Learning Attribute Relevance in Context in Instance-Based Learning Algorithms

David W. Aha
Dept. of Information & Computer Science
University of California, Irvine
Irvine, CA 92717
aha@ics.uci.edu

Robert L. Goldstone
Department of Psychology
University of Michigan
Ann Arbor, MI 48109
rob_goldstone@ub.cc.umich.edu

Abstract

There has been an upsurge of interest, in both artificial intelligence and cognitive psychology, in *exemplar-based* process models of categorization, which preserve specific instances instead of maintaining abstractions derived from them. Recent exemplar-based models provided accurate fits for subject results in a variety of experiments because, in accordance with Shepard's (1987) observations, they define similarity to degrade exponentially with the distance between instances in psychological space. Although several researchers have shown that an attribute's relevance in similarity calculations varies according to its *context* (i.e., the values of the other attributes in the instance and the target concept), previous exemplar models define attribute relevance to be invariant across all instances. This paper introduces the GCM-ISW model, an extension of Nosofsky's GCM model that uses *context-specific* attribute weights for categorization tasks. Since several researchers have reported that humans make context-sensitive classification decisions, our model will fit subject data more accurately when attribute relevance is context-sensitive. We also introduce a process component for GCM-ISW and show that its learning rate is significantly faster than the rates of previous exemplar-based process models when attribute relevance varies among instances. GCM-ISW is both computationally more efficient and more psychologically plausible than previous exemplar-based models.

1. Introduction

Several studies have shown that the Context Model (Medin & Schaffer, 1978) and the Generalized Context Model (GCM) (Nosofsky, 1986; 1987), two exemplar-based models of categorization, provide excellent fits for subject data from a wide variety of experiments. These models remove the assumption that attributes have equal relevance in similarity computations by introducing a parameter (attribute weight) for each attribute, whose value is determined by some exterior attention mechanism. Humans are hypothesized to selectively attend to attributes to optimize their classification behavior (Nosofsky, 1986). These models differ from previous exemplar models (e.g., Reed, 1972) in that they define similarity to decrease exponentially with psychological distance, in accordance with Shepard's (1987) numerous empirical observations on stimulus generalization. However, these models ignore evidence that attribute relevance varies depending on the context of the classification task (Tversky, 1977; Barsalou, 1982; Roth & Shoben, 1983; Medin & Edelson, 1988). Therefore, they cannot be expected to provide accurate fits when attribute relevance varies according to context, which occurs frequently in real-world classification tasks.

For example, consider the problem of predicting whether a pro-life politician will endorse proposed legislation on abortion rights. As with most real-world categorization tasks, some attributes should be given more attention than others. In this case, dimensions such as

"past voting record" should be weighted more than dimensions such as "height." Relative attribute relevance differs depending upon the prediction task (Aha & McNulty, 1989; Aha, 1989) (i.e., "past voting record" is far less relevant than "height" when predicting the ability to dunk a basketball). However, an attribute's relevance to a categorization task often also depends on its *context* – the values of the other attributes in an instance. For example, the relevance of the "past voting record" attribute will be low if the "percentage of pro-choice constituency" attribute has a high value (due to pressure from pro-choice political action groups). However, it will be high if the "seek re-election" attribute value is "false", which diminishes the influence of political action groups. Context sensitive attribute weights are required to derive an appropriate psychological space and satisfy the attention-optimization hypothesis when attribute relevance is context-dependent.

In this paper, we introduce the *GCM-ISW* (**I**nstance-**S**pecific **W**eights) model, an extension of the GCM that adds a set of attribute weight parameters for each instance for each target concept. Section 2 describes evidence that the GCM-ISW is more *computationally efficient* (i.e., records significantly faster learning rates) than previous exemplar-based process models when attribute relevance is context-dependent. Several researchers have reported evidence that humans make context-sensitive classification decisions when attribute relevance is dependent on context. In Section 3, we review this evidence and discuss alternative weighting schemes for exemplar-based models.

2. Instance-Based Learning Algorithms

This section describes a sequence of four, comprehensive *instance-based learning* (IBL) algorithms, which are exemplar-based process models. The attribute weights in the first (and simplest) model, named *GCM-NW* (**N**o **W**eights), are fixed to be equal. We describe evidence that a process model for the GCM, named *GCM-SW* (**S**ingle set of attribute **W**eights), learns significantly faster than the GCM-NW when the relevance to classification judgements varies among attributes. The third model, named *GCM-MW* (**M**ultiple sets of attribute **W**eights), employs a separate set of attribute weights per target concept and learns significantly faster than the GCM-SW when attribute relevance varies among target concepts. The final model, GCM-ISW, employs a separate set of attribute weights for each instance for each target concept. We present evidence that it learns significantly faster than GCM-MW when attribute relevance varies among instances.

2.1 GCM-SW: Learning Attribute Relevance

IBL algorithms input a sequence of training instances, drawn from an n-dimensional *instance space*, where n is the number of attributes used to describe each instance. A subset $p < n$ of these attributes, called *predictors*, are used to predict values for the remaining $(n - p)$ *targets*. In this paper, we assume that predictors have numeric values and target attributes have binary values: "positive" and "negative."[1] Positive target values are understood to be members of the *target concept*. For each target, IBL algorithms yield one *concept description* which contains the processed training instances and a set of attribute weight settings. Given an instance x and its similarity with each instance in target a's concept description, IBL algorithms can predict whether x is a member of concept a.

[1] Stanfill and Waltz (1986) describe an interesting IBL algorithm for symbolic-valued predictors. Kibler, Aha, and Albert (1989) address the issue of numeric-valued targets.

IBL algorithms use a *similarity function*, defined over the predictor attributes, to compute the similarity of the instance to be classified with the previously processed training instances. GCM-SW's similarity function is $\text{similarity}(x, y) = e^{-\text{distance}(x,y)}$, where

$$\text{distance}(x, y) = s\sqrt{\sum_{i=1}^{p} w_i(x_i - y_i)^2},$$

where parameter s (set to 10 in all our simulations) is GCM-SW's parameter that determines the slope of the exponential decay and w_i is GCM-SW's weight for attribute i. Values for attribute weights are always initialized to $\frac{1}{p}$, range in $[0, 1]$, and are normalized to sum to 1.

Given these similarities, a *memory updating function* modifies the attribute weights for the instances in a's concept description and, afterwards, always adds x to a's description.[2] The GCM-SW training algorithm is: (where cd_a is target a's concept description)

1. $\text{cd}_a \leftarrow \emptyset$
2. FOR EACH $x \in$ training set DO
   2.1 FOR EACH $y \in \text{cd}_a$: compute similarity(x, y)
   2.2 FOR EACH $y \in \text{cd}_a$: FOR EACH predictor attribute i: adjust_weight(i, x, y, a)
   2.3 $\text{cd}_a \leftarrow \text{cd}_a \cup \{x\}$

Attribute weights denote the estimated relevance of an attribute for a categorization task. Each predictor i's weight is computed using a function of the estimated conditional probability that two instances will have the same class, given that their similarity is high and the difference of their values for i is small. If we denote this probability at time t as $\text{Pr}_i(t)$, then the attribute weight for i after t training instances have been processed is $\text{Pr}_i(t) - (1 - \text{Pr}_i(t))$. Adjust_weight updates estimates of conditional probability as follows:

$$\text{Pr}_i(t+1) = \text{Pr}_i(t) + (r - \text{Pr}_i(t)) \times \text{similarity}(x, y) \times e^{-s|x_i - y_i|} \times \rho,$$

where Boolean variable r is 1 only if x_a equals y_a and ρ is a learning rate parameter, which is set to 0.01 for GCM-SW and GCM-MW in our simulations. The size of the update to i's conditional probability increases exponentially with linear decreases in both distance(x, y) and $|x_i - y_i|$. Therefore, attribute i's weight is most strongly influenced by highly similar instances with similar values for i.

The classification accuracy of our IBL algorithms is measured using a *classification function*, which inputs the computed similarities for target a and generates a class prediction (i.e., "positive" or "negative"). The probability that instance x will be a member of concept a is estimated as follows:

$$\text{Pr}(x_a = \text{"positive"}) = \frac{\sum_{y \in \text{cd}_a} \text{similarity}(x, y) \times (y_a = \text{"positive"})}{\sum_{y \in \text{cd}_a} \text{similarity}(x, y)}$$

Instance x is predicted to be a member of concept a only if this value is above 0.5. All the IBL algorithms in this paper use the following testing algorithm (for each target attribute

[2] Aha, Kibler, and Albert (in press) analyze IBL algorithms that significantly reduce storage requirements.

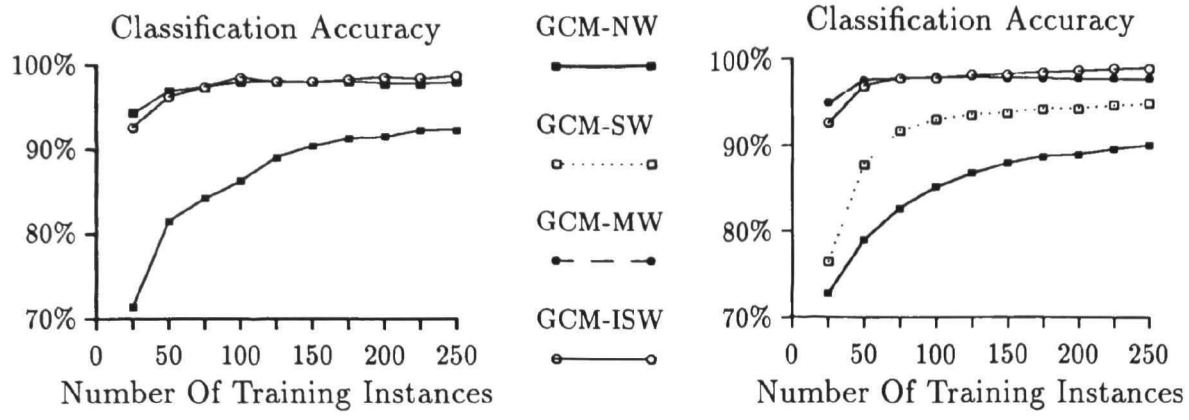

Figure 1: Learning curves for the four IBL algorithms. Left: GCM-NW learns slowly when attributes have different relevance. GCM-SW and GCM-MW behave identically in this simulation since there is only one target concept. Right: GCM-SW learns slowly when each attribute's relevance differs among target concepts. All our curves are averaged over 20 pairs of training and test sets with 250 and 100 instances respectively. Values for predictor attributes are selected randomly from [0, 1] according to a uniform distribution.

a): (1) compute the current training instance x's similarity to the instances in a's concept description, (2) compute the probability that x_a is "positive", and (3) output "positive" for this classification if this probability is above 0.5 (otherwise, output "negative").

GCM-SW's attribute weights are useful when attribute relevance varies among predictors. To show this, we compared its performance with the performance of GCM-NW, whose weights remain fixed with value $\frac{1}{p}$. GCM-NW learns slowly when attribute relevance differs among the predictors. The graph in the left of Figure 1 shows the average learning curves for a simulation with one target concept and ten predictors, only one of which was relevant. Target concept members were defined to be those whose relevant attribute's value was greater than 0.5. As expected, GCM-SW's average accuracy (measured across the ten applications to the test set per trial) is significantly greater than GCM-NW's ($t(19) = 4.54, p < 0.001$). However, since the GCM-SW model uses the same setting of attribute weights for all targets, it performs relatively poorly when the relative relevance of attributes differs greatly among target concepts (Aha & McNulty, 1989) or when relative attribute relevance varies among instances. The right-hand graph in Figure 1 shows the average learning curves when the artificial domain is extended to contain an additional three target concepts, where each of the four target concepts have a single (different) relevant predictor. GCM-SW's learning curve rises slowly because it is unable to learn concept-dependent attribute relevances: its weights for the four relevant attributes each converge to 0.25. GCM-SW's average classification accuracy is significantly lower than GCM-MW's ($t(19) = 5.33, p < 0.001$).

2.2 GCM-MW: Learning Concept-Dependent Attribute Relevance

GCM-MW's concept-dependent similarity function is $\text{similarity}(a, x, y) = e^{-\text{distance}(a,x,y)}$, where w_{a_i} denotes the weight of attribute i for target concept a in

$$\text{distance}(a, x, y) = s\sqrt{\sum_{i=1}^{p} w_{a_i}(x_i - y_i)^2}.$$

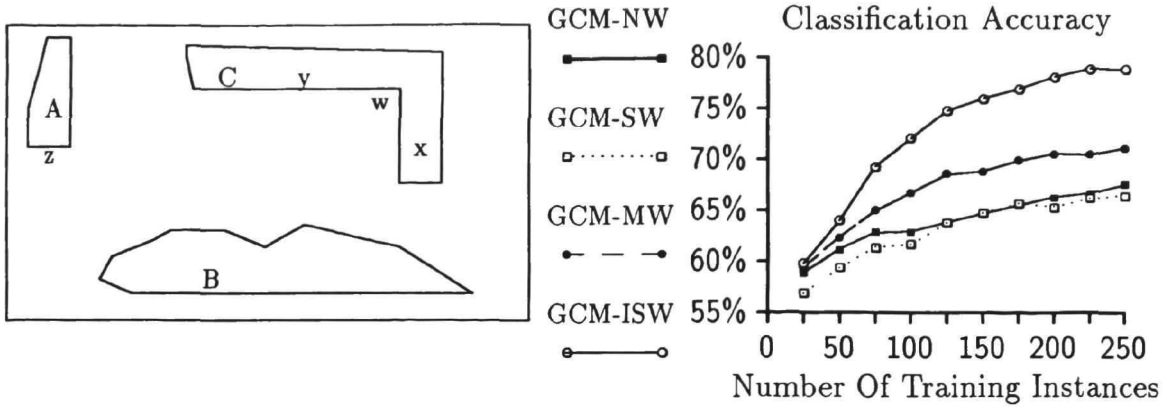

Figure 2: Left: Attribute relevance can vary among instances. Right: GCM-ISW works best in such situations.

The GCM-MW model will outperform the GCM-SW model when attribute relevance varies among target concepts. However, GCM-MW's assumption that an attribute's relevance is invariant across all instances is easily violated. For example, attribute relevance can vary among a concept's disjuncts. Furthermore, it can also vary *within* a disjunct. Figure 2 displays a two-dimensional domain containing three disjuncts of a single target concept. The horizontal attribute is more relevant than the vertical for disjunct A: small perturbations in the horizontal's values will more frequently change disjunct membership status than will perturbations in the vertical's values. The vertical attribute is more relevant for B while C's attributes are approximately equally relevant. However, attribute relevance differs greatly among instances. For example, although both attributes are relevant for classifications made by instance w, the horizontal attribute is more relevant for x and less relevant for y. Finally, z's vertical attribute is more relevant. GCM-MW's learning rate can be significantly reduced when attribute relevance varies among instances. Figure 2 displays the average learning curves when the learning task is changed so that each of the four concepts is defined by a set of five disjuncts. In this case, each disjunct is defined by a single relevant attribute and each attribute in the domain is relevant to exactly two disjuncts overall. The threshold values for inclusion in a disjunct were below 0.14 for the disjuncts of the first two target concepts and above 0.86 for the latter two target concepts. GCM-MW's average accuracy is significantly lower than GCM-ISW's ($t(19) = 3.85, p < 0.002$).

2.3 Learning Context-Sensitive Attribute Relevance

GCM-ISW differs from GCM-MW in that it learns *instance-specific* attribute weights, one for each ⟨attribute,instance,target⟩ triplet. This provides greater flexibility than found in GCM-MW: GCM-ISW removes the assumption that attribute relevance is invariant among a target concept's saved instances.

GCM-ISW's instance-specific weights can be easily misapplied. For example, if the only relevant attribute for instance z in Figure 2 is the vertical attribute, then z will appear to be very similar to x, which is located far from z in this instance space. Therefore, instance-specific weights should be used only when the instance being classified is highly similar to the classifying instance. GCM-ISW solves this problem by learning both concept-dependent weights

(as is done in GCM-MW) and a separate set of instance-specific weights. (Adjust_weight updates each saved instance's attribute weights when classifying each subsequently presented training instance.) GCM-ISW's similarity function then combines these two sets of weights to compute the *context*-specific similarity of two instances as follows:

$$\text{distance}(a, x, y) = s\sqrt{\sum_{i=1}^{p} \text{combine\_weights}(a, x, y, i) \times (x_i - y_i)^2}.$$

When computing the similarity of a new instance x to previously processed instance y, combine_weights calculates attribute i's *context-specific* weight as follows:

$$\text{combine\_weights}(a, x, y, i) = (w_{a_i}(y) \times \text{scale\_factor}) + (w_{a_i} \times (1 - \text{scale\_factor})),$$

where scale_factor $= (1 - |x_i - y_i|)^c$, $w_{a_i}(y)$ is i's attribute weight for saved instance y, and c is a combination parameter that determines the relative impact of the concept-dependent and instance-specific attribute weights in calculating the context-sensitive weight.[3] Combine_weights uses instance-specific weights more confidently when the difference of the values for i is small. This reduces the frequency with which instance-specific weights are used when the distance between instances is large. After GCM-ISW computes similarities, it updates the conditional probabilities and attribute weights for both its concept-dependent and its instance-specific weights. GCM-ISW performed significantly better than the other models in the third simulation and performed as well as GCM-SW and GCM-MW in the first and second simulations respectively.

3. Discussion: Supporting Evidence and Alternative Models

While formal psychological models involving context-specific weight learning do not exist, there is a plethora of psychological data suggesting the existence of such specific weighting systems. Nosofsky's (1986) GCM model treats attribute weights as parameters that can be assigned experimentally-derived values to accurately fit subject data. However, more flexible weighting systems, namely those that employ context-sensitive weights, are required to accurately fit subject data and increase learning rate when attribute relevance varies among instances. These weights can also be used to decrease storage requirements: attributes with low relevance can be discarded without sacrificing classification accuracy (Smith & Medin, 1981). Aha (1989) described simulations of IBL algorithms that drop both attributes *and* instances and, simultaneously, increase learning rates for real-world classification tasks.

Many researchers agree that models of categorization should be context-sensitive. Roth and Shoben (1983) and Barsalou (1982) argue that an instance's context influences its perceived typicality and determines which of its attributes receives attention. For example, Barsalou noted that, while some attributes of "basketball" (e.g., "round") are always salient, others (e.g., "floats") only become salient (i.e., quickly retrieved) in contexts involving water. This provides psychological support for the GCM-ISW model: people on a luxury liner attend more to the "floats" attribute when the ship is sinking (to judge whether objects are members of the "can support me in the water" category) than when it is in port. Goldstone, Medin,

[3] We used $c = 0.5$ for our simulations. We also set GCM-ISW's learning rate parameter to be higher (0.1) when it updates instance-specific weights. This is needed because, given any one training instance, few other training instances are highly similar to it. However, when updating concept-dependent weights, there will be several highly similar pairs of instances.

and Gentner (in press) argue that, when comparing instances, the influence that one attribute has depends on the other attributes that are shared by the instances. Medin and Edelson (1988) also suggested using context-specific attribute weights. When an instance is correctly classified, their proposed process model assigns high relative weights to the attributes shared by the classifying instance and the instance being classified. Misclassifications result in assigning higher weights to attributes that are *not* shared by these two instances.

Several alternative weighting schemes have been proposed for exemplar-based process models. Nosofsky, Clark, and Shin (1989) considered *value-specific* weighting algorithms. However, these are not as flexible as instance-specific weighting algorithms: value-specific weights for some attribute i will not work well when i's relevance varies over instances that have the same value for i. In another example, Medin and Shoben (1988) present examples that suggest an *instance-directed* attribute-weighting scheme, whereby the influence of one attribute depends on the other attributes that are present. For example, while "White" is more similar to "Gray" than is "Black" for the attribute "hair," exactly the opposite pattern emerges with the attribute "clouds." This suggests extending the instance-specific weighting method to distinguish between directions along attribute dimensions. For instances of hair, the gray-black distance is widened while the gray-white distance is reduced. In any case, a single predefined weight for the "color" dimension will not survive changes of context.

The GCM-ISW model adds an enormous number of parameters into the GCM model. Although GCM-ISW increases learning rate, its additional parameters are not needed when attribute relevance remains constant across the entire dimension. We are currently developing a more elaborate IBL algorithm that can learn which parameters should be permanently fixed without need for subsequent attention. The algorithm would initially assume that all dimensions are weighted equally for all categories. If this assumption does not yield sufficiently fast learning rates, then the system would relax its assumptions and allow an attribute's weight to vary across categories. The assumption that weights are fixed across instances could also be automatically relaxed. Shifts in the target concept description could lead to more or less specific weighting algorithms in attempts to maximize classification accuracy while minimizing the number of unique weights that are postulated.

IBL algorithms that learn context-specific attribute weights resemble rule-based learning algorithms. By weighting dimensions selectively on the basis of their category diagnosticity, the instance-based systems are qualitatively distinguished from the simple storage of instances in a "raw form." Although instance information is not discarded, it is selectively emphasized. This representation is similar to that used for rules. For example, consider the concept of legal-sized suitcases (i.e., those with lengths less than five feet). An instance-directed weighting algorithm could learn a high weight for $4'\,9''$ in the positive direction and a low weight in the negative direction for legal-sized suitcases. This is similar to the rule "if $4'\,9''$ or less, then legal-sized luggage, otherwise illegal."

4. Conclusion

Results from simulations suggest that previous exemplar models that selectively weight attribute dimensions, while better than no selective weighting at all, can be improved by representing context-sensitive attribute weights. We introduced GCM-ISW, an extension of Nosofsky's (1986) GCM model that learns context-sensitive weights by combining concept-

dependent and instance-specific attribute weights. Our results with simulations using a process model for the GCM-ISW show that its learning rate is significantly faster than the learning rates of previous process models for exemplar-based models (Aha & McNulty, 1989). We plan to show that the GCM-ISW model will fit subject data more accurately than will a process model for the GCM when attribute relevance varies among instances.

Acknowledgements

We would like to thank Marc Albert, Dale McNulty, Douglas Medin, and Mike Pazzani for providing comments on an earlier draft of this paper.

References

Aha, D. W. (1989). Incremental, instance-based learning of independent and graded concept descriptions. In *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 387–391). Ithaca, NY: Morgan Kaufmann.

Aha, D. W., Kibler, D., & Albert, M. K. (in press). Instance-based learning algorithms. *Machine Learning.*

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 Associates.

Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition, 10,* 82–106.

Goldstone, R. L., Medin, D. L., & Gentner, D. (in press). Attributes, relations, and the non-independence of features in similarity judgments. *Cognitive Psychology.*

Kibler, D., Aha, D. W., & Albert, M. (1989). Instance-based prediction of real-valued attributes. *Computational Intelligence, 5,* 51–57.

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

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

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 87–108.

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

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

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

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

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Stanfill, C., & Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM, 29,* 1213–1228.

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