A Theory of Diagnosis as Hypothesis Refinement

Peter Lucas
Department of Computer Science, Utrecht University
Padualaan 14
3584 CH Utrecht, The Netherlands
E-mail: lucas@cs.uu.nl

Abstract

In this paper, diagnosis is viewed as a two-stage process: domain knowledge is first interpreted in a diagnostic sense; next, observed findings are interpreted with respect to this interpreted knowledge and a given hypothesis, yielding a diagnosis. A set-theoretical framework is briefly discussed that captures this view on diagnosis; it is used to formalize various notions of diagnosis, those proposed in the literature included. Next, a theory of flexible diagnosis, called refinement diagnosis, is proposed and defined in terms of this framework. Relationships with notions of diagnosis known from the literature are investigated.

Keywords & Phrases: model-based diagnosis, theory of diagnosis, formal methods.

1 Introduction

In recent years, several theories of diagnosis have been developed, providing different foundations for diagnostic problem solving in intelligent systems. In particular, theories have been proposed which try to capture the structure of diagnosis. Diagnostic problem solving is variously described in terms of *abductive reasoning* (cf. [3, 7, 9, 10, 12]), as a specific form of *consistency-based reasoning* (cf. [13, 4, 6, 12]), or as *deductive reasoning* (cf. [1]). In the context of diagnosis, one usually speaks of *abductive diagnosis*, *consistency-based diagnosis* and *heuristic classification* [2], respectively. Abductive diagnosis is primarily used in systems incorporating causal models of normal or faulty behaviour. Consistency-based diagnosis appears especially suitable for models of normal structure and functional behaviour. Abductive and consistency-based diagnosis are often classified as *model-based* approaches to diagnosis. Heuristic classification is typically used in systems based on empirical knowledge.

Although the diagnostic frameworks mentioned above differ in several respects, in all of them diagnostic problem solving can be viewed as a special instance of hypothetical reasoning [11]. In solving a diagnostic problem, a hypothesis is first generated and next tested with respect to diagnostic knowledge and observed findings. If it passes the tests, it is accepted and called a diagnosis; when it fails to pass the tests, it is rejected. This view of diagnosis is quite general, but it is still unnecessarily restrictive. Instead of simply rejecting a hypothesis that does not comply with all requirements, it seems more natural to adjust or refine it, when possible. Then, a diagnosis obtained after refinement of a hypothesis may be viewed as the

best possible solution in a particular sense, given the domain knowledge, the set of observed findings and the hypothesis at hand. It, therefore, seems attractive to incorporate a principle of refinement into the basic definition of diagnosis, yielding notions of *refinement diagnosis*. The formalization of refinement diagnosis is the subject of this paper.

There are various reasons why refinement diagnosis may be a more appropriate basis for diagnostic problem solving than the more rigorous notions of diagnosis mentioned above:

- Real-world knowledge bases are, almost without exception, incomplete, i.e. the modelled problem domain has not been fully described. For example, knowledge of certain interactions among defects may be missing.
- Real-world knowledge bases are not completely accurate, e.g. the meaning of the domain knowledge may not have been captured sufficiently precisely, or may have been specified incorrectly.
- The findings that may be observed, and interpreted by an expert system, are only part of what might have been collected without limitations, such as available time and money.
- Part of the observed findings may be unreliable, due to impediments to the observation process, such as limited available time.

Although a model-based approach is often thought to shield the developer from such problems (cf. [13]), making simplifying assumptions will always be necessary in order to deal with real-world problems, whether a model-based approach is followed or not. It is often essential to establish a diagnosis, even when confronted with the imperfections mentioned above. In many domains, in particular medicine, it is usually better to arrive at a diagnosis that does not account for all observed findings, or that suggests findings that have not been observed, than to establish no diagnosis at all. It is sometimes said that such a diagnosis *underaccounts* or *overaccounts* for the set of observed findings.

The structure of this paper is as follows. In Section 2, a brief summary of a set-theoretical framework used to define notions of diagnosis is presented. This framework is employed to define notions of refinement diagnosis in Section 3. It is also suitable as a semantic framework for the notions of diagnosis mentioned above, as briefly discussed in Section 2. The paper is rounded off with a discussion of the achievements of the work presented in this paper.

2 A framework of diagnosis

In this section, we provide a brief overview of a set-theoretical framework of diagnosis that is used in the remainder of the paper (cf. [8]). Its underlying assumption is that diagnosis involves the interpretation of a knowledge base in terms of observable findings and possible defects. The type of knowledge represented in a knowledge base, the way in which this knowledge is interpreted in a diagnostic sense, as well as the interpretation of hypotheses and observed findings in the context of this knowledge determine the diagnoses for a problem.

2.1 The representation of interactions

Consider the following piece of medical knowledge:

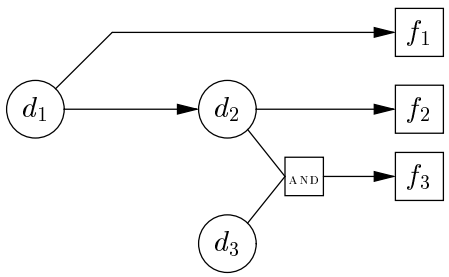

Figure 1: Causal net.

> "Influenza causes fever and infection of the trachea and bronchial tree, which causes a sore throat, but if the patient suffers from asthma, dyspnoea will occur as well."

In Figure 1, the directed graph representation of the causal knowledge embodied in this medical description is depicted, where an arc denotes a cause-effect relationship. The following meaning is ascribed to the elements of the causal graph:

$$\begin{array}{lcl} d_1 & = & \text{influenza} \\ d_2 & = & \text{tracheobronchitis} \\ d_3 & = & \text{asthma} \end{array}$$

$$\begin{array}{lcl} f_1 & = & \text{fever} \\ f_2 & = & \text{sore throat} \\ f_3 & = & \text{dyspnoea (shortness of breath)} \end{array}$$

Elements d_i are disorders; elements f_i are observable findings. Note that in the graph the disorders d_1 and d_2 are causally related to each other. Interactions among disorders can also be captured by means of a mapping of sets of disorders to sets of observable findings, yielding a diagnostic interpretation of this knowledge. Such a mapping will be called an evidence function. Since the term 'disorder' is not used in technical domains, where instead the term 'fault' is commonly employed to indicate device problems, the term '*defect*' will be used in the following to denote both disorders in medicine and faults in technical devices.

More formally, let $\Sigma = (\Delta, \Phi, e)$ be a *diagnostic specification*, where Δ denotes a set of defects, and Φ denotes a set of findings. Positive defects d (findings f) and negative defects $\neg d$ (findings $\neg f$) denote *present* defects (findings) and *absent* defects (findings), respectively. It is assumed that $\neg \circ \neg = \iota$, where ι is the identity function. If a defect d or a finding f is not included in a set, it is assumed to be unknown. Let a set X_P denote a set of positive elements, and let X_N denote a set of negative elements, such that X_P and X_N are disjoint. It is assumed that $\Delta = \Delta_P \cup \Delta_N$ and $\Phi = \Phi_P \cup \Phi_N$. The power set of a set S is denoted by $\wp(S)$. Now, an *evidence function* e is a mapping

$$e : \wp(\Delta) \to \wp(\Phi) \cup \{\bot\}$$

such that:

(1) for each $f \in \Phi$ there exists a set $D \subseteq \Delta$ with $f \in e(D)$ or $\neg f \in e(D)$ (and possibly both);

(2) if $d, \neg d \in D$ then $e(D) = \bot$;

(3) if $e(D) \neq \perp$ and $D' \subseteq D$ then $e(D') \neq \perp$.

If $e(D) \neq \perp$, it is said that $e(D)$ is the set of *observable findings* for D (D is *consistent*); otherwise, it is said that D is *inconsistent*.

According to the definition above, we may have that both $f \in e(D)$ and $\neg f \in e(D)$, which simply means that these findings may alternatively occur given the combined occurrence of the defects in the set D. In some domains it might hold that if $e(\{d\}) = e(\{d'\})$, it follows that $d = d'$, i.e. the defects d and d' are taken as synonyms for the same defect.

For the medical knowledge depicted in Figure 1, it holds, among others, that:

$$\begin{array}{llll} e(\{d_1\}) & = \{f_1, f_2\} & e(\{d_3\}) & = \varnothing \\ e(\{d_2\}) & = \{f_2\} & e(\{d_2, d_3\}) & = \{f_2, f_3\} \\ e(\{d_1, d_2\}) & = e(\{d_1\}) & e(\{d_1, d_2, d_3\}) & = \{f_1, f_2, f_3\} \end{array}$$

The property $e(\{d_i\}) \subseteq e(\{d_1, d_2\})$, $i = 1, 2$, formally expresses that the interaction between d_1 and d_2 is monotonic; the evidence function e is monotonically increasing. An evidence function may also be monotonically decreasing, or nonmonotonic. In particular, evidence functions describing functional behaviour of devices are monotonically decreasing (an example is given below).

Various semantic properties of a domain for which a diagnostic system must be built can be expressed precisely in terms of evidence functions. Local as well as global interactions between defects can be expressed readily. A typical global property of evidence functions encountered in the literature is interaction freeness (cf. [9]). A set of defects Δ is called *interaction free* iff

$$e(D) = \bigcup_{d \in D} e(\{d\})$$

for each consistent $D \subseteq \Delta$. This shows that an evidence function can be partially specified.

2.2 Notions of diagnosis

A specific evidence function provides a semantic interpretation of a knowledge base in terms of expected evidence for the combined occurrence of defects; yet, it does not yield a diagnosis. To employ an evidence function for the purpose of diagnosis, it must be interpreted with respect to actually observed findings. Such an interpretation of an evidence function and of observed findings can be viewed as a notion of diagnosis applied to solve a diagnostic problem at hand.

More formally, let $\mathcal{P} = (\Sigma, E)$ be a *diagnostic problem*, where $E \subseteq \Phi$ is the set of *observed findings*; it is assumed that if $f \in E$ then $\neg f \notin E$, i.e. contradictory observed findings are not allowed. Let R_Σ denote a notion of diagnosis R applied to Σ, then a mapping

$$R_{\Sigma, e_{|H}} : \wp(\Phi) \to \wp(\Delta) \cup \{u\}$$

will either provide a diagnostic solution for a diagnostic problem $\mathcal{P}$, or indicate that no solution exists, denoted by u (undefined). Here, H denotes a *hypothesis*, which is taken to be a set of defects ($H \subseteq \Delta$), and $e_{|H}$, called the *restricted evidence function* of e, is a restriction of e with respect to the power set $\wp(H)$:

$$e_{|H} : \wp(H) \to \wp(\Phi) \cup \{\perp\}$$

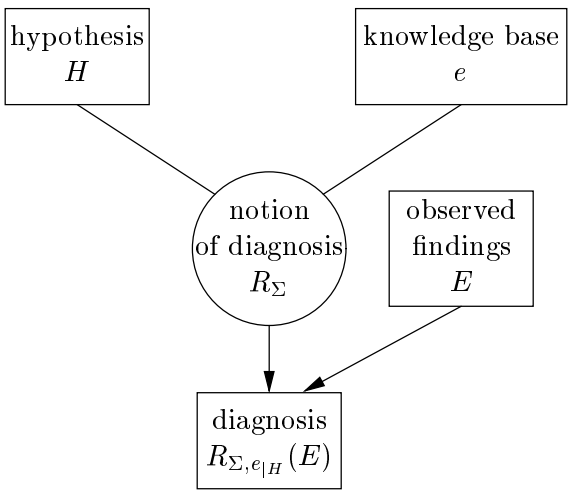

Figure 2: Schema of notion of diagnosis, diagnostic problem and solution.

where for each $D \subseteq H$: $e_{|H}(D) = e(D)$. A restricted evidence function $e_{|H}$ can be thought of as the relevant part of a knowledge base with respect to a hypothesis H. An *R-diagnostic solution*, or *R-diagnosis* for short, with respect to a hypothesis $H \subseteq \Delta$, is now defined as the set

$$R_{\Sigma,e_{|H}}(E), \text{ where } R_{\Sigma,e_{|H}}(E) \subseteq H \text{ if a solution exists.}$$

In Figure 2, the idea underlying the definition of a notion of diagnosis R and diagnostic solution to a diagnostic problem is illustrated schematically.

A notion of diagnosis R provides the possibility to express interactions among defects and observed findings at the level of diagnosis, which we call dependencies. We may also have that a hypothesis can be split up into two subhypotheses, that can be examined independently:

$$R_{\Sigma,e_{|H\cup H'}}(E) = R_{\Sigma,e_{|H}}(E) \cup R_{\Sigma,e_{|H'}}(E)$$

with $R_{\Sigma,e_{|H\cup H'}}(E) \neq u$. This means that the diagnostic solution with respect to the hypothesis $H \cup H'$ is obtained as the union of the solutions for the two separately examined hypotheses H and H'. This is called the *independence (or compositionality) assumption*. For many notions of diagnosis described in the literature, in particular for abductive diagnosis and consistency-based diagnosis, the independence assumption fails to hold.

To demonstrate how the definitions above can be employed, we consider a notion of diagnosis U, such that $U_{\Sigma,e_{|H}}(E) = H'$ if it holds that H' is the only subset of H such that $e_{|H}(H') \subseteq E$; otherwise, $H' = u$. This notion of diagnosis expresses that a diagnosis consists of a set of defects which, on the one hand, can account for at least part of all observed findings, and, on the other hand, every finding associated with the set of defects that is taken as a diagnosis has been observed. Furthermore, there is only one such subset of the given hypothesis H. Now, reconsider the medical example from Figure 1 with $H = \{d_2, d_3\}$. Some interesting diagnostic conclusions are: $U_{\Sigma,e_{|H}}(\{f_2\}) = \{d_2\}$, i.e. a patient with only sore throat has tracheobronchitis, $U_{\Sigma,e_{|H}}(\{f_2, f_3\}) = u$, i.e. there exists no unique diagnosis with respect to H accounting for both sore throat and dyspnoea as signs, and finally, $U_{\Sigma,e_{|H}}(\{f_3\}) = H$. In the first case, it is said that the hypotheses has been *adjusted*, in the second case, that the

hypothesis H is *rejected*, and in the last case, that the hypothesis H has been *accepted*. This example demonstrates the flexibility of the approach.

It is also straightforward to define notions of diagnosis proposed in the literature. For example, consider the following typical instances of notions of diagnosis:

- *Abductive diagnosis using 'must' relations (strong-causality diagnosis)* [3]:

$$\mathrm{SC}_{\Sigma,e_{|H}}(E) = \begin{cases} H & \text{if } e_{|H}(H) = E \\ u & \text{otherwise} \end{cases}$$

  i.e. it is necessary that all observable findings $e_{|H}(H)$ are observed in order to accept an hypothesis H as a diagnosis.

- *Abductive diagnosis using 'may' relations (weak-causality diagnosis)* [3, 9]:

$$\mathrm{WC}_{\Sigma,e_{|H}}(E) = \begin{cases} H & \text{if } e_{|H}(H) \supseteq E \\ u & \text{otherwise} \end{cases}$$

  i.e. all observed findings must be observable.

- *Consistency-based diagnosis* [13, 6]:

$$\mathrm{CB}_{\Sigma,e_{|H}}(E) = \begin{cases} H & \text{if } \forall f \in E : f \in e_{|H}(H) \vee \\ & \quad \neg f \notin e_{|H}(H) \\ u & \text{otherwise} \end{cases}$$

  i.e. observed findings may not contradict with observable findings.

These notions of diagnoses are the set-theoretic analogues of the corresponding notions of diagnosis mentioned above. They will be used below as reference points for notions of refinement diagnosis.

It is informative to relate these notions of diagnosis to each other in terms of a restriction relation $\sqsubseteq$; it holds that $R \sqsubseteq R'$ iff the diagnoses resulting from the notion of diagnosis R are a subset of those resulting from R' (for any legal diagnostic specification Σ). It is easily seen that: $\mathrm{SC} \sqsubseteq \mathrm{WC} \sqsubseteq \mathrm{CB}$.

3 Refinement diagnosis

The following question now arises: what can be taken as a basis for notions of diagnosis which incorporate certain principles of refinement? Obviously, there exists a wide range of possibilities. Which of the possible choices yields the most natural result depends, to a large extent, on the nature of the problem domain, which is partially expressed by the characteristics of the evidence functions e. Dependencies between a notion of diagnosis R, on the one hand, i.e. the interpretation of the set of observed findings given a specific knowledge base, and properties of a given evidence function e, on the other hand, are of importance in this respect.

Two classes of refinement diagnosis will be studied here. Firstly, the class of notions of refinement diagnosis, called *most general diagnosis*, is examined, where the least upper

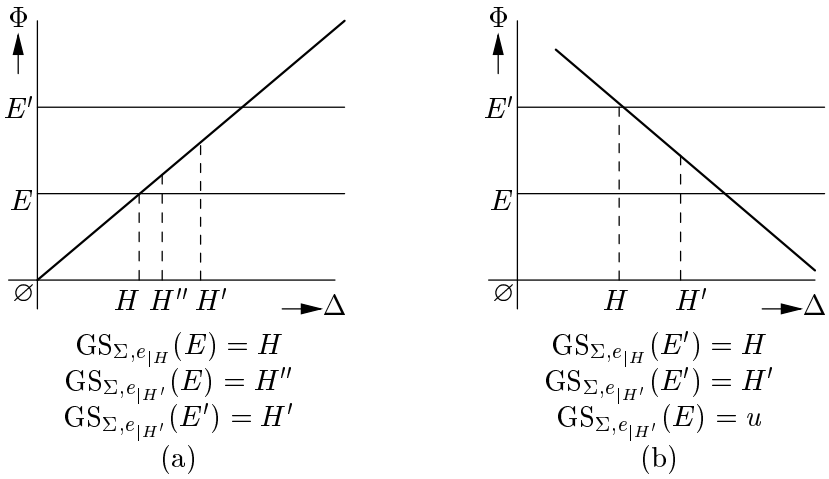

Figure 3: Monotonically increasing (a) and decreasing (b) evidence functions.

bound of accepted hypotheses (with respect to set inclusion) is taken as a diagnostic solution. Secondly, the class of notions of refinement diagnosis, called *most specific diagnosis*, based on taking the greatest lower bound of accepted hypotheses is studied. In most general diagnosis, the smallest set of defects that includes every accepted subhypothesis is considered most plausible; in contrast, in most specific diagnosis, the largest set of defects that is included in every accepted subhypothesis is considered most plausible.

3.1 Most general diagnosis

Notions of most general diagnosis capture the idea that if a specific diagnostic hypothesis is not accepted, then the 'nearest' subhypothesis should be taken instead. The least upper bound with respect to set inclusion of the set of accepted subhypotheses is an example of such a 'nearest' subhypothesis. The notions of most general diagnosis enforce independence or compositionality of diagnostic components in the sense of the previous section.

The notion of *most general subset diagnosis*, denoted by GS, is defined as follows:

$$\mathrm{GS}_{\Sigma,e_{|H}}(E) = \begin{cases} \bigcup_{\substack{H' \subseteq H \\ e_{|H}(H') \subseteq E}} H' & \text{if } H \text{ is consistent, and} \\ & \quad \exists H' \subseteq H : e_{|H}(H') \subseteq E \\ u & \text{otherwise} \end{cases}$$

Intuitively, a most general subset diagnosis is the smallest set of defects that includes all accepted subhypotheses of a given hypothesis, where an accepted subhypothesis concerns observable findings that all have been observed.

For the example in Figure 1 with $E = \{f_1, f_2\}$, we have that $\mathrm{GS}_{\Sigma,e_{|\{d_1,d_2\}}}(E) = \{d_1, d_2\}$, which is also an abductive diagnosis, because $\mathrm{SC}_{\Sigma,e_{|\{d_1,d_2\}}}(E) = \{d_1, d_2\}$. However, it holds that $\mathrm{GS}_{\Sigma,e_{|\{d_1,d_2\}}}(\{f_2\}) = \{d_2\}$, where $\mathrm{SC}_{\Sigma,e_{|\{d_1,d_2\}}}(\{f_2\}) = u$. Hence, $e(\{d_1, d_2\})$ predicts a finding that cannot be accounted for, causing the defect d_1 to be ignored. This may be a suitable approach to domains in which neglecting a particular defect may be dangerous.

In Figure 3, the relationship between diagnostic hypothesis H, the set of observed findings E and the resulting diagnosis $\mathrm{GS}_{\Sigma,e_{|H}}(E)$ is summarized by schematically depicting these sets as if they were real numbers and by taking set inclusion as the $\leq$ total order on the real

numbers. If most general subset diagnosis is applied to a monotonically decreasing evidence function, the resulting diagnosis is either undefined or equal to the given hypothesis H. This contrasts with GS applied to a monotonically increasing evidence function, which may also yield subsets of the hypothesis as a diagnosis. $\mathrm{GS}_{\Sigma,e_{|H'}}(E) = H''$ in Figure 3.(a) is intended to illustrate that $e(H'')$ may even be a superset of E. If the evidence function e is nonmonotonic, then the relationships between E and $e_{|H}(H')$ are investigated as before, but again, certain interactions between defects may be ignored.

Where most general subset diagnosis can be viewed as a more flexible version of strong-causality diagnosis SC, which for certain evidence functions is as little restrictive as consistency-based diagnosis, a similar, flexible notion of diagnosis can be designed for weak-causality diagnosis. This suggests replacing the subset relation in most general subset diagnosis by the superset relation, yielding the notion of most general superset diagnosis GO (the letter 'O' stands for 'cOntains').

The notion of *most general superset diagnosis*, denoted by GO, is defined as follows:

$$\mathrm{GO}_{\Sigma,e_{|H}}(E) = \begin{cases} \bigcup_{\substack{H' \subseteq H \\ e_{|H}(H') \supseteq E}} H' & \text{if } H \text{ is consistent, and} \\ & \quad \exists H' \subseteq H : e_{|H}(H') \supseteq E \\ u & \text{otherwise} \end{cases}$$

Most general superset diagnosis has much in common with weak-causality diagnosis WC defined in the previous section. If the notion of most general superset diagnosis is applied to evidence functions that are monotonically decreasing, or nonmonotonic, for the resulting diagnosis $\mathrm{GO}_{\Sigma,e_{|H}}(E) = H'$ it may even hold that $e(H') \subset E$, although for each of the diagnostic hypotheses $H'' \subseteq H$ that contribute to the diagnosis it holds that $e_{|H}(H'') \supseteq E$. Hence, the situation is the reverse of that for most general subset diagnosis discussed above, as might be expected from their respective definitions. In Figure 4, the various possibilities are schematically depicted.

As is true for weak-causality diagnosis WC, most general superset diagnosis restricted to monotonically increasing evidence functions is very unrestrictive, which is revealed by the fact that $\mathrm{GO}_{\Sigma,e_{|H}}(\varnothing) = H$ if $e(H) \neq \bot$, meaning that all defects constituting the hypothesis may have occurred, even if no findings have been observed. Note that the same diagnosis would have been produced by weak-causality diagnosis WC in this case. By adopting some criterion of parsimony, such as minimality according to set inclusion, the unrestrictiveness is alleviated; the empty diagnosis $\varnothing$ would then be produced.

An alternative to the definition of subset diagnosis is to consider all sets of defects D that have at least one finding f in common with the findings E observed. This leads to the following definition of the notion of *most general intersection diagnosis*, denoted by GI:

$$\mathrm{GI}_{\Sigma,e_{|H}}(E) = \begin{cases} \bigcup_{\substack{H' \subseteq H \\ (E = \varnothing \vee e_{|H}(H') = \varnothing \vee \\ e_{|H}(H') \cap E \neq \varnothing)}} H' & \text{if } H \text{ is consistent, and } (E = \varnothing \text{ or} \\ & \quad \exists H' \subseteq H : e_{|H}(H') = \varnothing \text{ or} \\ & \quad e_{|H}(H') \cap E \neq \varnothing) \\ u & \text{otherwise} \end{cases}$$

If the sets of observed and observable findings are nonempty, intersection diagnosis with respect to H stands for the least upper bound of subsets of defects of $H \subseteq \Delta$, where for each subset of defects H' admitted to the most general intersection diagnosis $\mathrm{GI}_{\Sigma,e_{|H}}(E)$, the associated set of observable findings $e_{|H}(H')$ is empty or has at least one finding in common with the set of observed findings E.

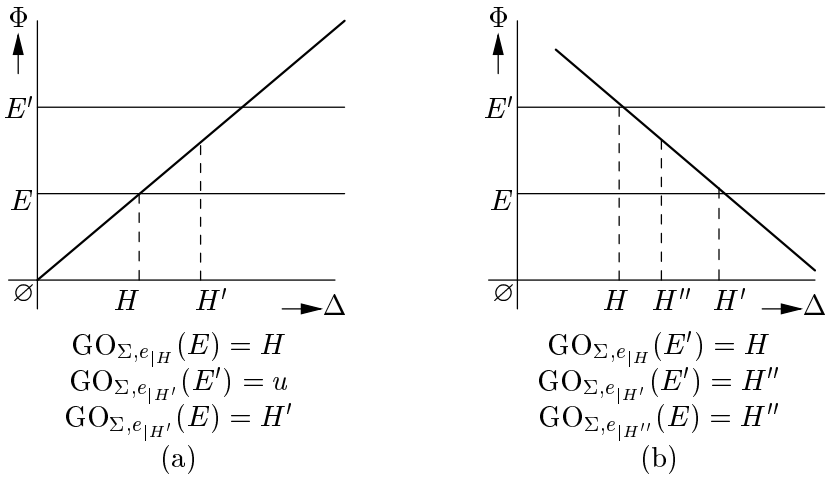

Figure 4: Monotonically increasing (a) and decreasing (b) evidence functions.

The advantage of most general intersection diagnosis over most general subset and superset diagnosis is that only defects that have at least one associated observable finding that has actually been observed, are included in a diagnosis. This will be an acceptable assumption in a domain where not all findings associated with a set of defects need be observed and not all observed findings need be accounted for. In representing a domain, it may be required to restrict to those observable findings that are in some way 'typical' for the defects.

Most general intersection diagnosis can be viewed as a refinement version of a mixture of the notions of weak-causality and strong-causality diagnosis.

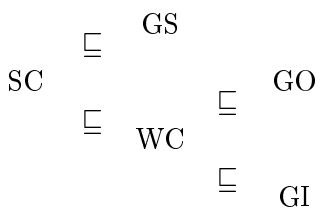

Figure 5: Restriction taxonomy of notions of diagnosis.

3.2 Comparison

Most general subset, superset and intersection diagnosis are three refinement approaches to diagnosis. The restriction relationships between these notions of diagnosis are shown in Figure 5. For most general subset diagnosis, all findings associated with a set of defects must be observed if the set of defects is to be included as part of a diagnosis. Most general superset diagnosis focusses on common findings of defects. For most general intersection diagnosis, at least one finding associated with a defect must be observed if the defect is to be included as part of a diagnosis. Notions of diagnosis can also be classified in terms of elements included in individual diagnoses using set inclusion; the subdiagnostic relation $\trianglelefteq$ does exactly this. The three notions of diagnosis discussed above stand in a subdiagnostic relation to each other:

$\mathrm{GS} \trianglelefteq \mathrm{GI}$

$\mathrm{GO} \trianglelefteq \mathrm{GI}$

This follows from the fact that if a set of observed findings is included in the set of observable findings associated with a set of defects, or vice versa, the intersection of the set of observed findings and observable findings is nonempty, given that neither the set of observed findings E, nor the set of observable findings $e_{|H}(H')$, is empty. For the empty cases, the most general intersection diagnosis is always equal to the largest result with respect to set inclusion of GO and GS. Hence, a most general intersection diagnosis will always contain at least as many elements as most general subset and superset diagnosis.

3.3 Most specific diagnosis

Rather than taking the least upper bound of a set of accepted subhypotheses of a given hypothesis, taking the greatest lower bound provides another approach to refinement diagnosis. We shall refer to notions of diagnosis based on taking the greatest lower bound as notions of *most specific diagnosis*. Where the concept of most general diagnosis formalizes notions of diagnosis that yield diagnoses that include *every* accepted subhypothesis, most specific diagnosis formalizes notions of diagnosis that yield diagnoses that are *common* to every accepted subhypothesis. In general it holds for a notion of most specific diagnosis S that if $S_{\Sigma,e_{|H}}(E) = \varnothing$ and $S_{\Sigma,e_{|H'}}(E) = H''$, then, by definition, $S_{\Sigma,e_{|H\cup H'}}(E) = \varnothing$. Hence, notions of most specific diagnosis are very restrictive.

As with the notion of most general subset diagnosis, in the notion of most specific subset diagnosis, subhypotheses are admitted to a diagnosis if their associated sets of findings are included in the set of observed findings of a diagnostic problem. However, of these accepted subhypotheses, only the defects the subhypotheses have in common constitute a diagnosis. Hence, the notion of *most specific subset diagnosis*, denoted by SS, is defined as follows:

$$\mathrm{SS}_{\Sigma,e_{|H}}(E) = \begin{cases} \bigcap_{\substack{H' \subseteq H \\ e_{|H}(H') \subseteq E}} H' & \text{if } H \text{ is consistent, and} \\ & \quad \exists H' \subseteq H : e_{|H}(H') \subseteq E \\ u & \text{otherwise} \end{cases}$$

This notion of diagnosis is extremely restrictive. For example, if an evidence function is interaction free, then the most specific subset diagnosis will almost always (with the exception when only one subhypothesis is accepted) be equal to the empty set.

If the evidence function is monotonically decreasing, then most specific subset diagnosis tries to construct the smallest diagnosis possible. It may be view as a flexible form of kernel, consistency-based diagnosis in the sense of [6]. The reason for the similarity between kernel diagnosis in consistency-based diagnosis and most specific subset diagnosis is that any hypothesis H' for which $e_{|H}(H') \subseteq E$ is also consistent with E.

The correspondence between kernel diagnosis and most specific subset diagnosis is illustrated by an example taken from [6]. Consider Figure 6, which depicts an electronic circuit with three multipliers, referred to as M_1, M_2 and M_3, and two adders, denoted by A_1 and A_2. Let $\Sigma = (\Delta, \Phi, e)$ be a diagnostic specification representing the circuit. The fact that some multiplier M_i is defective, is denoted by m_i; if it is nondefective, this is indicated by $\neg m_i$. A similar notational convention is adopted with regard to the two adders. It is convenient to assume that the input to the circuit is fixed (as assumed in [5] and [6]), as indicated in Figure 6. The normal output of the circuit, $O_1 = 12$ and $O_2 = 12$, is denoted by o_1 and o_2; abnormal output is denoted by $\neg o_j$, $j = 1, 2$.

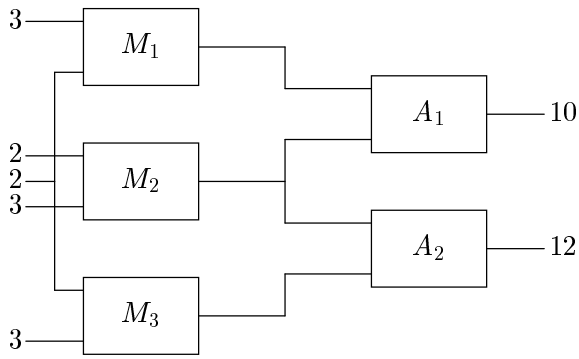

Figure 6: Multiplier-adder circuit.

The following values of the evidence function e are among those that correspond to the circuit's normal behaviour:

$$\begin{aligned}
e(\{\neg m_1, \neg m_2, \neg m_3, \neg a_1, \neg a_2\}) &= \{o_1, o_2\} \\
e(\{\neg m_1, \neg m_2, \neg m_3, a_1, \neg a_2\}) &= \{o_2\} \\
e(\{\neg m_1, \neg m_2, \neg m_3, a_1, a_2\}) &= \varnothing \\
e(\{\neg m_1, \neg m_2, \neg m_3, a_1\}) &= \{o_2\} \\
e(\{a_1\}) &= \{o_2\} \\
e(\{\neg m_1, \neg m_2, \neg m_3\}) &= \{o_1, o_2\} \\
&\vdots \\
e(\varnothing) &= \{o_1, o_2\}
\end{aligned}$$

($e(\varnothing)$ denotes that it is unknown whether defects are present or absent.) The most specific subset diagnosis with respect to the hypothesis $H = \{a_1\}$ is equal to

$$\mathrm{SS}_{\Sigma, e_{|\{a_1\}}}(\{\neg o_1, o_2\}) = \{a_1\}$$

which is indeed a kernel diagnosis for the diagnostic problem $\mathcal{P} = (\Sigma, E)$ using consistency-based diagnosis. Note that

$$\mathrm{SS}_{\Sigma, e_{|H}}(\{\neg o_1, o_2\}) = \{a_1\}$$

if $a_1 \in H$, for example, $H = \{\neg m_1, \neg m_2, \neg m_3, a_1, \neg a_2\}$.

As discussed above, most general superset diagnosis will often yield a diagnosis that contains too many defect elements, in particular when an evidence function is monotonically increasing. Most specific superset diagnosis is a more restrictive, and possibly more suitable, notion of diagnosis than most general superset diagnosis.

The notion of *most specific superset diagnosis*, denoted by SO, is defined as follows:

$$\mathrm{SO}_{\Sigma, e_{|H}}(E) = \begin{cases} \bigcap_{\substack{H' \subseteq H \\ e_{|H}(H') \supseteq E}} H' & \text{if } H \text{ is consistent, and} \\ & \quad \exists H' \subseteq H : e_{|H}(H') \supseteq E \\ u & \text{otherwise} \end{cases}$$

If the evidence function to which most specific superset diagnosis is applied, is monotonically increasing, the result may be intuitively attractive. The basic idea of most specific superset diagnosis is that the observed findings that are common to the accepted subhypotheses are due to the common defects of accepted subhypotheses.

Reconsider Figure 1. For $E = \{f_2, f_3\}$ (i.e. the patient has a sore throat and dyspnoea), the most specific superset diagnosis is equal to

$$\mathrm{SO}_{\Sigma, e_{|\{d_1,d_2,d_3\}}}(E) = \{d_3\}$$

because, it holds that $e_{|H}(\{d_1, d_3\}) \supseteq E$, $e_{|H}(\{d_2, d_3\}) \supseteq E$ and $e_{|H}(\{d_1, d_2, d_3\}) \supseteq E$, where $H = \{d_1, d_2, d_3\}$. All other subsets of H have associated sets of findings that are no supersets of E. The defect d_3 stands for asthma. While both d_1 and d_2 participate in subhypotheses that also account for E, only the defect d_3 occurs in all accepted subhypotheses, i.e. it turns out to be essential. It seems therefore intuitively right to accept d_3 as the most plausible diagnosis.

As the example above indicates, a most specific superset diagnosis need not account for all observed findings on the basis of the given evidence function. If an evidence function is interaction free, then most specific superset diagnosis is likely to produce a singleton set diagnosis for a given hypothesis that is very plausible if the associated sets of observed findings $e(\{d\})$ are mutually disjoint.

As discussed above, the notion of most general intersection diagnosis is very unrestrictive. All defects that, either individually or in combination with other defects, have findings in common with the set of observed findings, are included in a diagnosis. The notion of *most specific intersection diagnosis*, denoted by SI, is much more restrictive than most general intersection diagnosis; it is defined as follows:

$$\mathrm{SI}_{\Sigma, e_{|H}}(E) = \begin{cases} \bigcap_{\substack{H' \subseteq H \\ (E = \varnothing \vee e_{|H}(H') = \varnothing \vee \\ e_{|H}(H') \cap E \neq \varnothing)}} H' & \text{if } H \text{ is consistent, and } (E = \varnothing \text{ or} \\ & \quad \exists H' \subseteq H : e_{|H}(H') = \varnothing \text{ or} \\ & \quad e_{|H}(H') \cap E \neq \varnothing) \\ u & \text{otherwise} \end{cases}$$

If the evidence function e is monotonically increasing, the resulting diagnosis will be equal to the empty set if the function values $e(\{d\})$ have many observable findings in common.

3.4 Comparison

Although the notions of most specific diagnosis are very restrictive, they do not stand in a simple restriction relation to the other notions of diagnosis. However, it is easy to see that

$$\mathrm{SS}_{\Sigma, e_{|H}}(E) \subseteq \mathrm{GS}_{\Sigma, e_{|H}}(E)$$

holds for each consistent $H \subseteq \Delta$. Similar set inclusion relations hold for the other notions of diagnosis. We state without proof that:

$\mathrm{SS} \trianglelefteq \mathrm{GS}$
$\mathrm{SO} \trianglelefteq \mathrm{GO}$
$\mathrm{SI} \trianglelefteq \mathrm{GI}$

4 Discussion

In this paper, we introduced a general set-theoretical framework as a tool for the formalization of notions of diagnosis. As was shown, the particular properties of evidence functions to which a notion of diagnosis is applied, are important with respect to the appropriateness of a notion

of diagnosis. Several new notions of diagnosis have been proposed that are less rigorous in dealing with observed findings and evidence functions than common notions of diagnosis. These are certainly not the only notions of diagnosis that may be useful in certain domains.

There are a number of ways in which the notions of diagnosis discussed above might be enhanced. In the formalization of a diagnostic problem in Section 2, findings associated with a set of defects were just listed, without making an explicit distinction between those findings that are important and those that are not. However, the set of findings may be subdivided into subsets according to several, not necessarily mutually exclusive, criteria, taken as measures of the 'importance' or relevance of findings. Two examples of such criteria are:

- Frequency of occurrence: some findings may always be present given a set of present or absent defects, while others may only be observed occasionally.
- Discriminatory power: the observation of a finding associated with some set of defects, but not with other sets, makes the occurrence of that particular set of defects more likely than the occurrence of the other sets. In medicine, findings with high discriminatory power are known as *pathognomonic* findings.

Many other criteria are possible. These criteria could be incorporated into our notions of diagnosis by decomposing an evidence function into several different evidence functions with different meanings.

Finally, it is desirable to gather experimental evidence for the usefulness of refinement diagnosis with respect to real-world diagnostic applications. An experimental comparison of the results produced by the various notions of diagnosis in different applications will produce more insight into the applicability of the theory, and may suggest improvements.

References

[1] B.G. Buchanan and E.H. Shortliffe (1984). *Rule-based Expert Systems: the MYCIN Experiments of the Stanford Heuristic Programming Project.* Reading: Addison-Wesley.

[2] W.J. Clancey (1985). Heuristic classification. *Artificial Intelligence*, **27**, 289–350.

[3] L. Console, D. Theseider Dupré and P. Torasso (1989). A theory of diagnosis for incomplete causal models. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, pp. 1311–1317.

[4] L. Console and P. Torasso (1990). Integrating models of correct behaviour into abductive diagnosis. In *Proceedings of ECAI'90*, pp. 160–166.

[5] R. Davis and W. Hamscher (1988). Model-based reasoning: troubleshooting. In *Exploring Artificial Intelligence: Survey Talks from the National Conference on Artificial Intelligence* (H.E. Shrobe, ed.), pp. 297–346. San Mateo, California: Morgan Kaufmann.

[6] J. de Kleer, A.K. Mackworth and R. Reiter (1992). Characterizing diagnoses and systems. *Artificial Intelligence*, **52**, 197–222.

[7] K. Konolige (1994). Using default and causal reasoning in diagnosis. *Annals of Mathematics and Artificial Intelligence*, **11**, 97–135.

[8] P.J.F. Lucas (1996). Modelling interactions for diagnosis. In *Proceedings of CESA'96 IMACS Multiconference: modelling, analysis and simulation*, **1**, pp. 541–546, Lille, France.

[9] Y. Peng and J.A. Reggia (1990). *Abductive inference models for diagnostic problem solving.* New York: Springer-Verlag.

[10] D. Poole, R. Goebel and R. Aleliunas (1987). Theorist: a logical reasoning system for defaults and diagnosis. In *The Knowledge Frontier* (N. Cercone and G. Mc Calla, eds.), Berlin: Springer-Verlag.

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

[12] D. Poole (1994). Representing diagnosis knowledge. *Annals of Mathematics and Artificial Intelligence*, **11**, 33–50.

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