, \triangle_2$ such that $COMPS \supseteq \triangle_1 \supseteq \triangle_1'$, $COMPS \supseteq \triangle_2 \supseteq \triangle_2'$, $\triangle_1 \cup \triangle_2 = COMPS$.*

Thus, we may use the minimal refuted AB-hypotheses of $(SD, COMPS, OBS)$ to characterize the space of refuted diagnoses of $(SD, COMPS, OBS)$. The following theorem characterizes minimal refuted AB-hypotheses in terms of prime implicates.

Theorem 2
$\mathcal{D}(\triangle_1, \triangle_2)$ is a minimal refuted AB-hypothesis of $(SD, COMPS, OBS)$ iff
$\neg([\bigwedge_{c_1 \in \triangle_1} AB(c_1)] \wedge [\bigwedge_{c_2 \in \triangle_2} \neg AB(c_2)])$
is a prime implicate of $SD \cup OBS$ and $SD \cup OBS$ is satisfiable.

The following straightforward observation should be noted.

Remark 2 *The set of minimal refuted AB-hypotheses is equivalent to the set composed of the negation of those prime implicates of $SD \cup OBS$ which are AB-clauses.*

Relevance

As mentioned in the introduction, there are many applications within the context of hypothetical reasoning for which characterization of the space of refuted diagnoses is desirable. We provide a brief sampling of some such applications here and leave a more complete and rigorous discussion to an extended version of this paper.

The space of refuted diagnoses may be utilized in the context of diagnostic reasoning when we have no need to actually *generate* diagnoses. There are many instances when this is the case; the simplest example of which is differential diagnosis. When performing differential diagnosis, we are given a space of candidate diagnoses to be considered a priori. Our sole objective is to refute diagnoses, no further candidate generation is required. Differential diagnosis is viable when candidate diagnoses are substantially fewer than $2^{|COMPS|}$ and are easily enumerated.

As further illustration of the application of refuted diagnoses, we return to the example of administering first aid. There are many applications which combine diagnostic reasoning with some form of action, be it further testing, treatment or other. In such cases, actions may be preconditioned on the elimination of certain diagnoses. For example, there is no point in giving someone cardiopulmonary resuscitation until you have eliminated the diagnosis of $blocked - airway$. Differential diagnosis may be used in conjunction with such a system to define critical path/goal-directed diagnosis.

The task of selecting tests to obtain further measurements with the objective of discriminating diagnoses also falls within the above paradigm. In order to select a test action, we must know which diagnoses will be refuted by outcomes of the various actions.

Most importantly, we will see in the next sections that not only are the space of refuted diagnoses of interest in their own right, but they are at the computational core of both consistency-based and abductive diagnosis.

Consistency-based Diagnosis

In this section and the following section on abduction, we change tack and re-examine the characterization of diagnoses within the context of refuted AB-hypotheses and refuted diagnoses. In so doing, we show that whether characterizing the conjectured diagnoses or the eliminated diagnoses, the computation of refuted AB-hypotheses is vital. Furthermore, by recharacterizing diagnosis in terms of refutations, it arguably provides a more intuitive understanding of the process of generating diagnoses. By characterizing the source of computational complexity (the generation of prime implicates) in terms of the generation of refuted AB-hypotheses, useful insight is provided into addressing the complexity problems associated with model-based diagnosis.

Historically, the objective of consistency-based diagnosis was to determine why a correctly designed system was not functioning as intended. The observed faulty behavior was explained by noting that certain components were behaving in a manner which was contradictory to their designed behavior. The system description encoded the correct behavior of the artifact as per the design description, using the predicates AB and $\neg AB$ to denote abnormal and normal behavior of components. The space of diagnoses was succinctly characterized in terms of minimal diagnoses. In subsequent research, (e.g. Struss [7], de Kleer and Williams [2]) it was suggested that system descriptions be augmented with axioms describing the faulty behavior of components. The result was that minimal diagnoses were no longer guaranteed to be sufficient characterizations for the space of diagnoses, leading to the definition of kernel diagnosis found in [1].

In this section we specifically point out that the task of computing the set of conflicts for a consistency-based diagnosis of $(SD, COMPS, OBS)$ is equivalent to the task of generating the refuted AB-hypotheses for the

system. In particular, the set of minimal conflicts is equivalent to the set consisting of the negation of the minimal refuted AB-hypotheses. We use this more intuitive description to recharacterize consistency-based diagnosis in terms of refuted AB-hypotheses.

Again, we appeal to many of the definitions in [1] to provide a framework that is in keeping with previous work in this area. Some of the notation has been modified slightly for consistency (e.g. the use of AB-hypotheses and the substitution of $\mathcal{D}(\triangle_1, \triangle_2), \triangle_1 \cup \triangle_2 = COMPS$ for $\mathcal{D}(\triangle, COMPS - \triangle)$).

Definition 11 (Consistency-based diagnosis [1])
A consistency-based diagnosis of $(SD, COMPS, OBS)$ *is an AB-hypothesis* $\mathcal{D}(\triangle_1, \triangle_2)$ *such that* $\triangle_1 \cup \triangle_2 = COMPS$, *and* $SD \cup OBS \cup \{\mathcal{D}(\triangle_1, \triangle_2)\}$ *is satisfiable.*

In order to compute the consistency-based diagnoses for $(SD, COMPS, OBS)$, the conflicts must be generated. The diagnoses follow from the conflicts.

Definition 12 (Conflict [1])
A conflict of $(SD, COMPS, OBS)$ *is an AB-clause entailed by* $SD \cup OBS$. *A minimal conflict of* $(SD, COMPS, OBS)$ *is a conflict no proper subclause of which is a conflict of* $(SD, COMPS, OBS)$. *A positive conflict is a conflict all of whose literals are positive. A negative conflict is a conflict all of whose literals are negative.*

Theorem 3 ([1])
Suppose $(SD, COMPS, OBS)$ *is a system,* Π *is its set of minimal conflicts,* $\triangle_1, \triangle_2 \subseteq COMPS$ *and* $\triangle_1 \cup \triangle_2 = COMPS$. *Then* $\mathcal{D}(\triangle_1, \triangle_2)$ *is a consistency-based diagnosis iff*
$\Pi \cup \{\mathcal{D}(\triangle_1, \triangle_2)\}$
is satisfiable.

Remark 3 *The set of minimal conflicts for* $(SD, COMPS, OBS)$ *is equivalent to the set consisting of the negation of the minimal refuted AB-hypotheses for* $(SD, COMPS, OBS)$.

Furthermore, a positive conflict is a refutation of the normal behavior of some conjunction of components. A negative conflict is a refutation of the abnormal behavior of some conjunction of components.

The following theorem characterizes consistency-based diagnoses in terms of minimal refuted AB-hypotheses.

Theorem 4
Suppose $(SD, COMPS, OBS)$ *is a system and* R *the set of minimal refuted AB-hypotheses of the system,* $\triangle_1, \triangle_2 \subseteq COMPS$ *and* $\triangle_1 \cup \triangle_2 = COMPS$. *Then* $\mathcal{D}(\triangle_1, \triangle_2)$ *is a consistency-based diagnosis iff diagnosis iff* $\forall \mathcal{D}(\triangle_1', \triangle_2') \in R$, $\mathcal{D}(\triangle_1', \triangle_2') \not\subseteq \mathcal{D}(\triangle_1, \triangle_2)$.

It follows that the space of consistency-based diagnoses are simply all possible variations of $\mathcal{D}(\triangle_1, \triangle_2)$ $(\triangle_1 \cup \triangle_2 = COMPS)$ which do not contain the refuted AB-hypotheses.

Remark 4 *The space of consistency-based diagnoses for* $(SD, COMPS, OBS)$ *is equal to the initial space of* $2^{|COMPS|}$ *candidate diagnoses, minus the space of refuted diagnoses.*

The consistency-based diagnoses are merely the complement of the refuted diagnoses. The generation of consistency-based diagnoses from conflicts sets can be equivalently viewed as the removal of refuted AB-hypotheses from the current space of diagnoses. Rather than removing refuted AB-hypotheses from the space of possible diagnoses, conflicts are used to generate diagnoses. Recall that the minimal conflicts are the negation of the minimal refuted AB-hypotheses. Each refuted AB-hypothesis is composed of a conjunction of AB-literals. By ensuring that the negation of one AB-literal from each refuted AB-hypothesis (i.e. an AB-literal from the conflict disjunct) is contained in each diagnosis, the system ensures that no diagnosis contains a refuted hypothesis.

We now demonstrate the correspondence between kernel consistency-based (CB) diagnoses and refuted diagnoses.

Definition 13 (Partial CB diagnosis [1])
A partial consistency-based diagnosis of $(SD, COMPS, OBS)$ *is an AB-hypothesis* $\mathcal{D}(\triangle_1', \triangle_2')$ *such that for every AB-hypothesis* $\mathcal{D}(\triangle_1, \triangle_2)$ *where* $\triangle_1' \subseteq \triangle_1 \subseteq COMPS$ *and* $\triangle_2' \subseteq \triangle_2 \subseteq COMPS$, $\mathcal{D}(\triangle_1, \triangle_2)$ *is covered by* $\mathcal{D}(\triangle_1', \triangle_2')$, $SD \cup OBS \cup \{\mathcal{D}(\triangle_1, \triangle_2)\}$ *is satisfiable.*

Definition 14 (Kernel CB diagnosis [1])
A kernel consistency-based diagnosis of $(SD, COMPS, OBS)$ *is a partial consistency-based diagnosis with the property that the only partial consistency-based diagnosis which covers it is itself.*

The following theorem is a characterization of kernel consistency-based diagnosis.

Theorem 5 ([1])
The kernel consistency-based diagnoses of $(SD, COMPS, OBS)$ *are the prime implicants of the minimal conflicts of* $SD \cup OBS$.

It follows directly that:

Remark 5 *The kernel consistency-based diagnoses of* $(SD, COMPS, OBS)$ *are the prime implicants of the negations of the refuted AB-hypotheses of* $(SD, COMPS, OBS)$.

Remark 6 *Let* K *be the set of kernel consistency-based diagnoses of* $(SD, COMPS, OBS)$. *Then for every refuted diagnosis* $\mathcal{D}(\triangle_1, \triangle_2)$, $K \models \neg\mathcal{D}(\triangle_1, \triangle_2)$.

Contrast the calculation of consistency-based diagnoses with the calculation of refuted diagnoses. Both involve the generation of the prime implicates of $SD \cup OBS$, yet for the generation of consistency-based diagnoses, we must then generate the prime implicants of these prime implicates.

Abduction

In this section we show that the task of computing abductive explanations or diagnoses for (SD, $COMPS$, OBS) is equivalent to the task of generating the AB-hypotheses which are consistent with SD and which are refuted by the negation of OBS. This observation in itself is not surprising, but it assists in demonstrating the correspondence between abductive diagnosis, consistency-based diagnosis and refuted diagnoses. An important by-product of this section is a new characterization of abductive explanation and abductive diagnosis, not only in terms of AB-hypotheses, but also in terms of prime implicants/implicates,

Whereas the objective of *consistency-based diagnosis* is to find a labeling for *all* components that is consistent with the system description and observed behavior, the objective of *abduction* [4] is to find a labeling for a subset of components which, in conjunction with the system description, *explains* the observed behavior. In order to discuss the application of abduction to diagnostic problem-solving, we distinguish between an abductive explanation and an abductive diagnosis. Both are useful concepts within this framework.

Abductive Explanation

Definition 15 (Abductive Explanation)
An abductive explanation for $(SD, COMPS, OBS)$ *is any AB-hypothesis* $\mathcal{D}(\triangle_1, \triangle_2)$ *such that*

- $SD \cup \{\mathcal{D}(\triangle_1, \triangle_2)\} \models OBS$
- $SD \cup \{\mathcal{D}(\triangle_1, \triangle_2)\}$ *is satisfiable*
- $SD \not\models OBS$

Note in the above definition that $\triangle_1 \cup \triangle_2$ need not equal $COMPS$. The third criterion in the definition of an abductive explanation eliminates the null explanation $\mathcal{D}(\{\}, \{\})$. An observation that may be explained by the null explanation is assumed to be an invalid abduction problem in much of the abduction literature [6]. We have simply made this assumption explicit in the definition.

Theorem 6
If an abductive explanation exists for (SD, $COMPS$, OBS) *then* $SD \cup OBS$ *is satisfiable and* $SD \cup \neg OBS$ *is satisfiable.*

Remark 7
$\mathcal{D}(\triangle_1, \triangle_2)$ *is an abductive explanation for* (SD, $COMPS$, OBS) *iff*

- $SD \cup \neg OBS \models \neg\mathcal{D}(\triangle_1, \triangle_2)$
- $SD \cup \{\mathcal{D}(\triangle_1, \triangle_2)\}$ *is satisfiable*
- $SD \not\models OBS$

This follows directly from Theorem 6 and from the equivalence of $SD \cup \{\mathcal{D}(\triangle_1, \triangle_2)\} \models OBS$ and $SD \cup \neg OBS \models \neg\mathcal{D}(\triangle_1, \triangle_2)$.

The following theorem characterizes abductive explanations in terms of refuted AB-hypotheses.

Theorem 7
Suppose $(SD, COMPS, OBS)$ *is a system. Let*

- $R_{\neg OBS}$ *be the set of refuted AB-hypotheses of* $(SD, COMPS, \neg OBS)$
- $R_{\{\}}$ *be the set of refuted AB-hypotheses of* $(SD, COMPS, \{\})$

Then $\mathcal{D}(\triangle_1, \triangle_2)$ *is an abductive explanation for* $(SD, COMPS, OBS)$ *iff* $\mathcal{D}(\triangle_1, \triangle_2) \in R_{\neg OBS}$ *and* $\forall\mathcal{D}(\triangle_1', \triangle_2') \in R_{\{\}}, \mathcal{D}(\triangle_1', \triangle_2') \not\subseteq \mathcal{D}(\triangle_1, \triangle_2)$.

The abductive explanations are calculated as follows. Calculate the AB-hypotheses which would be refuted if we conjectured observing $\neg OBS$ instead of OBS; these refuted AB-hypotheses in $R_{\neg OBS}$ explain OBS. To ensure that they are satisfiable with SD, generate the AB-hypotheses which are refuted by SD alone ($R_{\{\}}$). These are the AB-hypotheses which are inconsistent with SD. Finally, perform a simple subset test to ensure that the inconsistent AB-hypotheses are not incorporated into the explanations.

Definition 16 (Minimal abductive explanation)
$\mathcal{D}(\triangle_1, \triangle_2)$ *is a minimal abductive explanation for* $(SD, COMPS, OBS)$ *iff for no proper non-empty subconjunct* $\mathcal{D}(\triangle_1', \triangle_2')$ *of* $\mathcal{D}(\triangle_1, \triangle_2)$ *is* $\mathcal{D}(\triangle_1', \triangle_2')$ *an abductive explanation for* $(SD, COMPS, OBS)$.

The following theorem characterizes minimal abductive explanations in terms of minimal refuted AB-hypotheses.

Theorem 8
Suppose $(SD, COMPS, OBS)$ *is a system. Let*

- $R^*_{\neg OBS}$ *be the set of minimal refuted AB-hypotheses of* $(SD, COMPS, \neg OBS)$
- $R^*_{\{\}}$ *be the set of minimal refuted AB-hypotheses of* $(SD, COMPS, \{\})$

Then $\mathcal{D}(\triangle_1, \triangle_2)$ *is a minimal abductive explanation for* $(SD, COMPS, OBS)$ *iff* $\mathcal{D}(\triangle_1, \triangle_2) \in R^*_{\neg OBS}$ *and* $\forall\mathcal{D}(\triangle_1', \triangle_2') \in R^*_{\{\}}, \mathcal{D}(\triangle_1', \triangle_2') \not\subseteq \mathcal{D}(\triangle_1, \triangle_2)$.

The following theorem characterizes minimal abductive explanations in terms of prime implicates.

Theorem 9
$\mathcal{D}(\triangle_1, \triangle_2)$ *is a minimal abductive explanation for* $(SD, COMPS, OBS)$ *iff* $\neg\mathcal{D}(\triangle_1, \triangle_2)$ *is a prime implicate of* $SD \cup \neg OBS$, *and for no non-empty subconjunct* $\mathcal{D}(\triangle_1', \triangle_2')$ *of* $\mathcal{D}(\triangle_1, \triangle_2)$, *is* $\neg\mathcal{D}(\triangle_1', \triangle_2')$ *a prime implicate of* SD.

This theorem follows directly from the equivalence of the prime implicates of $(SD, COMPS, OBS)$ and the negation of the minimal refuted AB-hypotheses of that system.

Abductive Diagnosis

The abductive *explanations* tell us which subset of AB-literals we can consistently conjoin to our system description to entail (explain) the observations. With an abductive *diagnosis*, we not only want to be able to explain the observations, but we also want to ascribe a labeling of AB or $\neg AB$ to each of the other components. As such, our abductive diagnoses must not only include labelings for those components which explain the observations, but also for those components whose behavior has been refuted by the observations and those components whose (combined) behavior is impossible given our system description. In this regard, the notion of abductive diagnosis is particularly useful for sequential abductive reasoning, because it encodes components that are refuted by the system description and observations as well as those that simply explain the observations to date.

Definition 17 (Abductive diagnosis)
Let $\triangle_1 \cup \triangle_2 = COMPS$. An abductive diagnosis for $(SD, COMPS, OBS)$ is an AB-hypothesis $\mathcal{D}(\triangle_1, \triangle_2)$ such that:

- $SD \cup \{\mathcal{D}(\triangle_1, \triangle_2)\} \models OBS$
- $SD \cup \{\mathcal{D}(\triangle_1, \triangle_2)\}$ *is satisfiable*
- $SD \not\models OBS$

Note that the distinction between this definition and the definition of an abductive explanation is that $\triangle_1 \cup \triangle_2 = COMPS$.

The following theorem characterizes abductive diagnoses with minimal refuted AB-hypotheses.

Theorem 10
Suppose $(SD, COMPS, OBS)$ is a system. Let

- $R^*_{\neg OBS}$ *be the set of minimal refuted AB-hypotheses of $(SD, COMPS, \neg OBS)$*
- R^*_{OBS} *be the set of minimal refuted AB-hypotheses of $(SD, COMPS, OBS)$*

Then $\mathcal{D}(\triangle_1, \triangle_2)$ is an abductive diagnosis for $(SD, COMPS, OBS)$ iff $\triangle_1 \cup \triangle_2 = COMPS$ and $\exists$ non-empty $\mathcal{D}(\triangle_i, \triangle_j) \in R^*_{\neg OBS}$ such that $\{\mathcal{D}(\triangle_i, \triangle_j)\} \cup \{\mathcal{D}(\triangle_1, \triangle_2)\}$ is satisfiable and $\forall \mathcal{D}(\triangle_1', \triangle_2') \in R^*_{OBS}$, $\mathcal{D}(\triangle_1', \triangle_2') \not\subseteq \mathcal{D}(\triangle_1, \triangle_2)$.

Note that whereas an abductive explanation is equal to the negation of a refuted AB-hypothesis of $(SD, COMPS, \neg OBS)$, (satisfiability condition not withstanding) an abductive diagnosis need only be *consistent* with the refuted AB-hypothesis from the same set.

One of the objectives of characterizing diagnosis is to find a succinct representation for the set of diagnoses. As with consistency-based diagnosis, the concept of a minimal abductive diagnosis which we do not define here, would be inadequate to characterize the space of all diagnoses. Consequently, we define the notion of a kernel abductive diagnosis. De Kleer, Mackworth and Reiter [1] also provided a definition of kernel abductive diagnosis, but their characterization differs from ours. Definitions 18 and 19 and Remark 8 below are equivalent to those defined by them with the notation changes mentioned previously.

Definition 18 (Partial abductive diagnoses[1])
A partial abductive diagnosis of $(SD, COMPS, OBS)$ is an AB-hypothesis $\mathcal{D}(\triangle_1, \triangle_2)$ such that for every non-empty AB-hypothesis $\mathcal{D}(\triangle_1', \triangle_2')$ covered by $\mathcal{D}(\triangle_1, \triangle_2)$, $SD \cup \{\mathcal{D}(\triangle_1', \triangle_2')\}$ is satisfiable and $SD \cup \{\mathcal{D}(\triangle_1', \triangle_2')\} \models OBS$.

Definition 19 (Kernel abductive diagnoses[1])
A kernel abductive diagnosis is a partial abductive diagnosis with the property that the only partial abductive diagnosis which covers it is itself.

Remark 8 ([1]) *The AB-hypothesis $\mathcal{D}(\triangle_1, \triangle_2)$, $\triangle_1 \cup \triangle_2 = COMPS$ is an abductive diagnosis iff there is a kernel abductive diagnosis which covers it.*

The following theorem characterizes kernel abductive diagnoses in terms of prime implicants/implicates.

Theorem 11
Suppose $(SD, COMPS, OBS)$ is a system. Let

- $\Pi_{\neg OBS}$ *be the set of AB-clauses which are prime implicates of $SD \cup \neg OBS$*
- Π_{OBS} *be the set of AB-clauses which are prime implicates of $SD \cup OBS$*
- C_{OBS} *be the set of AB-hypotheses which are prime implicants of Π_{OBS}.*

$\mathcal{D}(\triangle_1, \triangle_2)$ is a kernel abductive diagnosis iff $\mathcal{D}(\triangle_1, \triangle_2) \in \mathcal{KD}$, where $\forall \neg\mathcal{D}(\triangle_i, \triangle_j) \in \Pi_{\neg OBS}$, $\forall \mathcal{D}(\triangle_x, \triangle_y) \in C_{OBS}$, such that the sets $\triangle_i \cup \triangle_x$ and $\triangle_j \cup \triangle_y$ contain no complementary pairs of AB-literals [1]

$$\mathcal{KD} = \{\mathcal{D}(\triangle_i \cup \triangle_x, \triangle_j \cup \triangle_y)\}$$

Intuitively, the kernel abductive diagnoses can be viewed as the kernel consistency-based diagnoses with extra AB-literals conjoined where necessary to ensure that the kernel abductive diagnoses are not only consistent with $SD \cup OBS$, but that they entail OBS when conjoined with SD. Thus, the kernel abductive diagnoses are augmented by the minimal abductive explanations where the two can be consistently conjoined.

Discussion

This paper contains three main contributions, all relating to the theme of diagnosis as refutation. They may be summarized as follows:

- characterization of the space of refuted diagnoses,
- identification of an intuitive computational core for consistency-based and abductive diagnosis,
- a new characterization of abductive explanation and abductive diagnosis.

[1] this condition may rephrased as:
$\{\mathcal{D}(\triangle_i, \triangle_j)\} \cup \{\mathcal{D}(\triangle_x, \triangle_y)\}$ is satisfiable

Characterization of the space of refuted diagnoses is important because we maintain that many systems that employ diagnostic reasoning as a component, are more interested in the diagnoses that have been eliminated, than in the diagnoses that may be currently conjectured. Furthermore, we contend that the generation of refuted AB-hypotheses is at the computational core of model-based diagnostic reasoning.

The identification of an intuitive computational core for consistency-based and abductive diagnosis is important because previous analyses of the complexity of model-based diagnosis have defined the source of complexity in terms of computational notions such as implicates and implicants. By understanding that the source of complexity is in generating refuted AB-hypotheses, we may find ways of making our algorithms more efficient for specific applications. For example, rather than axiomatizing notions such as mutually incompatible component labelings, single hypothesis assumptions or the impossibility of certain diagnoses, and forcing our diagnostic machine to search for and generate them as prime implicates, we might cache them separately at the outset as refuted AB-hypotheses. Similarly, if certain common observations refute AB-hypotheses without the need for further inference, they might be placed in a look-up table, for easy generation of refuted AB-hypotheses.

Finally, our new characterization of abductive explanation and abductive diagnosis extends the work by de Kleer, Mackworth and Reiter [1].

References

[1] J. de Kleer, A.K. Mackworth, and R. Reiter. Characterizing diagnoses. In *Proceedings of the National Conference on Artificial Intelligence*, pages 324–330, 1990.

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

[3] S. McIlraith and R. Reiter. On tests for hypothetical reasoning. In *Readings in Model-based Diagnosis*. Morgan Kaufmann Publishers, Inc., to appear.

[4] D. Poole. Explanation and prediction: an architecture for default and abductive reasoning. *Computational Intelligence*, 5:97–110, 1989.

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

[6] R. Reiter and J. de Kleer. Foundations for assumption-based truth maintenance systems: Preliminary report. In *Proceedings of the National Conference on Artificial Intelligence*, pages 183–188, 1987.

[7] P. Struss and O. Dressler. Physical negation - integrating fault models into the general diagnostic engine. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1318–1323, 1989.

[8] I. Lebow T. Bartee and I. Reed. *Theory and design of digital machines*. McGraw Hill, 1962.