Generalization by humans and multi-layer adaptive networks

M. Pavel Mark A. Gluck Van Henkle

Stanford University

ABSTRACT

Generalization of a pattern categorization task was investigated in a simple, deterministic, inductive learning task. Each of eight patterns in a training set was specified in terms of four binary features. After subjects learned to categorize these patterns in a supervised learning paradigm they were asked generalize their knowledge by categorizing novel patterns. We analyzed both the details of the learning process as well as subjects' generalizations to novel patterns. Certain patterns in the training set were consistently found to be more difficult to learn than others. The subsequent generalizations made by subjects indicate that in spite of important individual differences, subjects showed systematic similarities in how they generalized to novel situations. The generalization performance of subjects was compared to those that could possibly be generated by a two-layer adaptive network. A comparison of network and human generalizations indicate that using a minimal network architecture is not a sufficient constraint to guarantee that a network will generalize the way humans do.

INTRODUCTION

Inductive learning is one of the most difficult and least understood aspects of cognition. During supervised learning an organism is exposed to a few examples of stimulus-response pairs (the training set) from which the organism infers how to to generate correct responses to many other stimuli. The theoretical problem arises from the fact that there usually are many rules that are consistent with the training set but which generate different responses to the novel stimuli. Unlike deduction, induction has no *a priori* normative procedure to decide which set of rules is the most appropriate.

Thus, induction problems can be considered ill-posed problems in that there too many very different solutions. Such problems can be solved by introducing additional constraints or objectives that are external to the original problem. One of the central problems for understanding induction in natural (human) or artificial systems is to determine useful constraints or regularization principles that convert the ill-posed problems into well posed problems.

In spite of the inherent difficulties with defining "good" inductions, people appear to be very good at rapidly learning to induce useful rules. Investigation of how people perform induction or generalization is, therefore, interesting not only to the students of cognition but also to builders of artificial learning machines. Although there have been many attempts to study this problem, most of previous research has been focused primarily on investigation and modeling of average performance (Medin & Schaefer, 1978; Medin, Dewey, & Murphy, 1983; Nosofsky, 1986).

Correspondence should be addressed to: M. Pavel, Bldg 420, Stanford, CA 94305. This research was supported by NSF Grant BNS-8618049, NSF Grant IST-8511589 and Grant from NASA Ames NCC-2-307.

One goal of study reported here was to examine how people generalize in a simple deterministic categorization task in which each pattern is characterized in terms of known binary features. While we expected certain similarities to emerge across human learners, we anticipated that the particular generalizations might be subject to considerable individual differences. To test this idea, we used an experimental paradigm that would permit us to to observe individual subjects during the learning of a categorization task on a set of training patterns and then allow us examine the types of categorizations they made on a set of novel test patterns. In the last section of this paper we compare human generalizations to those of a small adaptive network.

EXPERIMENT 1

The purpose of this study was to record subjects' progress in learning a deterministic categorization, analyze their generalizations, and compare their performance to that of small adaptive networks. The stimuli were similar to those used by Medin, Altom, Edelson, & Freko, (1982), but the procedure was designed to enable us to monitor the learning process in addition to evaluating subsequent generalizations.

Method

Seventy-eight Stanford undergraduates were run from a pool of subjects enrolled in an introductory psychology course. Each stimulus item was composed of four binary dimensions and was presented to subjects as a patient chart listing four different symptom types: Muscles (*tense* or *relaxed*), Insulin (*high* or *low*), Glands (*swollen* or *recessed*), and Sinus (*stuffy* or *runny*). The complete stimulus set consisted of the 16 possible patterns resulting from forming all combinations of the four binary dimensions. As shown in Table 1, with the alternate values of each dimension indicated by either "1" or "0", four of the 16 stimuli were designated as members of category A, four as members of category B, and the remaining 8 were presented as novel items to test for generalization.

TABLE 1: Category Structure from Experiment 1

Category	Item	Dimension			
		1	2	3	4
A	A1	1	1	1	1
	A2	1	1	0	0
	A3	0	1	1	1
	A4	1	0	0	0
B	B1	0	0	1	0
	B2	0	0	0	1
	B3	1	0	1	0
	B4	0	1	0	1
Novel	N1	0	0	0	0
	N2	0	0	1	1
	N3	0	1	0	0
	N4	1	0	1	1
	N5	1	1	1	0
	N6	1	1	0	1
	N7	0	1	1	0
	N8	1	0	0	1

The training stimuli were carefully selected so that the categorization could be performed perfectly by an exclusive-or (XOR) on the last two dimensions (3 & 4) while the first two dimensions (1 & 2) could be used to form a simpler but less effective rule.

Procedure. Throughout the experiment the two categories were referred to as "Turitis" and "Purosis", with the association of each name to a category randomized across subjects. The patient charts were presented on a computer screen with the symptoms arranged vertically (as above). For each individual subject, the order in which the symptom types were displayed on each chart was consistent throughout the entire experiment. However, across subjects the order of display was randomized. The particular symptom names associated with each dimension were also randomized across subjects.

Subjects were instructed to imagine that they were medical interns learning to diagnose patients suffering from one of two diseases. They were told that they would learn to make their diagnoses by attempting to diagnose individual patients: they would be shown a patient chart listing four symptoms, attempt to make a diagnosis, and then be given the correct diagnosis. Subjects were told that they would complete their training after correctly diagnosing approximately 32 patients in a row, at which point they would be tested on their ability to make diagnoses. After being given these instructions, the training phase of the experiment began.

The training phase consisted of successive presentations of the eight stimuli in categories A and B. On each trial one stimulus item was presented, the subject was prompted for a category judgement, and then the subject was given feedback specifying the correct categorization. The order of presentation for the training stimuli was randomized over blocks of 16 trials so that two instances of each item from category A and two instances of each item from category B occurred in each block. The training phase continued until a subject had either met the learning criterion of correctly categorizing all items in two successive blocks, or until the subject had completed 15 blocks without meeting the criterion.

Upon completing the training phase of the experiment the subjects were instructed that they would be tested on the knowledge they had gained in that phase by diagnosing 32 additional patients. They were then presented with 2 instances of each of the 16 stimuli in a random order. For each stimulus they were asked to make their diagnosis and then rate their confidence of the diagnosis on a scale of 1-7 (least to most confident). No feedback was given on these test trials. Following the experiment, subjects were asked to describe the methods they used to make their diagnoses. The entire experimental process took between 30 minutes and one hour, depending on how quickly the subject reached criterion during the training phase.

Results

A summary of the results for all subjects are shown in Figure 1. Each panel of Figure 1 represents the proportion of *A* responses; the first eight patterns represent the training and the last eight the transfer set. The first four patterns of the training set are from category A and the second four patterns from category B. The left panel shows data from the 40 subjects who reached the criterion together with the data of Medin at al (1982). *criterion* subjects learned the task better than those of Medin et al. (1982). Their average performance is almost identical to the results of Medin et al. (1982). In contrast, our *non-criterion* subjects who did not reached the criterion are more similar to Medin's data for the training patterns but differ considerably on the novel patterns. We conclude that our criterion subjects are most like those of Medin except for the more rigorous training given in this experiment.

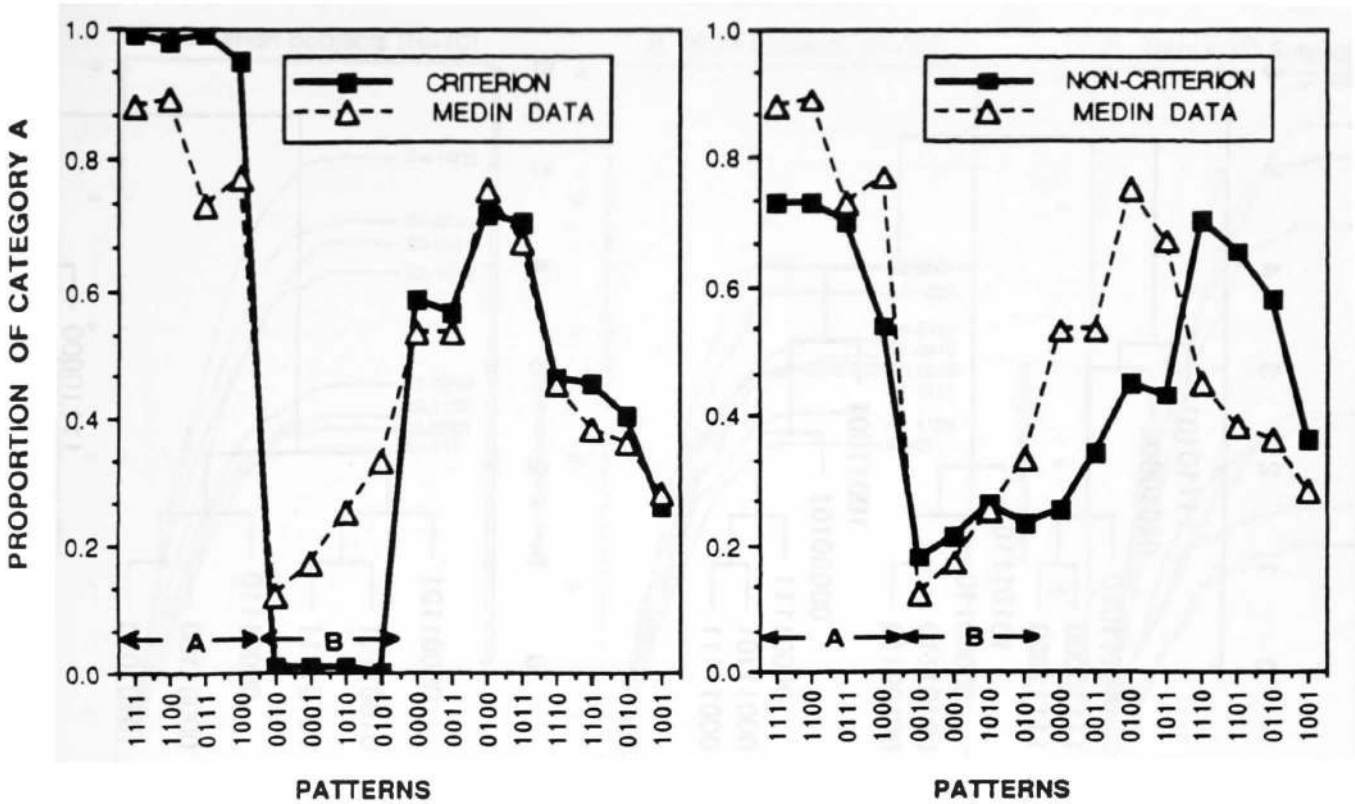
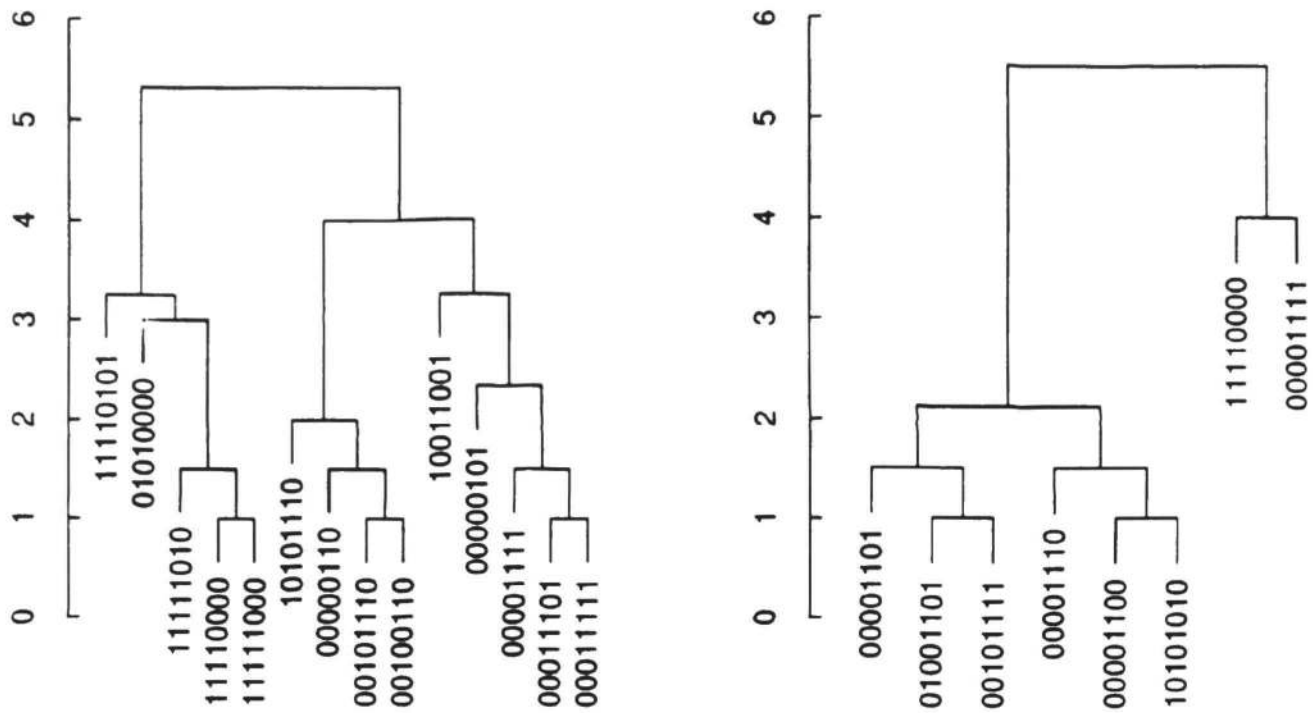


Figure 1. Generalization profiles.

Responses to the novel patterns represent the transfer of learning or generalization performed by subjects. There are several ways of interpreting the proportions of the *A* responses. According to one interpretation, the average responses arise from an ensemble of identically distributed subjects. That is, the probability of assigning a given pattern to category A is the same for each subject and is approximated by the graph on Figure 1. This interpretation is commonly assumed by investigators (e.g. Medin et al., 1982) who used such data to test exemplar-based models of categorization. An alternative way of interpreting these proportions, however, is in terms of a mixture of distributions corresponding to subjects who learned different rules during the training phase. We examined individual differences in order to distinguish these two interpretations.

The extreme version of the mixture hypothesis is that each subject learned a different set of rules. That model is unlikely because although there are 256 different possible generalizations for the eight test patterns, 14 different generalizations accounted for 85% of the subjects. In particular, generalizations of 38% of the subjects who reached criterion were consistent with the hypothesis that subjects based their categorization on the exclusive-or (XOR) of dimensions 3 and 4 (the graph of XOR performance, if plotted on Figure 1, would consist of alternation of four high and four low responses). On the other hand more than a half of the subjects who learned the task generalized differently. This supports the notion that individual subjects abstracted different set of rules during the training phase. The data in Figure 1 appear to represent a mixture of strategies and generalizations.

While subjects produced many different generalizations, it is possible that these generalizations are very similar to each other. The following analysis was performed to determine similarity among



A. Subjects' Generalizations **B. Network Generalizations**

Figure 2. Hierarchical clustering of generalizations

different generalizations. The generalization performed by each subject can be represented as an eight dimensional *generalization profile* vector, where each "1" bit corresponds to assignment of the corresponding pattern to category A. [1] Hence, similar generalizations would have similar profiles. To analyze the generalization profiles we computed Hamming distances between all pairs of the 14 most frequent generalizations and then used hierarchical clustering, based on average inter-cluster distances, to represent the similarities among generalizations. The resulting hierarchy is shown in Figure 2A. The distance between any two profiles, shown as terminal nodes of the tree, corresponds to the lowest common node on the tree. This analysis indicates that different subjects generalized in many, quite different ways.

In order to understand human categorization process it is important to determine how different subjects arrive at different rules. While a complete answer to this question is beyond the scope of this paper one can get some indications of the underlying processes by examining subjects' average performance during the training phase. Subjects' performance on each pattern during the learning phase was summarized by computing the average cumulative error for each training pattern. Three sets of such cumulative error learning curves are shown in Figure 3A for all subjects who reached criterion, for those that performed XOR (Figure 3B) and for the remainder (Figure 3C). The most important aspect of the cumulative error curves is that, more or less consistently over subjects, each pattern is learned with different difficulty. For example, the pattern 1111 from category A was very easy (few errors) while the pattern 1000 from the same category was very difficult. This regularity which was was true for for all

[1] The order of bits corresponds to the ordering of novel stimuli in Table 1,(N1,N2...N8).

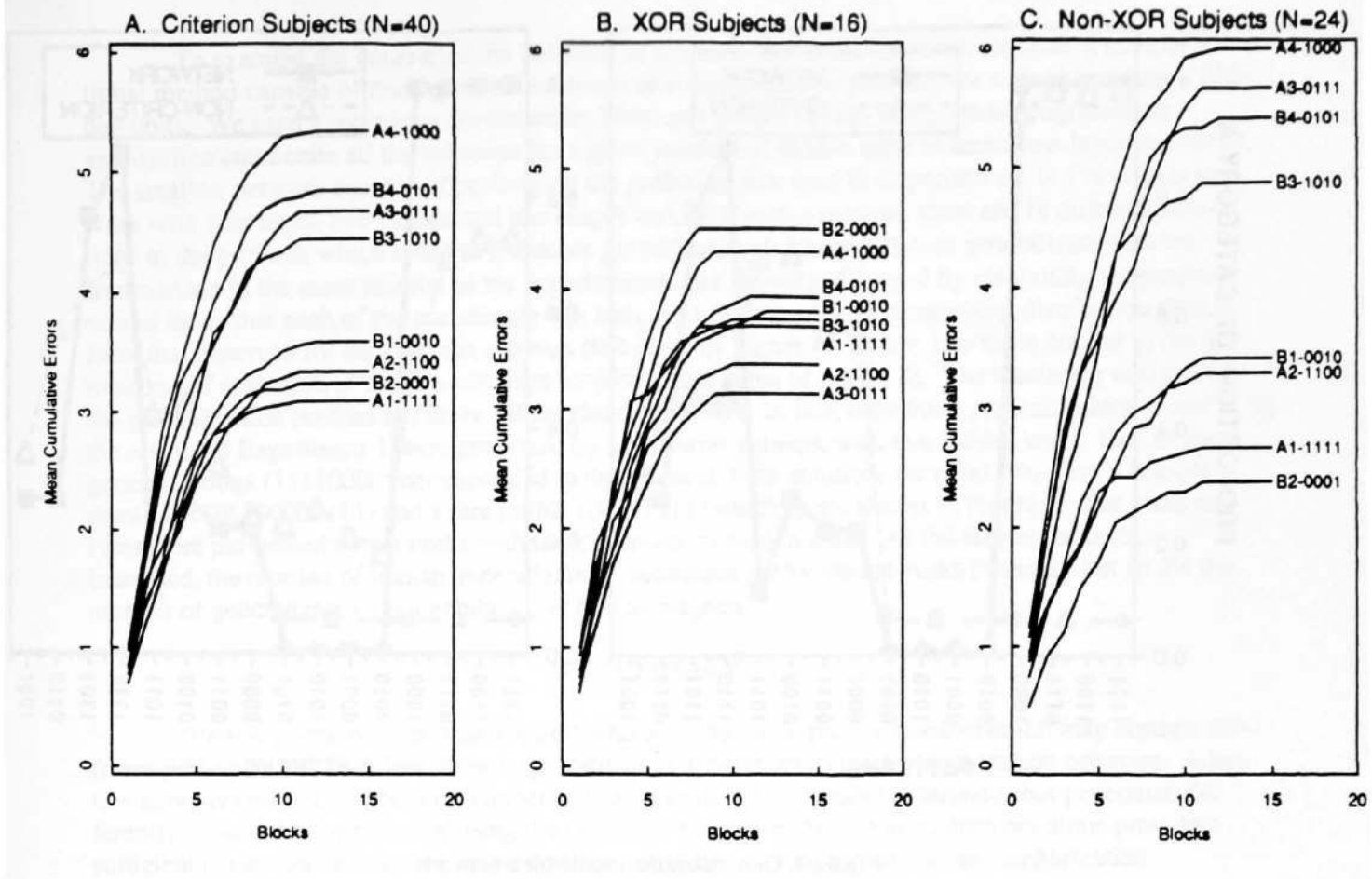


Figure 3. Cumulative error learning curves

subjects who reached the criterion performance is indicative of the type of rules abstracted by subjects. In particular, even those subjects who eventually used the XOR categorization were initially using the first two dimensions.

MODELING GENERALIZATIONS

The empirically observed subjects' generalizations can provide information about the constraints used by human beings. To discover these constraints frequently requires a model-based analysis of the data. Models of categorization can be used in two ways. One approach is based on those models that can represent any generalization and do not impose any prior constraints. Their utility is in remapping the data so that the constraints are easily observed and extracted.

Another way to discover the constraints imposed by the learner is to construct a model of a pattern categorization process that embodies some of those constraints. Such a model can then be used to predict the generalizations and its predictions can be compared to the data.

Generalizations by Networks

An interesting class of models to consider for categorization are multi-layered adaptive networks. Layered networks are acyclic (nonrecursive) directed graphs with defined starting (input) and terminating (output) nodes (units) in which each unit has a uniquely defined distance (in terms of arcs) from all the input units. Hidden units are those nodes that are labeled neither input nor output.

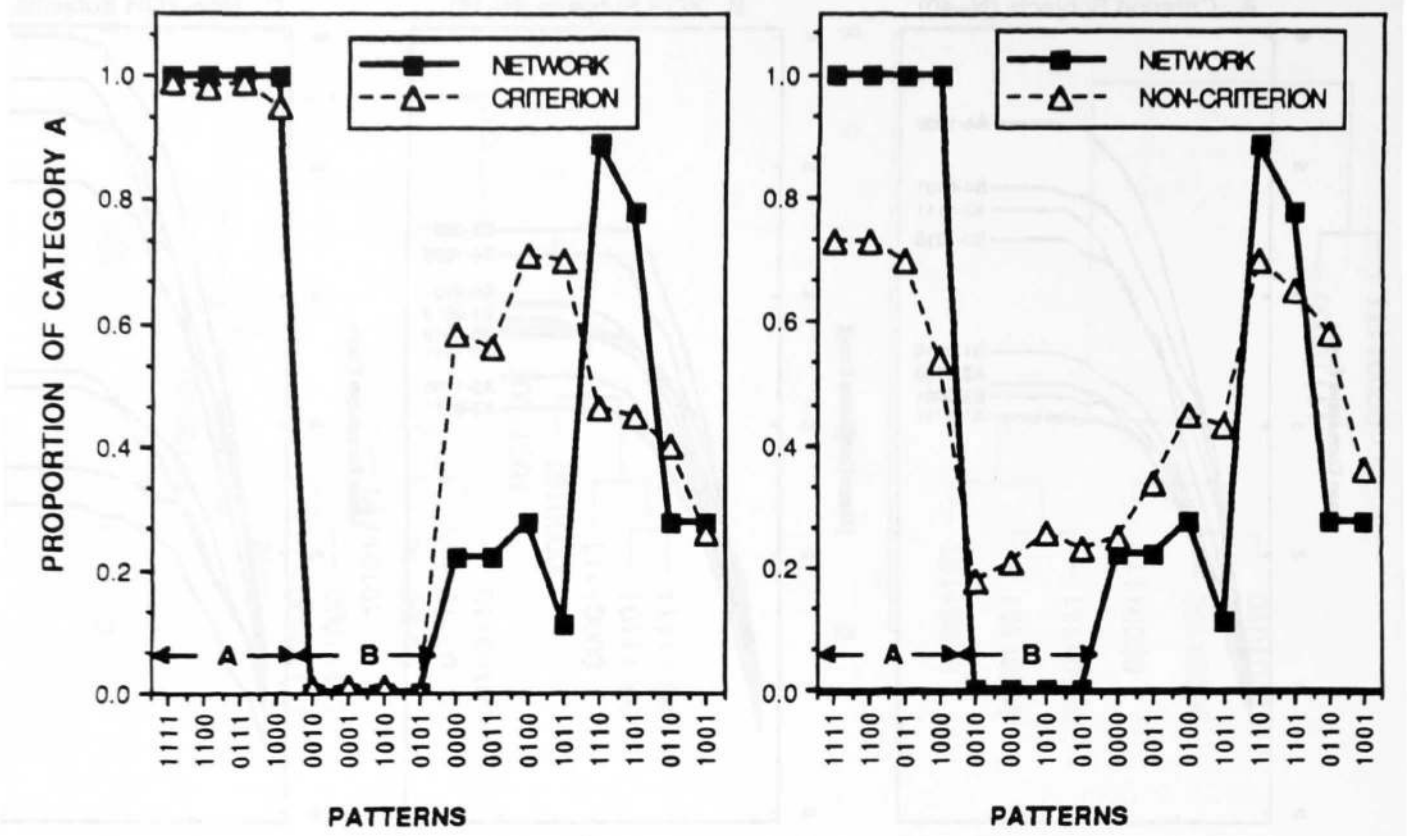


Figure 4. Generalization profile for a network.

Each directed arc is labeled by a real valued weight. A unit may, in general, be a dynamical system but in the current framework a unit is defined by a threshold function of the sum of incoming arcs; the value contributed by each arc is equal to the value of the originating unit multiplied by the weight of the arc. Each unit performs a linear threshold function which is the essential nonlinearity required for a pattern recognition mechanism.

A two-layer adaptive network consisting of an input, hidden and output layer with unlimited number of hidden units can represent any computable boolean function (Nilsson, 1965; Minsky and Papert, 1969). Therefore, such networks can be used to analyze the data by finding a set of weights that performs the same categorization as an individual subject and then examine the structure of such a network.

Because an unconstrained network can make any possible generalization, additional constraints must be imposed if an adaptive network is to predict human generalization performance. An important question to ask is whether or not a network with a specific set of constraints can predict a particular generalization. A complete theory would have to include a characterization of the effects of different constraints on generalization. Although such an analysis is beyond the scope of this paper we illustrate the approach using a particular constraint.

An example of one such constraint involves imposing a limit on the number of hidden units. In the extreme, most constraining case, this amounts to finding an adaptive network with the minimum number of hidden units that can perform the categorization on the training set. The motivation for such an approach is in the usual heuristic arguments for simplicity; a smaller network should generalize better.

To examine the generalization behavior of minimal networks, however, requires a computational method capable of finding all the solutions to a categorization problem for a given number of hidden units. We used a technique developed by Pavel and Moore (1988) using linear programming approach to enumerate all the solutions for a given number of hidden units in small two-layer networks. The smallest network capable of performing the particular task used in Experiment 1 is a two-layer network with four input, two hidden and one output unit. For such a network there are 18 different solutions to the problem which result in 8 distinct generalizations. These different generalizations were summarized in the same manner as the experimental data shown in Figure 4 by computing the proportion of times that each of the test stimuli was assigned to category A. The resulting distribution differs from that observed for the criterion subjects (left panel of Figure 4). In fact, it is more similar to the distribution of responses of the non-criterion subjects (right panel of Figure 4). The clustering analysis of the generalization profiles in Figure 3B are clearly different. In fact, only three generalizations found in the results of Experiment 1 were generated by a two-layer network with two hidden units. One of these generalizations (11110000) corresponded to the frequent XOR solution; the other two were a complement of XOR (00001111) and a rare profile (00101111) which is not shown in Figure 2. The same analyses were performed on networks with larger number of hidden units. As the number of units increased, the number of human generalizations accounted for by the networks increased but so did the number of generalizations not exhibited by human subjects.

SUMMARY

We have demonstrated that subjects who learn the same pattern categorization may abstract different principles and therefore show large individual differences in their generalization behavior. Adaptive networks with the minimum number of hidden units exhibit a similar behavior but generalize differently. Thus, the constraint of using the minimum number of hidden units does not alone provide a sufficient constraint on adaptive network models to allow them to model human categorization processes. Currently we are investigating the effects of other constraints.

REFERENCES

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 8*, 37-50.

Medin, D. L., Dewey, G. I., & Murphy, T. D. (1983). Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*, 607-625.

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

Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry.* Cambridge, MA: MIT Press.

Nilsson, N. J. (1965). *Learning machines.* New York: McGraw-Hill.

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

Pavel, M., & Moore, R. T. (1988). *Computational analysis of solutions of two-layer adaptive networks.* APL Technical Report, Dept. of Psychology, Stanford University.