Machine Learning, 10, 79-110 (1993)
© 1993 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

Induction Over the Unexplained: Using Overly-General Domain Theories to Aid Concept Learning

RAYMOND J. MOONEY MOONEY@CS.UTEXAS.EDU
Department of Computer Sciences, Taylor Hall 2.124, University of Texas, Austin, TX 78712

Editor: Thomas Dietterich

Abstract. This paper describes and evaluates an approach to combining empirical and explanation-based learning called *Induction Over the Unexplained* (IOU). IOU is intended for learning concepts that can be partially explained by an overly-general domain theory. An eclectic evaluation of the method is presented which includes results from all three major approaches: empirical, theoretical, and psychological. Empirical results show that IOU is effective at refining overly-general domain theories and that it learns more accurate concepts from fewer examples than a purely empirical approach. The application of theoretical results from PAC learnability theory explains why IOU requires fewer examples. IOU is also shown to be able to model psychological data demonstrating the effect of background knowledge on human learning.

Keywords. Combining explanation-based and empirical learning, knowledge-base refinement, theory specialization

1. Introduction

1.1. Motivation

Recent research in machine learning has primarily consisted of work on *explanation-based* or *empirical (similarity-based)* methods. Explanation-based methods (Mitchell, Keller, & Kedar-Cabelli, 1986; DeJong & Mooney, 1986) can learn accurate and efficient concept definitions from a single example; however, they require a complete and correct domain theory. Empirical methods (Quinlan, 1986; Michalski, 1983) can learn accurate and efficient concept definitions from large sets of examples; however, they generally fail to take adequate advantage of existing domain knowledge. Developing a learning method that combines explanation-based and empirical techniques has recently emerged as an important area of research in machine learning (Segre, 1989). The goal of this work is to develop a learning system that takes advantage of the strengths of both of these general approaches to concept acquisition.

This paper describes and evaluates such a system called *Induction Over the Unexplained* (IOU). Unlike other approaches to combining empirical and explanation-based techniques that use one method to bias or focus the other (Lebowitz, 1986; Pazzani, 1990; Flann & Dietterich, 1989), IOU uses each method to learn a different part of the final concept description. Many concepts have both explanatory and nonexplanatory aspects. For example, scripts for events such as a birthday party or a wedding have goal-directed as well as ritualistic actions. Concepts for artifacts such as a cup or a building have functionally important features as well as aesthetic or conventional ones. Animals have some attributes with clear survival value as well as more obscure features. Diseases have some symptoms that can be causally

explained by current biological theory as well as others that are simply known to be correlated with the condition. In IOU, explanation-based methods are used to learn the part of the concept definition that can be explained by an underlying domain theory. Empirical methods are used to learn the part of the concept definition consisting of unexplained regularities in the examples. Consequently, IOU is useful for acquiring concepts that can be partially explained by an existing overly-general domain theory. The background knowledge supplied by the domain theory allows the system to learn accurate concepts from fewer examples than a purely empirical approach. The empirical component of IOU allows it to overcome the incompleteness and over-generality of the domain theory and refine its background knowledge based on experience.

1.2. Evaluation methodology

Evaluation has become an increasingly important topic in machine learning research (Langley, 1989; Kibler & Langley, 1988). In general, three main methodologies have arisen:

Empirical Evaluation: Implemented learning systems are run on a set of benchmark problems and data are collected on their performance. Controlled experiments are conducted to determine the effect of independent variables such as learning algorithm, amount of training data, and amount of noise or missing values on dependent variables such as predictive accuracy and runtime (Kibler & Langley, 1988). Prototypical examples include Quinlan (1986), Fisher (1987), Minton (1988), and Shavlik, Mooney, and Towell (1991).

Theoretical Evaluation: Mathematical analysis is performed on learning algorithms to determine their computational complexity and sample complexity (i.e., the number of examples required to learn an accurate concept). Early theoretical work focussed on the *learning in the limit* model originated by Gold (1967). Recent research has focussed on the *PAC (probably approximately correct)* framework initiated by Valiant (1984), in which it is proven that if an algorithm is given a sufficient number of examples, with high probability it will learn a close approximation to the target concept (Haussler, 1988; Kearns, Li, Pitt, & Valiant, 1987).

Psychological Evaluation: Implemented learning algorithms and human subjects are given the same learning data and their performance is compared. Work under this approach generally tries to show that the system exhibits the same relative change in performance over different tasks or different amounts of training. Prototypical examples include Medin, Wattenmaker, and Michalski (1987), Skorstad, Falkenhainer, and Genter (1987), Roenbloom and Newell (1987), and Pazzani and Silverstein (1990).

For the most part, the researchers who pursue these different approaches to evaluation are disjoint and attend different academic conferences, although recently there have been a few researchers who have evaluated their work in multiple ways, such as employing both empirical and theoretical methodologies (Flann & Dietterich, 1989; Cohen, 1990; Pazzani & Sarrett, 1990).

This methodological division is unfortunate since the different evaluation methods are complementary rather than conflicting. Theoretical analysis generally provides information

on worst-case, asymptotic behavior and ignores average-case performance and constant factors. Also, theoretical analysis generally makes simplifying assumptions that may not be met by real problems. Empirical analysis attempts to address these issues by demonstrating successful performance of an implemented system on actual problems. However, empirical work is fundamentally limited by the assumption that results on specific problems are characteristic of general performance on a large class of problems. Psychological analysis, on the other hand, provides evidence that a system resembles the world's best-known learning system. *Homo sapiens*. Thousands of years of evolution have shaped the human mind to learn effectively in real-world situations and make a successful trade-off between accuracy and efficiency. Therefore, evidence that an algorithm exhibits aspects of human learning is evidence that it is a successful learning strategy. Psychological evaluation can also result in a greater understanding of human intelligence and contribute to the development of improved educational methods. However, human learning may also exhibit certain biological quirks and inefficiencies that we do not want our artificial learning systems to emulate.

Consequently, the best way to evaluate a learning system is to look at it from all of these angles. This paper presents an eclectic evaluation of the IOU method using empirical data, formal analysis, and the results of psychological simulation. The empirical evaluation shows that IOU can use an overly-general theory to learn more accurate concepts from fewer examples than standard empirical learning. The formal analysis shows that IOU runs in linear time and requires fewer examples to learn a *probably approximately correct* (PAC) concept (Valiant, 1984) than a purely empirical approach. Finally, the psychological simulation demonstrates that IOU can model the results of experiments demonstrating the effects of background knowledge on concept acquisition in humans.

2. The IOU method

The IOU method uses explanation-based techniques to determine as much as possible about a concept and then uses empirical methods to detect regularities in the unexplainable aspects of the examples. Features that can be explained by the existing domain theory are identified and removed from all of the examples. The reduced example descriptions are passed on to a standard empirical system, which finds additional commonalities and adds them to the concept definition. A test example must meet the requirements of the domain theory as well as the empirically learned definition in order to be considered a member of the concept. This section presents detailed information about the problem IOU addresses and the learning algorithm it uses.

2.1. Problem definition for IOU

The general problem IOU addresses is *theory-based concept specialization* as defined by Flann and Dietterich (1989). The system is assumed to have a domain theory for a generalization of the concept to be learned. A definition of the specific problem addressed by IOU is given in Table 1. The current implementation of IOU employs a feature-based description language that includes binary, discrete, and real-valued attributes. A domain theory

Table 1. The problem addressed by IOU.

Given:

- A set of positive and negative examples of an intended concept, C_i.
- A propositional Horn-clause domain theory for an explainable concept, C_e, that is a generalization of the intended concept, i.e., $C_i \rightarrow C_e$. An additional restriction is that the correct definition of C_i must be expressible in the form $C_e \wedge C_u$ where the features referenced in the definition of the unexplanable concept C_u are disjoint from the features referenced in the definition of C_e (i.e., the features used in the domain theory).

Find:

- A definition for the intended concept that is consistent with the examples and a specialization of the explainable concept.

is restricted to a set of propositional Horn-clause rules; however, propositions can include feature value pairs (e.g., COLOR = RED) and numerical thresholds (e.g., LENGTH < 3) as well as binary propositions (e.g., HAS-HANDLE).

Like other systems that employ overly-general theories (Flann & Dietterich, 1989; Cohen, 1990), IOU is restricted to a certain subclass of theory specialization problems. As described in Table 1, IOU assumes that the additional features needed to specialize the concept are disjoint from the features already explained by the domain theory. The relationship between this restriction and those of other theory specializers is discussed in Section 6.1. We believe this restriction is appropriate when learning concepts with separate explanatory and nonexplanatory constraints, such as those introduced as motivation for this work in Section 1.1. In addition, some psychological justification for this restriction is given in Section 5.

As an example of a problem suitable for IOU, consider the standard example in which the intended concept is CUP (Winston, 1983). The domain theory is the usual one with two exceptions. First, it has been made propositional. Second, the explainable concept has been renamed DRINKING-VESSEL, since the theory cannot actually distinguish between the concepts CUP, GLASS, MUG, SHOT-GLASS, etc.

STABLE ∧ LIFTABLE ∧ OPEN-VESSEL → DRINKING-VESSEL
HAS-BOTTOM ∧ FLAT-BOTTOM → STABLE
GRASPABLE ∧ LIGHTWEIGHT → LIFTABLE
HAS-HANDLE → GRASPABLE
WIDTH = SMALL ∧ INSULATING → GRASPABLE
HAS-CONCAVITY ∧ UPWARD-POINTING-CONCAVITY → OPEN-VESSEL

Assume the set of examples includes cups, shot-glasses, mugs and cans, as shown in Table 2. The problem is to use the domain theory to learn the explainable features of a cup (FLAT-BOTTOM, HAS-CONCAVITY, etc.) and to use empirical techniques to learn the nonexplanatory features (VOLUME = SMALL) that rule out shot glasses and mugs.

2.2. The IOU learning algorithm

A description of the basic IOU algorithm is shown in Table 3. Step 1 simply traverses back through the rules in the domain theory to find all of the features that can be explained

Table 2. Examples for learning CUP.

	CUP-1 (+)	CUP-2 (+)	SHOT-GLASS-1 (−)	MUG-1 (−)	CAN-1 (−)
HAS-BOTTOM	TRUE	TRUE	TRUE	TRUE	TRUE
FLAT-BOTTOM	TRUE	TRUE	TRUE	TRUE	TRUE
HAS-CONCAVITY	TRUE	TRUE	TRUE	TRUE	TRUE
UPWARD-POINTING	TRUE	TRUE	TRUE	TRUE	TRUE
LIGHTWEIGHT	TRUE	TRUE	TRUE	TRUE	TRUE
HAS-HANDLE	TRUE	FALSE	FALSE	TRUE	FALSE
WIDTH	SMALL	SMALL	SMALL	MEDIUM	SMALL
INSULATING	FALSE	TRUE	TRUE	FALSE	FALSE
COLOR	WHITE	RED	WHITE	COPPER	SILVER
VOLUME	SMALL	SMALL	TINY	LARGE	SMALL
SHAPE	CYLINDER	CYLINDER	CYLINDER	CYLINDER	CYLINDER

Table 3. The basic IOU algorithm.

(1) Examine the domain theory to determine the set of *explainable features*.
(2) Discard any negative examles that are not members of the explainable concept as determined by the domain theory.
(3) Remove the explainable features from the descriptions of the remaining examples.
(4) Give the "reduced" set of examples to a standard empirical learning system to compute the unexplainable component of the concept (C_u).
(5) Output: $C_e \wedge C_u$ as the final concept description.

by the domain theory. This set includes all features that are referenced by *any* proof supported by the domain theory. The set of explainable features for the CUP example are:

HAS-BOTTOM, FLAT-BOTTOM, HAS-CONCAVITY, UPWARD-POINTING-CONCAVITY, LIGHTWEIGHT, HAS-HANDLE, WIDTH, and INSULATING.

Step 2 eliminates negative examples that do not satisfy the domain theory. A standard Horn-clause theorem prover is used to determine if an example can be proven as a member of the explainable concept. Since the theory is assumed to be overly-general, it already accounts for the classification of these examples. In the CUP example, the negative CAN-1 instance can be discarded. Although CAN-1 is a stable open-vessel, it is not graspable, because it is not insulating nor does it have a handle. Therefore it cannot function as a drinking vessel for hot liquids.

Step 3 removes the explainable features of the remaining examples to allow the empirical component to focus on their unexplainable aspects. The resulting reduced set of data for the sample problem is shown in Table 4. It should be noted that since IOU removes all values of all explained features from the remaining examples, it is sensitive to how information is encoded as features. For example, if COLOR is a multi-valued feature and COLOR = RED is used in the domain theory, then all colors of all examples will be removed. On the other hand, if RED and BLUE are separate binary features and only one is used in the theory, then the other one will be treated as an unexplained feature and will be included in the reduced data. However, standard inductive systems (e.g., ID3, AQ) are also affected by such representational choices.

Table 4. Reduced examples for learning CUP.

	CUP-1 (+)	CUP-2 (+)	SHOT-GLASS-1 (−)	MUG-1 (−)
COLOR	WHITE	RED	WHITE	COPPER
VOLUME	SMALL	SMALL	TINY	LARGE
SHAPE	CYLINDER	CYLINDER	CYLINDER	CYLINDER

In step 4, the unexplained data are given to a standard empirical system for learning from examples. We currently use a version of ID3 (Quinlan, 1986) as the empirical component. ID3's decision tree is translated into a set of Horn-clause rules so that the final concept definition is in a uniform representation language. This step is similar to the translation process described by Quinlan (1987); however, IOU does not prune the resulting rules. For the sample problem, ID3 generates the simplest description: VOLUME = SMALL.

There are two important aspects to notice about IOU's use of the empirical component. First, it can use any empirical system that supports the description language used by the overall system. Even connectionist, genetic, or instance-based learning algorithms could be used; however, in these cases, the final concept would not be represented in a uniform language. Second, the amount of data (number of features and number of examples) given to the empirical component is reduced by the explanation-based component. This decreases computational complexity and helps focus the empirical component.

The final step of IOU simply combines the explanatory and nonexplanatory constraints into a final concept definition. For the example, this produces the following definition:

$$\text{DRINKING-VESSEL} \wedge \text{VOLUME} = \text{SMALL} \rightarrow \text{CUP}$$

New examples are classified by using the domain theory to determine if they meet the explanatory constraints and using the empirically learned definition to determine if they meet the nonexplanatory constraints. It is interesting to note that when ID3 is run on the data in Table 2, the extra negative example causes COLOR to be the most informative feature, and the system produces the following rule:

$$\text{COLOR} = \text{RED} \vee (\text{COLOR} = \text{WHITE} \wedge \text{HAS-HANDLE}) \rightarrow \text{CUP}$$

ID3 would clearly need many more examples to learn the correct concept.

There are two special cases that arise in step 5 of the algorithm. The first case occurs when all of the training examples are negative. Many empirical systems, like ID3, return the null concept in this case, i.e., all test examples are classified as negative. If the null concept is conjunctively combined with the explanatory concept, the result is the null concept. Since this would completely destroy the effectiveness of the domain theory, IOU does not alter the concept at all in this case, i.e., the learned definition is simply the explanatory concept. A better solution would be to use an empirical system that does not return the null concept when given all negative examples. One alternative would be to return the negation of the most-specific conjunctive generalization of the negative examples.

The second special case is when the reduced data passed to the empirical learner is inconsistent, i.e., there are both positive and negative examples with the same reduced description.

This can only happen when there is noise in the data or when the intended concept does not satisfy the assumption that the features in C_e and C_u are disjoint. In order to handle this case, the version of ID3 used in IOU includes a simple noise-handling technique. When ID3 has partitioned the data to the point that it contains only examples with the same description but different classes, it creates a leaf and labels it with the most common class in the partition. Other techniques for handling noise in ID3 could also be used (Quinlan, 1986; Mingers, 1989).

The algorithm in Table 3 is for learning a single concept definition. If the data consist of multiple, mutually-exclusive categories, then multiple trials are run to learn a different definition for each category. In each trial, the examples of one category are treated as positive examples and the examples of all other categories are treated as negative. These data are then given to the algorithm in Table 3 in order to learn a definition for the current category. Unfortunately, this approach may learn overlapping definitions (Dietterich, London, Clarkson, & Dromney, 1982). In IOU, this problem is resolved in the following manner. If the learned definitions classify a test example in more than one category, then the example is assigned to the matching category with the most examples in the training set. If an examples does not match the definition of any category, then it is assigned to the overall most common category in the training set.

In its current implementation, IOU is a batch learning system; however, the basic algorithm is easily made incremental if the empirical component is itself incremental. In incremental mode, each time a new example is encountered it is either discarded as an unprovable negative example or its explainable features are removed and the remaining features are passed along to the empirical component, which incrementally forms the unexplainable part of the definition. For example, an incremental system like ID5 (Utgoff, 1989) could be used as the empirical learner.

3. Theoretical analysis of IOU

This section presents a formal analysis of the sample complexity and computational complexity of IOU. First, we use existing results in learnability theory to show that the lower bound on the number of examples required to learn a PAC concept definition with IOU is less than with a purely empirical learner. Second, we show that the computational complexity of IOU is linear assuming the empirical component has linear complexity.

3.1. PAC analysis of IOU

This section presents theoretical evidence that IOU learns from fewer examples than a purely empirical approach. The analysis uses existing results in *PAC learnability*, a theoretical framework for analyzing learning algorithms originated by Valiant (1984). PAC learnability theory is primarily concerned with determining the number of examples required by a learning algorithm to guarantee that with probability $1-\delta$ the concept description produced by the algorithm has an error rate of at most ϵ. The basic approach taken in this section is to show that since IOU can assume that it already has part of the correct concept description,

it explores a smaller hypothesis space when learning the remaining part of the concept and therefore can learn a PAC concept description from fewer examples.

Since we do not want our analysis to depend on any particular empirical learning algorithm, we examine information-theoretic lower bounds on the number of examples required by any empirical learning algorithm to produce an answer that is PAC. In this regard, we use the following result:

THEOREM 1 (Ehrenfeucht, Haussler, Kearns, & Valiant, 1989). *Any PAC learning algorithm for a non-trivial*[1] *hypothesis space H must use sample size*

$$\Omega\left(\frac{1}{\epsilon}\left[\ln\frac{1}{\delta} + VCdim(H)\right]\right).$$

The *Vapnik-Chervonenkis dimension* (VC-dimension or VCdim) is a measure of the expressiveness of an hypothesis space (Haussler, 1988). Specifically, we say that a set of hypotheses H *shatters* a set of examples E if, for every possible way of labeling the elements of E positive or negative, there exists an hypothesis in H that will produce that labeling. The VC-dimension is defined to be the size $|E|$ of the largest set of examples such that H shatters E. It is easily shown that $VCdim(H) \leq \log_2|H|$ (Haussler, 1988).

For the featural description language used by IOU, let H_a denote the space of hypotheses considered by the empirical learner given *all* of the features used to describe examples and let ϵ_a represent the desired maximum error rate for the definition learned for the intended concept. By simple instantiation of Theorem 1, the total number of examples required by a purely empirical learning algorithm (m_{SBL}) is given by

$$m_{SBL} = \Omega\left(\frac{1}{\epsilon_a}\left[\ln\frac{1}{\delta} + VCdim(H_a)\right]\right).$$

With respect to IOU, let H_u denote the space of hypotheses considered by the empirical learner given only the unexplained features (i.e., features not referenced by the overly-general domain theory) and H_e denote the space of hypotheses over only the explained features. Recall that a critical assumption of IOU is that the concept can be represented in the form $C_e \wedge C_u$ where the explained features in C_e the unexplained features in C_u are disjoint. Since the domain theory is assumed to represent a correct description of the explanatory part $C_e \in H_e$, the goal of the empirical component of IOU is to find a description $C_u \in H_u$ for the unexplained aspects such that $C_e \wedge C_u$ is a PAC description of the intended concept. If the desired maximum error rate of C_u is ϵ_u, then the number of examples (m_u) required to learn C_u is given by

$$m_u = \Omega\left(\frac{1}{\epsilon_u}\left[\ln\frac{1}{\delta} + VCdim(H_u)\right]\right).$$

Since only those examples that satisfy the explanatory component, C_e, are actually passed along to the empirical component of IOU, only a fraction of the overall examples are available for learning C_u. Let α be the fraction of examples that satisfy the explanatory

component (i.e., the fraction of the examples that are members of the explainable concept). Then the total number of examples required by IOU (m_{IOU}) is given by

$$m_{IOU} = \frac{m_u}{\alpha}.$$

The overall error rate of $C_e \wedge C_u$ is given by the probability that an example is a member of the explainable concept times the error rate for members plus the probability that an example is *not* a member of the explainable concept times the error rate for nonmembers. Assuming the explanatory component is correct, all nonmembers are correctly, all nonmembers are correctly classified as negative examples since they do not meet the functional requirements. Members of the explainable concept, on the other hand, are correctly classified when they are classified correctly by the nonexplanatory component C_u. Therefore, the overall error rate is given by

$$\epsilon_a = \alpha\epsilon_u + (1 - \alpha)0 = \alpha\epsilon_u.$$

Substituting for m_u and ϵ_u in the above equation for m_u and solving for m_{IOU} we get

$$m_{IOU} = \Omega\left(\frac{1}{\epsilon_a}\left[\ln\frac{1}{\delta} + VCdim(H_u)\right]\right).$$

This simple calculation shows that although IOU passes along only a fraction of the training examples to the empirical component, this is canceled by the fact that the error rate for C_u is only a fraction of the overall error rate. If IOU did not reduce the number of examples used to learn C_u, even these simple calculations would be unnecessary and the result would follow directly from the fact that the empirical component of IOU explores a smaller hypothesis space.

The above bounds on m_{SBL} and m_{IOU} demonstrate that the minimum number of examples required by a purely empirical system and by IOU differ only by the VC-dimension of their hypothesis spaces (H_a for SBL and H_u fo IOU). Since for any non-trivial domain theory the hypothesis space explored by IOU uses only a subset of the features used by SBL, $VCdim(H_u) < VCdim(H_a)$ and therefore IOU requires fewer examples to learn a PAC concept description.

By examining specific hypothesis spaces, we can obtain tighter results. For purely conjunctive concepts on n binary features, $VCdim(H) \geq n$ (Ehrenfeucht et al., 1989). Since the lower-bound in Theorem 1 is also tight within a constant factor for the hypothesis space of purely conjunctive concepts (Ehrenfeucht et al., 1989), if the number of unexplained features is u, then

$$\frac{m_{IOU}}{m_{SBL}} = \Theta\left(\frac{VCdim(Hu)}{VCdim(H_a)}\right) = \Theta\left(\frac{u}{n}\right).$$

In other words, the ratio of the number examples required by IOU to the number of examples required by SBL is proportional to the percentage of unexplained features.

The hypothesis space kCNF allows for a limited amount of disjunction. A kCNF expression is a conjunction of any number of clauses, where each clause is a disjunction of at most k literals. It is important to notice that for kCNF (as for pure conjunctions) if $C_e \in H_e$ and $C_u \in H_u$, then $C_e \wedge C_u \in H_a$. For kCNF concepts on n features, $VCdim(H) = \Omega(n^k)$ (Ehrenfeucht et al., 1989). Since the lower-bound in Theorem 1 is also tight within a constant factor for kCNF (Ehrenfeucht et al., 1989),

$$\frac{m_{IOU}}{m_{SBL}} = \Theta\left(\frac{VCdim(H_u)}{VCdim(H_a)}\right) = \Theta\left(\left(\frac{u}{n}\right)^k\right).$$

Therefore, the advantage of IOU is more pronounced for kCNF. For example, even if 90% of the features are unexplained and $k = 5$, $(u/n)^k = 0.59$.

Finally, consider the case of learning simple conjunctive concepts where the number of atoms in the conjunction is at most s, where $s \ll n$. In other words, there are a large number of irrelevant features. For such concepts, $VCdim(H) = \Omega(s \log(n/s))$ (Haussler, 1988). For IOU, this means that some number e of the s relevant features are already known and explained by the domain theory and the goal of the empirical component is to learn the remaining $s - e$ relevant features from the $u = n - e$ unexplained features. This is simply a reduced instance of the problem of learning a simple conjunctive concept and therefore:

$$\frac{m_{IOU}}{m_{SBL}} = \Theta\left(\frac{VCdim(H_u)}{VCdim(H_a)}\right) = \Theta\left(\frac{(s - e)\log((n - e)/(s - e))}{s \log(n/s)}\right).$$

If s (and therefore e) is held constant while n increases, then this equation reduces to $m_{IOU}/m_{SBL} = log(n)/log(n) = O(1)$. In this case, the advantage of IOU is fixed and does not scale with increasing number of features. However, consider holding s to a fixed fraction of n ($s = \beta n$). This case represents scaling the entire problem instead of fixing the size of the concept. Under this assumption,

$$\frac{m_{IOU}}{m_{SBL}} = \Theta\left(\frac{(\beta n - e)\log((n - e)/(\beta n - e))}{\beta n \log(n/\beta n)}\right) = \Theta\left(\frac{n - e}{n}\right).$$

Since $n - e = u$, this is the same ratio obtained for pure conjunctive concepts.

3.2. Complexity analysis of IOU

This section analyzes the computational complexity of the IOU algorithm shown in Table 3. We simply demonstrate that each step in the algorithm effectively takes linear time and therefore the overall complexity is linear. Step 1 simply searches back through the rules in domain theory to find all of the features it references. This process is clearly linear in the size of the domain theory. Step 2 attempts to prove that each negative example is a member of the explanatory concept according to the domain theory. Since theorem proving with a propositional Horn-clause theory can be done in linear time (Dowling & Gallier,

1984), step 2 is also linear in the size of the domain theory and the number of examples. Step 3 simply removes a subset of the features from a subset of the examples; this is clearly linear in the size of the data set. Step 4 calls an empirical learning system (ID3) on the reduced data set. The worst-case time complexity of ID3 is $O(en^2)$ where e is the number of examples and n is the number of features (Utgoff, 1989). However, empirical evidence indicates that ID3's average-case complexity is actually linear in the number of features rather than quadratic (Shavlik et al., 1991). Therefore, the complexity of step 4 is effectively linear, although in general it depends on the complexity of the empirical learner. Step 5 simply combines the two parts of the final definition in constant time.

4. Empirical analysis of IOU

This section presents empirical evidence that IOU can refine overly-general domain theories and learn from fewer examples than a purely empirical system. Two types of experiments were conducted. First, artificial data were constructed for concepts that were specializations of concepts defined by simple domain theories. Second, overly-general theories were generated for natural data sets by randomly deleting parts of correct theories that were generated inductively. In both cases, learning curves were generated to compare the performance of a purely empirical learner (ID3) to IOU given an overly-general domain theory. Since IOU uses ID3 as its empirical component, it should be noted that this is the same as comparing IOU with and without an initial domain theory. The primary hypothesis being tested in these experiments is that, when IOU is given an initial overly-general domain theory, it learns a more accurate concept from fewer examples than a purely empirical method.

4.1. Experiments on artificial data

In this experiment, artificial data was used to test IOU's ability to specialize the theory of drinking vessels introduced in Section 2.1. The following specialized theory of cups was created and used to generate examples of the intended concept:

$$\text{DRINKING-VESSEL} \wedge \text{VOLUME} = \text{SMALL} \wedge \text{SHAPE} = (\text{HEMISPHERICAL} \vee \text{CYLINDRICAL} \vee \text{CONICAL}) \rightarrow \text{CUP}$$

The entire instance space was generated by forming the Cartesian product of the domains for all of the features. All of the features were binary except for WIDTH {SMALL, MEDIUM, LARGE}, SHAPE {HEMISPHERICAL, CYLINDRICAL, CONICAL, CUBICAL}, VOLUME {TINY, SMALL, LARGE}, and COLOR {CLEAR, WHITE, COPPER}. The total instance space consisted of 13,824 examples, 252 of which qualified as DRINKING-VESSELs and 63 of which qualified as CUPs. The theory for cups was used to separate these instances into positive and negative examples.

From this large set of examples, disjoint training and tests sets were created that controlled for the percentage of positive examples (e.g., cups) as well as the percentage of negative examples that satisfied the overly-general domain theory (e.g., drinking-vessels

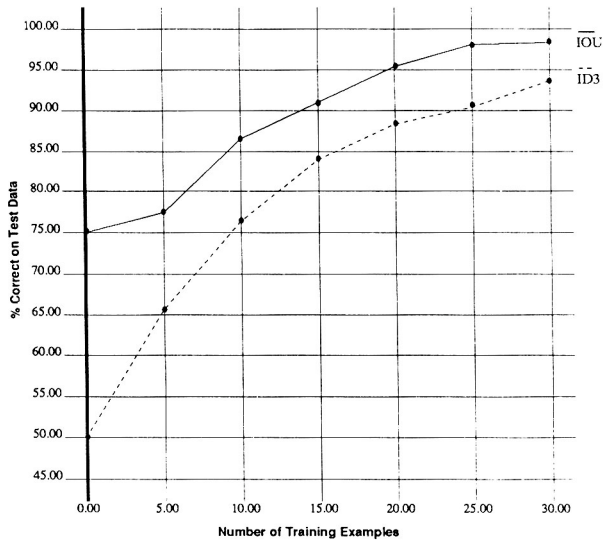

Figure 1. Learning curves for cup data.

that are not cups). Of course, the distribution was kept the same across training and test. Learning curves were generated by performing batch training on increasingly larger fractions of a set of training examples and repeatedly testing predictive accuracy on the same disjoint test set. The final results were averaged over 30 random selections of training and test sets.

Figure 1 shows learning curves for the cup data in which the training and test set contained 50% cups and 75% drinking-vessels (i.e., 50% of the negative examples are provable by the original theory). A test set consisted of 75 novel examples. When given zero training examples, IOU uses its initial theory and ID3 picks a class at random. The results show that IOU starts with a 25 percentage point advantage since it always correctly classifies the 25% of the examples that are not drinking-vessels. It maintains this advantage throughout training, although by 30 training examples the difference has narrowed to 5 percentage points. A separate one-tailed t-test for paired differences was run on each pair of plotted points and indicated that all differences between pairs of points on the learning curve were statistically significant ($p < .005$). The narrowing of the difference between ID3 and IOU is to be expected, since after enough training examples, both systems should converge to 100% correct, since both maintain consistency with all training examples and the hypothesis space is finite (Haussler, 1988). A similar experiment using an overly-general theory for PARTY to learn the specialized concept BIRTHDAY-PARTY produced similar results (Mooney, Ourston, & Wu, 1989).

The seemingly small (5%) difference in accuracy after 30 examples does not adequately reflect the difference between the results of the two systems. In one run chosen at random, IOU learned the completely correct concept after 30 training examples, while the concept learned by ID3 was:

(SHAPE = CYLINDRICAL ∧ VOLUME = SMALL) ∨ (HAS-CONCAVITY ∧ UPWARD-POINTING-CONCAVITY ∧ SHAPE = CONICAL ∧ VOLUME = SMALL) ∨ (SHAPE = HEMISPHERICAL ∧ VOLUME = SMALL) → CUP

Although this definition had an accuracy of 92%, it fails to reference many important features such as HAS-BOTTOM, LIGHTWEIGHT, HAS-HANDLE, INSULATING, etc., and most of the features it does reference are strangely distributed across different disjuncts.

The distribution of classes in the training and test data has certain predictable effects on the relative accuracy of IOU and ID3. If the percentage of examples that are drinking-vessels is increased, the results remain qualitatively the same; however, the accuracy of the initial theory, and therefore the gap between IOU and ID3, decreases. This is because the initial theory is guaranteed to correctly classify non-drinking-vessels as non-cups, and therefore decreasing the number of non-drinking-vessels decreases its accuracy. On the other hand, if the percentage of examples that are cups is increased while holding the percentage of drinking-vessels constant, the accuracy of the initial theory, and therefore the gap between IOU and ID3, increases. This is because the initial theory is overly-general and guaranteed to correctly classify all cups as positive examples.

4.2. Experiments on natural data

An ideal test of IOU would involve an actual domain theory acquired from an expert or a text-book and a large corpus of natural data. Unfortunately, there are currently very few natural data sets together with imperfect domain theories. The few that do exist have theories that are overly-specific (Towell, Shavlik, & Noordeweir, 1990) or otherwise unsuitable for IOU (Flann & Dietterich, 1989; Cohen, 1990) (see Section 6.1 for a discussion of different types of overly-general theories). Consequently, overly-general theories were created for two existing natural data sets by using inductive learning to learn an initial set of rules and then deleting antecedents from these rules to make them overly-general. The two data sets are briefly described below.

The *soybean* data set has 17 different soybean diseases described by 50 features, such as weather, time of year, and descriptions of leaves and stems. Each disease has 17 examples, for a total of 289 examples. This domain was popularized by Michalski and Chilausky (1980); however, the exact data are those used by Reinke (1984). The full soybean data set used here should not be confused with the much simpler, purely conjunctive, four-disease data used to test clustering systems (Stepp, 1984; Fisher, 1987).

The *audiology* data set (Porter, Bareiss, & Holte, 1990) consists of 226 cases of hearing disorders from the Baylor College of Medicine. There are 24 categories of hearing disorders and 68 features involving reported symptoms and the results of various hearing tests. Most of the features are binary and the remaining ones are discrete.

4.2.1. Experiment 1: Comparing learning rates

This first experiment tests the hypothesis that IOU learns a more accurate concept from fewer training examples when given an overly-general theory. Overly-general theories were

artificially created and the effect of these theories on IOU's learning rate was determined. ID3 was trained on the complete data sets for both the soybean and audiology domains and the resulting trees were translated into rules. The result is a "completely correct domain theory" with respect to the data. For the soybean data, this produced 116 Horn-clause rules with an average of 6.8 rules (disjuncts) per catagory and 3.2 antecedents per rule. For the audiology data, it produced 104 Horn-clause rules with an average of 4.3 rules (disjuncts) per category and 5.3 antecedents per rule. The rules for each category were then generalized by randomly deleting 20% of the features referenced in the antecedents of the rules for that category. If a feature was selected for deletion, any antecedent that referenced any of its values was removed. The accuracies of the resulting theories were 50% for soybean and 39% for audiology.

An example should help clarify the theory generalization process and the form of the final domain theory. Below are the initial rules for the audiology disorder OTITIS-MEDIA produced by ID3 when given all 226 examples.

BONE = UNMEASURED ∧ TYMP = C ∧ ¬AGE-GT-60 → OTITIS-MEDIA
SPEECH = VERY-POOR ∧ TYMP = B ∧ ¬AGE-GT-60 → OTITIS-MEDIA
SPEECH = NORMAL ∧ TYMP = B ∧ ¬AGE-GT-60 → OTITIS-MEDIA
O-AR-C = ABSENT ∧ AR-U = ELEVATED ∧ ¬FLUCTUATING ∧ AIR = MILD ∧ ¬NOISE ∧ TYMP = A ∧ ¬AGE-GT-60 → OTITIS-MEDIA

By randomly deleting two of the nine features appearing in the rules, TYMP and AR-U, these rules were generalized to:

BONE = UNMEASURED ∧ ¬AGE-GT-60 → OTITIS-MEDIA
SPEECH = VERY-POOR ∧ ¬AGE-GT-60 → OTITIS-MEDIA
SPEECH = NORMAL ∧ ¬AGE-GT-60 → OTITIS-MEDIA
O-AR-C = ABSENT ∧ ¬FLUCTUATING ∧ AIR = MILD ∧ ¬NOISE ∧ ¬AGE-GT-60 → OTITIS-MEDIA

Similar generalizations are performed on the rules for all of the other categories.

Although this process is guaranteed to create an overly-general theory, it will not necessarily create a theory meeting the constraint that the explainable and unexplainable features are disjoint. For example, suppose the correct theory is:

$X \wedge Y \rightarrow C$
$Z \rightarrow C$

If the feature X is deleted to produce the overly-general theory

$Y \rightarrow C_e$
$Z \rightarrow C_e$,

the theory cannot be corrected by conjoining X to C_e, since X occurs in only one disjunct of the original theory. However, this problem did not prevent IOU from being able to fit

the training data adequately. It was always able to receive 100% training accuracy on the soybean data, and training accuracy on the audiology data averaged at least 96% for all of the various sizes of training sets that were tested.

IOU was given the resulting overly-general theories and its performance was compared to ID3's. As in the experiments on artificial data, learning curves were generated by performing batch training on increasingly larger fractions of a set of training examples and repeatedly testing predictive accuracy on the same disjoint test set. Each soybean test set contained 89 examples, while the audiology test sets contained 76 examples. The final results were averaged over 30 random selections of training and test sets.

Figure 2 shows learning curves for the soybean and audiology data sets. In both cases, IOU substantially increased the accuracy of the original theory and maintained its advantage over ID3. At each point plotted on the learning curves, the difference between IOU and ID3 is highly significant according to a single-tailed t-test for paired differences ($p < .0005$).

On multi-category tasks like the soybean and audiology domains, there is an additional difference between IOU and ID3 that may be important. IOU learns separate rule sets for each category (see Section 2.2) while ID3 learns a single decision tree that discriminates all of the categories. Quinlan (1986) speculated that learning a separate decision tree for each category may give more accurate results. Therefore, we also tested a version called ID3-SC (single class) that learns a separate decision tree for each category using the same approach to multi-category data used in IOU. Figure 2 also shows the performance of this system. On the soybean data, ID3-SC does significantly better than ID3 ($p < .05$ for all points except 0) but still significantly worse than IOU ($p < .0005$ for all points). On the audiology data, ID3-SC does significantly better than ID3 for smaller amounts of training data ($p < .025$ for 10 and 20 examples); however, it always does significantly worse for larger amounts ($p < .025$ for 80, 110, and 150 examples); however, it always does significantly worse than IOU ($p < .0005$ for all points). With regard to decision-tree induction, these results support the benefits of learning a separate decision tree for each category, at least for small amounts of training data. It may also help explain why Shavlik et al. (1991) found that ID3 performed worse than connectionists methods on the soybean data and on small amounts of training data in general and suggests that ID3-SC would be more competitive in these situations.

Another way of measuring the advantage IOU gets from its initial overly-general theory is by examining the additional number of examples required by a purely empirical system to obtain the same level of performance. The learning curves for the soybean data show that IOU's accuracy after 100 examples is about the same as ID3-SC's accuracy after 200 examples. Therefore, IOU's initial theory is giving it a "100 example advantage" at this point. IOU also shows about a "100 example advantage" on the audiology data after 50 examples.

As an example of the refinements made by IOU, consider the unexplanatory component learned for OTITIS-MEDIA (the audiology category discussed earlier) after 150 training examples:

$$\text{AR-U} = \text{ELEVATED} \wedge \text{TYMP} = \text{C} \rightarrow \text{C}_u$$
$$\text{AR-U} = \text{NORMAL} \wedge \text{TYMP} = \text{C} \rightarrow \text{C}_u$$
$$\text{AR-U} = \text{UNMEASURED} \wedge \text{TYMP} = \text{C} \rightarrow \text{C}_u$$
$$\text{TYMP} = \text{B} \rightarrow \text{C}_u$$

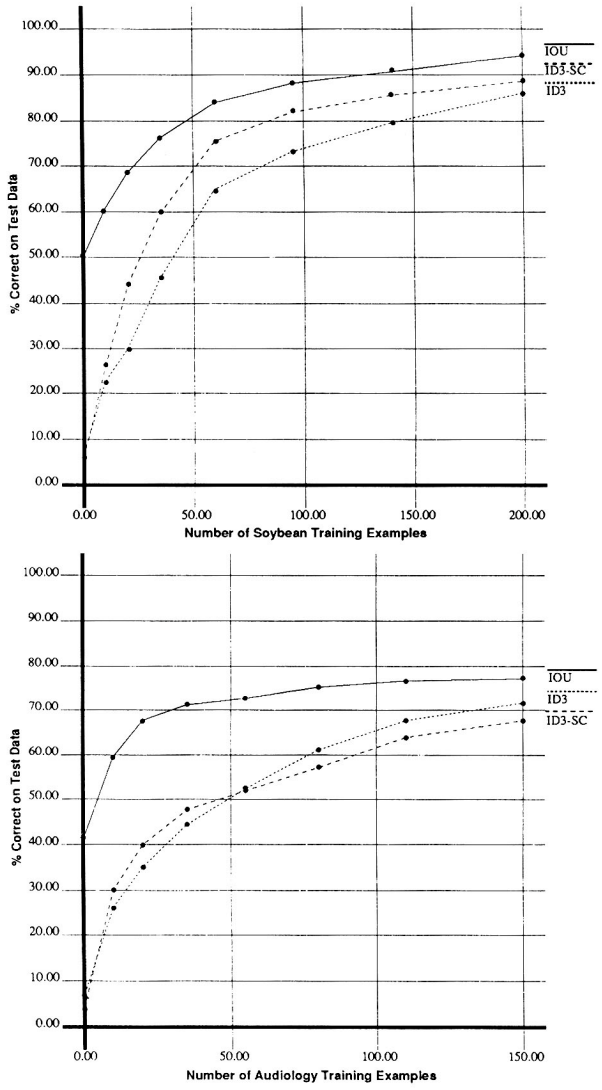

Figure 2. Learning curves for soybean and audiology data.

IOU quite nicely re-introduces the two features deleted from the original "correct" theory and even uses all of the original values; however, the actual logical combination is somewhat different. It should be noted that the only unmentioned value for AR-U is ABSENT, and that TYMP has four other values besides B and C.

4.2.2. Experiment 2: Varying the levels of over-generality and disjointness

A couple of issues raised by the first experiment concern IOU's sensitivity to aspects of the initial theory such as the degree of overly-generality and the degree to which the disjointness assumption is violated. This section presents empirical results on how these parameters affect IOU's predictive accuracy. The hypothesis is that IOU's performance degrades gracefully as the theory is made increasingly overly-general and as the disjointness assumption is increasingly violated.

First, the methodology from Experiment 1 was repeated while varying the percentage of features randomly deleted from the complete theory generated by ID3. The fraction of the features that are deleted is called the *level of over-generality*. The size of the training and test sets were fixed at 75 and 100 examples, respectively, and the results were averaged over 30 trials. Each trial used a different randomly-generated overly-general theory, training set, and test set.

The results are presented in Figure 3. As would be expected, IOU's accuracy starts at 100% with the complete theory and asymptotically approaches the performance of ID3-SC as the number of deleted features is increased. When all of the features are deleted from the theory (over-generality level of 1.0), IOU behaves exactly like ID3-SC, since there are no explainable features and all of the data for each category are passed along to ID3. The slight fluctuations in the performance of ID3 and ID3-SC are due to randomly breaking ties in the ID3 splitting criterion.

Second, the methodology from Experiment 1 was repeated while varying the amount of overlap between the explainable and unexplainable features. As in previous experiments, the features to be completed deleted from the theory were randomly selected given the desired level of over-generality. However, when each occurrence of a chosen feature was about to be deleted, with a certain probability (called the *level of overlap*) a random antecedent in the same rule was deleted instead. Consequently, when the level of overlap is 1.0, an equivalent number of antecedents are simply deleted at random from the initial theory. Therefore, changing this parameter allows one to measure IOU's sensitivity to increasing violations of the disjointness assumption. As in the experiments with level of over-generality, the size of the training and test sets were fixed at 75 and 100 examples, respectively, and the results were averaged over 30 trials. The level of over-generality was fixed at the intermediate value of 0.2.

The results are presented in Figure 4. On the soybean data, increasing the level of overlap only slightly degrades the performance of IOU. IOU consistently remains significantly better than both ID3 and ID3-SC. On the audiology data, on the other hand, the performance of IOU gradually degrades almost to the level of ID3 (it is not significantly better past an overlap level of 0.5). The difference between the soybean and audiology results can be explained by the fact that the soybean data have more redundant features. Results by Shavlik et al. (1991) show that randomly dropping up to half of the features in the soybean data does not significantly affect the performance of several inductive algorithms, while their performance on audiology degrades approximately linearly as more features are dropped. When there are redundant features, it is likely that any information present in the explanatory features removed by IOU is repeated in the non-explanatory features passed along to the

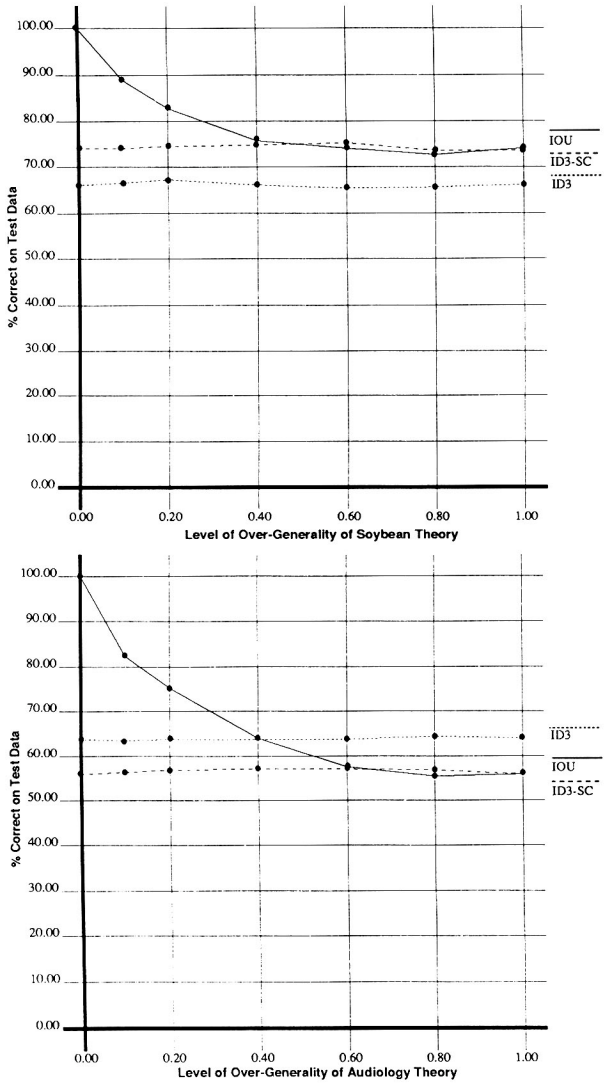

Figure 3. The effect of the level of over-generality.

inductive component. Therefore, when there are enough redundant features, IOU is extremely robust with respect to the disjointness assumption; otherwise its performance degrades gradually as this assumption is violated.

4.2.3. Experiment 3: Using an initial theory independent of the test data

One possible objection to the methodology in the previous experiments is that IOU is indirectly getting information about the test data since it is using a corrupted version of a theory

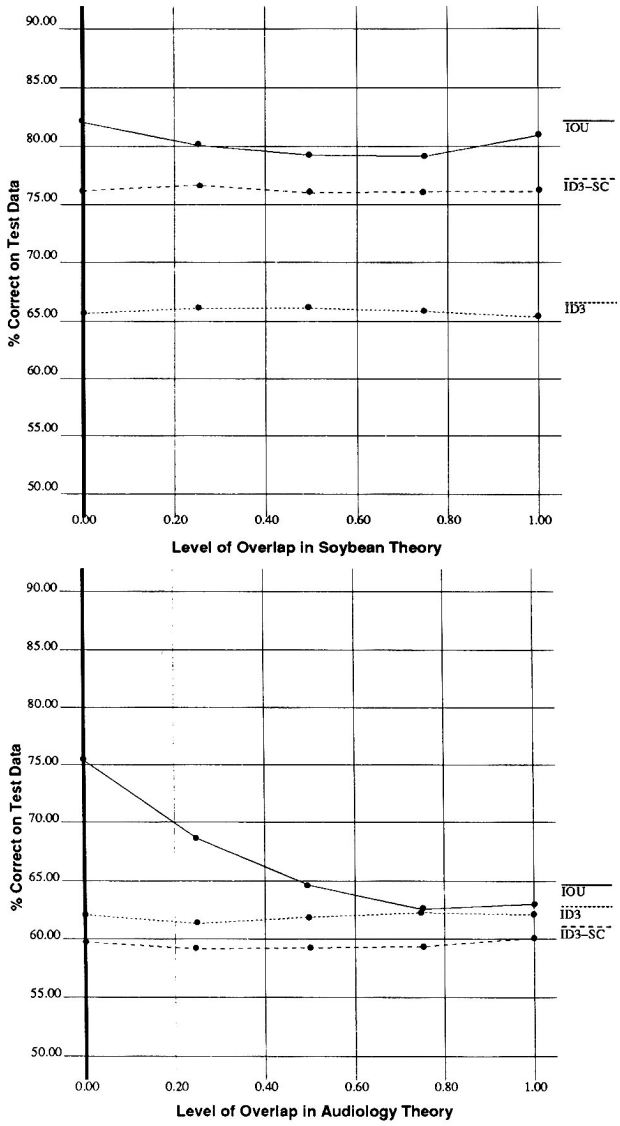

Figure 4. The effect of the level of feature overlap.

generated from the entire data set. Although this is true, it somewhat misses the point. The goal of knowledge-base refinement is to take advantage of information supplied in the form of approximate rules as well as specific data. If IOU can be shown to benefit from overly-general rules, it is demonstrating an ability to successfully perform knowledge-base refinement, regardless of where the rules came from. It is also important to realize that, compared to data, rules are a very compact and efficient way to transmit information from

one intelligent agent to another. The overly-general theory for audiology contains 428 feature-value pairs, less than 3% of the 15,368 feature-value pairs in the audiology data set.

However, in order to gauge the effect of this methodological bias, a final experiment similar to Experiment 1 was run with completely independent test data. The hypothesis is that using a theory independent of the test set will not invalidate the improved learning-rate of IOU observed in Experiment 1. In each trial, 10% of the examples were set aside purely for testing. The remaining 90% of the examples were used to create a theory using ID3. This theory was then randomly generalized as before using a level of over-generality of 0.2. Consequently, as in Experiment 2, a different overly-general theory was used in each of the 30 trials. Unfortunately, the resulting theories are not guaranteed to be overly-general with respect to the test examples, since they are completely independent. On average, the resulting theories were overly-specific on 5.7% of the soybean test examples and on 16.0% of the audiology test examples. Consequently, IOU is at a distinct disadvantage in this experiment, since it only specializes the theory and will never be able to classify some of the test examples correctly.

The results are presented in Figure 5. The results are qualitatively the same as before except that the over-specificity of the initial theory decreases the size of the gap between IOU and ID3/ID3-SC and it eventually disappears after a sufficient number of examples. By 95 examples, IOU is no longer significantly better than ID3-SC on the soybean data, and by 110 examples it is no longer significantly better than ID3 on the audiology data. Overall, the quality of the results is only affected by as much as might be expected by the degree of over-specificity of the initial theory. Consequently, this experiment verifies that the results in Experiment 1 were not simply an artifact of the dependence of the initial theory on the test data.

5. Psychological analysis of IOU

A number of recent psychological studies have focussed on the effect of prior knowledge on concept learning (Murphy & Medin, 1985). In particular, a couple of recent experiments are directly relevant to interpreting IOU as a model of human learning. First is an experiment by Ahn and Brewer (1988) that motivated the development of IOU by demonstrating that subjects learn explanatory and nonexplanatory aspects of a concept separately. Second is an experiment by Wisniewski (1989) demonstrating that subjects learn different concepts from the same examples depending on their background knowledge. The second part of this section shows that IOU can successfully model the results of this experiment.

5.1. Ahn and Brewer's Motivating Experiment

Some recent experiments by Ahn and Brewer (1988) were one of the original motivations behind the development of IOU. These experiments were designed to follow up some previous experiments by Ahn, Mooney, Brewer, and DeJong (1988) that investigated people's ability to use explanation-based learning to learn a plan schema from a single instance. The original experiments revealed that, like an explanation-based system, human subjects

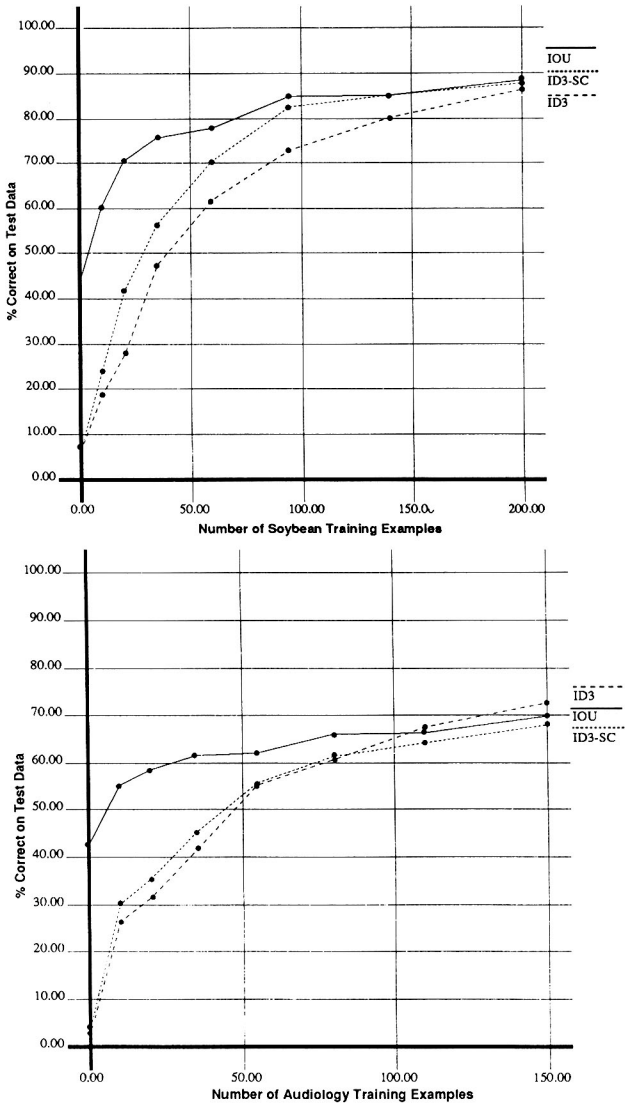

Figure 5. Learning curves with an independent initial theory.

could acquire a general plan schema from a single specific instance. The follow-up experiments explored subjects' ability to learn event schemata that contain both explainable and unexplainable (conventional) components after receiving only a single example, and after receiving multiple examples. For example, one of the schemata used in the experiments is the potlatch ceremony conducted by American Indian tribes of the Northwest. If one has the appropriate knowledge regarding the goals and customs of these Indians, many aspects of the potlatch ceremony can be explained in terms of a plan to increase the social

status of the host. However, there are also a number of ritualistic features of the ceremony that cannot be explained in this manner.

The results of this experiment indicated that the explainable aspects of the potlatch ceremony were acquired after exposure to only a single instance, while the nonexplanatory aspects of the ceremony were only acquired after multiple instances were presented. This supports the view that people use different learning mechanisms to acquire these different aspects of a concept, as in the IOU method. Subjects were also asked to rate their confidence in their assertions that a component is a part of the general ceremony. The subjects confidence ratings for explanatory components were significantly higher than for nonexplanatory ones after both one and two instances. Also, multiple examples increases subjects' confidence and accuracy with respect to nonexplanatory components but not with respect to explanatory ones. This suggests that, like IOU, people maintain separate explanatory and nonexplanatory components in their representation of concepts.

5.2. Modelling Wisniewski's experimental results with IOU

This section demonstrates IOU's ability to model the specific results of some additional psychological experiments exploring the effect of background knowledge on concept learning (Wisniewski, 1989). It is important to note that IOU was not specifically designed to simulate these results, but rather the basic method was already developed when the author learned of the results of this experiment. In Wisniewski's experiment, two groups of subjects performed a standard learning-from-examples task. Both groups received the same examples, but one group, the *function group*, was told the functions of the two categories to be discriminated and the other, the *discrimination group*, was not. For example, the function group was told that one category was used for killing bugs and the contrast category was used for wallpapering. Examples were described by a number of features. A particular feature value could be either *predictive* or *nonpredictive* of a particular category. In the training set containing 15 examples of each category, all examples containing a predictive feature value of a category were members of that category and 80% of the category members had the predictive feature value (the other 20% were missing a value for this feature). Nonpredictive feature values occurred equally often in both categories. A feature value was also *core* or *superficial*. A core feature value was relevant to a category's function, while a superficial one was not. For example, "contains poison" was a core feature value of the category whose function was "for killing bugs," while "manufactured in Florida" was superficial. Each category contained three superficial feature values (two predictive and one nonpredictive) and two core feature values (one predictive the other nonpredictive). The superficial-nonpredictive feature value of a category was the core-nonpredictive feature value of its contrast category. Each training example also had a couple of extra features with random values. Table 5 shows the different types of features for two contrasting categories used in the experiment.

After learning the training data, subjects were given 10 test examples of each category. *Superficial-core** test examples contained the two superficial-predictive feature values of the category and the two core feature values of the contrast category. *Core* examples contained just the core feature values of the category, while *superficial* examples contained

Table 5. Different feature types for experimental categories.

Mornek	Plapel
Function: for killing bugs	Function: for wallpapering
sprayed on plants **C-P**	sprayed on walls **C-P**
contains poison **C-NP**	contains poison **S-NP**
contains a sticky substance **S-NP**	contains a sticky substance **C-NP**
stored in a garage **S-P**	stored in a basement **S-P**
manufactured in Florida **S-P**	manufactured in Ohio **S-P**
C-P: core predictive	**C-NP**: core nonpredictive
S-P: superficial predictive	**S-NP**: superficial nonpredictive

Table 6. Examples of test items for Mornek.

superficial-core*	**core**
stored in a garage **S-P**	contains poison **C-NP**
manufactured in Florida **S-P**	sprayed on plants **C-P**
contains a sticky substance **C-NP***	best if used within 5 years **R**
sprayed on walls **C-P***	came in a 32-ounce container **R**
best if used within 1 year **R**	
superficial	**superficial-core**
stored in a garage **S-P**	stored in a garage **S-P**
manufactured in Florida **S-P**	manufactured in Florida **S-P**
best if used within 1 year **R**	contains a sticky substance **S-NP**
came in a 16-ounce container **R**	contains poison **C-NP**
	sprayed on plants **C-P**
	best if used within 1 year **R**
S-P: superficial predictive	**S-NP**: superficial nonpredictive
C-P: core predictive	**C-NP**: core nonpredictive
C-P*: core predictive (of contrast category)	**C-NP***: core nonpredictive (of contrast category)
R: random	

just the superficial-predictive feature values. *Core-superficial* examples contained all of the core and superficial feature values. Each test example also had a couple of random feature values. The test examples for the Mornek category are shown in Table 6.

Subjects in both groups were asked to rate their confidence in the category of each test example on a scale of 1 to 7, where 1 was most confident for the "wrong" category and 7 most confident for the "right" category. In general, the results demonstrated that subjects in the function group attributed more relevance to the core feature values while the discrimination group relied more heavily on superficial predictive features (see Table 7). However, the function group also made some use of superficial-predictive features values, indicating they were using a combination of empirical and explanation-based techniques.

In the simulation, IOU was used to model the performance of the function group and a standard empirical algorithm was used to model the discrimination group. Simple intuitive domain theories were constructed for connecting core feature values to category membership. For example, IOU's overly-general theory for Mornek is given below:

CONTACT-BUGS ∧ DEADLY → KILLS-BUGS
CONTAINS-POISON → DEADLY
ELECTRIC-SHOCK → DEADLY
SPRAYED-ON = PLANTS → CONTACT-BUGS
EMITS-LIGHT ∧ LOCATION = OUTDOORS → CONTACT-BUGS

Unfortunately, ID3 is not a good algorithm for modelling human empirical learning. For the purpose of this simulation, a particular problem is that it builds a minimum discriminant description rather than a characteristic one. Therefore, a standard *most-specific-conjunctive* (MSC) learning algorithm (Haussler, 1988) was used as the empirical system and as the empirical component of IOU. Early experiments by Bruner, Goodnow, and Austin (1956) indicated that subjects frequently use the MSC strategy when learning concepts from examples. This algorithm simply forms the conjunction of all feature-value pairs present in all of the positive examples. In order to accommodate missing features, only features that appear with different values in different positive examples are actually deleted from the most-specific conjunctive description. This has the same effect as replacing missing features with their most probable value given the class (Quinlan, 1986) before forming the MSC description.

In this simulation, IOU also uses standard explanation-based techniques to operationalize the explanatory component based on the given examples in order to make matching the resulting description easier. For example, KILLS-BUGS is operationalized to SPRAYED-ON = PLANTS ∧ CONTAINS-POISON. Consequently, both IOU and MSC form the most-specific conjunctive description for each category; however, IOU's definition is separated into explanatory and nonexplanatory features. For example, instances of Mornek have each of the features shown in Table 5 or they are missing a value for these features. However, they have differing values for the two other features. Therefore, the MSC description for this category is (explanatory features are in bold):

SPRAYED-ON = PLANTS ∧ CONTAINS-POISON ∧ CONTAINS-STICKY ∧
STORED-IN = GARAGE ∧ MANUFACTURED-IN = FLORIDA

The following equation was used to produce a confidence rating ($1 \leq C \leq 7$) for the test examples:

$$C = 4 + 1.5(M_1 - M_2).$$

M_1 and M_2 are match scores ($-1 \leq M_i \leq 1$) for the two categories computed by examining each feature-value pair in the most-specific-conjunctive description for the category and scoring as follows: +1 if the example had the same value, 0 if the feature was missing, and −1 if it had a conflicting value. The result was scaled by dividing by the maximum possible score. For IOU, explanatory (core) features were weighted more heavily by having them count twice as much (i.e., the match score was incremented or decremented by 2 instead of 1).

This scoring technique is a simple method for obtaining a confidence rating between 0 and 7 based on the degree to which an example matches the MSC description of each

Table 7. Average confidence ratings for test examples.

	Subjects		Simulation	
Item Type	Function	Discrimination	IOU	MSC
superficial-core*	4.00	5.02	3.79	4.60
core	6.16	5.93	5.07	4.60
superficial	6.04	6.36	4.86	5.20
superficial-core	6.43	6.54	5.71	5.80

of the two categories. Several other similar scoring mechanisms were tried without any significant effect on the qualitative results. The important factor is that the score is high when an example matches the description of the first category more than the second and that it is low when an example matches the description of the second category more than the first. The qualitative results are also insensitive to the exact additional weighting assigned to the explanatory features (a weighting factor of 1.5 or 3 works as well as 2).

Table 7 shows both the original experimental results and the results of the simulation. Although the exact confidence values of the simulation do not match the subjects, all of the important differences mentioned by Wisniewski (1989) are present. For the superficial-core* items, the IOU (function) scores are lower than the MSC (discrimination) scores. Although these items have the superficial features of the given category, they have the core features of the other category, causing the function group (and IOU) to rate them lower. IOU (function group) scores the core items higher than the superficial items higher than the core items. Finally, the IOU (function) scores are lower than the MSC (discrimination) scores for the superficial items but higher for the core items.

All of these correctly modeled effects stem from IOU's separation of concepts into explanatory and nonexplanatory components and its scoring procedure that weights the explanatory features more heavily. Since IOU is unique among current integrated learning systems in separating its concepts into explanatory and nonexplanatory components, it seems clear that other existing systems would be unable to model these results. However, the effects are not particularly dependent on the specific details of the IOU algorithm; and therefore other methods that include both explanatory and nonexplanatory features in their concepts and weight the former more heavily may also be able to model Wisniewski's results.

6. Discussion of related research

Due to the recent interest in combining empirical and explanation-based approaches, there is a significant amount of related research. All of these projects address slightly different problems and have different limitations and advantages. In particular, other approaches do not separate learned concepts into explanatory and nonexplanatory components and do not use empirical and explanation-based methods to learn different parts of a concept. Also, no other current system uses empirical learning as an independent sub-system. Therefore, only IOU has the flexibility of employing any empirical learning system and of easily switching to better ones as they are developed.

6.1. Other systems for overly-general theories

Several other existing systems attempt to use overly-general theories to aid concept learning. However, these systems handle different kinds of over-generality and cannot deal with the specific problem addressed by IOU. This section discusses three such systems: Induction Over Explanations (IOE), Incremental Version-Space Merging (IVSM), and Abductive Explanation-Based Learning (A-EBL).

IOE (Flann & Dietterich, 1989) assumes that a correct specialized theory may be obtained by inducing over the explanations of specific examples. The most specific explanation that covers all of the positive examples is constructed by pruning unmatched branches and replacing each occurrence of a constant by the same variable. However, IOE does not consider specializing by adding constraints on unexplained features, since these are assumed to be completely irrelevant. If this method is applied to the examples of drinking vessels given in Table 1, its inductive bias is inappropriate and it cannot produce a consistent description. The explanations for the two positive examples are identical except for how GRASPABLE is proved. Consequently, the resulting concept description is identical to the explanatory component learned by IOU, which is incapable of distinguishing between the positive and negative examples. For the same reason, IOE is incapable of handling overly-general theories like those used in Section 4.2.

IVSM (Hirsh, 1990) is a version-space learning algorithm that accepts abstract descriptions produced by explanation-based generalization as well as specific examples. IVSM can handle overly-general theories that produce multiple explanations of which at least one is correct. The algorithm computes the version-space of concept descriptions consistent with at least one explanation for each example. However, like IOE, it is incapable of handling problems suitable for IOU. If IVSM is applied to the examples in Table 1, it is also unable to produce a consistent description since the positive examples have only one explanation that is itself overly-general. For the same reason, IVSM is also incapable of handling overly-general theories like those used in Section 4.2.

A-EBL (Cohen, 1990) also attempts to deal with the multiple-explanation problem by using a greedy covering algorithm to finding a near-minimal set of explanations that covers all of the positive examples without covering any negative ones. Like IVSM, it is incapable of handling cases in which all explanations for an example are overly-general. Therefore, it is also unsuitable for the type of problems used to test IOU.

IOE, IVSM, and A-EBL all assume that the theory references all relevant features. These systems cannot specialize the theory using unexplained features. However, unlike IOU, they can all handle first-order Horn-clause theories. Finally, it should be noted that the specializations performed by these systems naturally complement the specialization performed by IOU. Any of these systems can be combined with IOU to produce a more robust theory specializer by using one of them to exclude as many negative examples as possible and then using IOU to add constraints on unexplainable features to remove the remaining negative examples. For IOE, this would involve first running IOE on the positive examples. If the most specific explanation covering the positive examples still covers negative examples, then the output theory of IOE and all of the examples can be passed to IOU for additional specialization. Combining IVSM with IOU would require updating the version space with all of the positive examples and then updating with any negative examples that do not cause

the version space to become empty. Next, pick an hypothesis from the resulting version space and give it to IOU as an initial theory together with all of the examples. IOU will attempt to specialize the theory to remove the negative examples unaccounted for by IVSM. One way of combining A-EBL with IOU would involve removing the constraint in A-EBL's greedy covering algorithm that chosen explanations be consistent with the negative examples. It will then generate a near-minimum set of explanations that covers all of the positive examples. Next, IOU can be used to further specialize this theory to avoid covering the negative examples. Better approaches would involve biasing A-EBL's covering algorithm towards explanations that cover few negatives while still covering plenty of positives.

6.2. Systems for overly-specific theories

Just as an overly-general theory is one whose concept extension is too large (a superset of the positive examples are classified as positive), an *overly-specific* theory is one whose concept extension is too small (only a subset of the positive examples are classified as positive). Missing rules or over-constrained rules are the underlying causes of over-specificity. A number of recent systems generalize an overly-specific theory by learning new rules that fill gaps in incomplete explanations. Unlike IOU, these systems cannot deal with overly-general theories; however, they are complementary to IOU and could be combined with it to produce a system that could both specialize and generalize an imperfect domain theory. Learning by Failing to Explain (LFE) (Hall, 1988) is such a method that has been applied to the domain of circuit design. If the system knows the function of a circuit and the function of all but one of its components, it infers what the function of the unknown component must be in order for the overall circuit to work. ODYSSEUS (Wilkins, 1988) is a learning apprentice in the domain of medical diagnosis that conjectures domain rules that will fill gaps in its explanations of a doctor's diagnostic actions. Both LFE and ODDYSSEUS learn new rules from a single example and do not perform induction over multiple examples. GEMINI (Danyluk, 1989), on the other hand, uses incremental conceptual clustering to find rules that complete the explanations of multiple examples. CIGOL (Muggleton & Buntine, 1988) uses a technique called *inverse resolution* to create a new rule that completes the proof of a single example. CIGOL uses stored negative examples to filter out candidate rules, but it does not directly form rules that cover multiple positive examples.

6.3. Systems for general imperfect theories

There are a number of recently developed systems that can, at least to some extent, deal with both overly-general and overly-specific domain theories. In this sense, they are more general than IOU. In principle, many of these systems could be run on problems suitable for IOU; however, I believe the differences mentioned below would allow IOU to perform better on such problems. Of course, detailed empirical studies are needed to verify this claim.

OCCAM (Pazzani, 1990) is a conceptual clustering system that combines explanation-based and empirical techniques. If an example can be explained, its generalized explanation

forms the basis of a new concept. If an example cannot be explained but is similar to a previous example, the common features of the two examples are used to form a new concept. In addition, empirically learned concepts can be used to help explain subsequent examples and specializations of explanation-based concepts can be empirically learned by noticing additional similarities among their examples. This last aspect is very similar to IOU; however, there are several differences. For one, OCCAM clusters unclassified examples rather than learning specific concepts from classified examples. Also, it cannot learn disjunctive concepts, does not reduce example descriptions before they are processed by empirical learning, and does not employ an independent empirical learner.

The ML-SMART system (Bergadano & Girodana, 1988) uses a possibly incomplete/incorrect first-order domain theory to guide the search for an inductive hypothesis that covers a set of positive examples but does not cover a set of negative examples. The system performs a general to specific search through the space of hypotheses trying to find a consistent operational concept description. The system prefers to specialize (operationalize) the current hypothesis using rules from the domain theory; but when this fails, it resorts to purely syntactic specialization rules. Unlike IOU, it does not immediately assume the relevance of explained features. IOU takes advantage of the stronger assumption that the domain theory is strictly overly-general to further focus the learning process. As illustrated by the cup example used by Bergadano and Giordana (1988), ML-SMART does not even guarantee that the learned concept will contain as many of the explainable aspects as consistently possible. That is, the learned definition does not even require a cup to have an upward-pointing concavity even though the theory can explain why having one is important and all of the examples are consistent with this requirement.

RTLS (Reduced Theory Learning System) (Ginsberg, 1988; Ginsberg, 1990) is capable of refining arbitrary propositional Horn-clause theories. RTLS first fully expands a theory into a completely operational disjunctive-normal-form (DNF) expression. Next, this expression is modified by a complicated procedure to make it consistent with a set of training examples. Finally, the resulting DNF formula is retranslated into a multi-level theory. The real disadvantage of RTLS is the inherent exponential complexity of reducing a domain theory and revising the resulting DNF formula.

KBANN (Knowledge-Based Artificial Neural Networks) (Towell et al., 1990) is also potentially capable of revising arbitrarily incorrect theories. KBANN translates a domain theory into an equivalent neural-network and then uses back-propagation (Rumelhart, Hinton, & Williams, 1986) to modify the weights in this network to make it consistent with a set of examples. Unlike a symbolic system, KBANN does not produce an easily comprehensible rule-based result. It also suffers from all of the computational and parameter-adjusting problems of back-propagation (Blum & Rivest, 1988; Shavlik et al., 1991).

6.4. Systems that use empirical learning to focus explanation

One of the first systems to integrate empirical and explanation-based learning was a version of UNIMEM (Lebowitz, 1986). Unlike other systems, UNIMEM does not attempt to deal with imperfect domain theories but rather uses empirical techniques to focus the explanation process. For example, when applied to Congressional voting records, UNIMEM's

clustering method detects that the members of Congress who voted for (or against) a particular bill are all similar in some respect (e.g., they are all Democrats from Southern states). The system then attempts to construct an explanation for how these similar features could account for their common position on the bill. The explanation is used to separate the causally relevant features shared by the examples from those shared features that are purely coincidental. Lebowitz claims that this approach focuses the explanation process and makes it more efficient while preventing the system from discovering spurious correlations to which a purely empirical system is susceptible.

7. Future research issues

The work reported in this paper has a number of shortcomings that need to be addressed by future research. First, IOU should be tested on an actual domain theory created by an expert for a realistic problem. Second, the incremental version of IOU discussed in Section 2.2 needs to be implemented and tested. Third, improved methods for handling noisy data and missing feature values need to be incorporated into the system. Finally, some of the more fundamental limitations of IOU need to be addressed. Several of these are listed below.

The current system uses a feature-based description language and needs to be extended to handle first-order predicate calculus. Many problems require relational descriptions and/or quantification, which the current system cannot handle. The recent development of effective inductive systems for first-order logic, such as FOIL (Quinlan, 1990), should be useful in extending the system in this manner.

IOU is also restricted by its assumption that the features referenced in the explanatory and nonexplanatory parts of the concept definition are disjoint. This prevents it from properly specializing certain types of overly-general theories. As discussed in Section 6.1, IOU could potentially be combined with other theory specialization techniques like IOE, IVSM, and A-EBL in order to handle a larger class of theory specialization problems. Another approach is simply not to remove the explained features of the examples before passing them along to the inductive learner (i.e., skip step 3 in the algorithm in Table 3). In this case, the empirical system can arbitrarily specialize the theory using all of the available features and IOU would be able to handle any overly-general theory. However, failing to focus the empirical system on unexplained features would undoubtedly reduce its effectiveness on problems in which explanatory and nonexplanatory features are disjoint. Also, the theoretical results in Section 3 rely on the disjointness assumption.

IOU's inability to handle overly-specific aspects of an imperfect domain theory is another obvious shortcoming. As mentioned in Section 6.2, this problem could potentially be addressed by combining IOU with a system that learns rules that complete partial explanations. In fact, we have recently developed a system called EITHER (Ourston & Mooney, 1990) that refines arbitrarily imperfect propositional Horn-clause theories by combining techniques for both learning and deleting rules and rule antecedents. The process EITHER uses to specialize existing rules by adding antecedents is derived from IOU.

8. Conclusion

This paper has presented and evaluated a learning method called Induction Over the Unexplained, which uses a combination of explanation-based and empirical methods to learn concepts with both explanatory and nonexplanatory aspects. IOU uses an overly-general domain theory to learn the explanatory part of a concept, and a generic inductive learning system to acquire the nonexplanatory part. Theoretical analysis was used to show that IOU can be expected to run in linear time and that the lower-bound on the number of examples required to learn a PAC concept definition is better for IOU than for a purely empirical approach. Empirical analysis was used to confirm IOU's faster learning rate by testing it on both artificial and natural data. Empirical results also demonstrate that IOU's performance degrades gracefully as the over-generality of its initial theory increases and its assumption of disjointness of explanatory and nonexplanatory features is violated. Simulation of the results of a recent psychology experiment on the effect of background knowledge on concept learning was used to demonstrate the ability of IOU to model aspects of human learning. IOU was also shown to complement rather than directly compete with other recently developed methods such as IOE, IVSM, A-EBL, GEMINI, ODDYSEUS, and Learning by Failing to Explain.

This paper has also demonstrated an eclectic approach to evaluation that combined theoretical, empirical, and psychological evidence to judge the effectiveness of IOU. The evidence provided by any one of these approaches is necessarily incomplete. By drawing upon a combination of evaluation methods, it is possible to form a more complete picture of the strengths and weaknesses of a particular learning method.

Acknowledgments

I would like to thank the editor, Tom Dietterich, and two anonymous reviewers who provided many insightful and detailed comments that greatly improved the final version of this paper. I would also like to thank Dirk Ourston for implementing an initial version of IOU, creating the cup data, and for numerous discussions about revising imperfect domain theories. Finally, thanks to Bob Stepp and Bob Reinke for providing the soybean data and Ray Bareiss, Bruce Porter, and Craig Wier for providing the audiology data, which was collected with the help of Professor James Jerger of the Baylor College of Medicine. This work was supported by the NASA Ames Research Center through grant number NCC 2-629.

Notes

1. An hypothesis space is trivial if it contains only one hypothesis or two disjoint hypotheses whose union covers the entire domain.

References

Ahn, W., & Brewer, W.F. (1988). Similarity-based and explanation-based learning of explanatory and nonexplanatory information. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 524-530). Hillsdale, NJ: Erlbaum.

Ahn, W., Mooney, R.J., Brewer, W.F., & DeJong, G.F. (1987). Schema acquisition from one example: Psychological evidence for explanation-based learning. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (pp. 50–57). Hillsdale, NJ: Erlbaum.

Bergadano, F., & Giordana, A. (1988). A knowledge intensive approach to concept induction. *Proceedings of the Fifth International Conference on Machine Learning* (pp. 305–317). San Mateo, CA: Morgan Kaufman.

Blum, A., & Rivest, R.L. (1988). Training a 3-node neural net is NP-complete. *Proceedings of the 1988 Workshop on Computational Learning Theory* (pp. 9–18). San Mateo, CA: Morgan Kaufman.

Bruner, J.S., Goodnow, J., & Austin, G.A. (1956). *A study in thinking*. New York: Wiley.

Cohen, W.W. (1990). Learning from textbook knowledge: A case study. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 743–748). Cambridge, MA. MIT Press.

Danyluk, A. (1989). Finding new rules for incomplete theories: Explicit biases for induction with contextual information. *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 34–36). San Mateo, CA: Morgan Kaufman.

DeJong, G.F., & Mooney, R.J. (1986). Explanation-based learning: An alternative view. *Machine Learning, 1*, 145–176.

Dietterich, T.G., London, B., Clarkson, K., & Dromey, G. (1982). Learning and inductive inference. In R. Cohen & E.A. Feigenbaum (Eds.), *Handbook of Artificial Intelligence: Volume III*. San Mateo, CA: Morgan Kaufman.

Dowling, W.F., & Gallier, J.H. (1984). Linear-time algorithms for testing the satisfiability of propositional Horn formulae. *Journal of Logic Programming, 3*, 267–284.

Ehrenfeucht, A., Haussler, D., Kearns, M., & Valiant, L. (1989). A general lower bound on the number of examples needed for learning. *Information and Computation, 82*, 267–284.

Fisher, D.H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning, 2*, 139–172.

Flann, N.S., & Dietterich, T.G. (1989). A study of explanation-based methods for inductive learning. *Machine Learning, 4*, 187–226.

Ginsberg, A. (1988). Theory revision via prior operationalization. *Proceedings of the Seventh National Conference on Artificial Intelligence* (pp. 743–748). San Mateo, CA: Morgan Kaufman.

Ginsberg, A. (1990). Theory reduction, theory revision, and retranslation. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 743–748). Cambridge, MA: MIT Press.

Gold, E.M. (1967). Language identification in the limit. *Information and Control, 10*, 447–474.

Hall, R. (1988). Learning by failing to explain: Using partial explanations to learn in incomplete or intractable domains. *Machine Learning, 3*, 45–78.

Haussler, D. (1988). Quantifying inductive bias: Artificial intelligence algorithms and Valiant's learning framework. *Artificial Intelligence, 36*, 177–221.

Hirsh, H. (1990). *Incremental version space merging: A general framework for concept learning*. Hingham, MA: Kluwer.

Kearns, M., Li, M., Pitt, L., & Valiant, L.G. (1987). Recent results on Boolean concept learning. *Proceedings of the Fourth International Machine Learning Workshop* (pp. 337–352). San Mateo, CA: Morgan Kaufman.

Kibler, D., & Langley, P. (1988). Machine learning as an experimental science. *Proceedings of the Third European Working Session on Learning* (pp. 81–92). London: Pitman.

Langley, P. (1989). Editorial: Toward a unified science of machine learning. *Machine Learning, 3*, 253–259.

Lebowitz, M. (1986). Integrated learning: Controlling explanation. *Cognitive Science, 10*, 219–240.

Medin, D.L., Wattenmaker, W.D., & Michalski, R.S. (1987). Constraints and preferences in inductive learning: An experimental study of human and machine performance. *Cognitive Science, 11*, 299–239.

Michalski, R.S. (1983). A theory and methodology of inductive learning. In R.S. Michalski, J.G. Carbonell, & T.M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*. Palo Alto, CA: Tioga.

Michalski, R.S., & Chilausky, R.L. (1980). Learning by being told and learning from examples: An experimental comparison 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*, 125–160.

Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine Learning, 4*, 227–243.

Minton, S. (1988). Quantitative results concerning the utility of explanation-based learning. *Proceedings of the Seventh National Conference on Artificial Intelligence* (pp. 564–569). San Mateo, CA: Morgan Kaufman.

Mitchell, T.M., Keller, R., & Kedar-Cabelli, S. (1986). Explanation-based generalization: A unifying view. *Machine Learning, 1*, 47–80.

Mooney, R., Ourston, D., & Wu, S.Y. (1989). Induction over the unexplained: A new approach to combining empirical and explanation-based learning (Technical Report AI89-110). Austin, TX: University of Texas, Artificial Intelligence Laboratory.

Muggleton, S., & Buntine, W. (1988). Machine invention of first-order predicates by inverting resolution. *Proceedings of the Fifth International Machine Learning Conference* (pp. 339-352). San Mateo, CA: Morgan Kaufman.

Murphy, G.L., & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316.

Ourston, D., & Mooney, R. (1990). Changing the rules: A comprehensive approach to theory refinement. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 815-820). Cambridge, MA: MIT Press.

Pazzani, M.J. (1990). *Creating a memory of causal relationships: An integration of empirical and explanation-based learning methods*. Hillsdale, NJ: Earlbaum.

Pazzani, M.J., & Sarrett, W. (1990). Average case analysis of conjunctive learning algorithms. *Proceedings of the Seventh International Machine Learning Conference* (pp. 339-347). San Mateo, CA: Morgan Kaufman.

Pazzani, M.J., & Silverstein, G. (1990). Feature selection and hypothesis selection models of induction. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 221-228). Hillsdale, NJ: Earlbaum.

Porter, B., Bareiss, R., & Holte, R.C. (1990). Concept learning and heuristic classification in weak-theory domains. *Artificial Intelligence, 45*, 229-263.

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning, 1*, 81-106.

Quinlan, J.R. (1990). Learning logical definitions from relations. *Machine Learning, 5*, 239-266.

Reinke, R. (1984). Knowledge acquisition and refinement tools for the ADVISE meta-expert system. Master's thesis, Department of Computer Science, University of Illinois, Urbana, IL.

Rosenbloom, P., & Newell, A. (1987). Learning by chunking: A production system model of practice. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production system models of learning and development*. Cambridge, MA: MIT Press.

Rumelhart, D.E., Hinton, G.E., & Williams, J.R. (1986). Learning internal representations by error propagation. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel Distributed Processing* (Vol. 1). Cambridge, MA: MIT Press.

Segre, A.M. (Ed.) (1989). Combining empirical and explanation-based learning. *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 2-92). San Mateo, CA: Morgan Kaufman.

Shavlik, J.W., Mooney, R.J., & Towell, G.G. (1991). Symbolic and neural learning algorithms: An experimental comparison. *Machine Learning, 6*, 11-143.

Skorstad, J., Falkenhainer, B., & Gentner, D. (1987). Analogical processing: A simulation and empirical corroboration. *Proceedings of the Sixth National Conference on Artificial Intelligence* (pp. 322-326). San Mateo: Morgan Kaufman.

Stepp, R.E. (1984). Conjunctive conceptual clustering: A methodology and experimentation. Doctoral dissertation, Department of Computer Science, University of Illinois, Urbana, IL.

Towell, G.G., Shavlik, J.W., & Noordewier, M.O. (1990). Refinement of approximate domain theories by knowledge-based neural networks. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 861-866). Cambridge, MA: MIT Press.

Utgoff, P.E. (1989). Incremental induction of decision trees. *Machine Learning, 4*, 161-186.

Valiant, L.G. (1984). A theory of the learnable. *Communications of the Association for Computing Machinery, 27*, 1134-1142.

Wilkins, D.C. (1988). Knowledge base refinement using apprenticeship learning techniques. *Proceedings of the Seventh National Conference on Artificial Intelligence* (pp. 646-651). San Mateo, CA: Morgan Kaufman.

Winston, P.H., Binford, T.O., Katz, B., & Lowry, M. (1983). Learning physical descriptions from functional definitions, examples, and precedents. *Proceedings of the Third National Conference on Artificial Intelligence* (pp. 433-439). San Mateo, CA: Morgan Kaufman.

Wisniewski, E.J. (1989). Learning from examples: The effect of different conceptual roles. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 980-986), Hillsdale, NJ: Earlbaum.

Received October 12, 1990
Accepted January 24, 1991
Final Manuscript January 27, 1992