# On Analytical and Similarity-Based Classification

**Marc Vilain, Phyllis Koton, and Melissa P. Chase**
The MITRE Corporation
Burlington Road, Bedford, MA 01730
Internet: {mbv, koton, pc}@linus.mitre.org

**Abstract**

This paper is concerned with knowledge representation issues in machine learning. In particular, it presents a representation language that supports a hybrid analytical and similarity-based classification scheme. Analytical classification is produced using a KL-ONE-like term-subsumption strategy, while similarity-based classification is driven by generalizations induced from a training set by an unsupervised learning procedure. This approach can be seen as providing an inductive bias to the learning procedure, thereby shortening the required training phase, and reducing the brittleness of the induced generalizations.

## Introduction

Classification is a central concern of knowledge representation and machine learning. At the heart of many knowledge representation systems is a classification procedure which determines where an individual fits within the knowledge base. These classification procedures are realized in various ways, for example in rule-based frameworks, or with the term subsumption strategy of KL-ONE. They are all, however, analytic methods. A major problem for these analytic methods is that they require a large knowledge base to guide classification; typically this knowledge base is constructed by hand.

The machine learning community's concern with classification addresses the above problem by automatically acquiring classification schemes from a collection of examples. Although various techniques have been developed, such as inductive learning and case-based reasoning, all can be thought of as statistical classification mechanisms. A major problem with these methods is their "example complexity," the large number of training examples required if one wants to induce a classification that distinguishes unusual classes while ensuring that typical classes are recognized as such. Further, if the classification method is incremental, it tends to be sensitive to the order in which the examples are presented; to recover from a poor classification typically requires a large number of "normal" cases.

The extent of this example complexity was measured by Aghassi (1990) in the context of the Heart Failure program, a model-based expert system that diagnoses patients with heart failure (Long et al. 1987). Aghassi estimates that in this moderately complex domain, a case-based classifier with no prior background knowledge would need to be trained with as many as 100,000 cases to provide reasonably accurate similarity classifications.

We believe that the approaches to classification taken by the knowledge representation and machine learning communities complement each other and can be fruitfully combined: express the normal cases within an analytical classification framework and use the statistical classification procedure to identify the exceptional cases. To accomplish this requires a hybrid knowledge representation that combines analytical and contingent languages.

In the following sections we describe a machine learning classification system and show how analytic knowledge can be incorporated within it, then we describe the hybrid knowledge representation, and discuss the relationship between this work and the machine learning community's concept of inductive bias.

## A Hybrid Learning System

### Background: Dynamic Memory

We have previously described a medical diagnosis program that used case-based reasoning to construct a knowledge base for classifying cases of heart failure (Koton 1988). The knowledge base, in the form of a *dynamic memory* (Kolodner 1983), was constructed entirely by the program, using the cases that had been presented to it, and using assistance from the Heart Failure program.

A dynamic memory records instances of *cases*, organizing them in a hierarchy. Cases are made up of attribute-value pairs (or *features*). The memory also applies a learning procedure to create *generalizations*, frames that record the similarities between a group of cases. Each generalization maintains a list of *norms*, the features that are common to most of its descendants in the hierarchy[1]. The descendant of a generalization need not share all of the generalization's norms, in which case it is said to be differentiated from the generalization by its distinguishing attributes.

[1] In many dynamic memory systems, "most" is implemented as ≥2/3.

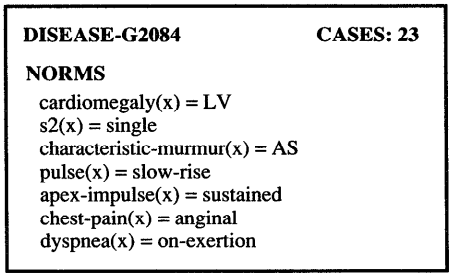


**Figure 1:** A generalization frame.

As a result of seeing many instances of patients with similar sets of symptoms, our program created generalization categories that associated certain symptoms with other symptoms. For example, the symptom *syncope on exertion* identifies a disease category containing 23 of the 45 cases presented to the program. Patients who exhibited syncope on exertion also had other features in common, for example, chest pain, a heart murmur, and an enlarged heart (Figure 1). Diagnosing these patients using the Heart Failure program revealed that 16 of them had the disease aortic stenosis. What our program had done, in effect, was to create a category describing the common symptoms of aortic stenosis, without having any previous knowledge of this disease.

However, since the memory construction was guided statistically, many "typical" cases of aortic stenosis had to be presented so that unusual cases would not produce an inaccurate classification scheme. As suggested above, this requirement could have been reduced by first defining some "normal" classifications analytically.

## Representing analytical and contingent knowledge

To allow for analytical knowledge, we have extended the dynamic memory representation with an analytic term-subsumption language in the style of KL-ONE. This language is used to define salient categories of a domain — for medical diagnosis, these consist of important combination of symptoms indicating the likelihood of a disease. The statistical associations induced by the case-based learning procedure are then used to extend the analytic framework with contingent knowledge, further refinements of the framework, and exceptions[2]. The representational challenge is in meaningfully combining the first-order analytical definitions with the contingent knowledge, which is neither analytical nor first-order.

As an example of this hybrid representation, consider the representation fragment shown in Figure 2. This fragment contains two preconstructed categories, *likely-cardiac-disease*, and *likely-valve-disease*. For the first,

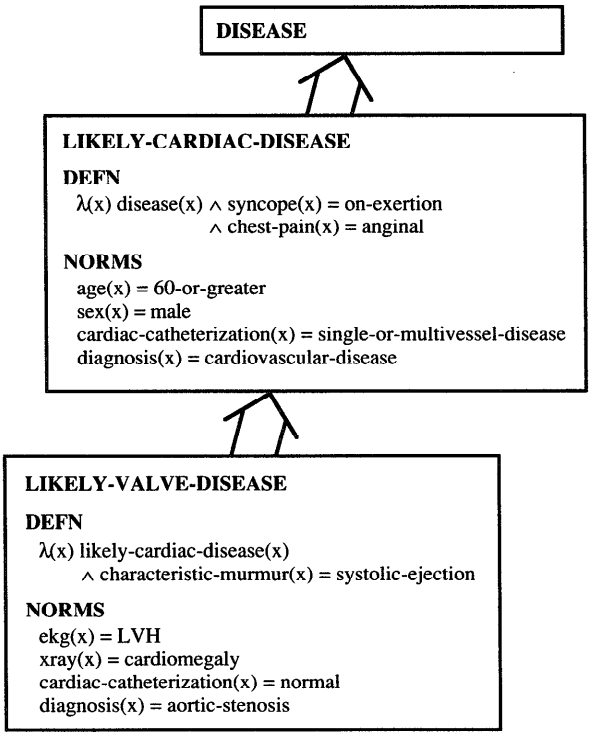


**Figure 2:** A representation fragment.

textbook knowledge tells us that symptoms of syncope on exertion and chest pain are indicators of the likelihood of heart disease. This is captured in the representation by giving the category *likely-cardiac-disease* the definition[3] $\lambda x$ (*syncope*($x$) = *on-exertion*)$\wedge$(*chest-pain*($x$) = *anginal*). In addition, the representation of this category indicates a number of *norms*, contingent properties which tend to co-occur or co-vary with syncope and chest pain (though not necessarily in the strict statistical sense), for example, age > 60, sex = male, and single- or multi-vessel disease on cardiac catheterization. In fact, the probability of syncope and chest pain being caused by cardiovascular disease is sufficiently high that the *likely-cardiac-disease* category includes that diagnosis as a norm. Norms are interpreted as defaults, so this knowledge structure does *not* imply that all patients with syncope and chest pain are males over 60 with cardiovascular disease, but that these symptoms are found primarily in older males and are caused by cardiovascular disease.

The default nature of norms is evidenced by the category *likely-valve-disease*, which is analytically differentiated from the category of patients with syncope and chest pain by the presence of the symptom *systolic ejection murmur*. Here, the default diagnosis of cardiovascular

[2]Much like a medical student at the beginning of the third year, whose book knowledge becomes extended by clinical experience.

[3]Properly speaking, this definition actually denotes something like *medical-case-whose-symptoms-are-likely-indicators-of-cardiac-disease*. We are using shorter names for their obvious legibility benefits!

disease is overridden by the default diagnosis of aortic stenosis. Also, we find a different set of symptoms that covary with syncope, chest pain, and murmur, for example, left ventricular hypertrophy on EKG and an enlarged heart on x-ray.

### Incorporating specific cases

The norms in the preceding examples could have been defined by the user as part of the knowledge acquisition process. More interestingly, they could equally have been induced by the dynamic memory learning procedure from a set of training cases, as with the generalization in Figure 1. The learning procedure can thus be seen as a mechanism for acquiring the contingent knowledge encoded in the representation. Additionally, by indexing training cases into the analytical hierarchy, the learning procedure extends the hierarchy with (contingent) refinements and exceptions.

The learning procedure indexes training cases in the hierarchy by matching them to existing categories. For example, suppose the system is presented with a case $\chi$ which is described by the features *syncope=on-exertion*, *chest-pain=anginal*, *murmur=systolic-ejection*, *sex=male*, and *age=30-or-less*. The values for attributes *syncope*, *chest-pain*, and *murmur* identifies this case as an exemplar of the category *likely-valve-disease*. However, the value *30-or-less* for attribute *age* distinguishes this case from the category *likely-valve-disease* because the majority of cases in that category have value *60-or-more* for *age* (a norm inherited from *likely-cardiac-disease*). The new case is thus inserted into the hierarchy directly underneath the *likely-valve-disease* frame, differentiating it by providing a different value to the *age* attribute.

When multiple cases are indexed to the same place in the hierarchy, the learning procedure creates a generalization capturing their common features as a set of norms. The cases are then indexed below the new generalization, differentiated from it by the features they fail to share. It is this process which inductively extends the analytical framework with contingent knowledge derived from the training cases.

## The Analytical Language

Our analytical language is an extremely simple definitional frame language, which is based on the term subsumption strategy of KL-ONE. The language provides three frame-forming operators which are used to form complex frame terms: definitions are performed by naming the resulting terms.

Our first operator, *AND*, simply conjoins frame terms. We interpret the expression (*AND* $\varphi_1\ \varphi_2\ \ldots\ \varphi_n$) as

$\lambda\, x\ \varphi_1(x) \wedge \varphi_2(x) \wedge \ldots \wedge \varphi_n(x)$

The second operator, *ATTR*, restricts a frame to take a certain value for an attribute. We thus intepret the expression (*ATTR* $\alpha\ \beta$) as

$\lambda\, x\ \alpha(x\ \beta)$

We will adopt a restriction common in the machine learning classification literature, and treat attributes as functions. Therefore, *ATTR* expressions such as the one above can additionally be read as $\lambda\, x\ \alpha(x) = \beta$

Additionally, it is often necessary to form the disjunction of attribute selections, for which we use our third frame-forming operator, *ATTR**. We thus interpret the expression (*ATTR** $\alpha_1\ \beta_1\ \ldots\ \alpha_n\ \beta_n$) as

$\lambda\, x\ (\alpha_1(x) = \beta_1) \vee \ldots \vee (\alpha_n(x) = \beta_n)$

Analytical definition are created by naming a complex term. In the heart disease domain, we might define likely cases of valve disease as follows

*LIKELY-VALVE-DISEASE* ≡
(*AND* *LIKELY-CARDIAC-DISEASE*
(*ATTR* *CHARACTERISTIC-MURMUR*
*SYSTOLIC-EJECTION*))

As usual, such definitions as these are interpretable as universally quantified biconditionals, in this case

$\forall x$ *LIKELY-VALVE-DISEASE* $(x) \Leftrightarrow$
*LIKELY-CARDIAC-DISEASE*$(x) \wedge$
*CHARACTERISTIC-MURMUR* $(x) =$
*SYSTOLIC-EJECTION*

As with other KL-ONE derivatives, classification in our analytical language is performed by term subsumption. That is, frames are organized in a hierarchy, with frame $\varphi_1$ placed below frame $\varphi_2$ just in case the meaning of $\varphi_1$ is entailed by that of $\varphi_2$, i.e., just in case the sentence $\forall x\ \varphi_1(x) \Rightarrow \varphi_2(x)$ is valid. We then say that $\varphi_1$ is subsumed by $\varphi_2$. In effect, $\varphi_1$ is (non-trivially) subsumed by $\varphi_2$ just in case $\varphi_1$ satisfies the definition of $\varphi_2$ and additionally possesses one or more attribute assignments not valid for $\varphi_2$. Classifying a frame $\varphi$ simply consists of finding those frames in the hierarchy that immediately subsume $\varphi$. Despite the extreme simplicity of this language, frame classification in general can be shown to be NP-complete, by a straightforward reduction from 3-SAT.

However, for the knowledge bases of interest to hybrid classification, we will be concerned only with definitions in the normal form

$\gamma_1 \equiv$ (*AND* $\gamma_2$ (*ATTR* $\alpha_1\ \beta_1$) ... (*ATTR* $\alpha_i\ \beta_i$)
(*ATTR** $\alpha_j\ \beta_j \ldots\ \alpha_n\ \beta_n$))

where $\gamma_1$ and $\gamma_2$ are categories, the $\alpha$ terms are attributes, and the $\beta$ terms are values. This normal form is of interest as it characterizes the generalizations induced by the learning procedure. As we shall see below, for knowledge bases of such definitions, analytical classification is no longer NP-complete, but tractable.

## The Contingent Language

The term hierarchy defined in the analytical language provides a framework within which to express the contingent associations derived by the case-based learning procedure. In essence, these associations describe the covariances of attribute-value assignments that hold for a particular generalization. These covarying feature assignments can be interpreted as mutually providing evidence for each other: if any single one of them holds, the others can be assumed to do so as well, at least in the context of the same generalization.

Consider, for example, two generalizations $\gamma_1$ and $\gamma_2$, in which $\gamma_2$ differs from $\gamma_1$ by assigning to some attributes $\alpha_1, \alpha_2 \ldots, \alpha_n$ the values $\beta_1, \beta_2, \ldots, \beta_n$. If we know that, for some entity $e$, $\gamma_1(e) \wedge \alpha_1(e) = \beta_1$, then if we additionally choose to believe $\gamma_2(e)$, we would assume $\alpha_2(e) = \beta_2, \ldots, \alpha_n(e) = \beta_n$ by default, as these feature assignments covary with $\alpha_1(e) = \beta_1$. The process of making assumptions in this way captures the essence of similarity-based classification. Knowing that $e$ is similar in some way to instances of $\gamma_2$ allows us to infer that $e$ might possibly be an instance of $\gamma_2$, and thereby possess other features common to the class.

In earlier papers ((Koton & Chase 1989), (Koton, Chase, & Vilain 1990)), we applied Reiter's default logic to model similarity-based classification in case memory. This treatment can be extended to encompass analytical classification as well. We do so by interpreting the generalizations formed by the case classification procedure with the following two axiom schemata, one analytical and the other contingent. Let $\gamma_1$ and $\gamma_2$ be generalizations that, as above, differ in $\gamma_1$'s assigning values $\beta_1 \ldots \beta_n$ to attributes $\alpha_1 \ldots \alpha_n$. Then

$$\gamma_2 \equiv (AND\ \gamma_1\ (ATTR*\ \alpha_1\ \beta_1\ \ldots\ \alpha_n\ \beta_n)) \qquad (link)$$

$$\frac{\gamma_2(x) : \alpha_1(x)=\beta_1}{\alpha_1(x)=\beta_1}, \ldots, \frac{\gamma_2(x) : \alpha_n(x)=\beta_n}{\alpha_n(x)=\beta_n} \qquad (norm)$$

The link schema creates an analytical definition for $\gamma_2$, requiring it to differ from $\gamma_1$ by at least one of the assignments to attributes $\alpha_1 \ldots \alpha_n$. The norm schema captures the covariance of the $\alpha_i$ by introducing a normal default rule (Reiter 1980) for each assignment of some $\beta_i$ to the corresponding $\alpha_i$.

It is easy to see how these axiom schemata enable similarity-based classification. For example, say that for some particular $\gamma_1, \alpha_1, \beta_1$, *etc*, an entity $e$ is described by the theory $\theta(e) = \gamma_1(e) \wedge (\alpha_1(e) = \beta_1) \wedge \ldots$. Then $\gamma_2(e)$ is true from the biconditional interpretation of the link schema, and the remaining $\alpha_i(e) = \beta_i$ become true by default, so long as they are consistent with $\theta(e)$.

The contingent language of our representation scheme is just the language of normal default rules that have the form of the *norm* schema. However, there is more to the representation of contingent reasoning than the specification of those normal defaults that express the case memory covariances induced by the learning procedure. The hierarchies induced by the learning procedure typically require the cancellation of feature assignments on the basis of specificity (as with the assignments to *diagnosis* in Figure 3). With a default encoding of property inheritance, this leads to case memory theories having multiple extensions, only the most *specific* of which is legitimate (e.g., $\theta_2$ in Figure 3). We must therefore indicate how to interpret this non-monotonic aspect of property inheritance.

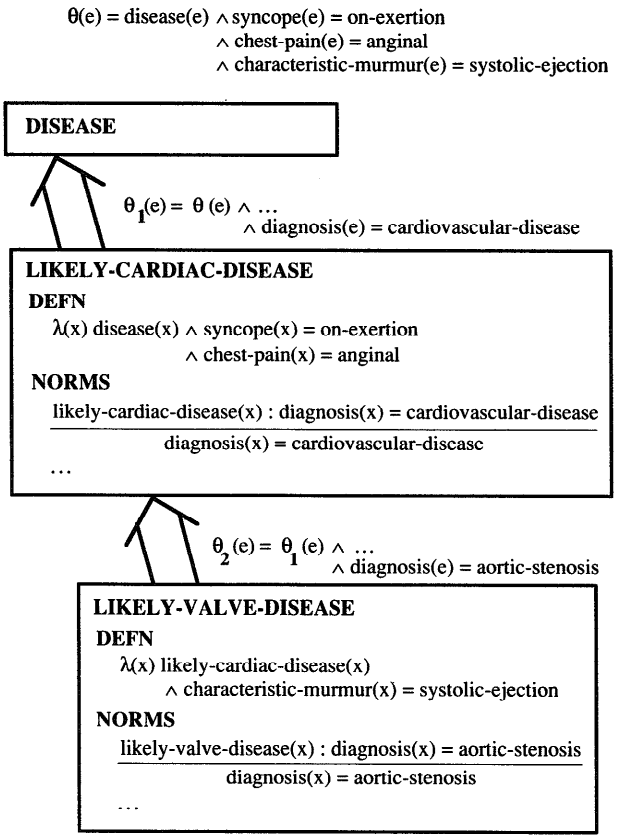


**Figure 3:** A memory fragment showing cancellation.

## Understanding Cancellation

In default logic, the traditional approach towards handling cancellation in non-monotonic inheritance is through semi-normal defaults ((Etherington & Reiter 1983), (Reiter & Criscuolo 1983)). This solution is unappealing for several reasons. For one, it potentially requires encoding the topology of the entire inheritance network into each norm default. The global reasoning task this presupposes is a poor model of inheritance with cancellation, which is typically understood as a local

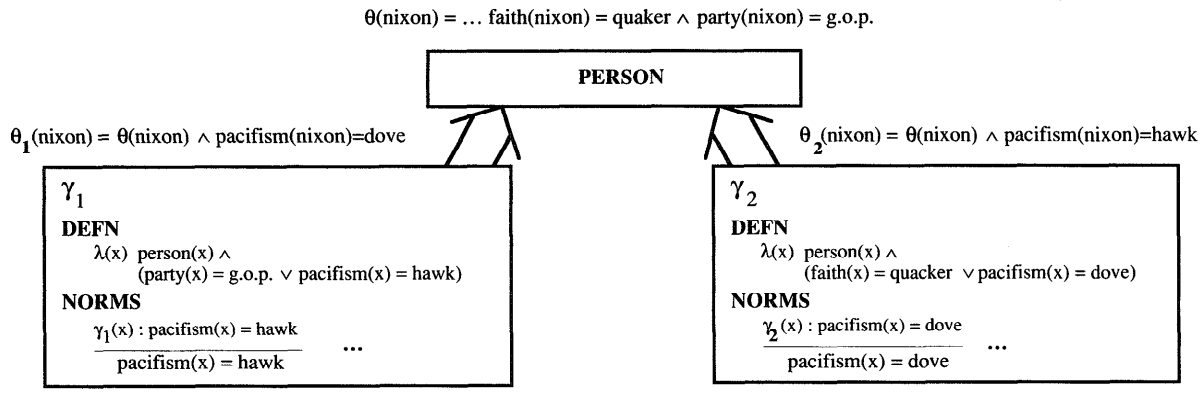


**Figure 4:** The infamous Nixon diamond.

reasoning process. Additionally, with a semi-normal encoding of cancellation, the default characterizations of memory structures induced by the learning procedure are not necessarily ordered in Etherington's sense (Koton, Chase, & Vilain 1990). This means that they can not be guaranteed an extension by existing default proof theories (Etherington 1988). As case memory theories actually always do have extensions, this makes a semi-normal encoding of cancellation even less appropriate.

We have chosen the alternative of separating inheritance from cancellation, expressing property inheritance with normal defaults (as above), and relying on an external criterion to determine cancellation. In (Koton, Chase, & Vilain 1990), we describe one such criterion based on Poole's notion of theory preference (Poole 1985). We can easily extend this criterion (and in fact simplify it) to encompass our analytical terms

Say $\theta(e)$ is some theory of some entity $e$, and say $\theta_1(e)$ and $\theta_2(e)$ are extensions of $\theta(e)$ that assign different values $\beta_1$ and $\beta_2$ to an attribute $\alpha$. Note that these values must have been respectively assigned by some default rules $\delta_1$ and $\delta_2$ such that

$$\delta_1 = \frac{\gamma_1(x) : \alpha(x)=\beta_1}{\alpha(x)=\beta_1} \text{ and } \delta_2 = \frac{\gamma_2(x) : \alpha(x)=\beta_2}{\alpha(x)=\beta_2}$$

We then say that $\theta_1 <_\alpha \theta_2$ just in case $\gamma_1$ (the precondition of $\delta_1$) subsumes $\gamma_2$ (the precondition of $\delta_2$). Subsumption in this case is simply the relationship of meaning-entailment defined over terms in the analytical language which is used to define the preconditions of defaults. In Figure 3, for example, $\theta_1 <_{\text{diagnosis}} \theta_2$ because the precondition of the default that assigned *cardiovascular-disease* to *diagnosis* in $\theta_1$ is subsumed by that of the default that assigned it *aortic-stenosis* in $\theta_2$.

To enforce cancellation on some feature $\alpha$, we simply select as our prefered extensions those maximal in $<_\alpha$. To extend this minimality criterion into a theory-preference criterion, say $\theta(e)$ is a theory of some entity $e$, with extensions $\theta_1(e)...\theta_n(e)$. From among these $\theta_i$, the prefered extensions are those which are maximal in $<_\alpha$ for all attributes $\alpha$. In Figure 3 for example, $\theta_2$ is prefered over $\theta_1$, thus cancelling the assignment of *cardiovascular-disease* to *diagnosis* which holds in $\theta_1$. In contrast, the $\theta_1$ and $\theta_2$ of Figure 4 are both maximal, since the defaults that lead to their incompatibility on *pacifism* have incomparable preconditions (neither one subsumes the other).

In essence, our preference criterion selects those extensions of a theory in which attribute values are only assumed using the most specific applicable default. What's especially appealing about this approach is that the determination of specificity is entirely cast in terms of the analytical language which naturally supports specificity through subsumption.

## Integrated Classification

The knowledge representation task begins by using the analytical language to define initial categories which will later be extended by the learning algorithm[4]. The representation hierarchy is then built by classifying these definitions with respect to the subsumption relation; as noted above, this is an NP-complete process. This NP-completeness must be taken in perspective, however. For the purpose of identifying salient diagnostic categories, one can limit definitions to conjunctions of feature assignments created by only using *AND* and *ATTR* expressions. In this case, classification becomes tractable with an algorithm such as that of (Schmolze & Lipkis 1983).

Next, the hierarchy is extended with the case memory learning procedure. This is accomplished by classifying training cases using both analytical and similarity criteria. Simultaneously, running statistics are maintained of the relative frequency of attribute-value assignments:

[4]These are similar to the T Box definitions of (Brachman, Fikes, & Levesque 1983) which are completed by (user-declared) A Box axioms.

these are used to establish norms, and reorganize the hierarchy should it prove to reflect a skewed presentation order. We will say little about the statistical aspects of the learning procedure, focussing on the classification process instead (see (Kolodner 1983) for details).

## The classification algorithm

The hierarchy is implemented as a graph in which categories are connected by two kinds of attribute-value links. *Necessary* links correspond to conjoined attribute-value assignments, such as those in the analytical definitions. *Contingent* links correspond to disjuncts from *ATTR** expressions, and are used to implement similarity-based classification. The key to the operation of the algorithm is that all of the contingent links connecting two nodes are interpreted as part of the same disjunction. This effectively restricts the analytical definitions of categories to be in the normal form.

$$\gamma_1 \equiv (AND\ \gamma_2\ \ (ATTR\ \alpha_1\ \beta_1)\ \ldots\ (ATTR\ \alpha_i\ \beta_i)\ (ATTR*\ \alpha_j\ \beta_j \ldots\ \alpha_n\ \beta_n))$$

The *ATTR* expressions in the normal form correspond to feature conjunctions specified in the category definitions predefined by the user. The *ATTRR** expression encodes the feature assignments by which the learning procedure determined the category to be differentiated from its parent in the hierarchy.

Turning to the details of the classification algorithm, to classify a case $\chi$, with initial description $\theta$:

(1) Let *known* be a list of properties known to be true of $\chi$, and let *defaults* be a list of properties conjectured to be true of $\chi$. The known properties are simply those which hold in $\theta$; the conjectured ones are added by the norm defaults.

(2) Starting with the root of the hierarchy, proceed down the hierarchy by following links which match propositions in *known*.

(3) *(Analytical classification)* If nodes $\gamma_1$ and $\gamma_2$ are linked by any necessary links, then to proceed from $\gamma_1$ to $\gamma_2$ all such links must be matched by properties in *known* .

(4) *(Similarity classification)* Separately, if nodes $\gamma_1$ and $\gamma_2$ are only linked by contingent links, then to proceed from $\gamma_1$ to $\gamma_2$ at least one such link must be matched by properties in *known*.

(5) Potentially, several paths could be followed from a given node $\gamma$, each leading to a different classification. If so, the algorithm follows each path independently.

(6) For each norm at each node, if the norms does not contradict some property in *known*, it is added to *defaults*, overriding any properties in *defaults* which it contradicts.

(7) The algorithm terminates (along each independent path) when the leaves of the hierarchy are reached, or no links can be followed from a node. The algorithm returns, $\chi_1 \ldots \chi_n$, the nodes at the end of each path, along with the values of *known* and *defaults*.

It can be shown that the nodes returned by this algorithm define the maximal extensions of the initial theory $\theta(\chi)$. The theories are simply defined by the properties which hold at these nodes (the union of *known* and *defaults* in the algorithm). It is also easy to show that the algorithm computes these extensions in polynomial time. For details see (Koton, Chase, & Vilain 1990).

In order to actually add $\chi$ to the hierarchy, $\chi$ must additionally be indexed to each $\chi_i$ in $\chi_1 \ldots \chi_n$, If $\chi_i$ is a generalization, $\chi$ is inserted below $\chi_i$ and linked to it by a contingent link for each property of $\chi_i$ which $\chi$ does not share. If $\chi_i$ is another case, a generalization $\chi'$ is first created, and given as norms those properties shared by $\chi_i$ and $\chi$. These two cases are then indexed below $\chi'$ with a contingent link for each property they don't share with $\chi'$. Indexing $\chi$ below the $\chi_i$ effectively joins the separate paths taken by the algorithm.

It is easy to show that indexing a case to the hierarchy with this strategy effectively provides the case with a normal form definition. It is also possible to show that the structure of the hierarchy encodes the subsumption relations between the definitions of its category nodes[5]. This leads directly to the tractability of computing the subsumption of the normal form definitions induced by the learning procedure.

## A classification example

To illustrate this classification process, say we had created a hierarchy by predefining the categories *likely-heart-disease* and *likely-valve-disease*. The definitions of these categories are reproduced in Figure 5 (below), along with some norms that might have been assigned at some point to these categories by the learning procedure (assume for now that *disease-g0017* has not been created yet). Say we are now shown a case $e$ of youthful aortic stenosis, a fictitious but illustratively useful disease, with $e$ characterized by the following feature assignments.

*syncope(e) = on-exertion*
*chest-pain(e) = anginal*
*characteristic-murmur(e) = on-ejection*
*age(e) = 30-or-under*
*diagnosis(e) = youthful-aortic-stenosis*

[5]In brief, this follows from the fact that the same attribute-value pair can never appear twice on a given path, thus imposing an ordering on the disjunctive components of definitions.

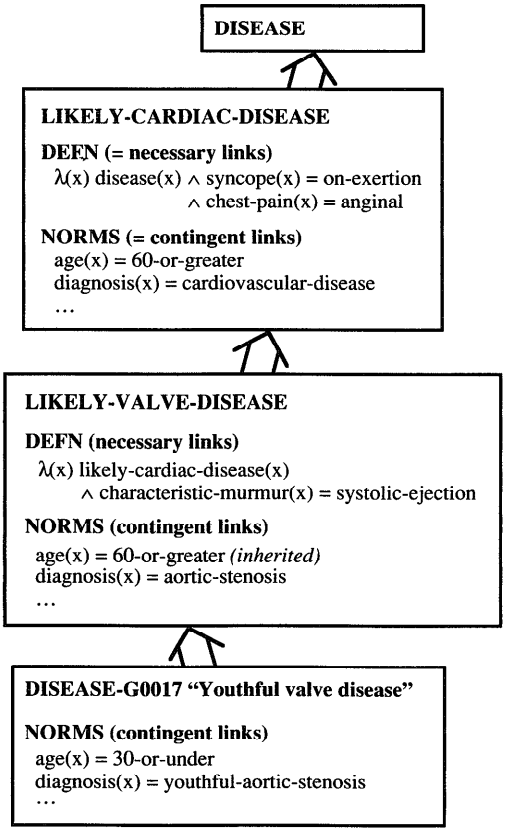


**Figure 5:** An induced category (DISEASE-G0017).

During classification, say *e* enters the hierarchy at the level of *disease*. It is then compared to the necessary links between *disease* and its descendant *likely-cardiac-disease*. These necessary links are simply the two explicit feature assignments in the latter category's definition: *syncope(x)=on-exertion* and *chest-pain(x)=anginal*. Since *e* has feature assignments to *syncope* and *chest-pain* that match these necessary links, it traverses them and is thus analytically classified as an instance of *likely-cardiac-disease*. Similarly, *e* matches the necessary links between *likely-cardiac-disease* and *likely-valve-disease*, allowing it to be analytically classified as an instance of the latter category. The case is then entered into the hierarchy by creating node *disease-g0017* and giving it as norms the feature assignments to *age* and *diagnosis* that differentiate *e* from *likely-valve-disease*.

These norms are in turn interpreted as contingent links between *likely-valve-disease and disease-g0017*; these can be exploited to perform similarity-based classification. For example, say we are now shown a new case *f* which shares *e*'s feature assignments for *syncope*, *chest pain*, *characteristic-murmur*, and *age*, but has no assignment to *diagnosis*. Like *e*, *f* will be classified analytically below *likely-cardiac-disease* and *likely-valve-disease*. From the latter, the contingent link for *age* can be followed down to *disease-g0017*, as *f* also assigns *30-or-less* to the *age* feature. This classifies *f* by similarity as an instance of *disease-g0017*; *f* then inherits this category's consistent norms by default, including an assignment of *youthful-aortic-stenosis* to *diagnosis*.

## Relation to Inductive Bias

The enterprise we have described above can be related to a concern of researchers in the machine learning community, *inductive bias*. The task of the statistical classification methods described earlier is to induce a classification upon being presented with a sequence of examples. That is, the learning program partitions the examples into a set of (not necessarily disjoint) classes. When the examples are labeled with their class (often just a binary labeling), the task is called "learning from examples" or "supervised learning," and the learning program produces an intensional description of the classes. When the examples are not labeled, and the learning program must induce the classes as well as the intensional descriptions of the classes, the task is called "concept formation" or "unsupervised learning." (Gennari, Langley, & Fisher 1989)

Both learning tasks may be viewed as a search through a space of hypotheses, in which each hypothesis represents a partition of the examples. To keep the learning task tractable, machine learning researchers early recognized the need to incorporate into their systems an inductive bias, namely, some mechanism for controlling the search of the hypothesis space (Russell & Grosof 1990). This bias controls the number of examples needed to induce a classification.

There are two major types of bias (Utgoff 1986): (1) restricting the hypothesis space, and (2) ordering the hypotheses. The first bias is usually imposed by limiting the concept description language; the second is often achieved through some general preference, such as, prefering a simpler to a more complex description. More recently, researchers have proposed a more general framework for inductive bias, namely, viewing it as prior knowledge that becomes part of the context within which the learning system operates (Russell & Grosof 1990). In particular, explanation-based learning, an analytic learning technique, can be viewed as using "background knowledge" as a type of bias; the generalizations produced through explanation-based learning are biased towards those that can explained in terms of this background knowledge (Ellman 1989).

Much of the research into inductive bias, particularly that which makes use of analytic learning techniques, has focused upon (supervised) learning from examples.

Our hybrid learning system, however, can be viewed as employing an analytical classification method as an inductive bias for an unsupervised concept formation task. In our system, the hypothesis space is implicitly represented; the current memory structure represents the current hypothesis regarding the classification scheme. The prior knowledge encoded in the memory structure is a declarative bias. When a new instance is presented to be incorporated into the memory structure (perhaps causing the memory to be restructured), the classification algorithm can be viewed as conducting a search of the hypothesis space, using the analytic knowledge as a bias focusing the search.

## Conclusion

In this paper we have presented a mechanism for combining the machine learning and knowledge representation approaches to classification, and described a hybrid knowledge representation appropriate for handling analytic and contingent knowledge. Our goal has been to use analytic knowledge as an inductive bias to focus the statistical learning procedure.

One obvious extension is to use prior *contingent* knowledge along with analytic knowledge to set up the initial memory. The analytic language was used to define the essential links of the hierarchical classification; we could also use the contingent language to define some of the well-known, textbook covarying features in advance. For example, a patient with general symptoms of heart disease who happens to be thirty or younger is almost certainly suffering of valve disease, despite the fact that the age norm for valve disease is greater than sixty. We expect that allowing for such non-analytic associations as these would further reduce the statistical learning method's sensitivity to exceptional cases and to poor presentation order.

We should note that our work to date has focussed primarily on theoretical considerations of representation and tractability. We intend to further validate our results by carrying out experiments to compare the performance of the learning method with and without the use of prior analytic knowledge. Finally, we would like to implement our work in other diagnostic domains and explore the feasibility of this approach to other application areas.

## References


Aghassi, D. 1990. Evaluating Case-based Reasoning for Heart Failure Diagnosis. SM thesis, Lab for Computer Science, Massachusetts Institute of Technology.

Brachman, R.J., Fikes, R.E., and Levesque, H.J. 1983. Krypton: A Functional Approach to Knowledge Representation. *Computer*, 16(10):67-73.

Ellman, T. 1989. Explanation-Based Learning: A Survey of Programs and Perspectives. *Computing Surveys*, 21:163-221.

Etherington, D.W. 1988. *Reasoning with Incomplete Information. Research Notes in Artificial Intelligence.* Morgan Kaufmann Publishers, Inc., San Mateo, CA.

Etherington, D.W., and Reiter, R. 1983. On Inheritance Hierarchies with Exceptions. In Proceedings of the Third National Conference on Artificial Intelligence, Washington, DC.

Gennari, J.H., Langley, P., and Fisher, D. 1989. Models of Incremental Concept Formation. *Artificial Intelligence*, 40:11-61.

Kolodner, J.L. 1983. Maintaining Organization in a Dynamic Long-Term Memory. *Cognitive Science*, 7:243-280.

Koton, P.A. 1988. Reasoning about Evidence in Causal Explanations. In Proceedings of the Seventh National Conference on Artificial Intelligence, St. Paul, MN.

Koton, P.A., and Chase, M.P. 1989. Knowledge Representation in a Case-Based Reasoning System: Defaults and Exceptions. In Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning, Toronto, Ontario, Canada.

Koton, P.A., Chase, M.P., and Vilain, M.B. 1990. Knowledge Representation in a Case-Based Reasoning System: An Extended Version. In preparation.

Long, W.J., Naimi, S., Criscitiello, M.G., and Jayes, R. 1987. The Development and Use of a Causal Model for Reasoning about Heart Failure. In Proceedings of the 11th Symposium on Computer Applications in Medical Care,

Poole, D. 1985. On the Comparison of Theories: Preferring the Most Specific Explanation. In Proceedings of the Ninth International Joint Conference on Artificial Intelligence, Los Angeles, CA.

Reiter, R. 1980. A Logic for Default Reasoning. *Artificial Intelligence*, 13:81-132.

Reiter, R., and Criscuolo, G. 1983. Some Representational Issues in Default Reasoning. *Computers and Mathematics with Applications*, 9:15-27.

Russell, S.J., and Grosof, B.N. 1990. Declarative Bias: An Overview. In D. Paul Benjamin, ed. *Change of Representation and Inductive Bias.* Kluwer Academic Publishers, Boston, Dordrecht, London.

Schmolze, J.G., and Lipkis, T.A. 1983. Classification in the KL-ONE Knowledge Representation System. In Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe, FRG.

Utgoff, P. 1986. *Machine Learning of Inductive Bias.* Kluwer Academic Publishers, Boston, Dordrecht, Lancaster.