Generalizations by Rule Models and Exemplar Models of Category Learning

Thomas J. Palmeri (tpalmeri@ucs.indiana.edu)
Robert M. Nosofsky (nosofsky@ucs.indiana.edu)
Psychology Department
Indiana University
Bloomington, IN 47401

Abstract

A rule-plus-exception model of category learning, RULEX (Nosofsky, Palmeri, & McKinley, 1992), and an exemplar-based connectionist model of category learning, ALCOVE (Kruschke, 1992), were evaluated on their ability to predict the types of generalization patterns exhibited by human subjects. Although both models were able to predict the average transfer data extremely well, each model had difficulty predicting certain types of generalizations shown by individual subjects. In particular, RULEX accurately predicted the prominence of rule-based generalizations, whereas ALCOVE accurately predicted the prominence of similarity-based generalizations. A hybrid model, incorporating both rules and similarity to exemplars, might best account for category learning. Furthermore, a stochastic learning rule, such as that used in RULEX, might be crucial for capturing the different types of generalizations patterns exhibited by humans.

Introduction

Two major theories have been advanced to describe category learning. Rule-based models posit that the category membership of a novel object is determined by the application of rules. Exemplar-based models posit that the category membership of a novel object is determined by how similar it is to previously stored exemplars.

This work was supported by Grant PHS R01 MH48494-01 from the National Institute of Mental Health to Indiana University.

In general, rule-based models have been limited to situations where the categories are well-defined. In this paper, we tested a new rule-plus-exception model, RULEX, which has been successful at learning ill-defined categories in addition to well-defined categories (Nosofsky, Palmeri, & McKinley, 1992). RULEX and an exemplar model, ALCOVE (Kruschke, 1992), are tested on their ability to predict the different types of generalization patterns that people make when categorizing novel objects.

In a typical category learning task, objects are presented one at a time and subjects are asked to decide whether an object is a member of category A or category B. During training, corrective feedback is provided about whether the correct response has been made. Following training, a transfer phase is given in which old objects as well as new objects are presented, without feedback. Subjects are required to judge whether the new objects are from category A or category B by generalizing from what they have learned during training.

Traditionally, category learning models have been judged by how well they predict average transfer data or learning data. A point has been reached in theory development where each of the major theories are able to quantitatively predict a large number of extant phenomena (Estes, in press). Clearly, additional data are needed to tease apart the predictions of each of the existing models.

An approach that we propose is to examine the types of generalization patterns made by individual subjects (see also Pavel, Gluck, & Henkle, 1988). It is reasonable to suggest that different subjects might be using very different strategies during a category learning experiment.

Average transfer data obscure the different patterns of generalization that might be found at the individual subject level (see also Martin & Caramazza, 1980). Our goal is to compare the observed generalization patterns to those predicted by RULEX and ALCOVE.

Table 1 displays the abstract category structure that was used (Medin & Schaffer, 1978). The categories were ill-defined in that there was not a simple rule which could be used to decide if a given stimulus was a member of category A or B. There were seven transfer stimuli, T1-T7. A generalization pattern AAABBBB reflects a subject who classified stimuli T1-T3 as an A and T4-T7 as a B.

We fit an exemplar model, ALCOVE, and a rule-plus-exception model, RULEX, to the resulting distribution of generalization patterns. Although both models were able to accurately predict the average transfer performance with high accuracy, each model was able to qualitatively predict only a portion of the distribution of generalization patterns found with individual subjects.

Method

Subjects. Subjects were 227 undergraduates at Indiana University who participated to receive credit in an introductory psychology course.

Stimulus		Observed	RULEX	ALCOVE
Category A				
A1	1112	0.86	0.86	0.90
A2	1212	0.90	0.89	0.90
A3	1211	0.93	0.93	0.93
A4	1121	0.66	0.71	0.61
A5	2111	0.65	0.70	0.64
Category B				
B1	1122	0.31	0.34	0.34
B2	2112	0.34	0.34	0.38
B3	2221	0.12	0.12	0.16
B4	2222	0.05	0.07	0.07
Transfer				
T1	1221	0.59	0.60	0.61
T2	1222	0.37	0.41	0.34
T3	1111	0.92	0.93	0.93
T4	2212	0.35	0.40	0.38
T5	2121	0.23	0.28	0.16
T6	2211	0.58	0.60	0.64
T7	2122	0.08	0.08	0.07

Table 1. Observed transfer data and predictions by RULEX and ALCOVE.

Stimuli. The stimuli were computer-generated line drawings of rocketships that varied along four binary-valued dimensions: shape of wing, nose, porthole, and tail. The abstract category structure is given in Table 1. Assignment of physical dimension to abstract dimension was randomized for every subject.

Procedure. There were 16 blocks of training trials. Each of the 9 training stimuli, A1-A5 and B1-B4, were presented once per block. The order of presentation was randomized for each subject. On every trial, the subject was presented with a rocketship and was asked to judge if it was from planet A or planet B. After responding, corrective feedback was provided.

During transfer, subjects were shown the 9 training stimuli as well as 7 new transfer stimuli, T1-T7, in random order. Subjects judged whether each of the rocketships was from planet A and planet B. There were three blocks of transfer trials. No feedback was provided.

Results and Discussion

A median split was conducted on the total number of errors made during the last four blocks of training. Overall transfer performance from the top median, the "learners," is shown in Table 1 as the probability of responding with category A.

Our primary interest was the distribution of generalization patterns which underlie the average transfer performance shown in Table 1.

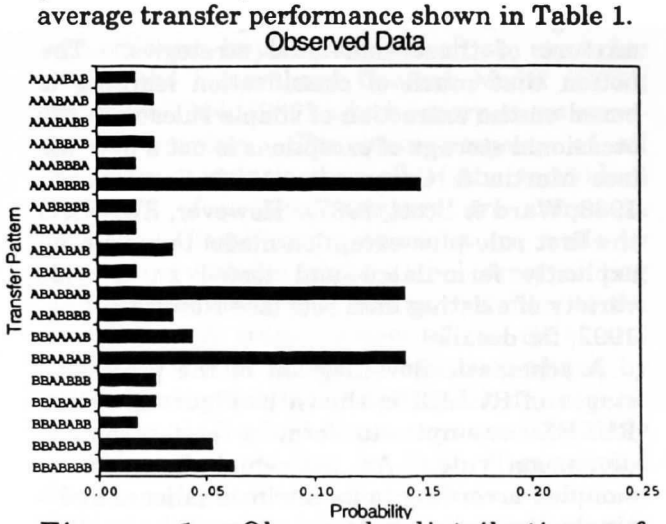

Figure 1. Observed distribution of generalization patterns.

Because there were 7 new transfer stimuli, each of which could be classified as an A or a B, there were 2^7=128 possible transfer patterns. Only 31 of the possible patterns were observed. Figure 1 displays a histogram of the 19 patterns of generalization that were observed in more than one subject.

As shown in Figure 1, there were three prominent generalization patterns observed, AAABBBB, ABABBAB, and BBAABAB. Pattern AAABBBB is consistent with a single dimension rule along the first dimension, where value 1 signals an A. Similarly, pattern BBAABAB is consistent with a single-dimension rule along the third dimension. We discuss the potential source of these and the other generalization patterns in more detail below.

Theoretical Analyses

RULEX

The Model. Nosofsky et al. (1992) introduced a rule-plus-exception model of classification learning called RULEX. The basis for classification learning is the acquisition of simple single-dimension rules or conjunctive rules supplemented by the partial storage of exceptions to those rules. One of the main properties of RULEX is that the behavior of individual subjects is highly idiosyncratic -- different subjects will form different rules and store different partial exceptions to those rules. Average classification data are presumed to be a mixture of these individual strategies. The notion that much of classification learning is based on the extraction of simple rules with the occasional storage of exceptions is not a new one (see Martin & Caramazza, 1980; Pavel et al., 1988; Ward & Scott, 1987). However, RULEX is the first rule-plus-exception model to have been explicitly formulated and tested on a wide variety of existing data sets (see Nosofsky et al., 1992, for details).

A schematic flow-diagram of the processing stages of RULEX is shown in Figure 2. First, RULEX attempts to form a perfect single-dimension rule. An individual dimension is sampled according to its intrinsic salience and a single-dimension rule is formed over that dimension. If the rule works perfectly for a certain number of trials, called the *upper test window* (equal to twice the number of training items by default), then it is permanently stored. If the rule fails to work perfectly, then it is discarded and a new dimension is selected according to its salience.

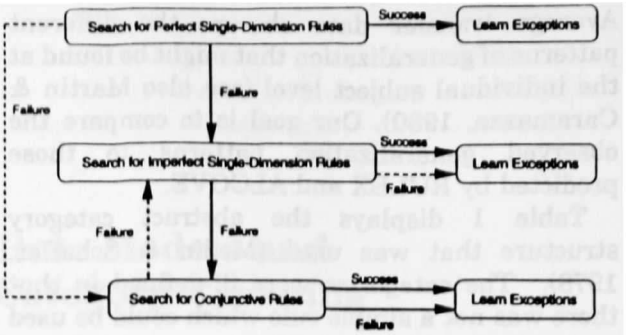

Figure 2. Schematic flow diagram of RULEX.

If no dimension yields a perfect rule then RULEX searches for imperfect single-dimension rules. A dimension is selected according to its intrinsic salience and an imperfect single-dimension rule is retained for a minimum number of trials, called the *lower test window* (equal to the number of training items by default). The imperfect single-dimension rule is maintained only if its performance exceeds a lax criterion (around 60% correct by default). Once the upper test window is reached, the imperfect rule is permanently stored if performance exceeds a strict criterion, *scrit* (in general, *scrit* can vary over some distribution of values). If this rule does not exceed the strict criterion then it is discarded and another dimension is selected. When all dimensions have been sampled a search for conjunctive rules begins in a manner similar to that for single-dimension rules.

After a single-dimension rule or a conjunctive rule is permanently stored, then RULEX begins the exception-storage process. If an item is encountered that contradicts the rule, RULEX probabilistically samples each of the dimensions of the item with probability *pstor*, which is a free parameter (in general, *pstor* can vary over some distribution of values); the dimensions that were part of the failed rule are sampled with probability one. Storage of the exceptions is also probabilistic. It is a function of the number of dimensions sampled and the number of exceptions already stored in memory.

Consider the category structure shown in Table 1. Suppose a subject formed a rule that value 1 on dimension 1 signals an A and value 2 on dimension 1 signals a B. Upon encountering stimulus 2111, the rule is applied, and an error

occurs. To form an exception, dimension 1 is sampled with probability one, and dimensions 2, 3, and 4 are sampled with probability *pstor*. If the sampled exception were 21** (where * represents a nonsampled dimension that can match any value), then RULEX would attempt to learn that the exception 21** signals category A. If this exception is stored and later produces an error, it is discarded from memory.

Classification decisions are made by first checking all of the exceptions stored in memory. If an exception applies to the given stimulus, then it is used to make a response. For example, the exception 21** applies to training stimuli 2111 and 2112. If no exceptions apply, then the rules are checked. If none of the rules apply, then a random guess is made.

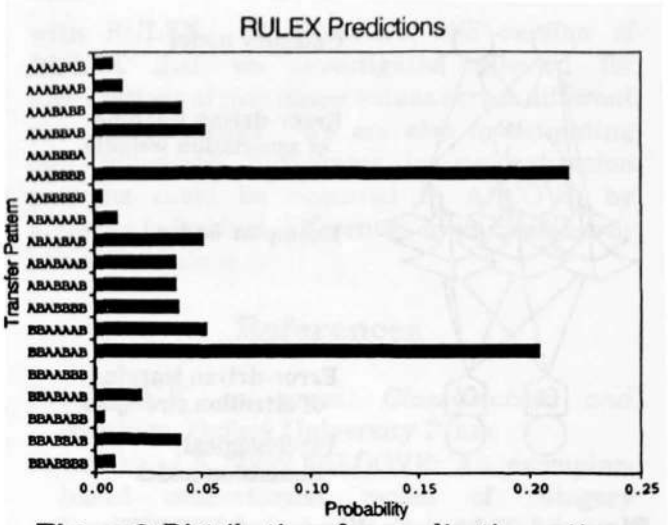

Figure 3. Distribution of generalization patterns predicted by RULEX.

Predicted Generalizations. RULEX is a simulation model which is inherently stochastic in terms of the rules and exceptions that are stored (although it uses a deterministic response rule). So, 5000 simulated "subjects" were run through 16 blocks of training and then were tested on the 9 old and the 7 new stimuli in a transfer block. RULEX was fitted to the averaged transfer data by minimizing the sum-of-squared deviations between the observed and predicted probabilities.[1] A four-parameter version of RULEX was fitted to the data, where *pstor* and *scrit* varied along an interval. The best fitting values were *pstor* varying between .30 and .65, and *scrit* varying between .60 and .80. The predicted average transfer performance is shown in Table 1 (using these same parameter values). As expected, the fit to the average transfer data was very good (RMSD=.029, %Var=99.1).

The predicted distribution of generalization patterns is shown in Figure 3. RULEX accounted for 68% of the variance in the distribution of generalization patterns (RMSD=.014). RULEX qualitatively predicts the prominence of patterns AAABBBB and BBAABAB, although it overpredicts their probabilities. As expected from these patterns, RULEX predicted that 40% of subjects would develop rules based on dimension 1 (pattern AAABBBB) and that 40% would develop rules based on dimension 3 (pattern BBAABAB). The remaining subjects developed single-dimension rules along dimension 4 or else stored only idiosyncratic, high-dimension exceptions. RULEX greatly underpredicts the probability of the generalization pattern ABABBAB. Although RULEX has some shortcomings, it is able to qualitatively predict some of the important generalization patterns exhibited by individual subjects. Recall too that it yielded excellent quantitative predictions of the averaged transfer data.

[1] Attempts to fit RULEX to the observed distribution of generalization patterns increased the fit only slightly.

ALCOVE

The Model. ALCOVE (Kruschke, 1992) is a connectionist implementation of an exemplar model, the Generalized Context Model (GCM; Nosofsky, 1984, 1986), which incorporates error-driven learning. The main property of all exemplar models is that all of the individual instances of a given category are stored in memory. Classification decisions are made by responding with the category label corresponding to the stored items that are most similar to the new item. A crucial aspect of the GCM and ALCOVE is that similarity can be modified by selective attention to the component dimensions of objects. Individual dimensions are selectively attended to depending on their diagnosticity. For example, in the category structure shown in Table 1, dimension 2 is relatively nondiagnostic because half of the items in category B have value 1 on that dimension and the other half have value 2.

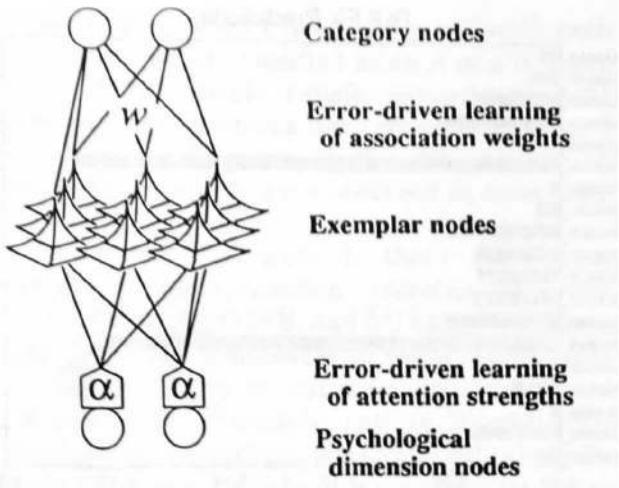

Figure 4. Schematic diagram of ALCOVE.

As shown in Figure 4, ALCOVE is a three-layer feed-forward network. The input layer has a single node for each dimension of the stimuli. The activation value of an input node is equal to the value of the stimulus on that dimension. Each hidden node represents a single training exemplar. The activation of a hidden node is a function of the similarity between the input stimulus and the exemplar that hidden node represents. The present stimuli were composed of binary-valued dimensions, so the activation of the jth hidden node is given by $a_j^{hid} = exp(-c \Sigma_i \alpha_i d_{ji})$, where $d_{ji}=0$ if the input stimulus and exemplar node j have the same value on dimension i, otherwise $d_{ji}=1$. The positive constant c is called the specificity of the node. Like the GCM, each dimension is weighted by a selective attention parameter α_i. Unlike the GCM, where these attention weights are free parameters, in ALCOVE these weights are learned via backpropagation.

Every hidden node in ALCOVE is connected to each category output node. The activation of output node k is given by $a_k^{out} = \Sigma_j w_{kj} a_j^{hid}$, where w_{kj} is the weight on the connection between hidden node j and output node k, and a_j^{hid} is the activation of hidden node j. Output activations are converted to response probabilities by an exponential form of Luce's choice rule,

$$p(K) = \frac{\exp(\phi a_K^{out})}{\sum_{out\ k} \exp(\phi a_k^{out})}$$

where $p(K)$ is the probability of responding with category label K and ϕ is a positive real-valued mapping constant.

The attention weights, α_i, and the weights between hidden nodes and output nodes, w_{kj}, are learned via backpropagation (Rumelhart, Hinton, & Williams, 1986), with

$$\Delta w_{kj}^{out} = \lambda_w (t_k - a_j^{out})\, a_j^{hid}$$

$$\Delta \alpha_i = -\lambda_\alpha \sum_{hid\ j} \left[\sum_{outk} (t_k - a_k)^{out} w_{kj}\right] a_j^{hid} c d_{ji}$$

where the teacher value, $t_k = +1$ if the stimulus belongs to category k, or $t_k = -1$ if the stimulus does not.

Predicted Generalizations. Unlike RULEX, ALCOVE has a deterministic learning rule and responses are probabilistic. Hence, simulations were not required, and the model could be fitted directly to the distribution of generalizations. The average predicted transfer performance is shown in Table 1 (using the parameters given below). As expected, the fit by ALCOVE is comparable to that of RULEX (RMSD=.012, %Var=98.6).

The predicted distribution of generalizations is shown in Figure 5 (RMSD=.017, %Var=50.9). The best fitting parameter values were ϕ=1.335, λ_w=0.055, λ_α=0.261, c=0.810. ALCOVE greatly underpredicts the probabilities for generalization patterns AAABBBB and BBAABAB. However, ALCOVE is able to predict the prominence prominence of pattern ABABBAB, unlike RULEX. A straightforward interpretation of this

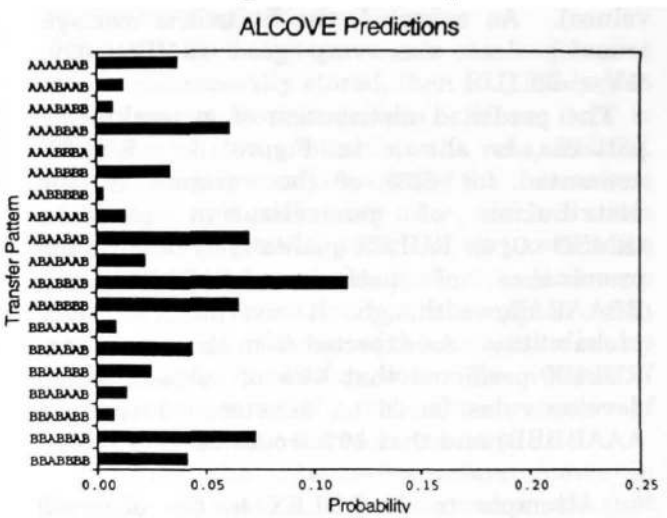

Figure 5. Distribution of generalization patterns predicted by ALCOVE.

pattern of generalization, therefore, is that it represents subjects who used a similarity-to-examples strategy for classification.

Conclusions

We compared a rule-plus-exception model, RULEX, and an exemplar model, ALCOVE, on their ability to account for generalization patterns observed by individual subjects. Although both models were able to accurately account for averaged transfer data quite accurately, both models had difficulty, qualitatively and quantitatively, accounting for the distribution of generalization patterns observed at the individual subject level. Our subjects exhibited three primary patterns. Two of these were best accounted for by RULEX and could easily be characterized by single-dimension rules along dimensions 1 and 3. The third pattern was best accounted for by ALCOVE and could easily be characterized by similarity to stored exemplars.

One possible explanation for our results is that some subjects used a rule-plus-exception strategy whereas other subjects used an exemplar strategy. Indeed, we have conducted preliminary exploration of such a mixed model and it produces greatly improved results. In addition, protocols extracted from subjects after the experiment revealed that some subjects reported using simple rules whereas other subjects reported merely memorizing the stimuli.

Another possible explanation is that category learning consists of both rule learning and storage of exemplars, not merely storage of exceptions to rules. Early in learning, subjects could easily pick up on the imperfect rules which underlie the categories. Later in learning, after each of the stimuli have been presented a number of times, category decisions are based on similarity to stored exemplars. We have preliminary evidence from another experiment showing a larger number of similarity-based generalization patterns when additional training blocks are given.

One problem with the current work is that a stochastic-learning model with a deterministic response rule, RULEX, was compared to a deterministic-learning model with a probabilistic response rule, ALCOVE. We are currently working on the development of a stochastic version of ALCOVE which might better compete with RULEX. Furthermore, the version of RULEX that we investigated allowed for distributions of parameter values across different simulated subjects. We are also investigating how individual differences in generalization patterns could be captured in ALCOVE by allowing individual differences in parameters or initial conditions.

References

Estes, W.K. in press. *Classification and cognition*. Oxford University Press.

Kruschke, J.K. 1992. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review* 99:22-44.

Martin, R.C., and Caramazza, A. 1980. Classification in well-defined and ill-defined categories: Evidence for common processing strategies. *Journal of Experimental Psychology: General* 109:320-353.

Medin, D.L., and Schaffer, M.M. 1978. Context theory of classification learning. *Psychological Review*, 85:207-238.

Nosofsky, R.M. 1984. Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10:104-114.

Nosofsky, R.M. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39-57.

Nosofsky, R.M., Palmeri, T.J., and McKinley, S.C. 1992. Rule-plus-exception model of classification learning. Cognitive Science Research Report No. 84, Indiana University: Bloomington.

Pavel, M., Gluck, M.A., and Henkle, V. 1988. Generalization by humans and multi-layer networks. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. 1986. Learning internal representations by error propagation. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel distributed processing: Exploration in the microstructure of cognition: Vol 1: Foundations*. Cambridge: Bradford Books/MIT Press.

Ward, T.B., and Scott, J. 1987. Analytic and holistic modes of learning family-resemblance concepts. *Memory & Cognition*, 15:42-54.