Feature Selection and Hypothesis Selection Models of Induction

Michael J. Pazzani & Glenn Silverstein
Department of Information and Computer Science
University of California
Irvine, CA 92717

Abstract

Recent research has shown that the prior knowledge of the learner influences both how quickly a concept is learned and the types of generalizations that a learner produces. We investigate two learning frameworks that have been proposed to account for these findings. Here, we contrast *feature selection* models of learning with *hypothesis selection* models. We report on an experiment that suggests that human learners use prior knowledge both to indicate what features may be relevant and to influence how the features are combined to form hypotheses. We present an extension to the POSTHOC system, a hypothesis selection model of concept learning, that is able to account for differences in learning rates observed in the experiment.

Introduction

There is a growing body of evidence that the prior knowledge of the learner influences the speed or accuracy of learning (e.g., Ahn, Mooney, Brewer, & DeJong, 1987; Ausubel & Schiff, 1954; Chapman & Chapman, 1967; Murphy & Medin, 1985; Nakamura, G, 1985; Pazzani, 1990; Schank, Collins, & Hunter, 1986; Wattenmaker, Dewey, Murphy, & Medin, 1986; Wisniewski,1989). In this paper, we contrast two learning frameworks that have been proposed to account for these findings. *Feature selection* models, such as that proposed by Lien & Cheng (1989), claim that prior knowledge influences learning by selecting a subset of the available features as potentially relevant. In the feature selection model, induction is accomplished by detection of covariation (Kelley, 1971; 1983) among the relevant features. In contrast, *hypothesis selection* models, exemplified by POSTHOC (Pazzani & Schulenburg,1989), claim that, in addition to selecting relevant features, prior knowledge influences how the relevant features are combined to form hypotheses and how hypotheses are revised when new data are encountered.

In this paper, we first present an experiment in which the feature selection and hypothesis selection frameworks make different predictions on learning rates. Next, we report on an extension to POSTHOC that provides it with the capability to reason about both positive and negative causal influences (i.e., factors that make an action more likely or less likely). Finally, we report on a simulation that indicates that POSTHOC can account for the learning rates observed in the experiment.

Drug Interactions: An Experiment

In order to differentiate between these two frameworks, we designed an experiment in which the their predictions differed. The experiment followed a 2x2 factorial design in which the factors were the type of concept acquired (internal or external disjunction) and the type of background information in the instructions. It has been reported (Wells, 1963), that in the absence of relevant prior knowledge, exclusive disjunctions concepts are more difficult for subjects to learn that inclusive disjunctive concepts. The feature selection model would predict that an exclusive disjunction of relevant features would take more trials to learn than an inclusive disjunction of the same relevant features. The reason for this prediction is that the feature selection model predicts that prior knowledge will influence only the selection of features. After this selection is made, one would expect the same degree of difficulty as in the case in which subjects have no relevant prior knowledge. In contrast, the hypothesis selection model predicts that an exclusive disjunction may be easier for subjects to learn than an inclusive disjunction, if the exclusive

disjunction of relevant features is consistent with the subjects prior knowledge, but the inclusive disjunction of these same features is not consistent.

In the experiment, subjects were told that they were to review records of patients brought to the emergency room of a hospital because of an overdose of sleeping pills and that the effect of the sleeping pills was to lower the patients heart rate. The patient records were described by the following five features and that each feature had one of two values:

• Gender:	The gender of the patient.	(Female or Male)
• Time:	The time of day.	(AM or PM)
• Oral:	Each patient is given a capsule to swallow.	(Drug-o or Sugar)
• Intravenous:	Each patient is given an injection.	(Drug-i or Saline)
• Doctor:	The attending physician.	(Ramsey or Jankins)

The subjects had to learn a way to predict whether or not the patient's heart rate will increase. Subjects had to learn either an inclusive or an exclusive disjunction of Drug-o and Drug-i (i.e., some subjects were presented with data that indicated that a patient's heart rate would increase only if a patient was given either Drug-i or Drug-o; other subjects were presented with data that indicated that the heart rate would increase only if a patient was given Drug-i or Drug-o, but not both).

The instructions also included background information. Two sets of instructions were prepared. The only difference between them is that one set of instructions omitted the item underlined below:

- The nurses in the PM shift receive a 10% higher salary than those in the AM shift.
- Female patients typically weigh less than male patients.
- Drug-o has been used by truck drivers to stay alert.
- Drug-i has been shown to increase aggressiveness in primate studies.
- When both Drug-o and Drug-i are given to laboratory animals, they result in a coma.
- Dr. Ramsey received his degree from Rutgers University in 1959.
- Dr. Jankins received his degree from Yale University in 1988.

Note that the background information contained items irrelevant to this particular task and that the relevant background information required plausible reasoning rather than purely deductive reasoning.

The hypothesis selection model makes several predictions about the outcome of the experiment. The predictions all follow from the thesis that concepts consistent with prior knowledge take fewer trials to learn than concepts that are not consistent with prior knowledge:

- Subjects learning the exclusive disjunction who were shown information on the drug interaction would learn more rapidly than subjects learning this concept without this information.

- Subjects learning the exclusive disjunction who were shown information on the drug interaction would learn more rapidly than subjects learning an inclusive disjunction with this information.

- Subjects learning the inclusive disjunction who were not shown information on the drug interaction would learn more rapidly than subjects learning this concept who were shown this information.

In contrast, the feature selection model would predict that including or omitting the information on the interaction between Drug-i and Drug-o would not affect the learning rate. It assumes that prior knowledge only focuses attention on features and covariation alone indicates how the features are combined.

Subjects. The subjects were 52 male and female undergraduates attending the University of California, Irvine who participated in this experiment to receive extra credit in an introductory psychology course. Subjects were randomly assigned to one of the four conditions. Subjects were tested in two groups of 26.

Stimuli. The stimuli consisted of patient records that were displayed on the monitor of a Macintosh computer. Since there are five two-valued features, a total of 32 records were constructed.

Procedures. Each subject was shown a patient record on the computer screen and asked to predict whether or not the heart rate would increase by clicking on a box containing the word Yes or a box containing the word No (i.e., using a mouse to move a pointer to the box and pressing a button on the mouse) . While still displaying the patient record, the computer indicated the correct answer by displaying the word Increase or Decrease. Next, the subject clicked on a box labeled Continue and the next patient record was shown. This process was repeated until the subjects were able to predict or classify correctly on 7 consecutive trials. The subjects were allowed as much time as they wanted to make their prediction and to view the record after the correct answer was shown. We recorded the number of the last trial on which the subject made an error. The records were presented in a random order. If the subject did not obtain the correct answer after 60 trials, we recorded that the last error was made on trial 60. The subjects were permitted to consult the instructions, containing information on operating the computer and background information at any time during the experiment.

Results. Table 1 displays the results of the four conditions. The results of this experiment confirmed the predictions of the hypothesis selection model ($p < .05$, level $F(1, 48) = 4.48$). A Tukey HSD finds a significant difference ($p < .05$) on the the following comparisons:

- Subjects learning an exclusive disjunction and provided with information on the drug interaction did learn more rapidly than subjects learning the same concept without this background knowledge (7.3 vs. 28.2). This difference suggests that the knowledge of the interaction between the drugs facilitates learning this concept.

- Subjects learning an exclusive disjunction and provided with information on the drug interaction would learn more rapidly than subjects learning an inclusive disjunction and provided with information on drug interaction. (7.3 vs. 16.2). In this case, the knowledge of the interaction interferes with learning an inclusive disjunction since this hypothesis is not consistent with the prior knowledge.

Although not statistically significant, the data do not contradict the third prediction of the hypothesis selection framework:

- Subjects learning an inclusive disjunction and provided with information on the drug interaction would learn less rapidly than subjects learning an inclusive disjunction without this extra misleading knowledge (14.9 vs. 16.2). If we ignore the score of a subject who failed to complete the experiment, this result is more in line with our expectations (11.3 vs. 16.2).

Table 1. Mean number of trials required by human subjects

	Inclusive	Exclusive
With knowledge of interaction	16.2	7.3
Without knowledge of interaction	14.9	28.2

The results of this experiment provide support for hypothesis selection models of concept learning. In this framework, the hypothesis space is reduced by eliminating those hypotheses that are not consistent with the prior knowledge of the learner. In contrast, feature selection models restrict the hypothesis space less than hypothesis selection models. All hypotheses composed of potential relevant features are considered consistent with prior knowledge. As a consequence, the feature selection framework does not account for the differences in learning rates observed in this experiment.

POSTHOC: A hypothesis selection model

POSTHOC (Pazzani & Schulenburg, 1989) is a hypothesis selection model of concept learning. Through the use of a set of productions and a background theory that represents prior knowledge, POSTHOC maintains a single hypothesis that summarizes the examples seen and classifies new examples. The productions are used to suggest hypotheses and to revise hypotheses that misclassify examples. Because some of the productions do not make use of background knowledge, the system has the ability to create hypotheses that are not consistent with its background knowledge (e.g., if the background knowledge is incomplete or incorrect). In this paper, we discuss an extension to the system that allows it to make use of negative as well as positive influences. Note that POSTHOC was not modified in anyway to make it learn exclusive disjunctions. Rather, hypotheses representing exclusive disjunctions are formed using background knowledge when there are two separate features that positively influence a result, but the combination of the features negatively influence the result. Without the background knowledge of the specific drug interaction, POSTHOC would create an initial hypothesis consistent with its theory and latter be forced to revise them using productions that ignore the background knowledge. If POSTHOC has no background theory at all, then it would create its hypotheses using only productions that ignore background knowledge.

In POSTHOC, examples are expressed as set of feature-value pairs and an outcome. For example, in the medical experiments described in the previous section, a patient record describing a male patient who was administered Drug-o orally and a saline solution intravenously after noon and whose heart rate increased would be represented as follows:

[gender male] [time pm] [oral drug-o] [intravenous saline] $\in$ increase

POSTHOC maintains a single hypothesis that consists of a DNF description (i.e., disjunction of conjunctions) of the concept being learned. One hypothesis that is consistent with the above example states that the heart rate will increase if Drug-o is given orally in the afternoon.

[oral drug-o] $\wedge$ [time pm] $\rightarrow$ increase

Of course, there are numerous other hypotheses consistent with this example that may or may not be consistent with future examples or with the prior knowledge of the learner.

The influence theory which comprises POSTHOC's prior knowledge consists of two components: a set of influences which describe tendencies that either facilitate or hinder the desired outcome (e.g. increasing the heart rate) and a set of inferences rules that indicate when these influences are present. The influence theory of POSTHOC for the medical experiment described in the previous section includes the following influences and inferences:

Influences:

```
(easier more-alert  increase)
(easier more-aggressive  increase)
```

Inferences:

```
(implies [oral drug-o]  more-alert)
(implies [intravenous drug-i] more-aggressive)
```

These influences and associated inferences suggest that making a person more alert or more aggressive can facilitate increasing that persons heart rate and taking Drug-o orally tends to make a person more alert and taking Drug-i intravenously tends to make a person more aggressive.

The influence theory described above includes only positive influences. However, some influences can only be expressed naturally as negative influences. One such example the knowledge of drug interactions which arises in the medical experiment described in the previous section. For instance, if we wished to include the knowledge that Drug-i and Drug-o tend to interact negatively in the patient to produce a coma (thus lowering the heart rate), then this knowledge is represented as the following influence and associated inference:

Influences:

```
(harder coma  increase)
```

Inference

```
(implies ([oral drug-o] ∧ [intravenous drug-i]) coma)
```

Adding negative influences to POSTHOC required extending the productions presented in Pazzani & Schulenburg (1989). There are three types of productions. One set deals with errors of commission in which a positive example is falsely classified as a negative example. These productions makes the hypothesis more general. The second set deals with errors of omission in which a negative example is falsely classified as a positive example. These productions makes the hypothesis more specific. The final set creates an initial hypothesis when the first positive example is encountered. For brevity only those productions which utilize the negative influences will be described. For a description of the remaining productions, the reader is referred to Pazzani & Schulenburg (1989):

Initializing Hypothesis.

```
IF there are features of the example that are indicative of the
inverse of a negative influence
THEN initialize the hypothesis to the negation of the conditions
indicative of the negative influence.
```

Errors of Omission.

```
IF the hypothesis is consistent with the influence theory
AND there are features that are indicative of the inverse of a
negative influence
THEN create a conjunction of the current hypothesis and the negation
of the conditions indicative of the negative influence.
```

Errors of Commission.

```
IF the  hypothesis is consistent with the background theory
AND for each true conjunction there are features not present in the
current example that would be necessary for the inverse of a negative
influence
THEN modify the conjunct by conjoining the negation of the conditions
indicative of the negative influence.
```

To illustrate the use of these productions, a trace of POSTHOC learning an exclusive disjunction is provided below. For brevity, we omit the "doctor" attribute from the examples. The first example that POSTHOC is presented with is an example of a treatment that successfully increases the patient's heart rate where a male patient is administered a sugar pill orally and Drug-i intravenously in the PM by Dr. Ramsey:

[gender male] [time pm] [oral sugar] [intravenous drug-i] ∈ increase

Since there is no initial hypothesis, POSTHOC uses an initialization production to create a hypothesis that accounts for the outcome of this example. A positive influence more-aggressive is present and POSTHOC creates the hypothesis that Drug-i leads to the increased heart rate:

[intravenous drug-i] → increase

This hypothesis is consistent with several more examples. Next, an example is presented where a patient is administered Drug-o orally and a saline solution intravenously. Here an error of omission occurs since POSTHOC predicts that the patient's heart rate will not increase but it does increase. The example encountered is:

[gender male] [time am] [oral drug-o] [intravenous saline] ∈ increase

The hypothesis is revised by an Error of Omission production for positive influences a multiple sufficient hypothesis is produced:

[intravenous drug-i] ∨ [oral drug-o] → increase

Again this hypothesis is consistent with several more examples. However, POSTHOC makes the wrong prediction when it encounters an example where the patient is administered both Drug-i intravenously and Drug-o orally. POSTHOC predicts that the patient's heart rate will increase, but the patients heart rate does not increase. This results in an error of commission. The example presented to POSTHOC is:

[gender male] [time am] [oral drug-o] [intravenous drug-i] ∉ increase

To correct its hypothesis, POSTHOC uses the Error of Commission production for negative influences. For the each disjunct, [oral drug-o] and [intravenous drug-i], the negation of the features indicative of the negative influence is conjoined with each conjunct, the results hypothesis is:

[intravenous drug-i] ∧ **not** ([oral drug-o] ∧ [intravenous drug-i]) ∨
[oral drug-o] ∧ **not** ([oral drug-o] ∧ [intravenous drug-i]) → increase

This hypothesis can be simplified to:

[intravenous drug-i] ∧ **not** ([oral drug-o]) ∨
[oral drug-o] ∧ **not** ([intravenous drug-i]) → increase

which is consistent with the remaining examples and represents the exclusive disjunction.

Simulation Results

We ran POSTHOC on 200 random orderings of the data on each of the same four conditions used in the experiment on human subjects. The results are shown in Table 2. The negative influence (i.e., the drug interaction) was not used to simulate the condition in which this item was not included in the instructions. The data show the same trends as the human experimental data: learning the exclusive disjunction of administering Drug-i intravenously and Drug-o orally was facilitated by the knowledge that Drug-i and Drug-o interact to put the patient in a coma (8.3 vs. 45.7) ; when provided with information on the drug interaction, learning the exclusive disjunction of the drugs was easier than learning the inclusive disjunction (6.0 vs. 8.3); learning an inclusive disjunction of the drugs when provided with misleading information on the drug interaction required more trials than the same concept without this extra misleading knowledge (6.0 vs. 3.7).

Table 2. Mean number of trials required by POSTHOC

	Inclusive	Exclusive
With knowledge of interaction	6.0	8.3
Without knowledge of interaction	3.7	45.7

In POSTHOC, we have focused on how prior knowledge influences learning rates and we have so far ignored other information used by human learners (e.g., perceptual salience of features, Bower & Trabasso, 1968). As a consequence, POSTHOC is not intended to make quantitative predictions on the number of training examples but rather predicts the relative difficulty of learning.

Future Directions

There are three possible direction in which we plan to extend the hypothesis selection model. First, we would like to be able to use the prior knowledge of the learner to influence the interpretation of ambiguous feature (Medin & Wisniewski, 1990). Second, we would like POSTHOC to be able to use more abstract knowledge. Currently, POSTHOC can represent the information that there there is a specific interaction between two drugs or that there is no drug interaction. In contrast, our subjects also appeared to have more general knowledge that indicates such things as drugs may interact and can use this knowledge to explain the specific interaction seen in the experiment in terms of the general knowledge of drug interactions. Finally, we plan to extend POSTHOC so that when it learns accurate hypotheses that are not consistent with its background knowledge, the background knowledge is revised to accommodate the new findings.

Conclusions

We have presented experimental evidence that provides support for hypothesis selection models of concept learning. We have extended POSTHOC to include negative influences and shown that with this extension alone, it is able to predict the relative order of difficulty of trials on inclusive and exclusive disjunctions. Recent work on the analysis of the limitations of inductive learning algorithms (Valiant, 1984; Dietterich, 1989) is in sharp contrast to the versatility demonstrated by human learners. We believe that approaches that make use of background knowledge to focus all aspects of learning are central to accounting for the generality of human learning.

Acknowledgements

The software used to run the experiment was designed by Francis Nguyen and Takeshi Tsubota. We would like to thank Kamal Ali, Cliff Brunk, David Foster, and Scott Truesdel for assistance in running the experiment and Gupi Silverstein and Caroline Ehrlich for commenting on an earlier draft of the paper. This research is supported in part by National Science Foundation Grant IRI-8908260.

Bibliography

Ahn, W., Mooney, R., Brewer, W., & DeJong, G. (1987). Schema acquisition from one example: Psychological evidence for explanation-based learning. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society.* Seattle, WA: Lawrence Erlbaum Associates.

Ausubel, D. M., & Schiff, H. M. (1954). The effect of incidental and experimentally induced experience on the learning of relevant and irrelevant causal relationships by children. *Journal of Genetic Psychology, 84,* 109-123.

Bower, G., & Trabasso, T. (1968). *Attention in learning: Theory and research.* New York: John Wiley and Sons.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous diagnostic observations. *Journal of Abnormal Psychology, 72*, 193-204.

Dietterich, T. (1989). Limitations on inductive learning. *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 124-128). Ithaca, NY: Morgan Kaufmann.

Kelley, H. (1971). Causal schemata and the attribution process. In E. Jones, D. Kanouse, H. Kelley, N. Nisbett, S. Valins & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior*. Morristown, NJ: General Learning Press.

Kelley, H. (1983). The process of causal attribution. *America Psychologist*, 107-128.

Lien, Y., & Cheng, P. (1989). A framework for psychological induction: Integrating the power law and covariation views. *The Eleventh Annual Conference of the Cognitive Science Society.* (pp. 729-733).Ann Arbor, MI: Lawrence Erlbaum Associates, Inc.

Medin, D., & Wisniewski, E. (1990). Paper presented at the Symposium on Computational Approaches to Concept Formation, Stanford, CA.

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

Nakamura, G. (1985). Knowledge-based classification of ill-defined categories. *Memory & Cognition. 13*, 377-384.

Pazzani, M. & Schulenburg, D. (1989). The influence of prior theories on the ease on concept acquisition. *The Eleventh Annual Conference of the Cognitive Science Society.* (pp. 812-819).Ann Arbor, MI: Lawrence Erlbaum Associates, Inc.

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

Schank, R., Collins, G., & Hunter, L. (1986). Transcending inductive category formation in learning. *Behavioral and Brain Sciences, 9*, 639-686.

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

Wattenmaker, W., Dewey, G., Murphy, T., & Medin, D. (1986). Linear severability and concept learning: Context, Relational properties and concept naturalness. *Cognitive Psychology, 18*, 158-194.

Wells, H. (1963). Effects of transfer and problem structure in disjunctive concept formation. *Journal of Experimental Psychology, 65*, 63-69.

Wisniewski, E. (1989). Learning from examples: The effect of different conceptual roles. *The Eleventh Annual Conference of the Cognitive Science Society.* (pp. 980-986).Ann Arbor, MI: Lawrence Erlbaum Associates, Inc.