Schema-based Categorization

Benjamin Martin
Department of Psychology
Stanford University
Stanford, CA 94305
ben@psych.stanford.edu

Abstract

Many theories of conceptual organization assume the existence of some form of mental similarity metric (Medin and Schaffer, 1978; Hintzman and Ludlum, 1980; Nosofsky, 1988; Shepard, 1987; Kruschke, 1992, among others.) In the domain of categorization, such theories have been called "similarity-based" (Murphy and Medin, 1985). Criticism of similarity-based theories has led to a call for "theory-based" models of categorization (Murphy and Medin, 1985; Rips, 1989; Barsalou, 1991; Medin, 1989). Theory-based views remain somewhat vague, however. In this paper I outline a schema-based theory of conceptual organization. The model depends on the notion of a mental similarity metric but makes use of connectionist learning principles to develop a conceptual organization that solves a problem faced by purely similarity-based models of categorization. I discuss the relationship of this theory to similarity-based and theory-based accounts.

Conceptual organization and categories

Psychologists study categorization to understand more than just which objects or events share a common name, or even how people decide whether perceptibly different objects are of the same type. The aim of studying categorization is to understand the nature of concepts. Behind this approach to studying concepts lies the intuition that concepts are the atoms of conceptual organization, categories are the atoms of categorization, and conceptual organization depends upon representing the relations among perceived or conceived objects in just the way that categorization depends upon representing the relations among classified entities. In short, when psychologists speak about categorization they aim to elucidate conceptual organization. For this reason, theories of categorization often underlie explanations of a range of behaviors far broader than mere classification. Behaviors such as generalization, prediction, communication, learning, and inference.

In adopting this broad notion of categorization as a ubiquitous but singular mental faculty underlying a great range of behaviors, many theories of conceptual organization make a tacit assumption that categorization has a unitary character. In other words, whether concepts are said to depend upon prototypes or exemplars, independent features, or relations among properties, theories, or regions in similarity space, they are thought to depend on them always and everywhere in more or less the same way.

While maintaining a belief in the unitary character of categorization, psychologists have sought to include an ever widening range of data in the explanatory scope of theories about categorization. At first theories such as that of Katz and Fodor (Katz and Fodor, 1963; Katz, 1972) were proposed to describe the relationship of words to concepts. There was a distinctly philosophical character to these views which Smith and Medin (1981) have called "the classical view". Later, largely as a result of the work of Rosch and her collaborators (e.g. Rosch and Mervis, 1975), theories of categorization *qua* conceptual organization were called upon to explain the graded structure of subjects' decisions about class membership. These and other data led to a class of models that Murphy and Medin (1985) have termed "similarity-based". In addition to analyzing similarity-based theories, Murphy and Medin asked whether existing theories of categorization could account for other important properties of conceptual organization. They reasoned that such theories would have trouble accounting for data concerning prediction, inference, and generalization when those behaviors depend upon high-level knowledge. They have called for a new view of conceptual organization based upon knowledge about the

relationships among the features of objects and the causal and functional properties relating members of a class.

In this paper I will outline a model of conceptual organization. This model depends upon many of the same theoretical constructs as traditional similarity-based models. I will argue that it can account nevertheless for data that have been difficult to explain on a purely similarity-based view.

Similarity-based categorization

Broadly speaking, similarity-based theories of concept formation assume that mental categories are formed through experience with the world, and reflect its structure. When we encounter an object or event, we register some information about it in terms of features. This information is then used to construct a stable representation of a concept, or to modify some existing concept. In either case, assimilation provides later access to information about an object's properties (seen and unseen), and it's relation to other objects we have encountered. The way in which we assimilate each new object or event depends upon its similarity to concepts we already possess. One way to imagine this process is to view similarity as a kind of internally represented space, Euclidean or otherwise. The features of objects are dimensions in this space and the closer two objects are in similarity-space, the more similar they are. In a similarity-based theory, this featural distance determines the likelihood of membership in a common class. Because features constitute the dimensions of similarity-space, probability of common class-membership is a function of common and distinctive features.

It is clear that such a general representation of the information that can be derived from an encounter with an object would be extremely useful for classifying or learning about objects, for generalizing from experience to predict class membership or unseen properties of new objects, or simply for remembering and organizing perceived objects.

A number of influential theories make use of these central assumptions (Rosch and Mervis, 1975; Posner and Keele, 1968; Medin and Schaffer, 1978; Hintzman and Ludlum, 1980; Sattath and Tversky, 1987; Nosofsky, 1988; Shepard, 1987; Anderson, 1991; Kruschke, 1992) and much of the work on these theories has involved the attempt to test their competing predictions. Unfortunately, even if the debate between prototype and exemplar-based theories is nothing more than an argument over certain possible restrictions on a similarity-based theory of categorization, there are problems that these similarity-based models all share.

A Problem with similarity-based categorization

Why is featural similarity a problematic basis for conceptual organization? Because knowledge about the detailed character, causes, and consequences of features as well relations among features give people a rich variety of explanatory and predictive knowledge on which to base their categorization judgments. Similarity based views of categorization simply leave out this knowledge, Therefore such theories cannot be adequate accounts of conceptual organization (Rips, 1989; Medin and Murphy, 1985; Barsalou, 1991; Medin, 1989). These considerations have led to theory-based models of categorization. According to such views, our mental categories (in other words our concepts) are formed in accordance with a large amount of prior knowledge about the regularities in the world and the plausible structure of experience. The behaviors that such theories take as central in a theory of categorization are quite broad, including not only prediction and generalization, but also inductive reasoning, and various judgments such as typicality and likelihood of membership in a class. The major criticism that emerges from such a view is that similarity is an insufficient basis for conceptual organization because it is too flexible to account for the highly constrained way that world knowledge can suggest structure in our experience and because it fails to predict some facts about the relationships among our subjective judgments about typicality, perceived similarity, and likelihood of membership in a category.

Schema-based categorization

The challenge from proponents of theory-based views is to account for judgments about category membership, typicality, generalization, inference, and prediction when those behaviors depend on knowledge that goes beyond mere similarity. If it is true that a singular system of conceptual organization underlies these diverse behaviors, what are its operative principles? While I do not hope to provide a complete answer,

I will to outline a model that tries to incorporate some additional knowledge about features and their functional and causal relations into conceptual organization.

Categories in the service of likelihood estimation

Several theorists have argued that categories exist in order organize knowledge so as to allow the best possible inference or prediction of unobserved features from those observed (Rosch and Mervis, 1975; Medin and Murphy, 1985; Bobick, 1987; Anderson, 1991; among others.) Taking this idea at face value, we might suppose that the job of categorization is to use experience with a set of objects or events, each represented by a vector of features (x), to find an optimal model (with parameters w) for constructing the internal representation (a vector t) of the important features of the object or event, x.

What are the important features of x? I believe that the answer lies in the domain of evolutionary psychology. It is through adaptation that some features come to be important and others negligible. Since an account of the adaptive significance of objects and events is outside the scope of this project, I will simply assume that all the experienced features in the vector x are important. As a result, the goal of the model will be veridical representation. This is a simplification but a useful one. Nevertheless, imposing relative importance on features may be one of the primary ways of representing high-level knowledge and it will be discussed further in section 5.

What is the optimal model for constructing the internal representation, t? It is the one that maximizes $P(t \mid x\&w)$, while at the same time finding a set of conceptual categories $(C_1 ... C_M)$ such that:

$$\forall x[\sum_{i=1}^{M} P(C_i \mid x) = 1] \quad (1)$$

In other words, we want to find categories that include all experienced objects or events and at the same time allow the best possible prediction of the features of those objects or events. Using (1) we can write:

$$P(t \mid x\&w) = \sum_{i=1}^{M} P(t \mid C_i\&x\&w)P(C_i \mid x\&w) \quad (2)$$

On the left side of (2) is the quantity we wish to maximize. We have rewritten it using our exhaustive category scheme, thus combining the two constraints we hope to satisfy. The game now is to express the two probabilities on the right side of (2) in such a way that we can compute them from information available through experience.

Taking $P(t \mid C_i\&x\&w)$ first, we must consider the nature of the features of x. For the sake of simplicity, let us consider them conditionally independent binary features denoting the presence or absence of a possible attribute of x.[1] In that case, we expect the features of t to be distributed binomially so we can write:

$$P(t \mid C_i\&x\&w) = \prod_{j=1}^{N} f_{ij}{}^{x_j}(1 - f_{ij})^{1-x_j} \quad (3)$$

Where the N is the number of features in the input (or the output or target; all are the same), and f_{ij} is the estimated probability of seeing attribute x_j in category C_i, prior to encountering the current object or event. The f_{ij}'s depend on the model w. How the model estimates these quantities will be described in the next section.

It remains to calculate the other part of (2), i.e. $P(C_i \mid x\&w)$. Using Bayes' rule, we can write:

$$P(C_i \mid x\&w) = \frac{P(x \mid C_i)P(C_i)}{\sum_{j=1}^{N} P(x \mid C_j)P(C_j)} \quad (4)$$

To compute this probability, we must find an expression for $P(C_i \mid x\&w)$. Now, as before we must make an assumption about the nature of objects and events. In this case we must decide how they are distributed within categories. Again for simplicity, let us imagine that categories form convex regions in similarity space.[2] In that case we may consider each category to be described by a multivariate gaussian distribution:

$$P(x \mid C_i) = Ke^{\sum_{j=1}^{N} \frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2}} \quad (5)$$

The priors $P(C_i)$, like the f_{ij}'s must be found by estimating the parameters of the model through an iterative learning procedure.

[1] Strictly speaking, this is unlikely to be true of conceptual organization. In my thesis (Martin, 1993) I discuss ways of relaxing this assumption.

[2] This issue, like the binomial assumption, is discussed in Martin (1993)

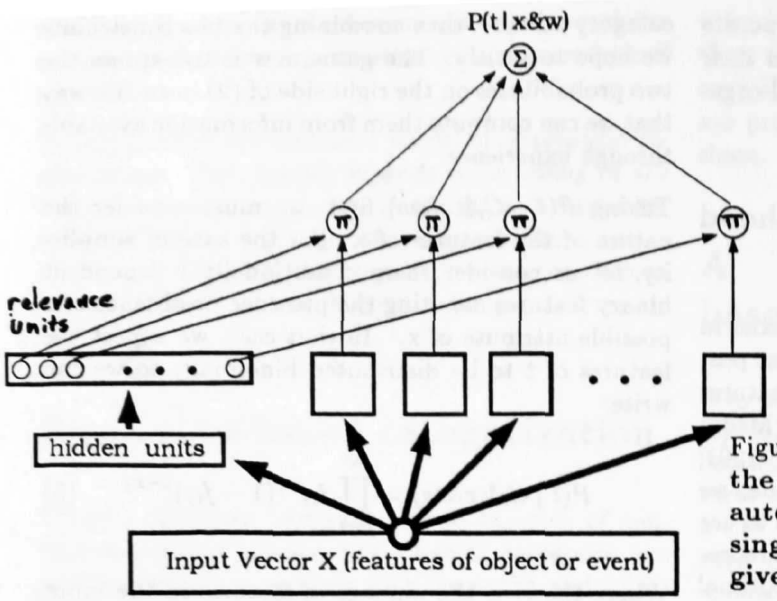

Figure 1: **The architecture of the model. Rectangles containing circles indicate vectors of units, thin arrows indicate links with a fixed weight of 1, and thick arrows indicate a pattern of adjustably weighted links in which the two connected structures are *fully* connected. The squares in figure 1 denote a structure that is shown in full in figure 2. The unit labeled Σ computes a sum, those labeled Π compute a product. The relevance layer actually contains two layers of units with fixed weights that compute** $P(C_i \mid x\&w) = \frac{P(x|C_i)P(C_i)}{\sum_{j=1}^{N} P(x|C_j)P(C_j)}$

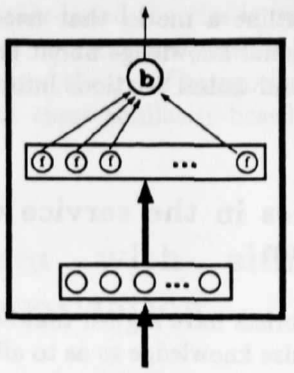

Figure 2: **An exploded view of a square from the previous figure. Each square acts as an auto-associator of the input, connected to a single "binomial" unit, b, whose activation is given by** $\prod_{j=1}^{N} f_{ij}{}^{x_j}(1-f_{ij})^{1-x_j}$

A connectionist approach to learning categories

Rumelhart, Durbin, Chauvin, and Golden have discussed an interpretation of connectionist networks as maximum likelihood estimators. In order to develop a model that learns iteratively from experience with a series of objects and events, I have used an architecture that incorporates several types of units and a learning rule that embody Bayesian principles. A more extensive analysis of these types of units can be found in Rumelhart et al. (In Preparation).

The model assumes that objects and events can be represented by vectors of binary, conditionally independent features. These features serve as the input to a group of auto-associators (AA's), denoted by squares in figures 1 and 2. the output of these AA's is a single activation value computed from the output vector by the unit to which they lead (see fig. 2.) That unit (there is one for each AA) computes the multivariate binomial likelihood of the feature vector (f) of a particular AA using the rule described in equation (3). Accordingly I call it a binomial unit. The activation of a binomial unit is nothing but $P(t \mid C_i\&x\&w)$.

The input also feeds into a hidden layer and thence to a layer of units whose activation is computed according to equations (4) and (5), with the $P(C_i)$'s initially equal and iteratively adjusted. These units compute $P(C_i \mid x\&w)$. I call them relevance units (a name suggested by David Rumelhart) because they compute the relevance of a given category to the internal representation of an input vector.[3] The next layer of units takes a product of the relevance units and the binomial units. That product is just the product in equation (2). Finally, the ultimate output unit of the network takes a sum of those products and computes $P(t \mid x\&w)$.

Simply put, this network exactly instantiates the Bayesian analysis described in the previous section. In order to iteratively train the model it is necessary only to maximize the probability of the internal representation, given the input and the parameters of the model. As a result, the error signal for training is simply the log of the output value of the final unit in the network. As the output approaches 1, the error approaches zero. The error signal is backpropagated just as in any backpropagation model, as described in Rumelhart et al. (1986).

[3] I believe these units are similar to the "gating units" described in Jacobs, Jordan, Nowlan, and Hinton (1991). Their role in the network is certainly the same.

Formalizing the notion of a schema

I have called this model "schema-based" because it provides a formal account of a number of characteristics of schemata.[4] Elsewhere (Martin, 1993), I have discussed the historical development of the term "schema" and identified eight critical properties of schemata. I have argued that each of these properties can be described in terms of the connectionist model I have outlined.

In the model, each conceptual category may be thought of as a schema. This means that each auto-associator in the network is a schema. Each schema assimilates the input vector through an informational bottleneck (created by its particular hidden layer.) Each schema does this differently because the process of learning causes a kind of specialization in the network. Every schema adapts to account for a particular conceptual category more effectively by discounting other inputs. The relevance units provide the mechanism for this discounting and the limited size of the hidden layers provides the impetus.

Incorporating knowledge

How does this model incorporate knowledge about the world that goes beyond featural similarity? Consider an example devised by Lance Rips (Rips, 1989). Subjects filling out a questionnaire report that a circular object with a diameter of 3 inches is more likely to be a pizza than a quarter, but subjects rated the circle more "similar" to a quarter. The reason for this is intuitively clear; we know that quarters show almost no variation in size, even when run over by a train. A model of categorization that depends on featural similarity thus encounters the difficulty that the overall similarity of a 3-inch circle to a quarter is higher than to a pizza while the object is extremely unlikely to be a quarter. This apparent dissociation between likelihood of category membership and judged similarity poses a difficulty for similarity-based theories.

The solution that the schema-based model offers depends on its ability to represent concepts separately from a decision about their relevance to a class-membership decision. What are subjects reporting when they judge the similarity of a 3-inch circle to a pizza or a quarter? They are reporting the likelihood of seeing those features as attributes of a pizza, based on their schema for pizza and their background knowledge. In other words, $P(t \mid pizza\&x\&w)$. When asked about class-membership, however, they are trying to find the larger of $P(quarter \mid x\&w)$ and $P(pizza \mid x\&w)$. In the schema-based model, it would be a simple consequence of having encountered only one size for quarters that $P(quarter \mid x\&w) < P(pizza \mid x\&w)$, when x includes the feature "3-inches around". Nevertheless, since 3-inches is more like the circumference of a quarter than a pizza, $P(t \mid quarter\&x\&w) > P(t \mid pizza\&x\&w)$.

In short, the schema-based model can account for cases in which there is a discrepancy between featural similarity and probability of co-clasification. This is because the space in which classes are bounded regions is not featural similarity space, but a transformation of the input-featural space that is developed over experience and reflects the distributional properties of all past experience, not simply the central tendency of a single category. Nevertheless, the binomial units in the model compute a measure of category-specific featural similarity. In this sense the model is very much similarity-based.

Because the schema based model transforms the input features through a connectionist hidden layer, it is possible to represent information about the relationships among features as well as representing complex information about the distribution of featural values. These higher order units are sensitive to intercorrelational properties of the input features. As I have discussed elsewhere (Martin, 1993), this allows for conceptual organization that depends upon correlational properties of features. Murphy and Medin (1985) have stressed that such correlations are a prime motivation for theory-based models.

Another way in which the model might represent knowledge about the relationships among features would be through the use of information about the importance of features both within and between categories. Such information might be incorporated into the model by changing the mapping computed by the schemata from an auto-associative mapping to some adaptively developed mapping. This would have the consequence of changing the space in which featural similarity is computed. If one could find the right expression to replace the multvariate binomial rule that enforces veridical representation, almost any form of featural contingency could be represented. In this way it might be possible to formalize very compli-

[4] The idea of formalizing schemata in terms of Neural networks was first proposed by Rumelhart, Hinton, McClelland, and Smolensky (1986). While my formalization bears little resemblance to their proposal, they first proposed describing schemata in terms of the dynamics of a network.

cated notions of theory-based conceptual organization.

References

[1991] Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409–429.

[1991] Barsalou, L. (1991). Deriving categories to achieve goals. In G. Bower (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory*. New York: Academic Press.

[1987] Bobick, A. (1987). *Natural Object Categorization*. Unpublished doctoral dissertation, MIT, Cambridge, MA.

[1980] Hintzman, D., & Ludlum, G. (1980). Differential forgetting of prototypes and old instances: Simulation by an exemplar based classification model. *Memory and Cognition, 8*, 378–382.

[1991] Jacobs, R., Jordan, M., Nowlan, S., & Hinton, G. (1991). Adaptive mixtures of local experts. *Neural Computation, 3*, 79–87.

[1972] Katz, J. (1972). *Semantic Theory*. New York: Harper and Row.

[1963] Katz, J., & Fodor, J. (1963). The structure of a semantic theory. *Language, 39*, 170–210.

[1992] Kruschke, J. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

[1993] Martin, B. (1993). *Categorization, Similarity, and Featural Information*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

[1989] Medin, D. (1989). Concepts and conceptual structure. *American Psychologist, 44*, 1469–1481.

[1978] Medin, D., & Schaffer, M. (1978). context theory of classification learning. *Psychological Review, 85*, 207–238.

[1988] Medin, D., & Shoben, E. (1988). Context and structure in conceptual combination. *Cognitive Psychology, 20*, 158–190.

[1985] Murphy, G., & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.

[1988] Nosofsky, R. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 54–65.

[1968] Posner, M., & Keele, S. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology: General, 77*, 353–363.

[1975] Rips, L. (1975). Induction about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14*, 665–681.

[1989] Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge, England: Cambridge University Press.

[1975] Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.

[1983] Roth, E., & Shoben, E. (1983). The effect of context on the structure of categories. *Cognitive Psychology, 15*, 346–378.

[1986] Rumelhart, D., & McClelland, J. (1986). *Parallel Distributed Processing: Volume 2*. Cambridge, MA: MIT Press.

[1986] Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in pdp models. In J. L. McCleland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (pp. 7–57). Cambridge, MA: MIT/Bradford Books.

[1987] Sattath, S., & Tversky, A. (1987). On the relation between common and distinctive feature models. *Psychological Review, 94*, 16–22.

[1987] Shepard, R. (1987). Towards a universal law of generalization for psychological science. *Science, 237*, 1317–1323.

[1961] Shepard, R., Hovland, C., & Jenkins, H. (1961). Learning and memorization of classifications. *Psychological Monographs, 75*.

[1981] Smith, E., & Medin, D. (1981). *Categories and Concepts*. Cambridge, Mass.: Harvard University Press.

[1984] Smith, E., & Medin, D. (1984). Concepts and concept formation. In M. Rosenzweig & L. Porter (Eds.), *Annual Review of Psychology*. Palo Alto, California: Annual Reviews Inc.