A Rational Analysis of Categorization

John R. Anderson
Department of Psychology
Carnegie-Mellon University
Pittsburgh, PA 15213

Michael Matessa
Department of Psychology
Carnegie-Mellon University
Pittsburgh, PA 15213

Abstract

A rational analysis tries to predict the behavior of a cognitive system from the assumption it is optimized to the environment. An iterative categorization algorithm has been developed which attempts to get optimal Bayesian estimates of the probabilities that objects will display various features. A prior probability is estimated that an object comes from a category and combined with conditional probabilities of displaying features if the object comes from the category. Separate Bayesian treatments are offered for the cases of discrete and continuous dimensions. The resulting algorithm is efficient, works well in the case of large data bases, and replicates the full range of empirical literature in human categorization.

A rational analysis (Anderson, 1990) is an attempt to specify a theory of some cognitive domain by specifying the goal of the domain, the statistical structure of the environment in which that goal is being achieved, and whatever computational constraints the system is operating under. The predictions about the behavior of the system can be derived assuming that the system will maximize the goals it expects to achieve while minimizing expected costs where expectation is defined with respect to the statistical structure of the environment. This approach is different from most approaches in cognitive psychology because it tries to derive a theory from assumptions about the structure of the environment rather than assumptions about the structure of the mind.

We have applied this approach to human categorization and have developed a rather effective algorithm for categorization. The analysis assumes that the goal of categorization is to maximize the accuracy of predictions about features of new objects. For instance, one might want to predict whether an object is dangerous or not. This approach to categorization sees nothing special about category labels. The fact an object might be called a tiger is just another feature one might want to predict about the object.

The Structure of the Environment

It is an interesting question what kind of structure we can assume of the environment in order to drive prediction. The theory developed rested on the structure of biological categories produced by the phenomenon of species. Species form a nearly disjoint partitioning of the natural objects because of the inability to interbreed. Within a species there is a common genetic pool which means that individual members of the species will display particular feature values with probabilities that reflect the proportion of that phenotype in the population. Another useful feature of species structure is that the display of features within a freely-interbreeding species is largely independent. Thus, there is little relationship between size and eye color in species where those two dimensions vary. Thus, the critical aspects of speciation is the disjoint partitioning of the object set and the independent probabilistic display of features within a species.

An interesting question is whether other types of objects display these same properties. Another common type of object is the artifact. Artifacts approximate a disjoint partitioning but there are occasional exceptions--for instance, mobile homes which are both homes and vehicles. Other types of objects (stones, geological formations, heavenly bodies, etc) seem to approximate a disjoint partitioning but here it is hard to know whether this is just a matter of our perceptions or whether there is any objective sense in which they do. One can use the understanding of speciation for natural kinds and understanding of the intended function in manufacture for artifacts to objectively assess the hypothesis of a disjoint partitioning.

We have taken this disjoint, probabilistic model of categories and used it as the understanding of the structure of the environment for doing prediction about object features. To maximize the prediction of features of objects we need to induce a disjoint partitioning of the object set into categories and determine what the probability of features will be for each category. The ideal prediction function would be described by the following formula:

$$Pred_{ij} = \sum_{x} P(x|F_n) Prob_i(j|x)$$

where $Pred_{ij}$ is the probability an object will display a value j on a dimension i which is not observed for that object, the summation is across all possible partitionings of the n objects seen into disjoint sets, $P(x|F_n)$ is the probability of partitioning x given the objects display observed feature structure F_n, and $Prob_i$ (j|x) is the probability the object in question would display value j on dimension i if x were the partition. The problem with this approach is that the number of partitions of n objects grows exponentially as the Bell exponential number (Berge, 1971). Assuming that humans cannot consider an exponentially exploding number of hypothesis we were motivated to explore iterative algorithms such as those developed by Fisher (1987) and Lebowitz (1987).

The following is a formal specification of the iterative algorithm:

1. Before seeing any objects, the category partitioning of the objects is initialized to be the empty set of no categories.
2. Given a partitioning for the first m objects, calculate for each category k the probability P_k that the $m+1st$ object comes from category k. Let P_o be the probability that the object comes from a completely new category.
3. Create a partitioning of the $m+1$ objects with the $m+1st$ object assigned to the category with maximum probability.
4. To estimate the probability of value j on dimension i for the $n+1st$ object calculate

$$Pred_{ij} = \sum_k P_k \; P(ij|k) \qquad \text{Equation 1}$$

where P_k is the probability the $n+1st$ object comes from category k and $P(ij/k)$ is the probability of displaying value j on dimension i.

The basic algorithm is one in which the category structure is grown by assigning each incoming object to the category it is most likely to come from. Thus, a specific partitioning of the objects is produced. Note, however, that the prediction for the new $n+1st$ object is <u>not</u> calculated by determining its most likely category and the probability of j given that category. This calculation is performed over all categories. This gives a much more accurate approximation to the ideal $Pred_{ij}$ because it handles situations where the new object is ambiguous between multiple categories. It will weight approximately equally these competing categories.

The algorithm is not guaranteed to produce the maximally probable partitioning of the object set since it only considers partitionings that can be incrementally grown. It also does not weight multiple possible partitionings as the ideal algorithm would. In cases of strong category structure, there will be only one probable partitioning and the iterative algorithm will uncover it. In cases of weak category structure, it will often fail to obtain the ideal partitioning, but still the predictions obtained by Equation 1 closely approximate the ideal quantity because of the weighting of multiple categories. We observe correlations about .95 between the predictions of our algorithm and the ideal quantities in cases of small data sets.

It remains to come up with a formula for calculating P_k and $P(ij/k)$. Since $P(ij/k)$ proves to be involved in the definition of P_k, we will focus on P_k. In Bayesian terminology P_k is a posterior probability $P(k/F)$ that the object belongs to category k given that it has feature structure F. Bayes formula can be used to express this in terms of a prior probability $P(k)$ of coming from category k before the feature structure is inspected and a conditional probability $P(F/k)$ of displaying the feature structure F given that it comes from category k.

$$P_k = P(k|F) = \frac{P(k)P(F|k)}{\sum_k P(k)P(F|k)} \qquad \text{Equation 2}$$

where the summation in the denominator is over all categories k currently in the partitioning including the potential new one. This then focuses our analysis on the derivation of a prior probability $P(k)$ and a conditional probability $P(F/k)$.

Prior Probability

With respect to prior probabilities the critical assumption is that there is a fixed probability c that any two objects come from the same category and this probability does not depend on the number of objects seen so far. This is called the coupling probability. If one takes this assumption about the coupling probability between two objects being independent of the other objects and generalizes it, one can derive a simple form for $P(k)$ (See Anderson, 1990, for the derivation):

$$P(k) = \frac{cn_k}{(1-c) + cn} \qquad \text{Equation 3}$$

where c is the coupling probability, n_k is the number of objects assigned to category k so far, and n is the total number of objects seen so far. Note for large n this closely approximates n_k/n which means that we have a strong base rate effect in these calculations with a bias to put new objects into large categories. Presumably the rational basis for this is apparent.

We also need a formula for $P(0)$ which is the probability that the new object comes from an

entirely new category. This is

$$P(O) = \frac{(1-c)}{(1-c) + cn} \qquad \text{Equation 4}$$

For large n this closely approximates $(1-c)/cn$ which is again a reasonable form--i.e., the probability of a brand new category depends on the coupling probability and number of objects seen. The greater the coupling probability and the more objects, the less likely it is that the new object comes from an entirely new category.

Conditional Probability

We can consider the probability of displaying features on various dimensions given category membership to be independent of the probabilities on other dimensions. Then we can write

$$P(F|k) = \prod_i P(ij|k) \qquad \text{Equation 5}$$

Where P(ij|k) is the probability of displaying value j on dimension i given that one comes from category k.

This independence assumption does not prevent us from recognizing categories with correlated features. Thus, we may know that being black and retrieving sticks are features found together in labradors. This would be represented by high probabilities of the stick-retrieving and the black features in the labrador category. What the independence assumption prevents us from doing is representing categories where values on two dimensions are either both one way or both the opposite. Thus, it would prevent us from recognizing a single category of animals which were either large and fierce or small and gentle, for instance. However, this turns out not to be a very serious limitation. What our algorithm does in this case is to spawn a different category to capture each two-feature combination--it would create a category of large and fierce creatures and another category of small and gentle creatures.

The effect of Equation (5) is to focus us down on an analysis of the individual $P(ij|k)$. Derivation of this quantity is itself an exercise in Bayesian analysis. We will treat separately discrete and continuous dimensions.

Discrete Dimensions

The basic Bayesian strategy for doing inference along a dimension is to assume a prior distribution of values along the dimension, determine the conditional probability of the data under various possible values of the priors, and then calculate a posterior distribution of possible values. The common practice is to start with a rather weak distribution of possible priors and as more and more data accumulates come up with a tighter and tighter posterior distribution.

In the case of a discrete dimension, the typical Bayesian analysis (Berger, 1985) is to assume that the prior distribution is a Dirichlet density. For a dimension with m values a Dirichlet distribution is characterized by m parameters α_j. We can define $\alpha_o = \sum_j \alpha_j$. The mean probability of the jth value is $p_j = \alpha_j/\alpha_o$. The value α_o reflects the strength of belief in these priors probabilities, p_j. The data after n observations will consist of a set of C_j counts of observations of value j on dimension i. The posterior distribution of probabilities is also a Dirichlet distribution but with parameters α_j+C_j. This implies that the mean expected value of displaying value j in dimension i is $(\alpha_j+C_j)/\sum(\alpha_j+C_j)$. This is P(ij|k) for Equation 5:

$$P(ij|k) = \frac{C_j + \alpha_i}{n_k + \alpha_0} \qquad \text{Equation 6}$$

where n_k is the number of objects in category k which have a value on dimension i and C_j is the number of objects in category k with the same value as the object to be classified. For large n_k this approximates C_j/n_k which one frequently sees promoted as the rational probability. However, it has to have this more complicated form to deal with problems of small samples. For instance, if one has just seen one object in a category and it has had the color red, one would not want to guess that all objects are red. If we assume there are seven colors and all the α_j were 1, the above formula would give 1/4 as the posterior probability of red and 1/8 for the other six colors unseen as yet.

Continuous Dimensions

Application of Bayesian inference schemes to continuous dimensions is more problematic but there is one approach that appears most tractable (Lee, 1989). The natural assumption is that the variable is distributed normally and the induction problem is to infer the mean and variance of that distribution. In standard Bayesian inference methodology we must begin with some prior assumptions about what the mean and variance of this distribution is. It is unreasonable to suppose we can know in advance what the precisely what either the mean and variance will be. Our prior knowledge must take the form of probability densities over possible means and variances. This is basically the same idea as in the discrete case where we had a Dirichlet distribution giving priors about probabilities of various values. The major complication is the need to state separately prior distributions for mean and variance.

The tractable suggestion for the prior distributions is that the inverse of the variance Σ^2 is distributed according to a chi-square distribution and the mean has a normal distribution. Given these priors, the posterior distribution of values x on a continuous dimension *i* for category k, after n observations has the following t distribution:

$$f_i(x|k) \sim t_{a_i}(\mu_i, \sigma_i \sqrt{1 + 1/\lambda_i}) \qquad \text{Equation 7}$$

The parameters a_i, μ_i, σ_i, and λ_i are defined as follows:

$$\lambda_i = \lambda_0 + n \qquad \text{Equation 8}$$

$$a_i = a_0 + n \qquad \text{Equation 9}$$

$$\mu_i = \frac{\lambda_0 \mu_0 + n\bar{x}}{\lambda_0 + n} \qquad \text{Equation 10}$$

$$\sigma_1^2 = \frac{a_0\sigma_0 + (n-1)s^2 + \frac{\lambda_0 n}{\lambda_0 + n}(\bar{x} - \mu_0)^2}{a_0 + n} \qquad \text{Equation 11}$$

where $\bar{x}$ is the mean of the n observations and s^2 is their variance. These equations basically provide us with a formula for merging the prior mean and variance, μ_0 and σ_0^2, with the empirical mean and variance, $\bar{x}$ and s^2, in a manner that is weighted by our confidences in these priors, λ_0 and a_0.

Equation 7 for the continuous case describes a probability density which serves the same role as Equation 6 for the discrete case which describes a probability. The product of conditional probabilities in Equation 5 will then be a product of probabilities and density values. Basically, Equations (5), (6), and (7) give us a basis for judging how similar an object is to the category's central tendency.

Conclusion

This completes our specification of the theory of categorization. Before looking at its application to various empirical phenomena a word of caution is in order. The claim is not that the human mind performs any of the Bayesian mathematics that fills the preceding pages. Rather the claim of the rational analysis is that, whatever the mind does, its output must be optimal. The mathematical analyses of the preceding pages serve the function of allowing us, as theorists, to determine what is optimal.

A second comment is in order concerning the output of the rational analysis. It delivers a probability that an object will display a particular feature. There remains the issue of how this relates to behavior. Our basic assumption will only be that there is a monotonic relationship between these probabilities and behavioral measures such as response probability, response latency, and confidence of response. The exact mapping will depend on such things as the subject's utilities for various possible outcomes, the degree to which individual subjects share the same priors and experiences, and the computational costs of achieving various possible mappings from rational probability to behavior. These are all issues for future exploration. What is remarkable is how well we can fit the data simply assuming a monotonic relationship.

Application of the Algorithm

We have applied the algorithm to a number of examples to illustrate its properties. The algorithm is quite efficient. A Franz LISP implementation categorized the 290 items from Michalski and Chilausky's data set on Soybean disease (each with 36 values) in 1 CPU minute on a Vax 780 or a MAC II. This is without any special effort to optimize the code. It also diagnosed the test set of 340 soybean instances with as much accuracy as apparently did the original system of Michalski and Chilausky.

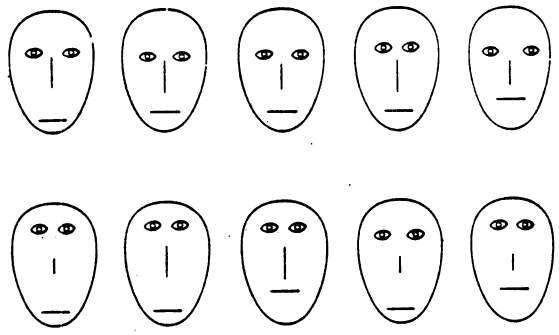
The algorithm has been applied to the full range of psychological experiments in categorization. Detailed discussions can be found in Anderson (in press) and Anderson & Matessa (in preparation). However, we will review here in varying detail the applications of the algorithm to 10 empirical phenomena. All these simulations were done with a constant setting of the parameters: c from Equation 3 and 4 at .3, α_i from Equation 6 at 1, λ_0 from Equation 8 at 1, a_0 from Equation 9 at 1, μ_0 from Equation 10 at the mean of the stimuli, and σ_0^2 from Equation 11 at the square of 1/4 the stimulus range. All of these are plausible settings and often correspond to conventions for setting Bayesian non-informative priors. The following are among the empirical phenomena we have successfully simulated:

1. Extraction of Central Tendencies, Continuous Dimensions The Bayesian model for continuous dimensions implies that categorization should vary with distance from central tendency. This enables the model to simulate the data of Posner & Keele (1968) on categorization of dot patterns and Reed (1972) on categorization of faces. Let us consider the experiment of Reed in a little detail:

Reed (1972) had subjects learn to categorize the 10

faces which are illustrated in Figure 1. The first row of faces are in one category and the second row of faces are in another category. The two sets of faces are deviations from underlying prototypes. After studying these faces subjects went to a test condition where they had to try to classify these and other faces. The critical data concerns the probability with which subjects assigned faces to conditions. As a general characterization, their categorization varied with distance of the face from the prototype.

Figure 1

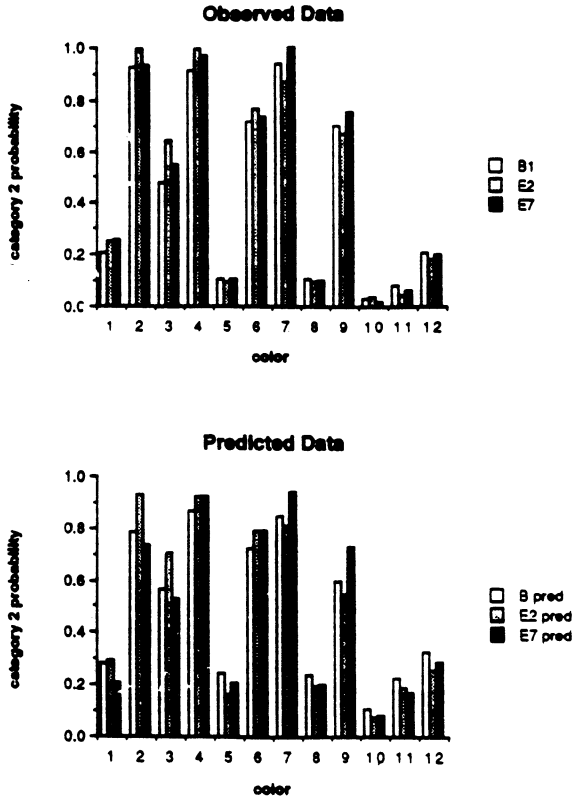


In our attempt to simulate these data we treated these faces as five-dimensional stimuli where the dimensions are height of the forehead which ranged from 54 to 88 mm, distance separation of the eyes which ranged from 20 to 55 mm, length of the nose which ranged from 32 to 64 mm, height of the mouth which ranged from 28 to 60 mm, and category label which was a binary-valued discrete dimension. Our rational model identified two or more internal categories, depending on presentation order, that corresponded to the experimenter's categories. That is, sometimes it subdivided the experimenter's categories into subcategories but it almost never merged items from the two experimenter categories into an internal category. Reed's subjects were asked to classify 25 test stimuli and the major test of our model was its classification of these test stimuli. Overall its confidence of category membership (calculated by Equation 1) correlated .90 with Reed's data.[1]

2. Extraction of Central Tendencies, Discrete Dimensions The model implies that stimuli should be better categorized if they display the majority value for a dimension. This enabled the model to simulate the data of Hayes-Roth & Hayes-Roth (1977), for instance.

3. Effect of Individual Instances If an instance is sufficiently different than the central tendency for its assigned category, the model will form a distinct category for it. This enables the model to account for the data of Medin & Schaffer (1978) on discrete dimensions and Nosofsky (1988) on continuous dimensions. Let us consider the experiment of Nosofsky:

Figure 2



Nosofsky trained his subjects on 12 stimuli that varied in brightness and saturation. The colors varied in brightness on the Munsell scale from 3 to 7 and in saturation from 4 to 12. In the base condition subjects had four trials on each item and were then tested. In the first experiment there was a condition E2 in which subjects saw stimulus 2 approximately 5 times as frequently and a condition E7 in which they saw stimulus 7 approximately 5 times as frequently. Part (a) of Figure 2 illustrated probability of classification in Category 2. As can be seen subjects are sensitive to the frequency manipulation. Part (b) of Figure 2 shows the probability our model assigned to a Category 2 response given the same experience. The overall correlation between data and theory is .98.

4. Linearly Separable versus Non-Linearly Separable Categories Unlike some categorization models this model is able to learn categories that cannot be separated by a plane in a n-dimensional

[1]We would like to thank Stephen Reed for making his data available.

hyperspace. This is because it can form multiple internal categories to correspond to an experimenter's category. This enables the model to account for the data of Medin & Schwanenflugel (1981) on discrete dimensions and Nosofsky, Clark, & Shin (1989) on continuous dimensions. Let us consider the experiment of Medin & Schwanenflugel. They performed an experiment where linearly non-separable categories were learned better than linearly separable categories.

Table 1 illustrates the material used by Medin & Schwanenflugel (1981). In the case of the linearly separable categories our model formed separate categories for each stimulus. In the case of linearly non-separable, it merged the first 2 in category A into an internal category, the second 2 in category A, and the first, second, and fourth in category B. Thus, only stimulus 3 in category B was in a singleton category and this was the stimulus that produced the highest error rate in the non-separable condition.

Table 1

LINEARLY SEPARABLE CATEGORIES

CATEGORY A					CATEGORY B				
	DIMENSION					DIMENSION			
EXEMPLAR	D_1	D_2	D_3	D_4	EXEMPLAR	D_1	D_2	D_3	D_4
A_1	1	1	1	0	B_1	1	0	1	0
A_2	1	0	1	1	B_2	0	1	1	0
A_3	1	1	0	1	B_3	0	0	0	1
A_4	0	1	1	1	B_4	1	1	0	0

CATEGORIES NOT LINEARLY SEPARABLE

CATEGORY A					CATEGORY B				
	DIMENSION					DIMENSION			
EXEMPLAR	D_1	D_2	D_3	D_4	EXEMPLAR	D_1	D_2	D_3	D_4
A_1	1	0	0	0	B_1	0	0	0	1
A_2	1	0	1	0	B_2	0	1	0	0
A_3	1	1	1	1	B_3	1	0	1	1
A_4	0	1	1	1	B_4	0	0	0	0

5. Basic-Level Categories The internal categories that the model extracts corresponds to what Rosch (1976) meant by basic-level categories.[2] Thus, it can simulate the data of Murphy & Smith (1982) and Hoffman & Ziessler (1983). We will describe the application to Murphy and Smith.

Murphy and Smith presented to their subjects 16 objects identified as examples of fictitious tools. The structure of the material, as encoded by Gluck and Corter (1985), is illustrated in Table 2. There were two superordinate categories which divided into 4 intermediate categories, which divided into 8 subordinate categories. Table 2 gives the attribute description of each category. Subjects were fastest to classify the material at the intermediate level which Murphy and Smith intended to be the basic level. Objects at this level had two attributes plus two labels in common. Only one additional feature and label was gained at the subordinate level, and all features were lost at the superordinate level except for their feature of being a pounder or a cutter.

Table 2

Gluck and Corter's Analysis of the Feature Structure of the Material from Murphy & Smith (1982)

	Categories			*Attributes*			
Item #	*Super-ordinate*	*Inter-mediate*	*Sub-ordinate*	*Handle*	*Shaft*	*Head*	*Size*
1.	Pounder	Hammer	Hammer1	2	2	0	0
2.				2	2	0	1
3.			Hammer 2	2	2	1	0
4.				2	2	1	0
5.		Brick	Brick1	0	3	4	0
6.				0	3	4	1
7.			Brick 2	1	3	4	0
8.				1	3	4	1
9.	Cutter	Knife	Knife1	3	4	2	0
10.				3	4	2	1
11.			Knife2	3	4	3	0
12.				3	4	3	1
13.		Pizza C.	P.C.1	4	0	5	0
14.				4	0	5	1
15.			P.C.2	4	1	5	0
16.				4	1	5	1

We modeled this material by encoding the stimuli as 7-dimensional objects with dimensions for the superordinate label (2 values), the intermediate label (4 values), the subordinate label (8 values), handle (5 values), shaft (5 values), head (6 values), and size (2 values). What category structure was obtained depended upon the value of the coupling probability. For $c > .96$ all were merged into one category; for $.95 > c > .8$ the two superordinate categories emerged; for $.8 > c > .4$ the model fluctuated between the superordinate and intermediate categories depending on presentation order; for $.4 > c > .2$ it extracted just the intermediate categories; for $.2 > c > .05$ it basically extracted the intermediate categories with an occasional singleton category or subordinate category; for $c < .05$ it extracted only singleton categories. In summary, the subordinate categories never emerged and only a very high levels of c did superordinate categories dominate. At the value of c used in the simulations of this paper ($c = .3$) only the basic level categories emerged. Thus, it seems fair to conclude that the analysis agrees with the subjects as to what the basic level is.

6. Probability Matching Faced with truly probabilistic categories and large samples of instances the model will estimate probability of

[2]Rosch's idea of a basic level is that there is a level in the generalization hierarchy to which we first assign objects. For instance, she argues we would first see an object as a bird not a sparrow or an animal.

features that correspond exactly to the empirical proportion. Thus, it predicts the data of Gluck & Bower (1988) on probability matching.

7. Base-Rate Effect Because of Equation 3 this model predicts that usually there will be a greater tendency to assign items to categories of large size. Thus, it handles the data of Homa & Cultice (1984). It also reproduces the more subtle interactions of Medin & Edelson (1988).

8. Correlated Features As noted earlier the model can handle categories with correlated features by breaking out separate internal categories for each feature combination. Thus, it handles the data of Medin, Altom, Edelson, & Freko (1982). They had subjects study the 9 cases in Table 3 which were all supposed to represent instances from one disease category, burlosis. This was simulated by presenting these 9 cases to the model with a sixth dimension, a disease label which was always burlosis. This was arbitrarily treated this as a binary dimension. Note that each of the five symptoms show a majority of ones associated with the disease.

Table 3

SYMPTOMS OF BURLOSIS
from Medin et al. (1982)

Case Study	*Blood Pressure*	*Skin Condition*	*Muscle Condition*	*Condition of Eyes*	*Weight Condition*
1. R.L.	0	1	0	1	1
2. L.F.	1	1	0	1	1
3. J.J.	0	0	1	1	1
4. R.M.	1	0	1	1	1
5. A.M.	1	1	1	1	1
6. J.S.	1	1	1	1	1
7. S.T.	1	0	0	0	0
8. S.E.	0	1	1	0	0
9. E.M.	1	1	1	0	0

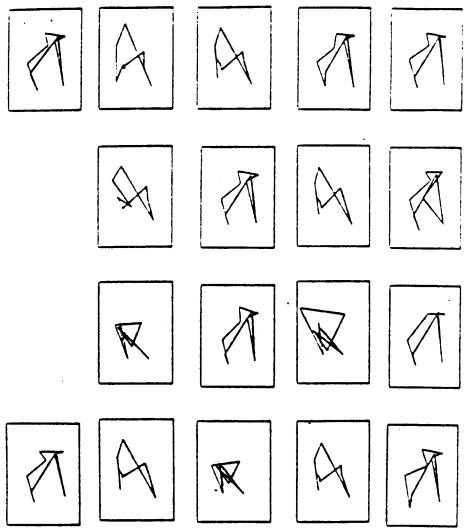
Note: Zero denotes absence of the symptom and 1 denotes presence.

The critical feature of these materials from the perspective of correlated features concerns the fourth dimension of conditions of eyes and the fifth dimension of weight. Values are either both 1 or both 0. The first six items in Table 3.5 have two 1's; the last three have two 0's. Subjects are sensitive to this correlation. When these stimuli were fed into the algorithm with $c=.3$, it typically extracted 3 categories--one to represent the first six items, one for the seventh, and one for the last two. Thus, the way it dealt with correlated features was to break out separate categories for the different possible values of the correlation.

9. Effects of Feedback If the category structure of the stimuli is strong enough the model can extract the categories without any feedback as to category identity. In the face of weak category structure, it is necessary to provide category labels to get learning. Thus, this model reproduces the data of Homa & Cultice (1984).

Figure 3 illustrates the stimulus material of Homa and Cultice. They are derived from the random 9-dot patterns introduced by Posner & Keele (1968) but Homa has introduced the feature of drawing lines to connect the dots. This makes it relatively cheap to write a computer program that will determine how to map the points of one into another in a way as to achieve maximal fit. Given such a mapping, we can describe each stimulus according to 18 ordered dimensions which are the x and y coordinates of each point. Then we can apply our categorization algorithm to these materials.

Figure 3

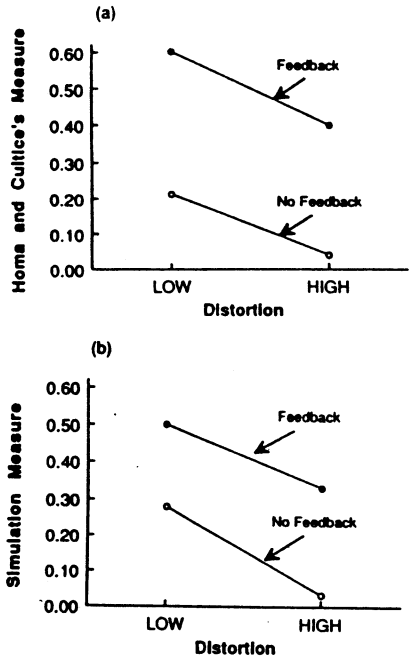


There are three categories in Figure 3--one category represented by 9 items, one by 6, and one by 3. In one condition of their experiment, subjects were given category labels and trained to sort the stimuli into three categories. in another condition they were free to sort the stimuli into whatever categories they wanted. Homa & Cultice were interested in determining how well subjects did at recovering the category structure without feedback. In the case of feedback, Homa & Cultice just measured accuracy of assignment in a final criteria test. In the case of no feedback, they tried to discover some way of assigning labels to the categories in the subjects' sort that made their categorization look optimal. It is hard to know how comparable the two measures are.

In our case, when there was feedback, we measured the probability of a category label according to Equation 1. When there was no feedback, we assigned labels to internal categories in such a way as to maximize probability of a correct label assignment when Equation 1 was used. Again

it is unclear how comparable our two measures were. In our case we corrected our measures for guessing. We ran a control condition where, rather than letting the algorithm decide which items go together, we randomly assigned items to internal categories and then proceeded to get performance scores in the same way as when the algorithm did the assignment. Thus, we got two measures--P, a mean probability of the correct category label when our algorithm did the clustering and G, a mean probability of category labeling in the control condition when we did the clustering randomly. Our final measure was (P - G) / (1 - G) which is a standard correction-for-guessing formula.

Figure 4



Homa and Cultice used a number of different training sets including a low distortion training set where the points were perturbated 1.1 units (the examples in Figure 3 are 1.1 distortions) and a high distortion set where they were perturbated 4.8 units. Figure 4 compares the performance of the subjects and the simulation for high and low distortion training stimuli in the presence of label feedback or not. In the case of Homa & Cultice, we used a correction for guessing measure to but set the guessing rate to be .33 since there were 3 categories. Both subjects and simulations show approximately additive effects of the two dimensions. Both the subjects and the simulation are nearly at chance in the presence of high distortion stimuli with no label feedback. However, our model does show greater sensitivity to feedback.

10. Effects of Input Order In the presence of weak-category structure, the categories the model forms is sensitive to presentation order. In this way we are able to simulate the data of Anderson (1990) and Elio & Anderson (1984).

Comparisons to Cheeseman, Kelly, Self, Stutz, Taylor, & Freeman (1988)

The Bayesian character of this classification model raises the issue of its relationship to the Autoclass model of Cheeseman et al. While it is hard to know how significant the differences are, there are a number of points of contrast:

Algorithm Rather than an algorithm that iteratively incorporates instances into an existing category structure, Cheeseman et al. use a parameter searching program that looks for the best fitting set of parameters. Not enough information is provided to compare the two algorithms with respect to efficiency or probability of identifying the optimal structure. Presumably, Autoclass is independent of the order of the examples.

Number of Classes Autoclass has a bias in favor of fewer classes whereas this bias is setable in the rational model according to the parameter c. Autoclass does not calculate a prior corresponding to the probabilities of various partitionings.

Conditional Probabilities It appears Autoclass uses the same Bayesian model as we do for discrete dimensions. The treatment of continuous dimensions is somewhat different although we cannot discern its exact mathematical basis. The posterior distribution is a normal distribution which will only be slightly different than the t-distribution we use. Both Autoclass and the rational model assume the various distributions are independent.

Qualitatively, the most striking difference is that AUTOCLASS derives a probability of an object belonging to a class whereas the rational model assigns the object to a specific class. However, Cheeseman et al. report that in the case of strong category structure the probability is very high that the object comes from a single category.

Acknowledgments

This research was supported by grant BNS 87-05811 from the National Science Foundation and Contract N00014-90-1489 from the Office of Naval Research.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought.* Hillsdale, NJ: Erlbaum.

Anderson, J.R. & Matessa, M. (In preparation). The Adaptive Nature of Human Categorization. .

Berge, C. (1971). *Principles of Combinatonics*. New York: Academic Press.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analyses*. New York: Springer-Verlag.

Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., & Freeman, D. (1988). A Bayesian Classification System. In *Proceedings of the Fifth International Conference on Machine Learning*. San Mateo, CA: , 54-64.

Elio, R., & Anderson, J. R. (1984). The effects of information order and learning mode on schema abstraction. *Memory and Cognition, 12*, 20-30.

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

Gluck, M.A., & Bower, G.H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 8*, 37-50.

Gluck, M.A., & Corter, J.E. (1985). Information and category utility. Unpublished Manuscript. Stanford University.

Hayes-Roth, B. & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior, 16*, 321-338.

Hoffman, J., & Ziessler, C. (1983). Objectidentifikation in kunstlichen Begriffshierarcchien. *Zeitschrift fur Psychologie, 194*, 135-167.

Homa, D., & Cultice, J. (1984). Role of feedback, category size, and stimulus distortion in the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 83-94.

Lebowitz, M. (1987). Experiments with incremental concept formation: UNIMEM. *Machine Learning, 2*, 103-138.

Lee, P.M. (1989). *Bayesian Statistics*. New York: Oxford.

Medin, D.L., & Edelson, S.M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117*, 68-85.

Medin, D.L., & Schwanenflugel, P.J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 355-368.

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

Michalski, R. S. & Chilausky, R. L. (1980). Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems, 4*, 125-161.

Murphy, G.L., & Smith, E.E. (1982). Basic level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior, 21*, 1-20.

Nosofsky, R. (1988). Similarity, frequency, and category represeentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 54-65.

Nosofsky, R.M., Clark, S.E., & Shin, H.J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 282-304.

Posner, M. I. & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353-363.

Reed, S.K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382-407.

Rosch, E., Mervis, C.B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 7*, 573-605.