Concept Learning and Flexible Weighting

David W. Aha
Applied Physics Laboratory
The Johns Hopkins University
Laurel, MD 20723
aha@cs.jhu.edu

Robert L. Goldstone
Department of Psychology
Indiana University
Bloomington, IN 47405
rgoldsto@ucs.indiana.edu

Abstract

We previously introduced an exemplar model, named GCM-ISW, that exploits a highly flexible weighting scheme. Our simulations showed that it records faster learning rates and higher asymptotic accuracies on several artificial categorization tasks than models with more limited abilities to warp input spaces. This paper extends our previous work; it describes experimental results that suggest human subjects also invoke such highly flexible schemes. In particular, our model provides significantly better fits than models with less flexibility, and we hypothesize that humans selectively weight attributes depending on an item's location in the input space.

We need more flexible models of concept learning

Many theories of human concept learning posit that concepts are represented by prototypes (Reed, 1972) or exemplars (Medin & Schaffer, 1978). Prototype models represent concepts by the "best example" or "central tendency" of the concept.[1] A new item belongs in a category C if it is relatively similar to C's prototype. Prototype models are relatively inflexible; they discard a great deal of information that people use during concept learning (e.g., the number of exemplars in a concept (Homa & Cultice, 1984), the variability of features (Fried & Holyoak, 1984), correlations between features (Medin *et al.*, 1982), and the particular exemplars used (Whittlesea, 1987)).

Exemplar models instead represent concepts by their individual exemplars; a new item is assigned to a category C if it is relatively similar to C's known exemplars. Exemplar representations are far more flexible than prototype representations since they retain sensitivity to all of the information listed above. This flexibility often translates to increased categorization accuracy. For example, unlike prototype models, humans and exemplar models can learn some non-linearly separable categories as easily as linearly separable categories (Medin & Schwanenflugel, 1981). This capability is not limited to flat learning architectures; several researchers capture this flexibility in radial basis networks (e.g., Kruschke, 1992; Hurwitz, 1991).

While existing exemplar models are more flexible than prototype models, they are still not sufficiently flexible. We argue that people represent categories not only with category exemplars, but also with a set of specific weights associated with each exemplar's (or set of exemplars) attributes. The subject experiments described in Section 2 suggest that the weight given to an attribute depends on its exemplar's "neighborhood" in psychological space, where exemplars are assumed to be describable by their attributes' values. Our claim is that concepts are not represented simply by a set of attribute weights. Rather, an attribute's importance in similarity calculations depends on its *context* – the other attributes that are true for a particular exemplar. For example, the relative importance of the "date of next deadline" attribute for predicting membership in the "will work this weekend" category varies depending on the "upcoming computer downtime" attribute's value (e.g., when a deadline exists for the middle of the following week, one might be more likely to work during the preceding weekend when it is known that the computers will not be functioning on the days immediately preceding the deadline). Moreover, people can learn the importance of an attribute in concept-learning situations even when they have little guidance for assigning attribute weight settings. Since

[1] Other summary information may also be stored by more advanced prototype models; our concerns primarily target problems with "pure" prototype models. More accurately, we are interested in supporting the learning behavior displayed by the advanced exemplar models described in Section 3 regardless of the models' representation for categories (Barsalou, 1989).

people have almost no background information on the artificial stimuli used in the experiments described in Section 2, and since the exemplars in those experiments exhibit somewhat arbitrary regularities, we can be confident that our subjects are actively attending to the stimuli's regularities rather than applying knowledge that they previously acquired.

The work presented here has several precursors. Medin and Schaffer (1978) developed an exemplar model for representing concepts that was subsequently generalized by Nosofsky (1984; 1986). In turn, Aha and McNulty (1989) created a learning algorithm for Nosofsky's model and extended its selective attention mechanism to be a function of the target concept. We further augmented this learning model to include exemplar-specific weights; each exemplar in each concept was given its own set of attribute weights (Aha & Goldstone, 1990). This new model, named GCM-ISW, achieved faster learning rates and higher asymptotic performance than other models on artificial categorization tasks whose concepts were best modeled by using context-sensitive settings for attribute weights.

In Section 2, we extend our previous work by showing that human subjects are highly flexible in that they can selectively weight an attribute differently depending on the region of the instance space in which it is located. In Section 3, we show that GCM-ISW can fit these subjects' predictions better than two concept-learning systems with less flexible weighting schemes for warping the instance space. Like humans, GCM-ISW can allow the importance of attributes to be a function of its region of instance space.

Experiments on weighting attributes

An experiment was conducted to determine whether human subjects can learn categories that require attributes to be weighted differently for different category exemplars. That is, this experiment investigates whether subjects are constrained to weight attributes equally regardless of their context. This experiment also investigates whether subjects subsequently generalize their categories according to the attribute weights that they have learned. First, the subjects learn to distinguish category A from category B exemplars until they can accurately classify a set of *training* exemplars. The subjects are then given a set of *test* exemplars to classify. We can indirectly ascertain the weights that subjects assigned to the attributes by observing how these test exemplars were classified.

During training, 40 undergraduate subjects were told to categorize picture items, corresponding to exemplars, into category A or category B. These pictures varied along two dimensions: size of square and position of line in square. Each dimension is defined over eight evenly-spaced values. Square size varied from 2.0 cm to 7.5 cm. Position of line in square varied from the far left side to the far right side.

Subjects were presented with twelve training items, where half belonged to each category. The particular items shown to the subjects in Experiments 1 and 2 are shown in Figure 1. The first matrix shows the two groups of items in Experiment 1. Each cell in this matrix represents a possible stimulus item. For example, the bottom-leftmost cell represents the item *very small square with line on the far left side of the square*. The twelve items that were shown in the training stage were labeled A or B according to their category. The cluster of items in the top-right of the first matrix is characterized by relatively large squares with lines relatively far to the right. The other cluster has relatively small squares with lines further to the left. Line position was the more important dimension for distinguishing category A from category B items for the first cluster; items with the value six for line position belonged in category B whereas items with the value seven belonged in category A. Conversely, size was the more important dimension for the other cluster of items; items with a value of seven on the size dimension were exemplars of category B, while items with a value of six belonged in category A.

During training, after the twelve items' ordering was randomized, they were subsequently presented to the subjects on a Macintosh SE. For each item, the subject pressed A or B to indicate their category prediction. Subjects were told whether their classification was correct immediately after their response. Training continued until the subject performed four error-free classifications of the complete set of training items.

During testing, all 64 possible combinations of line position and square size were displayed to subjects in a random order. For each item, subjects indicated whether they believed the item belonged in category A or B. Only twelve of these items were previously shown to the subjects; the remaining 52 were novel items, and their placement in category A or B represent generalizations of these categories.

The results from the test stage of Experiment 1 are displayed in Figure 2. The number in each cell indicates the percentage of times that subjects placed the item into category B during testing. The percentages indicate fairly good retention of the items that were presented during training and widespread generaliza-

Experiment #1:

	1	2	3	4	5	6	7	8
1	.	.	.	.	.	.	.	.
2	.	.	.	.	.	B	Z	.
3	.	.	.	.	.	B	A	.
4	.	.	.	.	.	B	A	.
5	.	.	.	.	.	Y	A	.
6	.	A	A	A	X	.	.	.
7	.	W	B	B	B	.	.	.
8	.	.	.	.	.	.	.	.

Experiment #2:

	1	2	3	4	5	6	7	8
1	.	.	.	.	.	.	.	.
2	.	.	.	B	B	B	Z	.
3	.	.	.	Y	A	A	A	.
4	.	A	X	.	.	.	.	.
5	.	A	B	.	.	.	.	.
6	.	A	B	.	.	.	.	.
7	.	W	B	.	.	.	.	.
8	.	.	.	.	.	.	.	.

Figure 1: Training sets and critical test items for the two experiments. The horizontal and vertical axes denote (increasingly right) line positions and (decreasing) square size dimensions respectively. The categories of the training items are shown as A and B. The four critical test items per experiment are marked with one of $\{W, X, Y, Z\}$.

Average of the Subjects' Predictions:

	1	2	3	4	5	6	7	8
1	50	45	45	40	50	65	10	30
2	65	60	80	80	85	90	20	25
3	55	50	55	60	70	85	0	20
4	40	55	45	40	65	85	0	15
5	20	20	15	25	30	70	0	20
6	10	0	5	10	10	5	5	15
7	85	90	90	95	90	85	75	80
8	55	60	65	60	55	60	55	60

GCM-ISW's Predictions:

	1	2	3	4	5	6	7	8
1	67	74	77	78	78	76	34	31
2	50	63	72	76	78	78	33	30
3	31	43	58	69	74	74	29	28
4	20	26	37	52	64	69	26	24
5	17	19	23	32	46	57	23	21
6	21	21	22	27	37	48	26	25
7	74	75	77	78	78	74	53	61
8	78	78	78	78	76	72	42	49

Figure 2: The subjects' averaged predictions and GCM-ISW's probabilistic guess that the test items in Experiment 1 belong to category B.

tion of the training knowledge to the new items.

Particular test items of interest to us are labeled by the letters W, X, Y, and Z in Figure 1. These items were not presented during training. Their pattern of classification seems to confirm that subjects generalized their categories by differentially weighting the attributes for different items. For example, item W was categorized as an exemplar of category B by 90% of the subjects in Experiment 1, although it is as close to category A items as it is to category B items. Similarly, item X in Experiment 1 was categorized as a member of category A by 90% of the subjects. These results indicate that subjects strongly weight the size dimension in these categorizations. It is as if the subjects are *stretching* the vertical axis in this area of the space, so that the A and B items become separated by a greater psychological distance. However, the entire vertical axis is not stretched. Instead, it is selectively stretched in this single region of the space (i.e., the lower-left). Similarly, the horizontal axis is selectively stretched in the upper-right region; item Y was categorized as a B by 70% of the subjects while item Z was categorized as an A by 80% of the subjects, indicating that subjects considered line position to be more important than square size for categorizing items in this region. In summary, subjects generalized their concepts on the basis of the square size dimension for one cluster and on the basis of the line position dimension for the other cluster.

Experiment 2 replicates Experiment 1 with a relocation of the training items. One possible explanation of Experiment 1's results is that, perceptually, there was a bigger difference between size six and size seven squares than there is between size two and three squares and/or a relatively large perceptual difference between lines in positions six and seven. If this were true, then our generalization results could be explained without requiring that subjects learned to selectively weight dimensions in particular regions of the space. Experiment 2's results refute this possible explanation; if one assumed that there is a large perceptual difference between size six and size seven squares, then precisely the wrong prediction would be

Average of the Subject's Predictions:

	1	2	3	4	5	6	7	8
1	65	50	65	55	55	60	65	60
2	70	75	85	90	90	90	90	80
3	15	10	15	15	10	10	0	10
4	15	5	85	35	15	15	5	25
5	20	5	90	60	35	40	50	35
6	20	0	85	70	50	60	50	60
7	30	15	90	80	75	75	60	60
8	30	10	70	55	55	50	55	50

GCM-ISW's Predictions:

	1	2	3	4	5	6	7	8
1	47	39	69	75	77	78	78	78
2	59	50	72	77	77	77	76	74
3	24	25	46	36	27	23	21	21
4	22	22	58	49	36	28	23	21
5	25	26	68	65	55	42	32	25
6	28	30	75	74	70	61	48	36
7	30	31	76	77	76	72	65	54
8	29	31	74	77	77	76	74	68

Figure 3: The subjects' averaged predictions and GCM-ISW's probabilistic guess that the test items in Experiment 2 belong to category B.

GCM-SW's Predictions:

	1	2	3	4	5	6	7	8
1	55	64	71	74	72	67	59	51
2	44	53	64	73	69	63	47	39
3	33	38	49	51	38	31	25	27
4	28	27	43	47	39	32	27	25
5	31	33	59	57	50	41	34	30
6	37	40	66	67	62	53	45	38
7	48	59	76	74	70	63	56	48
8	58	66	74	76	74	69	63	57

Figure 4: GCM-SW's probabilistic guess that the test items in Experiment 2 belong to category B.

made for Experiment 2, where item W is now placed in category A based on its line position. More specifically, 85% of the subjects categorized both items W and Y as members of category A, whereas items X and Z were predicted to belong to category B by 85% and 90% of the subjects respectively. The results for Experiment 2 are summarized in Figure 3.

Protocols were also obtained from the subjects. In Experiment 1, the modal protocol, given by 15 out of the 20 subjects, can be expressed by the following subject's statement:

> I looked at the size of the square. If it was big, then I looked at where the bar was. If it was a little further to the right, then I put it in A. Otherwise, I put it in B. If the square was small, I looked carefully at its size. A squares were slightly bigger than B squares.

This protocol reveals a two-step process whereby a subject (1) determines the region in which an item belongs and (2) focuses on the particular dimension that is important for that region.

Simulations on weighting attributes

Three exemplar-based process models were evaluated in simulations for their ability to fit the subjects' responses. These models, GCM-NW, GCM-SW, and GCM-ISW, were previously described in (Aha & Goldstone, 1990), are all derived from Nosofsky's (1986) Generalized Context Model (GCM), and differ only in how they weight attribute dimensions. The least flexible, GCM-NW, weights all attributes equally. GCM-SW instead uses a single set of attributes and, in keeping with Nosofsky's *attention-optimization hypothesis*, tunes attribute weights so as to optimize categorization performance. Finally, GCM-ISW is an extension of GCM-SW that maintains a separate set of attribute weight settings with each stored exemplar.

These models process training items incrementally and, for each item x, compute an estimate of the probability that x is a member of each category C as follows:

$$\text{Probability}(x \in C) = \frac{\sum_{y \in S_C} \text{Similarity}(x, y)}{\sum_{y \in S} \text{Similarity}(x, y)},$$

where S_C is category C's stored exemplars and S is the set of all stored exemplars. Similarity is defined as:

$$\text{Similarity}(x, y) = e^{-c\text{Distance}(x,y)},$$

where

$$\text{Distance}(x, y) = \sqrt{\sum_i f(i, x, y) \times (x_i - y_i)^2},$$

and where i ranges over the set of attributes used to describe the exemplars, parameter c's setting (fixed at 10 in our experiments) determines the slope of the

exponential decay, and function f determines the normalized weight for attribute i (i.e., $\sum_i f(i,x,y) = 1$ and $\forall i\{0 \leq f(i,x,y) \leq 1\}$).

Function f is a constant function for GCM-NW. For GCM-SW, $f(i,x,y) = w_i$, which is an estimate of the conditional probability that two exemplars will be in the same category given that they have high similarity and highly similar values for attribute dimension i. Weights are initially equal and their settings are updated after each training item is presented via a strategy akin to the delta rule (Rumelhart, McClelland, & the PDP Research Group, 1986).[2] Finally, GCM-ISW's function f combines the category-specific weight settings learned by GCM-SW with a separate set of weight settings stored with exemplar y. Exemplar-specific weight settings are updated in the same manner as category-specific weights except that they are only updated for similarity computations involving their exemplar. More specifically, $f(i,x,y)$ interpolates between the category-specific weight for attribute i and y's exemplar-specific weight for i. This value is more similar to the exemplar-specific setting when $|x_i - y_i|$ is small and more similar to the category-specific setting when this difference is high.

GCM-NW, GCM-SW, and GCM-ISW have three, four, and six free parameters respectively. Informal manual searches were used to find values for these parameters that allowed the models to perform well: 10 for c, which determines the slope of the exponential decay defining similarity; 1 for the GCM's concept bias parameters; 0.01 for GCM-SW's and GCM-ISW's learning rate parameter for updating category-specific weights; 0.1 for GCM-ISW's similar parameter for exemplar-specific weights; and 0.5 for GCM-ISW's parameter for combining exemplar- and category-specific weights in function f. GCM-ISW's additional parameters certainly contributed to its superior performance. However, alternative values for the other models' parameters would not affect their relative behavior because the concepts were equally probable during training and different slopes would still not allow GCM-NW and GCM-SW to locally warp the instance space.

These models were trained and tested in the same way as the subjects except that their items were represented as two-dimensional vectors and they yield estimates of the *probability* that items are members of category B rather than a binary categorization prediction.

[2] Briefly, the magnitudes of weight changes are a decreasing function of Similarity(x,y) and an exponentially decreasing function of $|x_i - y_i|$. Weight settings are increased when x and y are in the same category and otherwise are decreased.

The testing results for GCM-ISW in Experiment 1 are summarized earlier in Figure 2 alongside the subjects' average predictions. Fisher's method for converting correlations (r) to Z-scores was used to evaluate the fits of each model to the subject data. The correlation between GCM-ISW's results and the averaged subject data for the 64 test items was 0.81 and 0.85 for Experiments 1 and 2 respectively. GCM-SW's was 0.66 for both experiments and GCM-NW's was 0.65 and 0.68. GCM-ISW's results correlated significantly better with the subject data from the first experiment than did GCM-NW ($Z = 2.75, p < 0.01$) and GCM-SW ($Z = 2.61, p < 0.01$). This is also true for Experiment 2's results (i.e., ($Z = 3.62, p < 0.0005$) and ($Z = 3.36, p < 0.002$) respectively). For example, visual inspections help to confirm that GCM-SW's predictions for Experiment 2, shown in Figure 4, are not as similar to the subjects' predictions as are GCM-ISW's, as shown in Figure 3.

GCM-ISW's correlations with the subjects' averaged responses for the four critical test items were significantly better than GCM-SW's and GCM-NW's for both experiments (i.e., $Z(1) = 1.92, p < 0.1$; $Z(1) = 2.53, p < 0.025$ and $Z(1) = 3.05, p < 0.0025$; $Z(1) = 2.28, p < 0.025$ respectively). More specifically, GCM-ISW's correlations for these two sets of four test items were 0.97 and 0.95 respectively. GCM-SW's respective correlations were 0.17 and -0.84 while GCM-NW's were -0.42 for both experiments. GCM-ISW's categorization predictions matched the predictions made by the majority of subjects on all eight critical test items, whereas GCM-SW agreed on only two and GCM-NW on only four.

In summary, GCM-ISW provides a better fit to the subject data than do the other models. Its combination of category-specific and exemplar-specific attribute weights captures the context sensitivity of attribute importance in these experiments. Thus, these results support our claim that *a psychologically plausible learning algorithm's selective attention processes must be a context-dependent function*; a simple strategy of using one weight per attribute will not necessarily provide optimal fits to subject data.

Discussion

Many other exemplar models of human concept formation can "locally" stretch the input space. For example, Nosofsky, Clark, and Shin (1989) described a model that associates a weight with each value of each dimension. However, this strategy is less flexible than GCM-ISW's; it constrains items sharing an attribute's value to also share its weight set-

ting. Medin and Edelson (1988) proposed a process model similar to GCM-ISW that uses exemplar-specific attribute weights to account for subjects' context-specific sensitivity to base rate information during categorization tasks. However, their model does not ensure that exemplar-specific weights are used only in a local region of the instance space; they may be used to help classify dissimilar items. This constraint should always be applied to models with localized weighting schemes. Medin and Shoben (1988) investigated an *exemplar-directed* attribute-weighting scheme that distinguishes between directions along numeric-valued attribute dimensions. We plan to evaluate an extension of GCM-ISW that incorporates this increased flexibility. We also plan to study models with *region-specific* weighting schemes, in which a region's weights are abstracted so as to specify the relative importance of attributes for similarity decisions within a small region of the instance space. Such models blur the distinction between rule- and exemplar-based models since they use both exemplars and rule-like abstractions derived from them to guide categorization decisions. Furthermore, our model will vary the degree to which abstraction is performed in a region-specific manner, thus increasing its flexibility to represent complex concepts.

Several other researchers have also advocated that psychologically plausible process models should categorize items in a context-sensitive manner (e.g., Barsalou & Medin, 1986; Tversky, 1977). We believe that many future models will incorporate a context-sensitive categorization capability and that they will continue to fit subject data significantly better than models that do not support this flexibility.

References

Aha, D. W., & Goldstone, R. L. (1990). Learning attribute relevance in context in instance-based learning algorithms. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 141–148). Cambridge, MA: Lawrence Erlbaum.

Aha, D. W., & McNulty, D. (1989). Learning relative attribute weights for independent, instance-based concept descriptions. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 530–537). Ann Arbor, MI: Lawrence Erlbaum.

Barsalou, L. W., & Medin, D. L. (1986). Concepts: Static definitions or context-dependent representations? *Cahiers de Psychologie Cognitive, 6*, 187–202.

Barsalou, L. W. (1989). On the indistinguishability of exemplar memory and abstraction in category representation. In T. K. Srull & R. S. Wyer (Eds.), *Advances in social cognition*. Hillsdale, NJ: Lawrence Erlbaum.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 234–257.

Homa, D., & Cultice, J. (1984). Role of feedback, category size, and stimulus distortion on acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learming, Memory and Cognition, 8*, 37–50.

Hurwitz, J. B. (1991). Learning rule-based and probabilistic categories in a hidden pattern-unit network model. Unpublished manuscript. Harvard University, Department of Psychology, Cambridge, MA.

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

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.

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

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 15*, 39–57.

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.

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

Rumelhart D. E., McClelland, J. L., & The PDP Research Group (Eds.), (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

Whittlesea, B. W. A. (1987). Preservation of specific experiences in the representation of general knowledge. *Journal of Experimemtal Psychology: Learning, Memory, and Cognition, 13*, 3–17.