Chapter 2

An Incremental Bayesian Algorithm for Categorization

John R. Anderson
Michael Matessa

1. Introduction

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 the computational constraints under which the system is operating. 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 an 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 will be dangerous or not. This approach to categorization sees nothing special about category labels. The fact that an object might be called a *tiger* is just another feature one might want to predict about the object.

2. The Structure of the Environment

Predictions are driven by assumptions about the structure of the environment. The theory we have developed rests on the structure of natural kind categories produced by the phenomenon of *species*, which 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. In summary, the critical aspects of speciation are the disjoint partitioning of the object set and the independent probabilistic display of features within a species.

Other types of objects may display these same properties. Artifacts are another common type of object that approximate a disjoint partitioning, but there are occasional exceptions — for instance, mobile homes are both homes and vehicles. Other classes 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 this holds in some objective sense. One can use the understanding of speciation for natural kinds and the understanding of the intended function in manufacture in the case of artifacts to objectively assess the hypothesis of a disjoint partitioning.

2.1 Algorithms for Prediction

We have used this disjoint, probabilistic model of categories to understand the structure of the environment and to make predictions about object features. To maximize the prediction of object features, we must induce a disjoint partitioning of the object set into categories and determine the probability of features for each category. The ideal prediction function would be described by the formula

$$Pred_{ij} = \sum_x P(x|F_n) Prob_i(j|x) \quad , \tag{1}$$

where $Pred_{ij}$ is the probability that 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 that the object in question would display value j in 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 hypotheses, we were motivated to explore incremental algorithms such as those developed by Fisher (1987) and Lebowitz (1987).

The following steps give a formal specification of our incremental algorithm's learning and performance on each object it encounters:

1. If no previous object has been seen, initialize the category partitioning of the objects to be the empty set (i.e., no categories).

2. Given a partitioning for the first m objects, calculate for each category k the probability P_k that the $m+1$st object comes from category k. Let P_0 be the probability that the object comes from a completely new category.

3. Create a partitioning of the $m+1$ objects in which the $m+1$st object is assigned to the category with maximum probability.

4. To predict value j on dimension i for the $n+1$st object calculate

$$Pred_{ij} = \sum_k P_k P(ij|k) \quad , \tag{2}$$

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

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

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. Neither does it weight multiple possible partitionings as the ideal algorithm would. In cases of strong category structure, there will be only one probable partitioning and the incremental 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 2 closely approximate the ideal quantity because of the weighting of multiple categories. As we will see, the correlations are about 0.95 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 this latter term. 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:

$$P_k = P(k|F) = \frac{P(k)P(F|k)}{\sum_i P(i)P(F|i)} \quad , \tag{3}$$

where the summation in the denominator is over all categories i 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 $P(F|k)$.

2.2 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 that 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 (Anderson, 1990) a simple form for the prior probability

$$P(k) = \frac{cn_k}{(1-c)+cn} \quad , \tag{4}$$

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. The rational basis for this bias should be 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

$$P(0) = \frac{(1-c)}{(1-c)+cn} \quad . \tag{5}$$

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

The impact of the coupling parameter c will be to influence the number and size of categories formed. The larger the value, the fewer and larger the categories that will be produced. Since computation costs are linearly related to number of categories and not to size of categories, there might be some pressure to set c larger than its true value in the environment.

One consequence noted of Equations 4 and 5 is that there is a bias to put objects into large categories. Some have questioned the rationality of this strategy. However, we should stress that Equation 4 just sets the priors and must be combined with conditional probabilities for Equation 3. If an instance much better matches a smaller category, the conditional probabilities for the smaller category will be much higher and the instance will be assigned to that category. Thus, the bias in Equations 4 and 5 does not mean that such evidence will be ignored. However, if such feature-matching evidence is equivocal, the system will assign the instance to the larger category, which is the sensible action.

This base rate effect contributes to the order sensitivity of our algorithm. Suppose we have an instance that is ambiguous between two categories. If by chance we have seen more instances of one category before the instance, we will be biased to assign it to that category. This will make that category larger and increase our tendency to assign instances to the category. In some cases of ambiguous stimuli, this process can snowball.

2.3 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

$$P(F|k) = \prod_i P(ij|k) \quad , \tag{6}$$

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 one from recognizing categories with correlated features. Thus, one 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. The independence assumption does prevent one from representing categories in which values on two dimensions are either both one way or both the opposite. For instance, it would prevent one from recognizing a single category of animals that were either large and fierce or small and gentle. However, this is not a very serious limitation. In such cases, our algorithm spawns 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 6 is to focus our attention 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.

2.4 Discrete Dimensions

The basic Bayesian strategy for making inferences 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 accumulate, come up with a tighter and tighter posterior distribution.

In the case of a discrete dimension, the typical Bayesian analysis (Berger, 1985) assumes 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 $p_j = \alpha_j/\alpha_o$, and the value α_o reflects the strength of belief in these prior 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 $\alpha_j + C_j$. 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 6:

$$P(ij|k) = \frac{C_j + \alpha_j}{n_k + \alpha_0} \quad , \tag{7}$$

where n_k is the number of objects in category k that 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 must have this more complicated form to deal with problems of small samples. For instance, if one has just seen a single object in a category and it 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.

2.5 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 the mean and variance of this distribution. It is unreasonable to suppose we can know precisely in advance either the mean or the variance. 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:

$$f_i(x|k) \sim t_{a_i}\left(\mu_i, \sigma_i\sqrt{1 + 1/\lambda_i}\right) \quad . \tag{8}$$

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

$$\lambda_i = \lambda_0 + n \tag{9}$$

$$a_i = a_0 + n \tag{10}$$

$$\mu_i = \frac{\lambda_0\mu_0 + n\overline{x}}{\lambda_0 + n} \tag{11}$$

$$\sigma_i^2 = \frac{a_0\sigma_0^2 + (n-1)s^2 + \frac{\lambda_0 n}{\lambda_0+n}(\overline{x} - \mu_0)^2}{a_0 + n} \quad , \tag{12}$$

where $\overline{x}$ 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, $\overline{x}$ and s^2, in a manner that is weighted by our confidences in these priors, λ_0 and a_0.

Equation 8 for the continuous case describes a probability density that serves the same role as Equation 7 for the discrete case, which describes a probability. The product of conditional probabilities in Equation 6 will then be a product of probabilities and density values. Basically, Equations 6, 7, and 8 give us a basis for judging the similarity of an object to the category's central tendency.

2.6 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.

3. Application of the Algorithm

We have applied the algorithm to a number of examples to illustrate its properties. The predictions of this algorithm are potentially order sensitive in that different partitionings may be uncovered for different orderings of instances. In the presence of a strong categorical structure, the algorithm picks out the obvious categories and, as we will discuss later, there usually is little practical consequence to the different categories it extracts in the case of weak category structure. The incremental algorithm is also extremely fast. A Franz Lisp implementation categorized the 290 items from Michalski and Chilausky's (1980) data set on soybean disease (each with 36 values) in one CPU minute on a Vax 780 or on a MAC II. This was 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 hand-crafted diagnostic system of Michalski and Chilausky (1980).

The first experiment in Medin and Schaffer (1978) is a nice one for illustrating the detailed calculations of the algorithm. They had subjects study the following six instances, each described with binary features:

1 1 1 1 1	0 0 0 0 0
1 0 1 0 1	0 1 0 0 0
0 1 0 1 1	1 0 1 1 0

The first four binary values were choices in visual dimensions of size, shape, color, and number. The fifth dimension reflected the category label. They then presented these six objects without their category

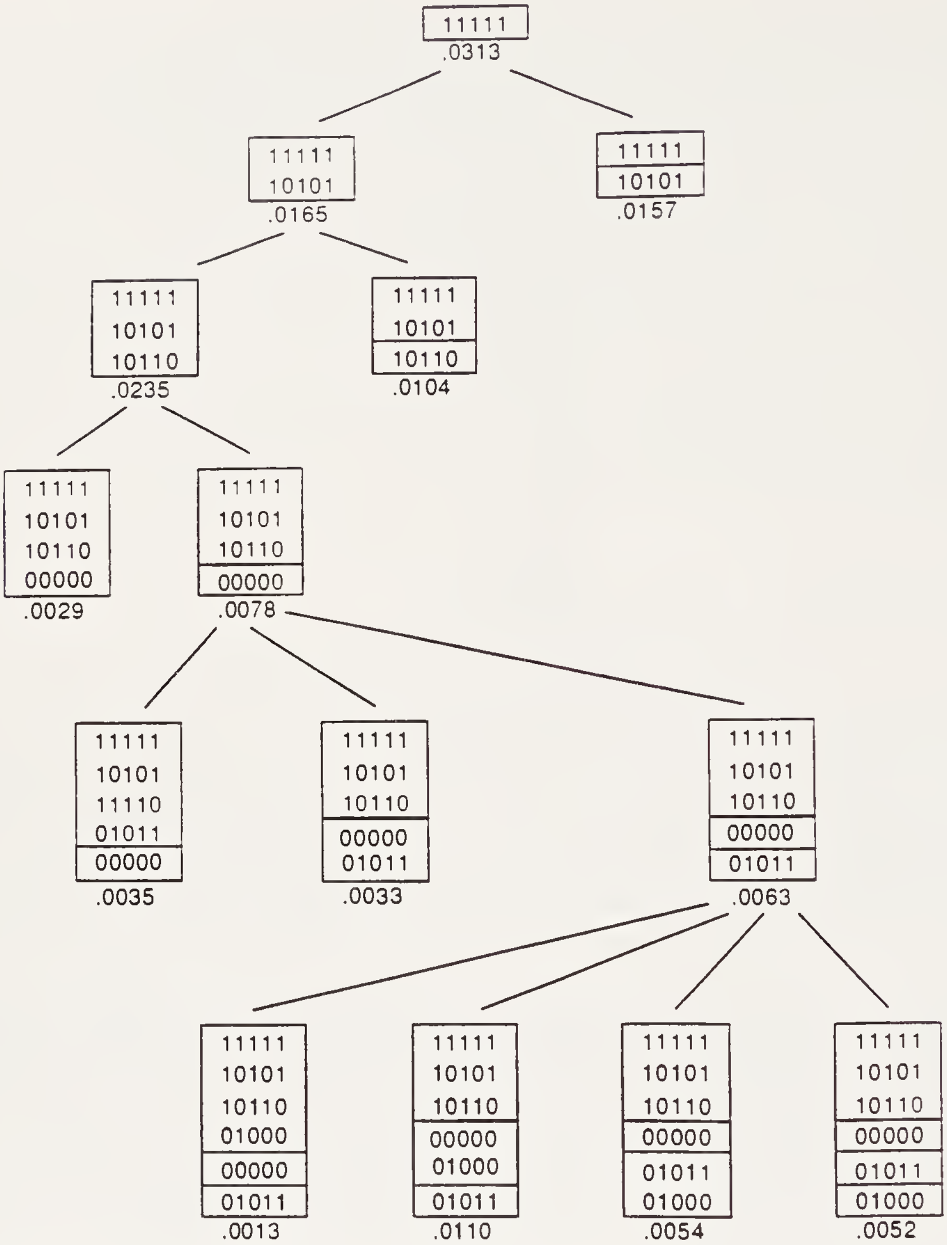

Figure 1. An illustration of the operation of the incremental algorithm in the material from the first experiment of Medin and Schaffer (1978).

label plus six new objects without a label: 0111_, 1101_, 1110_, 1000_, 0010_, and 0001_. Subjects were to predict the missing category label.

We derived simulations of this experiment by running the program across various random orderings of the stimuli and averaging the results. Figure 1 shows one simulation run in which we used the order 11111, 10101, 10110, 00000, 01011, 01000 and had the coupling probability

$c = 0.5$ (see Equations 4 and 5) and set all $\alpha_j = 1$ (see Equation 7). Figure 1 illustrates the search behavior of the algorithm as it considers various possible partitionings. The numbers associated with each partition are measures of how probable the new item is given the category to which it is assigned in that partition. These are the values $P(k)P(F|k)$ calculated by Equations 4 through 11. Thus, we start out with categorizing 11111 in the only possible way — that is, assigning it to its own category. The probability of this is the prior probability of a 1 on each dimension or $(0.5)^5 = 0.0313$. Then we consider the two ways to expand this to include 10101 and choose the categorization that has both objects in the same category because that is more likely. Each new object is incorporated by considering the possible extensions of the best partition so far. We end up choosing the partition {11111, 10101, 10110}, {00000, 01000}, {01011}, which has three categories. Note that the system's categorization does not respect the organization given by Medin and Schaffer.

Having come up with a particular categorization, we then tested the algorithm by presenting it with the 12 test stimuli and assessing the probabilities it would assign to the two possible values for the fifth dimension (the label). Figure 2 relates our algorithm to their data. Plotted along the abscissa are the 12 test stimuli of Medin and Schaffer in their rank order determined by subjects' confidence that the category label was a 1. The ordinate gives the algorithm's probability that the missing value was a 1. Figure 2 illustrates three functions for different ranges of the coupling probability. The best rank order correlation occurred for coupling probabilities in the range 0.2 to 0.3.

This coupling probability gave a rank order correlation of 0.87, and a coupling probability of 0.3 produced correlations of 0.98 and 0.78 for two slightly larger experimental sets used by Medin and Schaffer. These rank order correlations are as good as those obtained by Medin and Schaffer with their many-parameter model. It also does better than the Act simulation reported in Anderson, Kline, and Beasley (1979). We have set the coupling probability c to 0.3 throughout our applications.

The reader will note that the actual probabilities of category labels estimated by the model in Figure 2 only deviate weakly above and below 0.5. This reflects the very poor category structure of these objects. Better structured material gives much higher prediction probabilities.

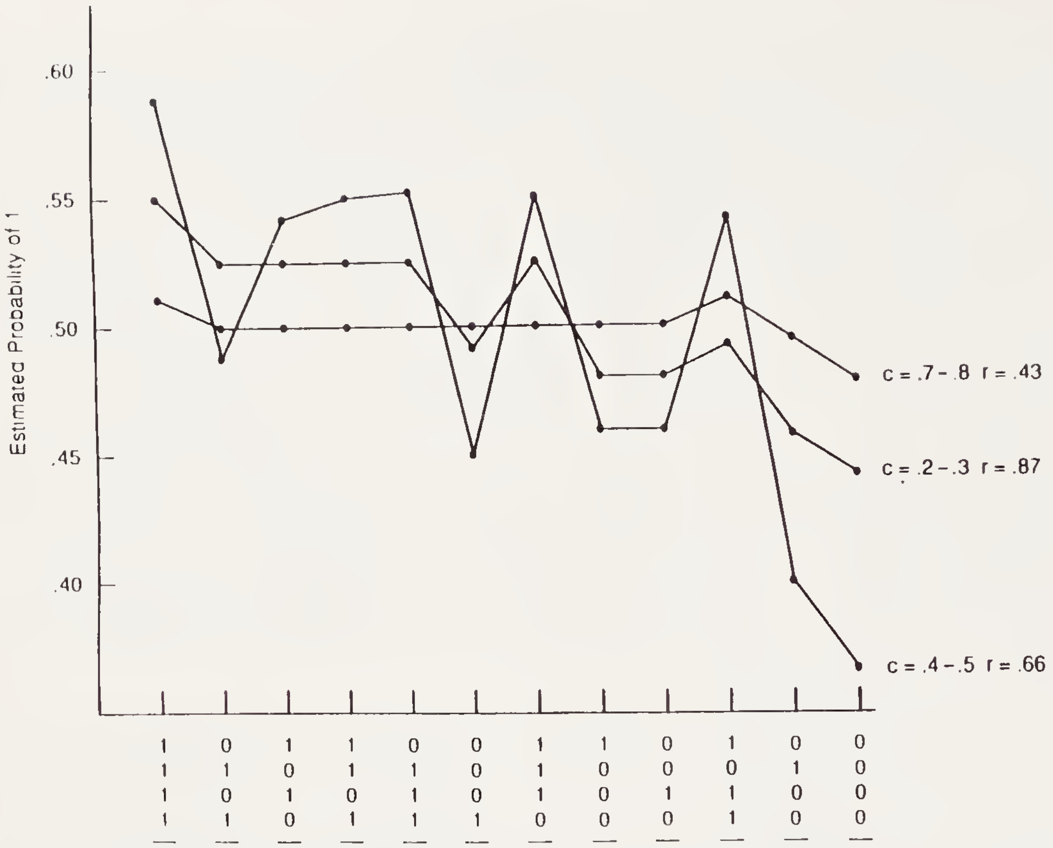

Figure 2. Estimated probability of category 1 for the 12 test stimuli in the first experiment of Medin and Schaffer (1978). Different functions are for different ranges of the coupling probability.

Detailed descriptions of the application of the algorithm to particular experiments can be found in Anderson (1990) and Anderson (in press). However, we will briefly review the applications of the algorithm to a number of empirical phenomena. The following are among the empirical phenomena we have successfully simulated:

1. *Extraction of central tendencies for continuous dimensions*. For continuous dimensions, the Bayesian model implies that categorization should vary with distance from central tendency. This enables the model to simulate the data of Posner and Keele (1968) on categorization of dot patterns and Reed (1972) on categorization of faces.

2. *Extraction of central tendencies for discrete dimensions*. The model implies that stimuli should be better categorized if they display the majority value for a dimension. This enables the model to simulate the data of Hayes-Roth and Hayes-Roth (1977).

3. *Effect of individual instances.* If an instance is sufficiently different from 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 and Schaffer (1978) on discrete dimensions and Nosofsky (1988) on continuous dimensions.

4. *Linearly separable vs. non-linearly separable categories.* In contrast to some categorization models, this model is able to learn categories that cannot be separated by a plane in an n-dimensional 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 (1983) on discrete dimensions and Nosofsky, Clark, and Shin (1989) on continuous dimensions.

5. *Basic-level categories.* The internal categories that the model extracts correspond to what Rosch (1976) meant by basic-level categories. Thus, it can simulate the data of Murphy and Smith (1982) and Hoffman and Ziessler (1983).[1]

6. *Probability matching.* Faced with truly probabilistic categories and large samples of instances, the model will estimate probability of features that correspond exactly to the empirical proportion. Thus, it predicts Gluck and Bower's (1988) data on probability matching.

7. *Base-rate effect.* Because of Equation 4, this model predicts that usually there will be a greater tendency to assign items to categories of large size. Thus, it handles Homa and Cultice's (1984) data and reproduces the more subtle interactions of Medin and 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, and Freko (1982).

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, one must provide category labels to get successful learning. Thus, this model reproduces the data of Homa and Cultice (1984).

[1] For a similar application, see Gluck and Corter (1985).

10. *Effects of input order*. In the presence of weak category structure, the categories that the model forms are sensitive to presentation order. In this way it is able to simulate the data of Anderson and Matessa (see Anderson, 1990) and Elio and Anderson (1984).

All these simulations were done with a constant setting of the parameters: c from Equations 4 and 5 at 0.3, α_j from Equation 7 at 1, λ_0 from Equation 9 at 1, a_0 from Equation 10 at 1, μ_0 from Equation 11 at the mean of the stimuli, and σ_0^2 from Equation 12 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.

4. Comparison to Autoclass

The Bayesian character of this classification model raises the issue of its relationship to the Autoclass system of Cheeseman, Kelly, Self, Stutz, Taylor, and Freeman (1988). Although it is hard to know the significance of the differences, there are a number of points of contrast.

Algorithm. Rather than an algorithm that incrementally incorporates instances into an existing category structure, Cheeseman et al. use a parameter-searching program that tries to find the best-fitting set of parameters. The Cheeseman program is quite fast and is not sensitive to the order of the examples. On the other hand, it does not easily generate predictions that can be incrementally updated with each example.

Number of classes. Autoclass has a bias in favor of fewer classes, whereas this bias can be set in the rational model according to the parameter c. The system does not calculate a prior corresponding to the probabilities of various partitionings.

Conditional probabilities. It appears that 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 from 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.

5. Order Sensitivity and Hierarchical Algorithms

The categorization algorithm that we have described is order sensitive and this has been a point of criticism of the model (Ahn & Medin, 1989). If critical examples have accidental similarities, the model will create pseudo-categories around these. If the initial examples have exaggerated differences, the algorithm will fail to identify the true categories but will split them into lower-level categories. The basic problem is that the algorithm is unable to split categories that it has formed into subcategories or to merge existing categories into larger categories, as occurs in Fisher's (1987) approach. In Anderson (1990) we showed that subjects display some sensitivity to order, but much less than our algorithm.

An interesting question concerns the consequences of this order sensitivity from the rational analysis goal of maximizing predictive accuracy, where maximal accuracy is defined with respect to the ideal algorithm (Equation 1). It is usually impossible to calculate the predictions of the ideal algorithm but the first experiment of Medin and Schaffer (1978, Figures 1 and 2) used a sufficiently small stimulus set to make such calculation possible. We calculated the ideal probabilities for the test stimuli in Figure 2 using $c = 0.5$ and $\alpha_j = 1$. At $c = 0.5$, depending on ordering, the incremental algorithm selects one of the following three partitionings:

A: (01011) (00000, 01000) (10101, 10110, 11111)
B: (11111, 01011) (00000, 01000) (10101, 10110)
C: (10101, 10110, 11111) (01011, 00000, 01000)

Figure 1 illustrates the partitioning A, which has log probability −25.77 and occurs 22% of the time.[2] Similarly, B has log probability −26.52 and occurs 16% of the time, whereas C has log probability −25.64 and occurs 61% of the time. The latter is the most probable

[2] We are calculating the product of $P(k|F)$ (from Equation 3) for all the instances. This represents the likelihood of the data given parameters c and α.

Table 1. Correlation among various bases for predicting the stimuli in Figure 2.

	Ideal	A	B	C
Partition A	0.89	×	×	×
Partition B	0.98	0.86	×	×
Partition C	0.80	0.49	0.81	×
Average	0.96	0.78	0.96	0.92

partitioning of all. By comparison, a partitioning that merges all into one category has log probability −26.50, one that breaks them up into single categories has log probability −27.37,[3] and something awful like (11111, 00000) (01000, 10101), and (10110, 01011) has log probability −32.07. The median probability of the 203 partitions expressed in log terms is −28.66, or about 5% the probability of the most probable. The Medin and Schaffer stimulus set has weak category structure and the algorithm does not always find the most probable partition. In the case of strong category structure, the program extracts the most probable interpretations independent of order of presentation. Fisher (1987) reports a similar result for his Cobweb program.

The critical issue is how well the various partitions do at predicting features. Therefore, we looked at various partitions with respect to predicting the fifth dimension of the 12-stimuli illustration in Figure 2. We looked at the correlations among the predictions of various procedures. Table 1 reproduces the correlation matrices among the predictions of the three partitionings A, B, and C, their weighted average (as produced by the incremental algorithm), and the weighted prediction from the ideal algorithm (Equation 1). As can be seen, they all are relatively highly correlated with the ideal and, except for A and C, with each other. The weighted average of A through C is very highly correlated ($r = 0.96$) with the ideal. This suggests that there is relatively little cost associated with using the incremental algorithm.

It is interesting that the predictions made by the most probable partition are not particularly good. This reflects that the most probable

[3] A singleton category structure is less likely than a single category because of the high value of c. At $c = 0.3$, the log probability of the single category becomes −28.11 and the log probability of the singleton categories is −22.95.

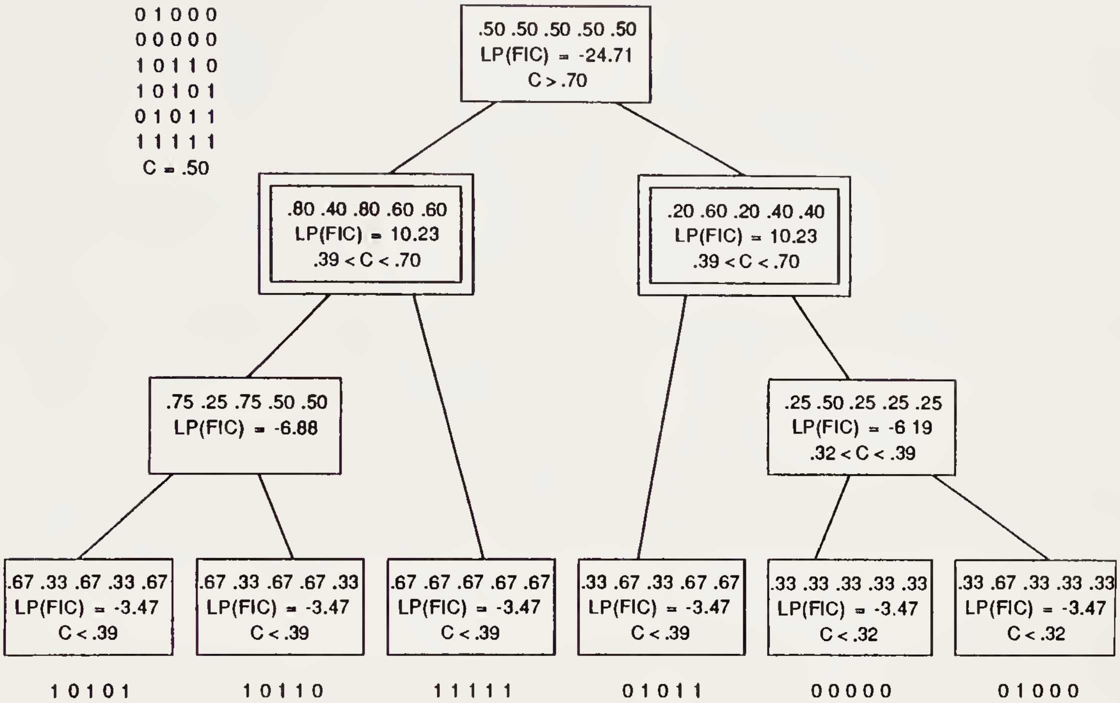

Figure 3. Hierarchical organization of stimuli from first experiment of Medin and Schaffer (1978).

partition has only 5% of the probability of the 203 possible partitions of the six stimuli. As the set size gets larger or as the category structure improves, the most probable partition will tend to dominate the prediction. This suggests that it makes sense for the system to strive for the most probable partition but that prediction from some other highly probable partition may be as good or better.

5.1 A Hierarchical Algorithm

Consideration of the order sensitivity of the algorithm has led us to consider other incremental algorithms that are less order sensitive. We were also interested in producing a hierarchical category structure and exploring the issue of whether other levels in the hierarchy, besides the basic level, might be useful for prediction.

We have developed another algorithm that is somewhat more successful at identifying the maximally probable partition but avoids considering all possible partitions as does the ideal algorithm. This algorithm organizes the data into a hierarchical structure. Figure 3 illustrates a hierarchical organization for the stimuli from the first experiment of Medin and Schaffer. Having built such a hierarchical representation of the stimulus set, our algorithm tries to determine which partitioning within the hierarchy offers the optimal decomposition of the stimulus set. This will depend on the setting of the coupling parameter c. The higher the value of c, the larger the categories that the algorithm will tend to select. Given the structure in Figure 3, the algorithm will select the top-level node as the single category for values of c greater than 0.7. For values of c in the range 0.39 to 0.70, it will select the two subnodes. For values of c from 0.32 to 0.39, it will select the bottom-level nodes except for 00000 and 01000, which it will merge into a single category. Below 0.32 it selects singleton categories.

At any point in time, the algorithm will have a hierarchical organization for the observed instances and, given a value of c, it will have identified a set of categories. The basic algorithm for growing this network consists of the following steps:

1. As before, given a new instance, determine a category K to associate with this instance.
2. If K is an existing category, sort the new instance to a location below that category node.
3. If K is a new category, sort that category to a location below the top node for the hierarchy, with the constraint that one category cannot be placed under another.
4. Search upward from where the new item was inserted to see whether some change in the category structure is warranted. Note that this does not reorganize the hierarchy, but only changes which nodes in the hierarchy might be considered category nodes.

Figure 4 illustrates the basic logic for sorting and inserting a new instance into the hierarchical structure. We have an existing hierarchical structure consisting of a node a with subnodes b and c. We have a new instance d that we want to incorporate somewhere in the hierarchical structure under a. There are three possibilities: (i) d will be associated with the hierarchical structure dominated by b; (ii) d will be associated with the hierarchical structure dominated by c; or (iii) a binary branch will be created with b and c in one and d in the other. The way to choose among these is to identify the branching that will yield the maximally probable pair of categories to cover all the items under b, c, and d. For

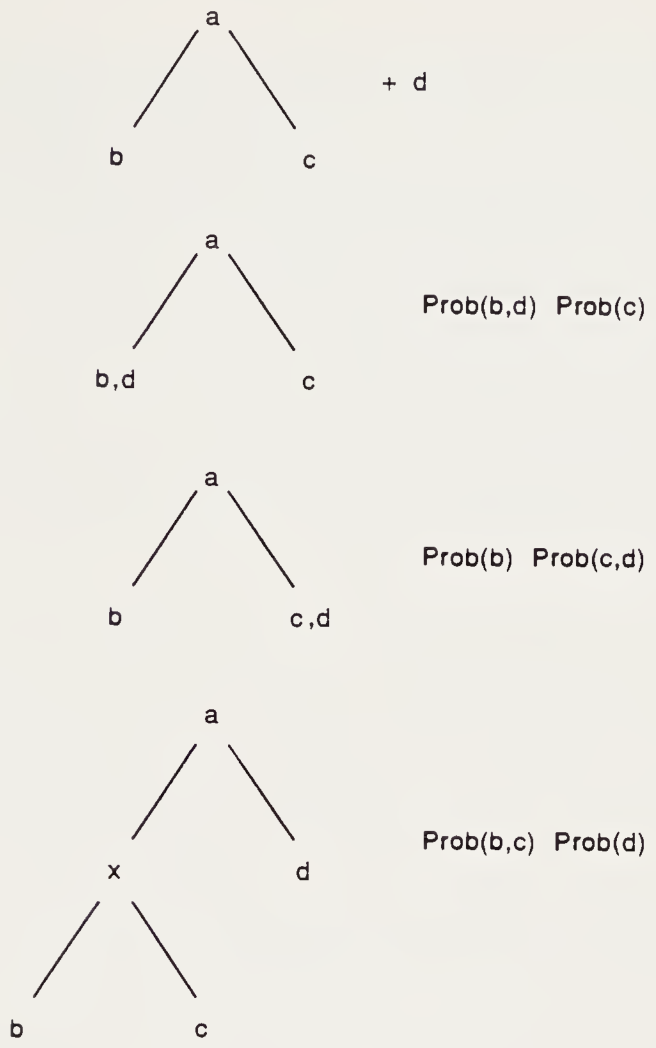

Figure 4. Choose structure to maximize probability. Should it depend on prior probability?

(i) we consider the product of the probability of the category consisting of {b d} and the category consisting of {c}. For (ii) we consider the {b} category and the {c d} category. For (iii) we consider the {b c} category and the {d} category.

It is worthwhile comparing the performance of this hierarchical algorithm with that of the previous algorithm. We used the material of Medin and Schaffer for this purpose. There are 720 possible orderings of the six stimuli. With $c = 0.5$, the old algorithm identified three different categories and identified the optimal categorical structure 61% of the time, producing categorical structures with average log probability of -25.81. The hierarchical algorithm identified six different categories, but identified the optimal category structure 80% of the time, and produced categorical structures with average log probability of −25.74. Thus, there is not much difference in average goodness, but the hierarchical algorithm is somewhat more successful at finding the optimal structure. With larger stimulus sets, we have not been able to test the hierarchical algorithm exhaustively, but it does appear more stable and does more often select the ideal structure.

We have also explored hierarchical algorithms that are not order sensitive. Basically, they used classical clustering techniques (Annenberg, 1973) to create hierarchies of stimuli sorted according to similarity where this was measured by Equations 6–8. Such algorithms are more expensive computationally because they must perform all pairwise comparisons. We have not found that they yield notably better results.

A good case for illustrating the problems of these algorithms is the iris data base of Fisher (1936). According to Fisher there are three underlying types of irises. Our algorithms, whether hierarchical or not, and whether incremental or not, fail to identify this category structure. They always identify one of the categories, Iris Setosa, but either fail to separate the other two (Iris Versicolor and Iris Virginica) or split them up inappropriately (as defined by Fisher). Cheeseman et al. claim that their algorithm is successful on these data, but we have observed that it produces an inappropriate splitting into three categories that do not correspond to the original ones. We also have observed that human subjects extract two categories or produce an inappropriate splitting into three. This notwithstanding, one can show that the original categorization produced by Fisher is more probable than the two-category solution or the various inappropriate three-category solutions. However, it is apparently impossible to find this more probable partitioning given the various approaches, artificial or human. However, prediction of features is not enhanced by the more probable partitioning. Thus, it is not clear that we should consider the behavior of these various algorithms as failures.

5.2 Genus-Level Identification

Although it does not seem that the hierarchical approach produces substantially better categorical organization, we think there may be some significance to levels of the hierarchy. There are at least two other levels that are significant for purposes of prediction — the *genus*[4] (at the higher level) and the *individual* (at the lower level). We will discuss the significance of each for prediction.

The genus level offers a level of aggregation above the species, which corresponds to a group of biologically related species that are more similar to one another than are arbitrary pairs of species. The significance of the genus level does not come in making predictions about known properties of known species. For instance, we are much better off predicting the cat-chasing propensity of Fido knowing that he is a dog than knowing that he is a mammal. The genus level's significance comes in making predictions about unknown properties of a known species (e.g., whether Fido has a spleen) and in making predictions about unknown species.

In Bayesian terms, this means that genus level can be used to set more informed priors for the species under the genus. This helps in making predictions about new species and about unexperienced properties of existing species. The interesting complication is that these priors themselves depend on estimates of the parameters for the existing species, which in turn depend on the priors. Thus, it might seem that we have a difficult joint estimation problem. The typical Bayesian approaches to such estimation problems are known as *hierarchical* methods (Berger, 1985, Section 4.6). The technical development of such methods can be quite complex and is not justified here, since we have not yet gathered data that require such complex quantitative analysis. We will simply note that, for our purposes, they provide a rationale for making estimates of the mean and variance within a species sensitive to estimates of the mean and variances for other species within a genus.

There certainly is evidence that people have this sensitivity. Even young children have expectations about the properties of new animals based on animals that they have seen (Carey, 1985). They also have expectations that certain dimensions are less variable for certain types of categories. Thus, they expect that animals within a category will have the same constitution, whereas artifacts within a category will have

[4] We use the term *genus* in its more general sense to refer to a kind and not to imply the precision that is involved in the distinction among genus, family, order, class, and phylum in biology. We suspect that the level useful in prediction might be considerably above the biological genus level and actually closer to the phylum level.

the same function (Gelman, 1988). Moreover, these expectations show developmental trends to more accurate forms as experience accumulates.

The experiment of Nisbett, Krantz, Jepson, and Kunda (1983) also illustrates differential sensitivity to variance in categories of different kinds. They asked subjects to suppose they had a sample of a new mineral, a new bird, or a new tribe of people from a new island. They were given samples of different size and told that all the objects within the sample had some property. Subjects were willing to extrapolate from a single observation for some dimensions, like conductivity of the mineral or color of the tribe of people, whereas they required 20 observations before they extrapolated with any confidence for other dimensions, like the obesity of the people.

This ability to show sensitivity to variance is one thing that distinguishes this hierarchical Bayesian approach to categorization from most others. Many approaches (e.g., instance-based models) would predict that subjects would be biased in their estimate of the mean of a new species by the mean of existing species. However, these other approaches do not have the mechanisms for showing a similar sensitivity to variance.

5.3 Individual-Level Identification

The individual provides a much lower level of aggregation below the category. For purposes of prediction, there is a real advantage to identifying a repetition of an individual and making predictions from the individual rather than the category. This is because the individual may reliably deviate from the mean of the category and because many features are much more certain at the individual level than at the category level.

Retrieving an individual and making a prediction based on it corresponds to a memory retrieval. From this perspective the difference between retrieval and categorization concerns whether prediction is being made at the individual level or the category level. It is basically the same logic of prediction; however, it must be parameterized differently:

1. To reflect the fact that individuals repeat themselves much less often than categories, we use a lower value of c.
2. We need to capture the fact that the features are much less likely to change. This requires lower values of the α_i for discrete dimensions and much lower values of σ_0^2 for continuous ones.

There has been much speculation as to how categorization behavior relates to memory behavior. The instance-based models (Medin & Schaffer, 1978; Nosofsky, 1986) would argue that everything is really memory-based, whereas connectionist models (McClelland, Rumelhart, & Hinton, 1986) would argue that there are no separate representations of instances and everything is merged. Both frameworks try to account for differences between categorization and memory by arguing that a single representation is processed in different ways. The current model offers a representation that distinguishes the two levels but uses the same Bayesian logic at both levels. Of course, the rational representation is only an acknowledgement of the fact that there are individuals and categories in the real world. It does not really make any claims about how they are processed in the head. Anderson (in press) presents experimental evidence that people make different predictions when they operate at the category level from those they make at the individual level.

6. Summary

In summary, we have identified an incremental Bayesian algorithm that is fast, yields near-optimal predictions about stimulus dimensions, and corresponds with uncanny accuracy to the behavior to humans. We have explored the potential of some hierarchical variations of the algorithm, but these produce marginal improvements at best in the prediction behavior. However, there is reason to suppose that human categorization behavior has sensitivities to levels both above and below the basic level category.

Our interest in the original incremental algorithm began as an attempt to approximate the ideal, computationally impossible, prediction specified by Equation 1. It was not intended as a serious model for human cognition or as an artificial intelligence application. However, after more than two years of exploration, we have failed to find a real improvement and continue to be surprised at its success.

Acknowledgements

The research reported in this chapter was supported by Grant N00014-90-1489 from the Cognitive Science Program, Office of Naval Research.

References

Ahn, W., & Medin, D. L. (1989). A two-stage categorization model of family resemblance sorting. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 315–322). Ann Arbor, MI: Lawrence Erlbaum.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R. (in press). The adaptive nature of human categorization. *Psychological Review.*

Anderson, J. R., Kline, P. J., & Beasley, C. M. (1979). A general learning theory and its applications to schema abstraction. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 13). New York: Academic Press.

Annenberg, M. R. (1973). *Cluster analysis for applications.* New York: Academic Press.

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

Berger, J. O. (1985). *Statistical decision theory and Bayesian analyses.* New York: Springer-Verlag.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., & Freeman, D. (1988). A Bayesian classification system. *Proceedings of the Fifth International Conference on Machine Learning* (pp. 54–64). Ann Arbor, MI: Morgan Kaufmann.

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.

Fisher, R. A. (1936). Multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179–188.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology*, *20*, 65–95.

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, uncertainty, and the utility of categories. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* (pp. 283–287). Irvine, CA: Lawrence Erlbaum.

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). Obectidentifikation 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.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing*. Cambridge, MA: MIT Press.

Medin, D. L. (1983). Structural principles of categorization. In B. Shepp & T. Tighe (Eds.), *Interaction: Perception, development, and cognition*. Hillsdale, NJ: Lawrence Erlbaum.

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.

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., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

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.

Nisbett, R. E., Krantz, D. H., Jepson, D., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*, 339–363.

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

Nosofsky, R. M. (1988). Similarity, frequency, and category representation. *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.