Learning from an Approximate Theory and Noisy Examples

Somkiat Tangkitvanich and Masamichi Shimura
Department of Computer Science, Tokyo Institute of Technology
2-12-1 Oh-Okayama, Meguro, Tokyo 152, JAPAN
sia@cs.titech.ac.jp

Abstract

This paper presents an approach to a new learning problem, the problem of learning from an approximate theory and a set of noisy examples. This problem requires a new learning approach since it cannot be satisfactorily solved by either inductive, or analytic learning algorithms or their existing combinations. Our approach can be viewed as an extension of the minimum description length (MDL) principle, and is unique in that it is based on the encoding of the refinement required to transform the given theory into a better theory rather than on the encoding of the resultant theory as in traditional MDL. Experimental results show that, based on our approach, the theory learned from an approximate theory and a set of noisy examples is more accurate than either the approximate theory itself or a theory learned from the examples alone. This suggests that our approach can combine useful information from both the theory and the training set even though both of them are only partially correct.

Introduction

Previous machine learning approaches learn either empirically from noise-free [Mitchell, 1978] or noisy examples [Quinlan, 1983], or analytically from a correct theory and noise-free examples [Mitchell *et al.*, 1986; DeJong and Mooney, 1986], or empirically and analytically from an approximate theory and noise-free examples [Richards, 1992; Pazzani *et al.*, 1991; Cohen, 1990; Wogulis, 1991; Ginsberg, 1990; Cain, 1991; Bergadano and Giordana, 1988].

This paper discusses the problem of learning from an approximate theory and a set of noisy examples, a new learning problem which cannot be satisfactorily solved by the previous approaches. In this problem, it is harmful to place full confidence in either the given theory or the training set. Thus, an analytic approach will not learn successfully since it places full confidence in the theory which is only approximately correct. An inductive approach does not satisfactorily solve the problem since it cannot take advantage of the theory for biasing the learning. The approaches that combine analytic with inductive techniques, and modify the given theory to fit the examples, will not learn correctly since they place full confidence in the noisy training set. Consequently, we require a new learning approach that can balance the confidence in the theory against that in the examples.

Our approach, presented in this paper, can take advantage of the given theory for biasing the learning and can tolerate noise in the training examples. In our approach, the given theory is refined as little as necessary while its ability to explain the examples is increased as much as possible. The amount of the refinement and the ability of the theory to explain the examples are judged by the encoding lengths required to describe the refinement and the examples with the help of the theory, respectively. By keeping a balance between the amount of refinement and the ability to explain the examples, our approach places full confidence in neither the given theory nor the examples.

Although our approach can be applied to the learning of a theory represented in any language, we demonstrate its application to the learning of a relational theory. The prototype system, which we call LATEX (Learning from an Approximate Theory and noisy EXample), has been tested in the chess endgame domain in both knowledge-free and knowledge-intensive environments. The results show that a theory learned by LATEX from an approximate theory and a set of noisy examples is remarkably more accurate than the approximate theory itself, a theory learned by LATEX from the examples alone, or a theory learned by the FOIL learning system [Quinlan, 1990]. This suggests that our system can combine useful information from the theory and the training set even though both of them are only partially correct.

Description of the Problem

Our learning problem can be defined as follows:

- Given:
  - a prior knowledge in form of an approximate theory,

- a set of noisy positive and negative examples.

- Find:
  - a theory which is similar to the given theory but is expected to be more accurate in classifying the unseen examples.

The fundamental assumption of learning from an initial theory is that although the theory is flawed, it is still relatively "close" or "approximates" to the target theory [Mooney, To appear]. Intuitively, an approximate theory is supposed to facilitate the learning rather than hinder it. Such a theory can be obtained from a human expert, a prior learning [Muggleton *et al.*, 1992], a textbook [Cohen, 1990] or other sources. It has been pointed out that the accuracy of a theory is not a good criterion to quantify its approximateness [Mooney, To appear; Pazzani and Kibler, 1992]. Unfortunately, there has been no satisfactory criterion yet. In the next section, we show that the notion of an approximate theory can be precisely defined based on the encoding length of the refinement.

Our Approach

We present a new learning approach based on encoding a refinement and examples. The approach can be viewed as an extension of the minimum description length (MDL) principle [Wallace and Boulton, 1968; Rissanen, 1978], a general principle applicable to any inductive learning task involving sufficient examples. According to the principle, the simplest theory that can explain the examples well is the best one. Simplicity of a theory is judged by its length under an encoding scheme chosen subjectively by the experimenter. The ability of a theory to explain the given examples is judged by the length of the examples encoded with the help of the theory, with shorter length indicating more ability to explain. From the MDL perspective, learning occurs only when the sum of these encoding lengths is less than the explicit encoding length of the examples, that is when there is a compression.

General as it is, the MDL principle in its original form cannot take advantage of a prior knowledge in the form of an initial theory. This is a weakness since the information in the theory may be essential for an accurate learning when sufficient examples are not available. Our approach extends the MDL principle so that it can take advantage of an initial theory.

Extending MDL to Learn from an Approximate Theory

In our approach, the theory that is most similar to the given theory and can explain the examples well is the best theory. Similarity between a theory T' and the given theory T is judged by the length under some encoding scheme to describe the *refinement* required to transform T into T'. The ability of a theory to explain the examples is judged by the length of the examples encoded with the help of the theory, with shorter length indicating more ability to explain. Qualitatively, the best theory is the theory with the minimum description length calculated from the sum of

1) the description length of the refinement, and

2) the description length of the examples encoded with the help of the refined theory.

The refined theory in 2) can be obtained from the initial theory and the description of the refinement. When the bias for a similar theory is appropriate, as in the case of learning from an approximate theory, a learning algorithm based on our approach will achieve a higher accuracy than an algorithm that cannot learn from a theory. It should be noted that when there is no initial theory, the bias reduces to one that prefers the simplest theory that can explain the examples, and our approach degenerates to the MDL approach.

The emphasis on encoding refinement is a unique feature of our approach. It has the following advantages for our learning problem.

1. It can balance the confidence in the theory against that in the training set. Consequently, our approach can take advantage of the information in both the theory and the training set while being able to avoid the pitfalls of placing full confidence in either of them. From a Bayesian perspective, our approach can be interpreted as assigning a prior probability to each theory in the hypothesis space, favoring the ones which are similar to the given theory, updating the probability by using the training examples, and selecting the theory that has a maximum posterior probability. However, in comparison with a Bayesian approach, an approach based on the MDL principle provides the user with the conceptually simpler problem of computing code lengths, rather than estimating probabilities [Quinlan and Rivest, 1989].
2. It provides a precise way to define an approximate theory. Intuitively, an approximate theory is a theory that facilitates the learning rather than hinder it. Since learning is related to producing a compression in the encoding length, an approximate theory can be judged based on the help it provides in shortening the description length of the target theory. Given an approximate theory, the encoding of the target theory as a sequence of refinements of that theory should be shorter than a direct encoding of the target theory. This leads to the following definition.

Definition 1 Approximate Theory
A theory T_0 is an approximate theory of T_* under an encoding scheme E iff $l_E(T_0, T_*) < l_E(\phi, T_*)$, where $l_E(T_i, T_j)$ is the length required to encode the transformation from T_i into T_j, and ϕ is an empty theory.

How can such a theory facilitate learning? Within the PAC learnability framework [Valiant, 1984], the following theorem shows that it reduces the sample

complexity of any algorithm that accepts an initial theory T_0 and i) never examines a hypothesis h_1 before another hypothesis h_2 if $l_E(T_0, h_2) < l_E(T_0, h_1)$, and ii) outputs a hypothesis consistent with the training examples. Such an algorithm reflects important elements of a number of existing algorithms (e.g., [Ginsberg, 1990; Cain, 1991]) that modify the initial theory to fit the examples.

Theorem 1 *Let L be any algorithm that satisfies i) and ii). For a finite hypothesis space, L with an approximate theory T_0 has less sample complexity than L with an empty theory ϕ.*

The above theorem is applicable when the examples are noise-free. The proof of the theorem and an analogous theorem for learning from noisy examples is given in [Tangkitvanich and Shimura, 1993].

Learning a Relational Theory

In learning a relational theory, the examples and the theory are represented in form of tuples and a set of function-free first-order Horn clauses, respectively.

Encoding Training Examples

From the MDL perspective, learning occurs when there is a compression in the encoding length of the examples. In learning a relational theory, although both the positive and negative examples are given, it is a common practice to learn a theory that characterizes only the positive examples [Quinlan, 1990; Muggleton and Feng, 1990]. In encoding terms, this suggests that it is appropriate to compress only the encoding length of the positive examples.[1] One way to produce a compression is to encode the examples with the help of an approximate theory.

From Shannon's information theory, the optimal code length for an object e that has a probability p_e is $-\log_2 p_e$ bits. Without the help of an initial theory, the optimal code length required to indicate that an example is positive is thus $-\log_2 p_0$ bits, where p_0 is the probability that an example left in the training set is a positive example. In contrast, with the help of Cl, a clause in an approximate theory that covers the example, the optimal code length becomes $-\log_2 p_{cl}$ bits, where p_{cl} is the probability that an example covered by Cl is a positive example. Consequently, there is a compression produced by Cl in encoding a positive example if $p_{cl} > p_0$, that is when the positive examples are more concentrated in the clause than in the training set. The total compression obtained in encoding n positive examples covered by Cl is

$$Compress(Cl) = n \times (\log_2 p_0 - \log_2 p_{cl}). \quad (1)$$

[1] On the contrary, if a theory that characterizes both types of examples (e.g., a relational decision tree [Watanabe and Rendell, 1991]) is to be learned, it would be appropriate to compress the encoding length of both types.

The compression produced by a theory is the sum of the compressions produced by all the clauses in the theory. By using a more accurate theory to encode the examples, we can obtain further compressions. Such a theory can be obtained by means of refinement. However, the compression is not obtained without any cost since we have to encode the refinement as well.

Encoding Refinements

We assume that the head of each clause in the theory is identical. This limits the ability to learn an intermediate concept but simplifies the encoding of a refinement. With this assumption, it is possible to encode any refinement by using only two basic transformations: a literal addition and a literal deletion. Other forms of transformation can be encoded by using these. For example, literal replacement can be encoded by using a combination of a literal addition and a literal deletion. Clause addition can be encoded by using a combination of literal additions. The head of a new clause need not be encoded since it is identical to that of an existing clause. Clause deletion can be encoded in a similar way by using a combination of literal deletions. Thus, the overall refinement can be encoded by the following self-delimiting code:

$$\begin{array}{c} |n_1|\, refine_{1,1}|\, refine_{1,2}|...|refine_{1,n_1}| \\ \hline |n_2|\, refine_{2,1}|\, refine_{2,2}|...|refine_{2,n_2}| \\ \hline \vdots \quad \vdots \quad \vdots \\ \hline |n_k|\, refine_{k,1}|\, refine_{k,2}|...|refine_{k,n_k}| \end{array}$$

In the above encoding scheme, n_i is the encoding of an integer indicating the number of refinements applied to clause i (with $n_i = 0$ indicating no refinement), $refine_{i,j}$ is the encoding of the j-th refinement to clause i. $refine_{i,j}$ is composed of a one-bit flag indicating whether the refinement is a literal addition or a literal deletion, and the encoding of the content of the refinement. For a literal addition, the encoding of the content contains the information required to indicate whether the literal to be added is negated or not, and from which relation and variables the literal is constructed. For a literal deletion, the encoding of the content contains the information required to indicate which literal in the clause is to be deleted. Note that the encoding scheme is natural for our learning problem in that it requires a shorter encoding length for a refinement that has a littler effect on the theory. For example, adding a literal requires a shorter length than adding a clause.

We now quantify the relationship between the refinement and its effect on the compression in the encoding length. Let Cl_i be a clause in the initial theory, $Length(refine_{i,j})$ be the length required for $refine_{i,j}$, and Cl'_i be the refined clause obtained from Cl_i and $refine_{i,j}$. The compression produced by the refinement can be estimated by

$$Compression(refine_{i,j}) = Compress(Cl'_i) - Compress(Cl_i) - Length(refine_{i,j}). \quad (2)$$

The Learning Algorithm

The algorithm of LATEX is very simple. In each iteration, the theory is refined by using all the possible applications of the following refinement operators: a clause-addition operator, a clause-deletion operator, a literal-addition operator and a literal-deletion operator. The literal-addition operator adds to an existing clause a literal in the background knowledge. The literal must contain at least one existing variable and must satisfy constraints to avoid problematic recursion. The clause-addition operator adds to the theory a new clause containing one literal. Among all the possible refinements, LATEX selects one that produces maximal positive compression. The system terminates when no refinement can produce a positive compressions. The refined theory is then passed to a simple post-processing routine which removes clauses that cover more negative than positive examples.

Admittedly, the current greedy search strategy and the simple refinement operators prevent LATEX from learning some theories, e.g., those contain literals that do not discriminate positive from negative examples. We are now incorporating a more complex search strategy and better operators to overcome this limitation.

Experimental Evaluations

LATEX has been experimented on the king-rook-king board classification domain described in [Quinlan, 1990; Richards, 1992]. In this domain, there are two types of variable, representing row and column. For each type, there are three relations given as the background knowledge: *eq(X,Y)* indicating that X is the same as Y, *adj(X,Y)* indicating that X is adjacent to Y and *lessthan(X,Y)* indicating that X is less than Y.

In our experiments, the training sets are randomly generated and noise are randomly introduced into them. To introduce noise, we adopt the classification noise model used in [Angluin and Laird, 1988]. In this model, a noise rate of η implies that the class of each example is replaced by the opposite class with probability η. The test set is composed of 10,000 examples selected independently to the training set.

The experiments are made on 5 initial theories which are the operationalized version of those used by FORTE [Richards, 1992]. The theories are generated by corrupting the correct theory with six operators: clause-addition, clause-deletion, literal-addition, literal-deletion, literal-replacement and variable-replacement operators. Each corrupted theory is an approximate theory according to our definition and is averagely 45.79% correct. The average number of clauses in an operationalized theory is 14.2, and the average number of literals in a clause is 2.8.

Figure 1 compares the learning curves of LATEX with an initial theory (LATEX-Th), LATEX without an initial theory (LATEX-NoTh), and FOIL [Quinlan, 1990], for $\eta = 10\%$. The curves demonstrate many interesting points. First, throughout the learning session, the theory learned by LATEX from an initial theory and the examples is significantly more accurate than that learned by LATEX from the examples alone. Further, although the training examples are considerably noisy, the theory learned from the initial theory and the examples is much more accurate than the initial theory itself. This means that LATEX can extract useful information from the noisy examples to improve the accuracy of the theory.

In other words, the experiments show that by combining the information in the theory and the examples, LATEX achieves a higher accuracy than it could with either one alone. Both are beneficial to the system. This suggests a dual view of our approach: as a means of refining an initial theory using examples, or as a means of improving the learning from examples using an initial theory.

It is also interesting to compare the learning curve of LATEX with that of FOIL. Without an initial theory, LATEX degenerates to an ordinary inductive learning system based on the MDL principle. Throughout the training sessions, the theory learned by LATEX is significantly more accurate than that learned by FOIL. However, another experiment which is not reported here shows that there are no significant differences in the accuracies achieved by the two systems when there is no noise in the training set. Hence, the differences can be attributed to the differences in the noise-handling mechanisms of the two systems. Investigation reveals that, when the examples are noisy, the theories learned by FOIL contain more literals and require longer encoding lengths than those learned by LATEX. In other words, the theories learned by FOIL are much more complex.

Related Work

In this section, we discuss three related approaches for learning a relational theory from noisy examples. Other approaches (e.g., [Towell *et al.*, 1990; Drastal *et al.*, 1989; Ginsberg, 1990]) will be discussed in the full paper.

- FOIL

  Unlike LATEX, FOIL [Quinlan, 1990] is a relational learning system that cannot take advantage of an initial theory. However, it is informative to compare the two systems from an inductive learning perspective. FOIL uses an information-based heuristic called Gain to select a literal and uses another information-based heuristic as its stopping criterion to handle noise. In contrast, LATEX uses a single compression-based criterion for both tasks. When used to select a literal, our criterion and Gain are

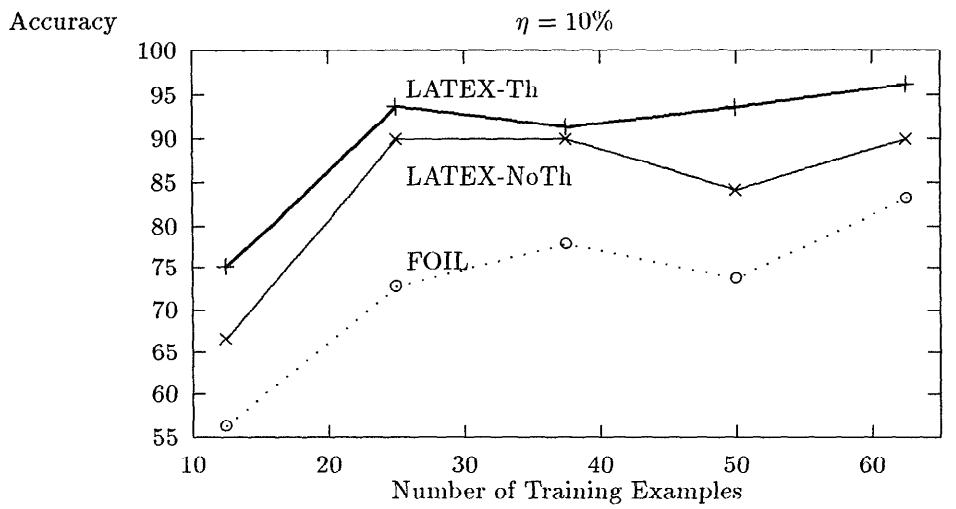

Figure 1: Learning curves of LATEX and FOIL

similar in that they suggest selecting a literal that discriminates positive from negative examples. However, when used to handle noise, our criterion and FOIL's stopping criterion have different effects. The experiments reveal that a theory learned by FOIL is much more complex than that learned by LATEX. This is because FOIL's stopping criterion allows the building of long clauses to cover a small number of examples. Former study [Dzeroski and Lavrac, 1991] also arrived at the same conclusion.

- FOCL

  FOCL extends FOIL to learn from an initial theory in an interesting way. However, the original algorithm of FOCL [Pazzani and Kibler, 1992] is unsuitable for learning from noisy examples since it refines the given theory to be consistent with them. Two extensions of FOCL are designed to deal with noise: one with FOIL's stopping criterion, the other with a pruning technique [Brunk and Pazzani, 1991].

  Currently, we are not aware of any experimental results of testing any FOCL algorithms in learning from an initial theory and noisy examples. However, it should be noted that while FOCL with the pruning technique requires another training set for pruning, and both extensions of FOCL use separate mechanisms for selecting literals and handling noise, LATEX requires a single training set and uses a single mechanism for both tasks.

- Muggleton et. al.'s Learning System

  Recently, Muggleton et. al. [Muggleton *et al.*, 1992] proposed an approach to learn a theory represented as a logic program from noisy examples. The system based on their approach receives as an input an overly general theory learned by the GOLEM system [Muggleton and Feng, 1990]. It then specializes the theory by using a technique called closed-world specialization [Bain and Muggleton, 1990]. If there are several possible specializations, the one that yields maximal compression is selected. Since the system attempts to minimize the encoding length, it can be viewed as incorporating the MDL principle.

  From the point of view of learning from an initial theory, there is an important difference between the theory acceptable by their system and that acceptable by LATEX. While their system assumes an overly general theory produced by a prior learning of GOLEM, LATEX requires no such assumptions. LATEX can accept a theory that is overly general, overly specific, or both. The theory can be obtained from an expert, a textbook, a prior learning and other sources.

Conclusion

We presented an approach for learning from an approximate theory and noisy examples. The approach is based on minimizing the encoding lengths of the refinement and the examples, and can be viewed as an extension of the MDL principle that can take advantage of an initial theory. We also demonstrated the applicability of our approach in learning a relational theory, and showed that the system based on our approach can tolerate noise in both the knowledge-free and knowledge-intensive environment. In the knowledge-free environment, our system compares favorably with FOIL. In the knowledge-intensive environment, it combines useful information from both the theory and the training

set and achieves a higher accuracy than it could with either one alone. Consequently, our approach can be viewed either as a means to improve the accuracy of an initial theory using training examples, or as a means to improve the learning from examples using an initial theory.

Directions for future work include experimenting with other models of noise in the examples and comparing an approximate theory according to our formalization with a theory obtained from a knowledge source in a real-world domain.

Acknowledgments

We would like to thank Hussein Almuallim for his insightful comments. We are also indebted to Boonserm Kijsirikul for his excellent help in implementing LaTeX, Tsuyoshi Murata and Craig Hicks for their help in preparing the paper.

References

Angluin, D. and Laird, P. 1988. Learning from noisy examples. *Machine Learning* 2:343–370.

Bain, M. and Muggleton, S. 1990. Non-monotonic learning. In *Machine Intelligence12*. Oxford University Press.

Bergadano, F. and Giordana, A. 1988. A knowledge intensive approach to concept induction. In *Proc. the Fifth International Conference on Machine Learning*. Morgan Kaufmann. 305–317.

Brunk, C. and Pazzani, M. 1991. An investigation of noise-tolerant relational concept learning algorithms. In *Proc. the Eighth International Workshop on Machine Learning*. Morgan Kaufman. 389–393.

Cain, T. 1991. The DUCTOR: A theory revision system for propositional domains. In *Proc. the Eighth International Workshop on Machine Learning*. Morgan Kaufman. 485–489.

Cohen, W. 1990. Learning from textbook knowledge: A case study. In *Proc. the Eighth National Conference on Artificial Intelligence*. AAAI Press/MIT Press. 743–748.

DeJong, G. and Mooney, R. 1986. Explanation-based learning: An alternative view. *Machine Learning* 1(2):145 – 176.

Drastal, G.; Czako, G.; and Raatz, S. 1989. Induction in an abstract space: A form of constructive induction. In *IJCAI 89*. Morgan Kaufman. 708–712.

Dzeroski, S. and Lavrac, N. 1991. Learning relations from noisy examples: an empirical comparison of LINUS and FOIL. In *Proc. the Eighth International Workshop on Machine Learning*. Morgan Kaufman. 399–402.

Ginsberg, A. 1990. Theory reduction, theory revision, and retranslation. In *AAAI 90*. Morgan Kaufman. 777–782.

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

Mitchell, T. M. 1978. Version spaces: An approach to concept learning. Technical Report HPP-79-2, Stanford University, Palo Alto, CA.

Mooney, R. To appear. A preliminary PAC analysis of theory revision. In Petsche, T.; Judd, S.; and Hanson, S., editors, *Computational Learning Theory and Natural Learning Systems*, volume 3. MIT Press.

Muggleton, S. and Feng, C. 1990. Efficient induction of logic programs. In *Proc. the First Conference on Algorithmic Learning Theory*. OHMSHA. 368–381.

Muggleton, S.; Srinivasan, A.; and Bain, M. 1992. Compression, significance and accuracy. In *Proc. the Ninth International Conference on Machine Learning*. Morgan Kaufmann. 339–347.

Pazzani, M. and Kibler, D. 1992. The utility of knowledge in inductive learning. *Machine Learning* 9:57–94.

Pazzani, M.; Brunk, C.; and Silverstein, G. 1991. A knowledge-intensive approach to learning relational concepts. In *Proc. the Eighth International Workshop on Machine Learning*. Morgan Kaufman. 432–436.

Quinlan, J.R. and Rivest, R.L. 1989. Inferring decision trees using the minimum description length principle. *Information and Computation* 80:227–248.

Quinlan, J. R. 1983. Learning from noisy data. In *Proc. the 1983 International Machine Learning Workshop*. 58 – 64.

Quinlan, J.R. 1990. Learning logical definitions from relations. *Machine Learning* 5:239–266.

Richards, B. 1992. *An Operator-based Approach to First-order Theory Revision*. Ph.D. Dissertation, The University of Texas at Austin. AI92-181.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica* 14:465–471.

Tangkitvanich, S. and Shimura, M. 1993. Theory refinement based on the minimum description length principle. Technical Report 93TR-001, Department of Computer Science, Tokyo Institute of Technology.

Towell, G. G.; Shavlik, J. W.; and Noordewier, M. O. 1990. Refinement of approximate domain theories by knowledge-based neural networks. In *AAAI 90*. AAAI Press / The MIT Press. 861–866.

Valiant, L. 1984. A theory of the learnable. *CACM*. 27:1134–1142.

Wallace, C. and Boulton, D. 1968. An information measure for classification. *Computer J.* 11:185–194.

Watanabe, L. and Rendell, L. 1991. Learning structural decision trees from examples. In *IJCAI 91*. Morgan Kaufman. 77–776.

Wogulis, J. 1991. Revising relational theory. In *Proc. the Eighth International Workshop on Machine Learning*. Morgan Kaufman. 462–466.