An Analysis of Representation Shift In Concept Learning

William W. Cohen
Computer Science Department
Rutgers University
New Brunswick, NJ 08903

Abstract

In spite of the importance of representation in learning, little progress has been made toward understanding what makes representations work. This paper describes a framework for knowledge-level analysis of changes in the representation of training examples in concept learning. This a very fundamental sort of representation change; such a change alters the very space over which learning occurs, and hence necessitates selection of a new hypothesis space and (probably) a new learning algorithm. The goals of this paper are first, to provide a framework for analysis of representation shifts; second, to make explicit the assumptions implicit in representation shifts that have actually been used in learning systems; and third, to suggest a procedure for finding the most appropriate representation shift, given some background knowledge about a learning problem. The analytic framework is used to analyze a class of hybrid EBL/SBL systems by characterizing the sorts of domain theories that can be used with these systems.

1 Introduction

Few will take issue with the claim that the choice of an appropriate representation is crucial to success in solving AI problems. This is particularly true for learning problems. However, in spite of the importance of representation in learning, little progress has been made towards *understanding* what makes representations work: determining whether a representation will or will not help in solving a learning problem is still by and large a black art. In particular, there is no accepted procedure for answering questions about representation change such as: when is a representation appropriate for a learning problem? what assumptions are being implicitly made when a particular representation is being used? given some background knowledge about a learning problem, which representations are good, which are bad, and which are optimal?

The subject of this paper is *representation change in learning*; in particular, changes in the representation of training examples in concept learning problems. This is in contrast to previous work in representation change in learning, which has primarily addressed the problem of changing the representation of hypotheses (*i.e.*, shift of inductive bias.) Since a hypothesis is a set of examples, a change in the representation of training examples requires change in inductive bias; however, being able to change the language used to describe examples *as well as* the language used to describe hypotheses makes possible more radical representation shifts.

The problem of automatically performing representation shifts of this kind is outside the scope of this paper. The goals of this paper are:

1. to provide a framework for analysis of representational choices made by humans,
2. to uncover the assumptions implicit in certain representation shifts that are used in real learning systems, and
3. to suggest a methodology for selection of an appropriate representation shift, given some background knowledge about the learning problem.

All analysis is done at the "knowledge level" [Newell, 1982]; consideration is given only to when learnability is made possible or impossible, not to when it is made easy or difficult. This enables our results to be independent of particular learning algorithms and particular learnability criteria; however, it also restricts the analysis to representation shifts that are not information-preserving.

Proofs of theorems are included in an appendix.

2 Motivation and Overview

Consider a learning program with the architecture shown in figure 1. First, examples are mapped from an initial space X_I into the representation space X_R; then, a learning algorithm is run in the representation space; finally, the concept learned in the representation space is translated back to the initial space. Such a learning system is shifting the representation of examples from X_I to X_R using the function f_R.

Several learning systems of this sort have been built [Cohen, 1988; Cohen, 1990b; Flann and Dietterich,

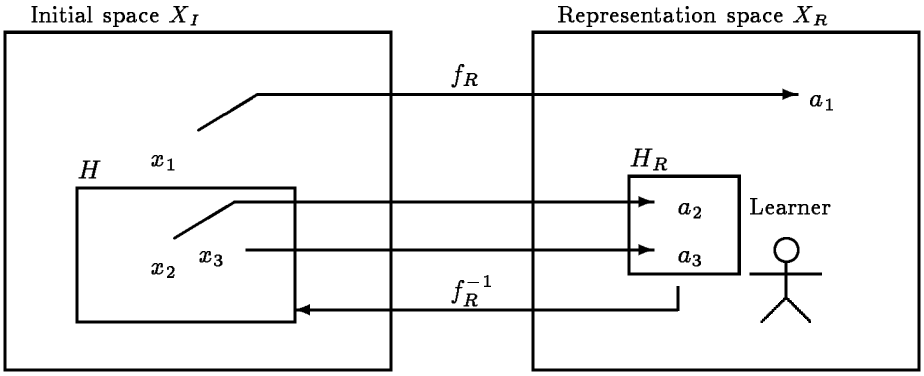

Figure 1: A learning system using representation shift

1989; Hirsh, 1989]. One advantage of this architecture is that it allows background knowledge to be applied to a learning problem in a highly modular way. Background knowledge is used only to *select the space over which learning will occur*, via the representation-shifting function f_R; standard concept learning techniques can then be applied in the new space X_R.

A small notational issue should be clarified at this point. In this paper, we will adopt the convention that if f is a function with domain X and if $S \subseteq X$, then $f(S)$ denotes the set $\{f(x) : x \in S\}$. This will greatly simplify the notation in situations where there are two parallel functions that must be considered, one which maps instances to instances, and one which maps concepts to concepts. An example of this appears in Figure 1, which actually contains two representation-shifting functions: f_R, which maps examples from the initial space to the representation space, and an "inverse" of f_R, denoted f_R^{-1} in the figure, which maps *concepts* in the representation space back to *concepts* in the initial space. The function f_R^{-1} is of course not a true inverse of f_R; rather, it it the inverse of f_R extended to concepts using the convention described above.

The effectiveness of such a learning system depends on the representation shift. The first issue addressed in this paper is: when is a representation shift appropriate for a learning problem? In short, what are the *principles* that underlie representation shift? We establish a correspondence called *knowledge-level equivalence* between a representation and the assumptions about the learning problem that are implicitly made when that representation is selected. This correspondence is the basis of our analytic framework.

The correspondence can be used in three ways.

- Given a representation shift, it is possible to identify the background knowledge to which a representation shift corresponds. In other words, it is possible to uncover the assumptions implicit in the use of a specific representation shift.
- Conversely, given some background knowledge about a learning problem, it is possible to test whether a given representation shift is necessarily appropriate, by testing whether the knowledge which corresponds to the representation shift is implied by the background knowledge.
- Finally, if one assumes that it is desirable to use a representation shift which corresponds to the strongest possible knowledge, it is possible to determine if a representation shift is optimal with respect to some given background knowledge about a learning system. This suggests a procedure for constructing a representation shift given some background knowledge about a learning problem. Such a representation shift is a means of incorporating background knowledge into a concept learning problem.

Two applications are given of our framework. In section 4.2, we find the assumptions implicit in the use of a common representational shift: a shift from examples to the "explanation structures" generated by the explanation based generalization (EBG) algorithm [Mitchell *et al.*, 1986]. This shift is found to correspond to an assumption about the correctness of EBG. In section 4.3, we relax this assumption of correctness, and construct a new representation shift which is knowledge-level equivalent to the relaxed assumption. This representation shift could be used as the basis of a learning system that makes a weaker assumption about the correctness of EBG.

3 A framework for analyzing representation shift

The question addressed in this section is: when is a representation shift appropriate to a particular class of learning problems? The meaning of this question is intuitively clear. However, in order to formulate and answer the question above precisely, it is necessary to formally define the terms involved: *representation shift*, *learning problem*, and *appropriate*.

3.1 What is a representation shift?

A *representation shift* will be formalized as a function f_R from an original *initial space* X_I into another space

X_R, called the *representation space.*

$$\text{Representation shift} \equiv \text{a function } f_R : X_I \rightarrow X_R$$

In a typical case, X_I is a set of potential training examples in some initial description language, which is inappropriate for learning. X_R contains representations of the elements of X_I, in what is hopefully a better format for learning. For example X_I might be a set of digitized photographs of soybean plants, and X_R might consist of feature-vector representations of these plants using certain features known to be significant for the learning task at hand, as in [Michalski and Chilausky, 1980]; or X_I might be a set of chess positions, and X_R might be proofs that white has a forced exchange from that position, as in [Flann and Dietterich, 1989].

3.2 What is a learning problem?

A definition of a "class of learning problems" must now be given. It is assumed that the problem faced by the learner is to correctly identify some unknown *target concept* T given some "partial information" about T. Typical examples of the sorts of partial information available might be a set of answers to queries issued by the learner, or a set of labeled examples of members and nonmembers of T.

We would like to *exclude* from our formalization of a learning problem anything specific to a particular learning algorithm or a particular model of learnability. One way to do this is to assume that the partial information is requested by the learner (for example, via a set of calls to oracles) and is not an explicit input to the learner. (It seems reasonable to assume that for every learner that takes partial knowledge as an explicit input, there is a learner that requests this input from an oracle.) In this case, a *learning problem* is simply a possible target concept T, and a class of learning problems is simply a set of such concepts.

$$\text{Class of learning problems} \equiv \text{a set } \mathcal{P} = \{T_1, T_2, \ldots, \}$$

3.3 When is a representation appropriate?

Given these definitions, it is now possible to state precisely when a representation is appropriate. *A representation f_R is appropriate for the class of learning problems $\mathcal{P}$ if and only if each learning problem is potentially solvable in the representation space.* A learning problem T is *potentially solvable* if it is possible to exactly describe a target concept T by an image of T under f_R. If we let $APPROP(f_R, \mathcal{P})$ denote the proposition "f_R is appropriate for $\mathcal{P}$", and let X_R be the range of f_R, then the definition of appropriateness can be stated in symbols:

$$\begin{aligned} &APPROP(f_R, \mathcal{P}) \equiv \\ &\quad f_R \text{ is a well-defined function} \wedge \\ &\quad \forall T \in \mathcal{P},\ \exists T_R \subseteq X_R : T = f_R^{-1}(T_R) \end{aligned}$$

An alternative definition which some readers may find more intuitive is

$$\begin{aligned} &APPROP(f_R, \mathcal{P}) \equiv \\ &\quad f_R \text{ is a well-defined function} \wedge \\ &\quad \forall T \in \mathcal{P},\ \forall x, y \in X_I, \\ &\qquad (f_R(x) = f_R(y) \Rightarrow \mathcal{X}_T(x) = \mathcal{X}_T(y)) \end{aligned}$$

where $\mathcal{X}_T(x)$ denotes the characteristic function of T on x.

The definition of potentially solvable assumes that the goal of learning is exact identification of the target concept. This fits many formal definitions of learnability, such as Gold's learnability in the limit [Gold, 1967], Littlestone's mistake-bounded learnability [Littlestone, 1988], and Angluin's definitions of learning from queries [Angluin, 1988]. An exception is the Valiant criterion of probably approximately correct (pac) learnability [Valiant, 1984], which requires only approximate identification of the target concept; this suggests that a probabilistic extension of this analytic framework may be worth investigating.

Notice that the appropriateness of f_R does not say anything about how easy or difficult it is to solve the learning problem in the representation space X_R; it merely says that it is still *possible* to solve the learning problem.

3.4 Representations and background knowledge

Let $\mathcal{A}(\mathcal{P})$ denote a first order sentence which is a statement about $\mathcal{P}$; $\mathcal{A}(\mathcal{P})$ will alternatively be viewed as an assumption about the class of learning problems, or as background knowledge about the class of learning problems. We can now formulate precisely and answer the questions given in the introduction:

Question 1: Given a representation f_R what assumptions are implicitly made when f_R is used on the class of learning problems $\mathcal{P}$?

The answer to this question is obvious: *any learning system that uses a representation f_R makes the assumption that f_R is appropriate for the learning problem*; *i.e.*, that $APPROP(f_R, \mathcal{P})$ is true. A particular learner may (and probably will) make other assumptions as well in the process of generating a hypothesis; however, since *every* learning system that uses the representation f_R assumes $APPROP(f_R, \mathcal{P})$, then it seems reasonable to characterize $APPROP(f_R, \mathcal{P})$ as the assumptions made in choosing the representation f_R.

Question 2a: Given some background knowledge $\mathcal{A}(\mathcal{P})$ about a class of learning problems $\mathcal{P}$, what representations f_R can be used?

Question 2b: In the same circumstances, what representations f_R *should* be used?

The answer to question 2a is also obvious: if

$$\mathcal{A}(\mathcal{P}) \Rightarrow APPROP(f_R, \mathcal{P}) \tag{1}$$

then the representation f_R *could* be used for the learning problem $\mathcal{P}$, given the background knowledge $\mathcal{A}(\mathcal{P})$. The answer to 2b is not obvious, however. There may be many representations that have this property; which of these *should* be used? In short, are there any grounds for preferring one of these representations over others?

One context in which f_{R_1} is preferable to f_{R_2} is when the appropriateness of f_{R_1} *implies* the appropriateness of f_{R_2}, *i.e.*

$$APPROP(f_{R_1}, \mathcal{P}) \Rightarrow APPROP(f_{R_2}, \mathcal{P}) \qquad (2)$$

The rational for preferring f_{R_1} is the following: consider a learning system L_1 that uses f_{R_1}, and another learning system L_2 that uses f_{R_2}. L_1 "knows" only that f_{R_1} is appropriate, and L_2 "knows" only that f_{R_2} is appropriate. Then if equation 2 holds, *everything known by L_2 is deductively implied by what is known by L_1*, and can, at least in principle, be derived from knowledge available to L_1. At the knowledge level, L_1 is "better informed about the learning problem" than L_2.

This rule of preference means that logical implication imposes a partial order on the desirability of representations. The best representations are those f_R that satisfy equation 1 and are maximal with respect to this partial order. It is easy to show that if f_R has the property

$$\mathcal{A}(\mathcal{P}) \iff APPROP(f_R, \mathcal{P})$$

then it fulfills this requirement. Such a representation will said to be *knowledge-level equivalent* (KL-equivalent) to the assumption $\mathcal{A}(\mathcal{P})$. The answer to question 2b is thus: *given $\mathcal{A}(\mathcal{P})$, the most desirable representation is some f_R such that f_R is KL-equivalent to $\mathcal{A}(\mathcal{P})$.*

Notice that f_R is not unique: that is, given an assumption $\mathcal{A}(\mathcal{P})$, there is more than one f_R that is KL-equivalent to $\mathcal{A}(\mathcal{P})$. In particular, if g is any one-to-one function and f_R is KL-equivalent to $\mathcal{A}(\mathcal{P})$, then $f_R \circ g$ is also KL-equivalent to $\mathcal{A}(\mathcal{P})$. This is to be expected: if g is one-to-one, then it is an "information-preserving" operation and should make no difference at the knowledge level. However, the following theorem shows that this is the only sort of variation of f_R which is possible: *i.e.*, that the KL-equivalent representation for an assumption $\mathcal{A}(\mathcal{P})$ is unique up to composition with one-to-one functions.

Theorem 1 *If f_{R_1} and f_{R_2} are both KL-equivalent to $\mathcal{A}(\mathcal{P})$, then there is a one-to-one function g such that $f_{R_1} = f_{R_2} \circ g$.*

The analytic framework presented in this section is summarized in table 1.

4 Applications of the framework

The true value of an analytic framework lies in its usefulness in the analysis of meaningful problems. We will consider two uses of the framework. First, in section 4.2, we find the assumption implicit in the use of a common representational shift. This is a simple analysis: it merely requires finding a meaningful condition which is logically equivalent to $APPROP(f_R, \mathcal{P})$. In section 4.3, we consider the related problem of constructing a representation shift which is optimal given a particular assumption about a learning problem. Here, given an assumption about the learning problem, the problem is to construct a representation shift which is KL-equivalent to this assumption.

The representation shifts and assumptions which we consider are closely connected to the operation of explanation based generalization (EBG), a procedure that uses a logical theory to generalize a single positive example of some concept. A short discussion of EBG precedes our analyses.

4.1 Explanation based generalization

To avoid unnecessary detail, explanation based generalization will be discussed at a rather abstract level in this paper. Readers requiring more detail are referred to [Kedar-Cabelli and McCarty, 1987; Mitchell *et al.*, 1986]. Let Θ be a theory, and for $x \in X_I$, let $PROOFS_\Theta(x)$ denote the set of proofs of x in Θ. This assumes that x is a logical goal, for example *cup(obj1)*; such an example is implicitly tagged with the "target concept" to which it is relevant. Let $\mathcal{O}$ denote an operationality predicate; in this context, an operationality predicate tells if a subproof is a "detail" not relevant to the concept to be learned. *Explanation based generalization* is an operation defined on a proof $p_x \in PROOFS_\Theta(x)$ and denoted by the function $EBG(\Theta, \mathcal{O}, p_x)$. This function always returns a set that generalizes x; *i.e.*, $\forall x \in X_I,\ \{x\} \subseteq EBG(\Theta, \mathcal{O}, p_x) \subseteq X_I$.

Notice that the usual assumptions about Θ — *e.g.*, that it is complete and correct, that it is defined only for positive examples — have not been made. The reason is that the assumptions which an EBL system makes about Θ depend in part on *how generalizations are used.* The remainder of this section will consider several possible schemes for using EBG, and analytically determine which assumptions about the domain theory are being made.

There are several known algorithms for explanation based generalization. Generally, a proof p_x is generalized in two ways: operational subtrees are pruned in accordance with the $\mathcal{O}$ predicate, and some of the constants appearing in the proof are replaced with variables. This generalized proof is called an *explanation structure*, and will be denoted $e_\mathcal{O}(p_x)$. A rule is then extracted from the explanation structure by conjoining the leaves of the explanation structure to form the antecedent, and using the root of the explanation structure as the consequent. The set of goals provable directly from this rule is the generalization of x.

As an example of explanation based generalization, [DeJong and Mooney, 1986] describes a simple theory of social interactions, under which the goal *kill(john,john)* is generalized to form the rule

$$\mathrm{kill(X, X) \leftarrow depressed(X) \wedge buy(X, W) \wedge gun(W)}$$

which can be paraphrased as saying "X will kill himself if he is depressed and has bought a gun." Here the example is the goal *kill(john,john)* and the generalization is the set of goals provable by the rule above, *i.e.* the set of goals of the form *kill(X,X)* such that

$$\mathrm{depressed(X) \wedge buy(X, W) \wedge gun(W)}$$

is provable. In this case, the generalization corresponds loosely to the concept of "suicide".

Table 1: Summary of the analytic framework

Intuitive Notion	Formalization
representation shift	a function $f_R : X_I \rightarrow X_R$
learning problem	a target concept $T \subseteq X_I$
class of learning problems	a set $\mathcal{P} = \{T_1, T_2, \ldots\}$
background knowledge of learning problem	a first order formula $\mathcal{A}(\mathcal{P})$
f_R is appropriate for $\mathcal{P}$	$APPROP(f_R, \mathcal{P})$
assumption implicitly made in using f_R	$APPROP(f_R, \mathcal{P})$
f_R can be used given $\mathcal{A}(\mathcal{P})$	$\mathcal{A}(\mathcal{P}) \Rightarrow APPROP(f_R, \mathcal{P})$
f_R should be used given $\mathcal{A}(\mathcal{P})$	$\mathcal{A}(\mathcal{P}) \iff APPROP(f_R, \mathcal{P})$ (*i.e.*, f_R is KL-equivalent to $\mathcal{A}(\mathcal{P})$)

Later analysis will not depend on the precise algorithm used for EBG; however, the following property is assumed to hold.[1]

Assumption 1 *For all Θ, $\mathcal{O}$, p_x, and y,*

$$\begin{array}{ll} & y \in EBG(\Theta, \mathcal{O}, p_x) \\ \iff & \exists p_y \in PROOFS_\Theta(y) : e_\mathcal{O}(p_y) = e_\mathcal{O}(p_x) \end{array}$$

4.2 Analysis of a representation shift

An important topic in machine learning research is integration of explanation based and similarity based learning techniques. Several integrated learning systems have been built that have the architecture of figure 1, and that use as a representation-shifting function some close variant of the function $f(x) \equiv e_\mathcal{O}(p_x)$ These systems *learn from a set of explanation structures* using SBL techniques; they differ primarily in the learning methods used. (See [Flann and Dietterich, 1989] for a discussion of some of these systems.)

What assumptions are made by a learning program that that uses this representation shift? The answer $APPROP(f, \mathcal{P})$ is correct, but not very meaningful. However, it can be shown that $APPROP(f, \mathcal{P})$ is logically equivalent to a meaningful assumption about the behavior of EBG.

Theorem 2 *Let $f_{\Theta,\mathcal{O}}$ be defined as $f_{\Theta,\mathcal{O}}(x) \equiv e_\mathcal{O}(p_x)$. Then*

$$\begin{array}{l} APPROP(f_{\Theta,\mathcal{O}}, \mathcal{P}) \equiv \\ \quad \forall x \in X_I, |e_\mathcal{O}(PROOFS_\Theta(x))| = 1 \ \wedge \\ \quad \forall T \in \mathcal{P},\ \forall x \in T,\ \forall p_x \in PROOFS_\Theta(x), \\ \qquad EBG(\Theta, \mathcal{O}, p_x) \subseteq T \end{array}$$

[1] This assumption should hold for any reasonable implementation of EBG. The "only if" direction requires only that the output of EBG be determined by Θ,$\mathcal{O}$, and $e_\mathcal{O}(p_x)$: in this case, if $\exists p_y \in PROOFS_\Theta(y) : e_\mathcal{O}(p_y) = e_\mathcal{O}(p_x)$, then $EBG(\Theta, \mathcal{O}, p_y) = EBG(\Theta, \mathcal{O}, p_x)$. We know that $y \in EBG(\Theta, \mathcal{O}, p_y)$, and hence $y \in EBG(\Theta, \mathcal{O}, p_x)$ as well. The "if" direction is also intuitively clear, if the conjunction that describes the set $EBG(\Theta, \mathcal{O}, p_x)$ represents preconditions under which the "abstract version" $e_\mathcal{O}(p_x)$ of the proof p_x is applicable: if y is in this set, those preconditions are true, and some instantiation of the "abstract proof" $e_\mathcal{O}(p_x)$ should succeed for y as well as for x.

This restatement of $APPROP$ shows that it is equivalent to two assumptions about EBG. The first assumption is that every instance has at most a single proof, up to the level of detail specified by the explanation structure. Since this constraint is usually enforced by ensuring that there is at most one proof for every instance in the domain theory, we will call this assumption the *unique proof assumption.* The second assumption is about the "validity" of generalizations of positive examples. Intuitively, the generalization of any $x \in T$ is *valid*: *i.e.*, it is a correct generalization of x with respect to T. Notice that this assumption is much weaker than the assumption, commonly associated with EBL systems, that the theory is a complete and correct definition of the target concept T.

Theorem 2 has the interesting corollary that f is KL-equivalent to the assumption above (*i.e.*, the conjunction of the unique proof assumption and the validity assumption.) This means that if L' is *any* hybrid SBL/EBL concept learning system which makes these assumptions, and which assumes nothing else about the learning problem, then a learner that uses the representation shift $f_{\Theta,\mathcal{O}}$ will have access to exactly the same knowledge that is available to L'. This is true regardless of the architecture of L'.

4.3 Constructing a representation shift

Given the analysis above, it is natural to ask: in what ways can these assumptions about EBG be relaxed? One class of theory that violates these assumptions and that appears to occur often in practice is the class of "abductive" theories: theories that make assumptions in the process of theorem proving. These theories often violate the unique proof assumption. A common subclass of abductive theories is the class of theories that involve plan recognition; often, any of several assumptions could be introduced that would suffice to explain an action, but only explanations based on the right assumption will be correct.

The crucial property of an abductive theory seems to be that *for every true observation, there is some correct explanation.* It seems reasonable to require that a generalization associated with a valid explanation is valid. This "abductive assumption" about the correctness of EBG can be precisely stated as:

$$\begin{aligned}&\mathcal{AB}_{\Theta,\mathcal{O}}(\mathcal{P}) \equiv \\ &\quad \forall T \in \mathcal{P},\ \forall x \in T, \\ &\qquad \exists p_x \in PROOFS_{\Theta}(x) : EBG(\Theta, \mathcal{O}, p_x) \subseteq T\end{aligned}$$

A reasonable application of our framework would be to find a representation shift which is optimal under this assumption: that is, a representation shift which is KL-equivalent to $\mathcal{AB}_{\Theta,\mathcal{O}}(\mathcal{P})$. A learning algorithm using this representation would have the advantages of systems which use the representation shifting function f, but would be useful even when the unique proof assumption does not hold.

In order to find a representation that is KL-equivalent to this assumption, it appears necessary to make some assumptions about $\mathcal{P}$ as well. Let us say that a set $S \subseteq 2^X$ is a *hitset* iff $\exists S_0 \subseteq X : S = \{R : \exists r \in R \cap S_0\}$. A class of learning problems $\mathcal{P}$ will be said to have *hitset images under* f_R iff $\forall T \in \mathcal{P},\ f_R(T)$ is a hitset.

Theorem 3 *Let the function* $f^*_{\Theta,\mathcal{O}}(x)$ *be defined as* $f^*_{\Theta,\mathcal{O}}(x) \equiv \{e_{\mathcal{O}}(p_x) : p_x \in PROOFS_{\Theta}(x)\}$. *If* $\mathcal{P}$ *has hitset images*[2] *under* $f^*_{\Theta,\mathcal{O}}$, *then* $f^*_{\Theta,\mathcal{O}}$ *is KL-equivalent to* $\mathcal{AB}_{\Theta,\mathcal{O}}(\mathcal{P})$.

Discussion of learning algorithms for the representation space which is the range of f^* is outside the scope of this paper; however, the A-EBL technique described in [Cohen, 1989; Cohen, 1990a; Cohen, 1990b] can be viewed such a learning algorithm.

A few remarks are in order about the theorem above, in particular about the requirement that $\mathcal{P}$ have "hitset images". First, this property generally holds if $\mathcal{P}$ contains only concepts that can be expressed using a set of rules produced by EBG; in particular, it can easily shown that if assumption 1 holds, then the image under f^* of the set $EBG(\Theta, \mathcal{O}, p_x)$ is always a hitset.

Second, the proof of the theorem shows that the property of having hitset images is *not* necessary to establish that

$$\mathcal{AB}_{\Theta,\mathcal{O}}(\mathcal{P}) \Rightarrow APPROP(f^*_{\Theta,\mathcal{O}}, \mathcal{P})$$

The property is only necessary for the other side of the logical equivalence which is the definition of KL-equivalence; that is, having hitset images is only necessary to show that

$$APPROP(f^*_{\Theta,\mathcal{O}}, \mathcal{P}) \Rightarrow \mathcal{AB}_{\Theta,\mathcal{O}}(\mathcal{P})$$

In other words, the function f^* will be an appropriate representation shift *whenever* $\mathcal{AB}(\mathcal{P})$ holds; the condition that $\mathcal{P}$ has hitset images is only needed to show that f^* is the best representation shift.

4.4 Discussion of the results

Any system which uses EBG must make an assumption about the domain theory which is used; the assumption that is commonly associated with EBG is that the domain theory is a complete, correct, and tractable definition of the target concept. In recent years, several hybrid learning systems have been built that have the architecture of figure 1, and which use as a representation-shifting function a mapping from instances to the explanation structures associated with those instances. Presumably, these systems make a weaker assumption about the correctness of the domain theory.

The analysis above confirms this intuition. It shows that the domain theory used by such a learning system must satisfy two conditions: every instance must have exactly one associated explanation structure, and EBG must never overgeneralize a positive example. This is a strong assumption, but it does *not* require the domain theory to be complete and correct; in particular, the domain theory might produce an (incorrect) proof for a negative example of the concept. In other words, the domain theory can be overly general with respect to the target concept. However, the theory must be sufficiently detailed that EBG does not overgeneralize any positive examples.

The second result shows that some of the requirements on the domain theory can be relaxed by using a slightly different abstraction function. By mapping an instance to a *set* of explanation structures, rather than to a single explanation structure, a somewhat weaker requirement is placed on the domain theory. The domain theory must still be complete; however, the theory may produce multiple inconsistent explanation structures for each instance, rather than a single explanation structure. The theory must still be sufficiently detailed that EBG does not overgeneralize on *all* explanations of a positive example; however, EBG may overgeneralize on *some* of the possible explanations of an instance.

The second result is also of interest because it shows that it is possible to relax assumptions about the domain theory in a principled way, and thus to expand the range of competence of a representation-shifting learning system.

5 Conclusions

In this paper, we have developed a framework for knowledge-level analysis of the changes in representations of training examples in learning. There have been several learning systems implemented that use changes of representation of the kind considered in this paper; some good examples are [Cohen, 1990b] and the systems discussed in [Flann and Dietterich, 1989].[3] The

[2] Of course, the assumption that $\mathcal{P}$ has hitset images could be simply added to $\mathcal{AB}(\mathcal{P})$. This presentation emphasizes the fact that this assumption is motivated by the desire to construct a KL-equivalent representation shift rather than by a natural assumption about the learning problem.

[3] It is important to keep in mind the difference between these systems and work on constructive induction, for example [Drastal *et al.*, 1989]: constructive induction is concerned with *information-preserving* transformations of training instances.

framework was used to uncover the assumptions implicit in a commonly-used representation shift, and to derive a representation shift which is appropriate for a weak assumption about the correctness of EBL. The first result is important because it helps us to understand the assumptions made by the learning systems which use this representation shift; the second, because it shows how to relax these assumptions in a principled way.

Most prior analyses of representation change have focused on automatic reformulation of problem-solving strategies [Amarel, 1981; Korf, 1981; Riddle, 1989; Subramanian, 1989]. Some of the formal models of representation change in problem solving, however, are general enough to serve as a model for representation change in learning. In particular, our model of representation change in learning and notion of appropriateness is closely related to the models presented in [Holte and Zimmer, 1989] and [Lowry, 1988]; in fact, it is a special case, in which the problem to be solved is required to be a learning problem. However, the earlier models do not develop the notion of a "most appropriate" representation, and have not been applied to analysis of representation shifts in learning.

Work in representation change in learning has generally focused on automatic change or derivation of inductive bias [Chrisman, 1989; Keller, 1989; Russell and Grosof, 1989; Utgoff, 1984]. The representation changes considered in this paper – changes in the representation of training examples – are more fundamental than shift of bias; such a change alters the very space over which learning occurs, and hence necessitates selection of a new hypothesis space and (probably) a new learning algorithm. Our goals, however, are more modest: the goal of this paper is to provide a framework for analysis of representational choices made by humans, and to suggest a methodology for humans to follow in constructing new representational shifts, given background knowledge about a learning problem.

These goals have been at least partially met, however, many problems remain for future research. One difficult problem is automation (or partial automation) of the methodology suggested for constructing representational shifts. Another set of problems is suggested by the observation that there is a second possible interpretation of a the architecture of figure 1: X_I might be the space in which training examples are "naturally found" (for instance, X_I might be the set of all soybean plants). In this case, f_R is a noncomputable function that denotes an *initial choice* of representations. Our basic framework applies to initial-choice representations as well as to representation-shifting functions, although obviously proof techniques for showing the appropriateness of initial-choice representations will be different from the proof techniques used in this paper. Finally, the definition of the appropriateness of a representational shift assumes that the goal of learning is exact identification of the target concept; this is a more stringent identification criterion than the Valiant criterion of probably approximately correct learnability [Valiant, 1984], which requires only approximate identification of the target concept. This suggests that a probabilistic extension of the analytic framework may be worth investigating.

6 Acknowledgements

I am grateful to Alex Borgida, for his advice and encouragement; Chun Liew, Patricia Riddle, and Haym Hirsh, for valuable comments on a draft of the paper; and to many other members of the Rutgers AI community. The paper was also much improved by the comments of an anonymous reviewer (Rob Holte.) The author is supported by an AT&T Fellowship. Initial stages of this research were supported by a Marion Johnson Fellowship.

A Proofs of theorems

A.1 Proof of theorem 1

If $\mathcal{X}_T(x)$ is the characteristic function of T on x then it can be easily verified that

$$APPROP(f_A, \mathcal{P}) \equiv \tag{3}$$
$$\forall T \in \mathcal{P},\ \forall x, y \in X_I, \quad (f_R(x) = f_R(y) \Rightarrow \mathcal{X}_T(x) = \mathcal{X}_T(y)) \tag{4}$$

So if

$$\begin{aligned} & \mathcal{A}(\mathcal{P}) \\ \iff & APPROP(f_{R_1}, \mathcal{P}) \\ \iff & APPROP(f_{R_1}, \mathcal{P}) \end{aligned}$$

then for $i = 1, 2$

$$\forall x, y \in X_I, \quad f_{R_i}(x) = f_{R_i}(y) \iff (\forall T \in \mathcal{P} \mathcal{X}_T(x) = \mathcal{X}_T(y))$$

and hence

$$f_{R_1}(x) = f_{R_1}(y) \iff f_{R_2}(x) = f_{R_2}(y)$$

and the mapping $g(x) = f_{R_2}(f_{R_1}^{-1}(x))$ is well-defined and one-to-one. ■

A.2 Proof of theorem 2

To prove the theorem, it is necessary to show both sides of the logical equivalence. The proposed simplification of $APPROP(f_{R1}, \mathcal{P})$ will be referred to as $\mathcal{A}(\mathcal{P})$.

To show that ($f_{\Theta,\mathcal{O}}$ is appropriate for $\mathcal{P}$) $\Rightarrow \mathcal{A}(\mathcal{P})$. First, note that if $f_{\Theta,\mathcal{O}}$ is a well-defined function over X_I, then the unique proof assumption holds. If $f_{\Theta,\mathcal{O}}$ is also appropriate for $\mathcal{P}$, then

$$\forall T \in \mathcal{P},\ \forall x, y,\ \neg(f_{\Theta,\mathcal{O}}(x) = f_{\Theta,\mathcal{O}}(y) \wedge \mathcal{X}_T(x) \neq \mathcal{X}_T(y))$$

again using $\mathcal{X}_T(x)$ to denote the characteristic function of T on x. Now, if $f_{\Theta,\mathcal{O}}(x) = f_{\Theta,\mathcal{O}}(y)$ then there is some explanation structure e such that $f_{\Theta,\mathcal{O}}(x) = f_{\Theta,\mathcal{O}}(y) = e$, and

$$\begin{aligned} & \forall T \in \mathcal{P}, \forall x \in T,\ (e_{\mathcal{O}}(p_x) = e_{\mathcal{O}}(p_y) \Rightarrow y \in T) \\ \Rightarrow\ & \forall T \in \mathcal{P}, \forall x \in T,\ (y \in EBG(\Theta, \mathcal{O}, p_x) \Rightarrow y \in T) \quad (5) \\ \Rightarrow\ & \forall T \in \mathcal{P}, \forall x \in T,\ EBG(\Theta, \mathcal{O}, p_x) \subseteq T \end{aligned}$$

where p_x is any proof of x, and where p_y is any proof of y; hence the second conjunct holds. Line 5 follows from assumption 1.

To show that $\mathcal{A}(\mathcal{P}) \Rightarrow$ *($f_{\Theta,\mathcal{O}}$ is appropriate for $\mathcal{P}$).* First, note that the unique proof assumption implies that $f_{\Theta,\mathcal{O}}$ is a function. Second, observe that for any $\mathcal{P}$,

$$\begin{aligned} & \forall T \in \mathcal{P},\ \forall x \in T,\ (EBG(\Theta, \mathcal{O}, p_x) \subseteq T) \\ \Rightarrow\ & \forall T \in \mathcal{P},\ \forall x \in \overline{T},\ (EBG(\Theta, \mathcal{O}, p_x) \subseteq \overline{T}) \end{aligned}$$

where p_x again denotes a proof of x. To see that this is true, imagine that the antecedent holds but

$$\begin{aligned} & \exists T \in \mathcal{P}, x \in \overline{T}, p \in PROOFS_\Theta(x) : \\ & \quad EBG(\Theta, \mathcal{O}, p_x) \not\subseteq \overline{T} \end{aligned}$$

Then there is some $y \in EBG(\Theta, \mathcal{O}, p_x) - \overline{T}$; note that this implies that y is in T. But by assumption 1

$$\begin{aligned} & y \in EBG(\Theta, \mathcal{O}, p_x) \\ \Rightarrow\ & e_\mathcal{O}(p_y) = e_\mathcal{O}(p_x) \\ \Rightarrow\ & EBG(\Theta, \mathcal{O}, p_y) = EBG(\Theta, \mathcal{O}, p_x) \end{aligned}$$

and so

$$\exists T \in \mathcal{P}, y \in T : EBG(\Theta, \mathcal{O}, p_y) \not\subseteq T$$

a contradiction. The preceding argument means that $\mathcal{A}(\mathcal{P})$ implies

$$\begin{aligned} & \forall T \in \mathcal{P},\ \forall x, y \in X_I, \\ & \quad (y \in EBG(\Theta, \mathcal{O}, p_x) \Rightarrow \mathcal{X}_T(x) = \mathcal{X}_T(y)) \end{aligned}$$

Since by assumption 1

$$\begin{aligned} & y \in EBG(\Theta, \mathcal{O}, x) \\ \iff\ & e_\mathcal{O}(p_y) = e_\mathcal{O}(p_x) \\ \iff\ & f_{\Theta,\mathcal{O}}(x) = f_{\Theta,\mathcal{O}}(y) \end{aligned}$$

it follows that

$$\begin{aligned} & \forall T \in \mathcal{P},\ \forall x, y, \\ & \quad (f_{\Theta,\mathcal{O}}(x) = f_{\Theta,\mathcal{O}}(y) \Rightarrow \mathcal{X}_T(x) = \mathcal{X}_T(y)) \end{aligned}$$

Hence $f_{\Theta,\mathcal{O}}$ is appropriate by equation 3. ■

A.3 Proof of theorem 3

Again, both sides of the equivalence must be shown.

To show $\mathcal{AB}(\mathcal{P}) \Rightarrow f^*_{\Theta,\mathcal{O}}$ *is appropriate.* The contrapositive will be shown. If $f^*_{\Theta,\mathcal{O}}$ is not appropriate, then

$$\exists T \in \mathcal{P}, x, y : f^*_{\Theta,\mathcal{O}}(x) = f^*_{\Theta,\mathcal{O}}(y) \land \mathcal{X}_T(x) \neq \mathcal{X}_T(y)$$

Assume without loss of generality that $x \in T$ and $y \in \overline{T}$. Now, by assumption 1, if $f^*_{\Theta,\mathcal{O}}(x) = f^*_{\Theta,\mathcal{O}}(y)$ then $\forall p_x \in PROOFS(x), y \in EBG(\Theta, \mathcal{O}, p_x)$. But then

$$\begin{aligned} & \exists T \in \mathcal{P}, x \in T, y \in \overline{T} : \\ & \quad \forall p_x \in PROOFS(x),\ y \in EBG(\Theta, \mathcal{O}, p_x) \end{aligned}$$

contradicting $\mathcal{AB}(\mathcal{P})$.

To show $f^*_{\Theta,\mathcal{O}}$ *is appropriate* $\Rightarrow \mathcal{AB}(\mathcal{P})$. This direction of the proof uses the easily-verified proposition that for any appropriate representation f_R for $\mathcal{P}$

$$\forall T \in \mathcal{P},\ \forall x \in X_I,\ (x \in T \iff f_R(x) \in f_R(T))$$

If $f^*_{\Theta,\mathcal{O}}$ is appropriate for $\mathcal{P}$ then

$$\forall T \in \mathcal{P}, \forall x \in X_I,\ (x \in T \Rightarrow f^*_{\Theta,\mathcal{O}}(x) \in f^*_{\Theta,\mathcal{O}}(T))$$

Now, since $f^*_{\Theta,\mathcal{O}}(T)$ is a hitset, there is some E_0 such that $f^*_{\Theta,\mathcal{O}}(T) = \{E : \exists e_\mathcal{O}(p_{xi}) \in E \cap E_0\}$ and so $f^*_{\Theta,\mathcal{O}}(x) \in f^*_{\Theta,\mathcal{O}}(T) \Rightarrow \exists e_\mathcal{O}(p_{xi}) \in f^*_{\Theta,\mathcal{O}}(x) \cap E_0$

Let e' denote the explanation structure in $f^*_{\Theta,\mathcal{O}}(x) \cap E_0$. We claim that any y that has a proof p_y with an explanation structure equivalent to e' will be in T. If this is true, then by assumption 1, $EBG(\Theta, \mathcal{O}, p_x) \subseteq T$ for the proof $p_{xi'}$ of x with explanation structure e'; and since nothing has been assumed except that $x \in T$ and $T \in \mathcal{P}$, $\mathcal{AB}(\mathcal{P})$ will be established and the proof will be complete.

But the claim must be true, because for any y,

$$\begin{aligned} & e' \in f^*_{\Theta,\mathcal{O}}(y) \\ \Rightarrow\ & e' \in f^*_{\Theta,\mathcal{O}}(y) \cap E_0 \\ \Rightarrow\ & f^*_{\Theta,\mathcal{O}}(y) \in f^*_{\Theta,\mathcal{O}}(T) \\ \Rightarrow\ & y \in T \end{aligned}$$

the last step following because $f^*_{\Theta,\mathcal{O}}$ is appropriate. ■

References

[Amarel, 1981] Saul Amarel. On representations of problems of readings about actions. In *Readings in Artificial Intelligence*. Morgan Kaufmann, 1981.

[Angluin, 1988] Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4), 1988.

[Chrisman, 1989] Lonnie Chrisman. Evaluating bias during pac-learning. In *Proceedings of the Sixth International Workshop on Machine Learning*, Ithaca, New York, 1989. Morgan Kaufmann.

[Cohen, 1988] William W. Cohen. Generalizing number and learning from multiple examples in explanation-based learning. In *Proceedings of the Fifth International Conference on Machine Learning*, Ann Arbor, Michigan, 1988. Morgan Kaufmann.

[Cohen, 1989] William W. Cohen. Abductive explanation based learning: A solution to the multiple explanation problem. Technical Report ML-TR-26, Rutgers University, 1989.

[Cohen, 1990a] William W. Cohen. An analysis of representation shift in concept learning. In *Proceedings of the Seventh International Conference on Machine Learning*, Austin, Texas, 1990. Morgan Kaufmann.

[Cohen, 1990b] William W. Cohen. Learning from textbook knowledge: A case study. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, Boston, Massachusetts, 1990. MIT Press.

[DeJong and Mooney, 1986] Gerald DeJong and Raymond Mooney. EBL: An alternative view. *Machine Learning*, 1(2), 1986.

[Drastal *et al.*, 1989] George Drastal, Gabor Czako, and Stan Raatz. Induction in abstraction spaces: A form of constructive induction. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, Michigan, 1989. Morgan Kaufmann.

[Flann and Dietterich, 1989] Nicholas Flann and Thomas Dietterich. A study of explanation-based methods for inductive learning. *Machine Learning*, 4(2), 1989.

[Gold, 1967] Mark Gold. Language identification in the limit. *Information and Control*, 10, 1967.

[Hirsh, 1989] Haym Hirsh. Incremental version space merging: A general framework for concept learning. PhD Thesis, Stanford University Department of Computer Science, 1989.

[Holte and Zimmer, 1989] Robert Holte and Robert Zimmer. A mathematical framework for studying representations. In *Proceedings of the Sixth International Workshop on Machine Learning*, Ithaca, New York, 1989. Morgan Kaufmann.

[Kedar-Cabelli and McCarty, 1987] Smadar Kedar-Cabelli and L. Thorne McCarty. Explanation-based generalization as resolution theorem proving. In *Proceedings of the Fourth International Workshop on Machine Learning*, Irvine, California, 1987. Morgan Kaufmann.

[Keller, 1989] Rich Keller. Compiling learning vocabulary from performance description. In *Proceedings of the Sixth International Workshop on Machine Learning*, Ithaca, New York, 1989. Morgan Kaufmann.

[Korf, 1981] Richard Korf. Toward a model of representation change. *Artificial Intelligence*, 14:41–78, 1981.

[Littlestone, 1988] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4), 1988.

[Lowry, 1988] Michael Lowry. STRATA: Problem reformulation and abstract data types. In *Change of Representation and Inductive Bias*. Kluwer Academic Publishers, 1988.

[Michalski and Chilausky, 1980] R.S. Michalski and R. L. Chilausky. Learning from being told and learning from examples: an experimental comparision of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *Policy Analysis and Information Systems*, 4, 1980.

[Mitchell *et al.*, 1986] T. Mitchell, R. Keller, and S. Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine Learning*, 1(1), 1986.

[Newell, 1982] Allen Newell. The knowledge level. *Artificial Intelligence*, 18:87–127, 1982.

[Riddle, 1989] Patricia Riddle. Reformulation from state space to reduction space. In *Proceedings of the Sixth International Workshop on Machine Learning*, Ithaca, New York, 1989. Morgan Kaufmann.

[Russell and Grosof, 1989] Stuart Russell and Benjamin Grosof. Declarative bias for structural domains. In *Proceedings of the Sixth International Workshop on Machine Learning*, Ithaca, New York, 1989. Morgan Kaufmann.

[Subramanian, 1989] Devika Subramanian. Representational issues in machine learning. In *Proceedings of the Sixth International Workshop on Machine Learning*, Ithaca, New York, 1989. Morgan Kaufmann.

[Utgoff, 1984] Paul Utgoff. Shift of bias for inductive concept learning. Technical Report CBM-TR-145, Rutgers University, 1984.

[Valiant, 1984] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11), November 1984.