Learning Overlapping Categories*

Joel D. Martin
School of Information and Computer Science
Georgia Institute of Technology
Atlanta, Georgia 30332
joel@gatech.edu

Abstract

Models of human category learning have predominately assumed that both the structure in the world and the analogous structure of the internal cognitive representations are best modeled by hierarchies of disjoint categories. Strict taxonomies do, in fact, capture important structure of the world. However, there are realistic situations in which systems of overlapping categories can engender more accurate inferences than can taxonomies. Two preliminary models for learning overlapping categories are presented and their benefit is illustrated. The models are discussed with respect to their potential implications for theory-based category learning and conceptual combination.

1 Introduction

The natural world can be neatly organized into a hierarchy of disjoint categories. A platypus is a mammal, not a bird; and mammals are animals, not plants. Artificial categories can also be placed in a strict hierarchy. Something that is a balpeen hammer is not a claw hammer. Similarly, it is a hammer, not a saw; and a hand tool, not a power tool. Each category, whether natural or artificial, can be distinguished from its alternatives by its distinct intensional description, which might be represented as lists of feature frequencies or a list of instances. Taxonomies such as these are elegant, economical, and seem to be the most suitable representation for many natural and artificial categories.

Presumably, if humans learn and use taxonomic structures, they would efficiently and accurately characterize the regularities in the world. The assumption that they do so has been an extremely fruitful seed for generating models of human category learning. So fruitful, in fact, that alternate structures of categories have been neglected until recently. Structures of overlapping categories can actually be superior to taxonomies when there are multiple equally good ways to partition a set of experiences. In these circumstances, non-disjoint categories permit faster learning and more economical storage than is possible with disjoint categories.

If one assumes the world contains such domains and that humans are optimally adapted to their environment as suggested by Anderson (1988), then overlapping category structures should motivate better models of human behavior. Indeed, some structures of overlapping categories suggest preliminary methods for modeling theory-based concept learning (Murphy & Medin, 1985) and conceptual combination (Smith & Osherson, 1984).

*This research was supported in part by a Georgia Institute of Technology Stelson grant and NIH grant 7R23HD20522-03.

2 Taxonomies

The purpose of taxonomies is to maximize the probability of correct inference without requiring the retrieval of large sets of past experiences for every prediction. They compactly summarize predictive relationships. However, there are domains for which taxonomies not only do not maximize the probability of correct inferences, but also do not compactly represent the regularities. These are domains that permit multiple different, but equally predictive taxonomies. Each of these taxonomies allows roughly the same number of correct inferences, but they differ with respect to which attributes or aspects of experiences about which they are most informative.

2.1 Disjoint categories

A set of objects or events can be split into disjoint subsets in many ways. Given a collection of a dog, a cat, a goldfish, and a whale, one split may group the dog with the fish and the cat with the whale. Another may assign each individual to a separate subset, and yet another may divide the mammals from the fish. Of these many possibilities the one that permits the most correct inferences is preferred. Any collection of categories is only useful to the extent that category membership permits accurate predictions. For instance, knowing that a particular object is a bird allows probable predictions about body covering, food, and habitat.

This notion has been formalized for a single set of disjunctive categories by both Gluck and Corter (1985) and Anderson (1990). For example, Gluck and Corter derive that the probability of making correct inferences is equal to the sum, over all categories, of the probability of the category multiplied by the square of the probability of a particular value given the category:

$$P(correct) = \sum_k P(C_k) \sum_j \sum_i P(A_j = V_{ij} \mid C_k)^2 \qquad (1)$$

This measure could be applied by performing every possible partitioning and choosing the one with the largest $P(correct)$. This would be horrendously expensive, though, because the number of possible partitionings grows exponentially with the number of categories. A more practical approach accepts a set of experiences in a particular order and deals with each instance in turn. Each instance can be used to update the intension of a category. It is added to the category that allows the greatest rise in the probability of correct inference. This is a hill-climbing search toward the optimal partitioning (Fisher, 1987).

Gluck and Corter (1985) originally intended their measure to be a predictor of the basic level in a taxonomy (eg. Rosch et al., 1976). However, it or a similar measure can be used recursively to form complete taxonomies, as was suggested by Fisher (1987) and Anderson (1990).

Gluck and Corter's measure will be used below as a paradigmatic example of a model for learning disjoint categories. As well, Fisher's (1987) method for generating hierarchies will be assumed. The generated hierarchies, unlike the disjoint category measure, are not necessarily close to optimal. The technique produces performance that is below the theoretical maximum in some situations because it attempts to maximize the predictive work done at each layer. However, there are no formal theories for acquiring optimally predictive hierarchies, so this heuristic method will suffice.

2.2 The trouble with taxonomies

Fisher's (1987) COBWEB system, which is rooted in Gluck and Corter's (1985) measure, has been quite successful at acquiring taxonomies that allow accurate inferences. COBWEB has been applied to both real and artificial domains. Given this success, it is valuable to identify what

shortcomings COBWEB and related models might have, if any. As well, if shortcomings are found, it is essential to demonstrate that the conditions in which taxonomies fail occur frequently in human experience. However, if such conditions are very rare, humans would do well enough with taxonomies and not require any alternative model.

2.2.1 The problems

All of the difficulties with taxonomies that will be discussed here occur when the domain to be learned allows many different, but equally useful taxonomies. These domains will be referred to as *cross-classification* domains. For example, a list of athletes may specify their height, weight, eye color, and hair color. For these athletes, the values for height and weight are mutually predictive, as are the values for eye and hair color. Height though, is only randomly related to eye and hair color, and similarly for weight. Taller athletes are heavier and blue-eyed ones tend to be blond, but tall athletes do not have a characteristic hair color.

A system that learns disjoint categories in this domain has several choices. First, it could choose to partition the instances on the basis of height and weight or on the basis of eye and hair color. This, however, limits the probability of correct inferences by completely ignoring one of the predictive relationships. Alternatively, it could try to balance the two relationships without storing every instance. This again loses some predictive information. In order for the categories to capture one relationship, it must jeopardize the accuracy of the other.

A third possibility is that the system could simply create a category for each combination of values for the domain. This results in no timing or storage benefit for categories over individual instances, but does achieve better prediction than the above proposals. This method will be generally inefficient for domains in which there are multiple clusters of mutually predictive values, because there is potentially an exponential number of categories that must be stored. More importantly, though, each of these categories must be separately learned. This requires that every combination must be seen before the domain is well learned. This is likely too restrictive, because humans, at least, do not have to have seen a striped apple to be able to make inferences about it.

A final possibility is that the system could choose one relationship for each level of the hierarchy, such that, for example, height and weight would initially classify the instance, and then hair and eye color would classify it further. This possibility shares the criticisms above, but is somewhat more economical. Instead of storing every combination of values as a separate category, this possibility allows each partitioning to be a decision about the values of some set of correlated attributes, thereby apparently reducing the order of the number of categories. However, each decision except the one at the top of the tree must be replicated many times throughout the hierarchy. This requires excessive storage and, as in the last section, learning rate and accuracy will be adversely affected. One additional difficulty with this approach is that it fixes the order of the decisions. If some instance is encountered that happens not to have values for the decision that is highest in the tree, then the information in the hierarchy may be inaccessible.

A single set of disjoint categories or a strict hierarchy, when applied in cross-classification domains, will either require excessive storage, extended learning, or will produce less than optimal prediction behavior.

2.2.2 Cross-classification domains

These difficulties are hardly significant if humans only rarely encounter cross-classification domains. On the other hand, if cross-classification domains are common and humans are optimally adapted to their world(Anderson, 1988), then humans must learn overlapping categories.

Linnaeus' taxonomy of the plant and animal kingdom was one of the most significant early indications that hierarchies of disjoint categories reflect the structure of the world. An important source suggesting the existence of overlapping categories in the world is artificial intelligence and cognitive science work in knowledge representation. As structured representations for knowledge appeared, it was acknowledged that the same event or object would need to be multiply classified (Minsky, 1975). Minsky argued that a generator from a car is an instance of both a mechanical system and an electrical one. These different points of view would likely be useful for different tasks. Most proposals for generic knowledge representation languages include the ability to identify multiple classes for an instance (eg. *ConjGeneric* in KRYPTON, Brachman, Fikes, & Levesque, 1983). A particular person can be classified as both a male and a parent and can inherit inferences from each.

The primary reason cited in the representation literature for the use of multiple inheritance is that it limits duplication of information and eliminates unnecessary subclasses (Touretzky, 1986). For example, different animals can have different roles in human life, some are pets, some are circus performers, and some are work animals. A strict hierarchy might require a node for each animal-role pair. With ten different animals and ten different roles, the hierarchy would have at most 100 nodes. The memory requirements could be greatly reduced if circus elephants, for example, could inherit properties from both *circus-performer* and *elephant.* If such overlapping categories were permitted, memory would only require 20 nodes to represent most of the same information.

The world does seem to present cross-classification domains. There are many simple examples of combined concepts, such as pet fish or striped apple (see Osherson & Smith, 1982). Additionally, as mentioned above, researchers concerned with representing knowledge are able to adopt overlapping categories to simplify their task, suggesting that the structure is available. Both of these pieces of evidence argue that because English words express overlapping categories, there must be such structure in the world. Anderson (1990), however, cautions against assuming that category labels correspond to the actual category structure of a domain. The labels may be merely attributes like any other. Two categories may or may not share a particular label. Therefore, the apparent existence of overlapping categories might be the result of some name or label attributes that overlap across categories, just as the extension of *red thing* and the extension of *round thing* overlap.

A category label, however, is assumed to be something special, in that it is extraordinarily useful for predicting other attributes. This is true for the above examples. The category, *circus performer*, indicates much about the lifestyle of the creature, and *elephant* indicates much about the size and shape of the creature. Hence, even if the categories are only labels, there are important predictive relations between the label and other attributes. A strict hierarchy would require that the predictive relations about, for example, circus performers would be duplicated for all types of circus performer. Overlapping categories could permit the clusters of predictive relations to be encapsulated and reused, rather than duplicated. Again, systems of disjoint categories produce inferior performance when compared with systems of overlapping categories.

3 Models of Overlapping Categories

Overlapping categories, by encapsulating predictive relationships, produce more accurate inferences and reduce duplication of information. For this to be true, the various categories must complement each other. In particular, different categories that apply to the same object must differ with respect to the attributes that they are best able to predict. If something is a *beanbag* and a *chair*, the *beanbag* category is best for predicting shape, whereas the *chair* category is best

for predicting size and function.

There are two general approaches to learning sets of complementary, overlapping categories. They be learned either one at a time or simultaneously. Suppose the task is to learn, among other things, the concepts *young* and *mallard*. A learner might at one time hear about a young mallard and only care about predicting body covering and means of locomotion. It would then begin to learn the category, *mallard*. At another time, it might care about predicting degree of coordination or source of food and hence will begin to learn the category *young* for animals. Over time, the two overlapping categories would become more entrenched as young and mallard occurred in other contexts. Both *young* and *mallard* will still participate in sets of disjoint categories. For instance, young things contrasts with old things and mallard contrasts with geese. The resulting structure for the example can then be viewed as an interwoven set of alternate hierarchies.

A simultaneous strategy for learning overlapping categories would produce a similar structure, but would do it with fewer instances. Instead of starting with particular prediction goals, the simultaneous strategy begins with the general goal of extracting as much predictive structure as possible. With each instance, it attempts to maximize its ability to predict all attributes. It would begin to learn about both *young* and *mallard* simultaneously.

3.1 Model 1: Learning overlapping categories across presentation

One possible model for learning overlapping categories across instance presentations uses a single set of categories that at any one time are considered disjoint. An instance, however, can be classified into different categories depending on what the prediction goal is. In this model, an instance is learned by first highlighting which attribute or attributes will be most useful to predict. Then, with this in mind, the instance is incorporated into the best category just as in Fisher's (1987) model. Prediction proceeds in the same manner and results in prediction of the most probable values for the goal attributes.

To actually implement this idea, Fisher's basic algorithm was adopted. However, the Gluck and Corter (1985) measure was inadequate, because it always assumed that all attributes are equally important to predict. Their measure was modified (see Martin & Billman, 1990) to produce the following:

$$Q = \sum_k \sum_l \frac{1}{N_I} P(C_k \mid matchfnc \wedge I_l) \sum_j \sum_i P(a_{A_j}) P(A_j = V_{ij} \mid C_k)^2 \qquad (2)$$

This metric is fairly complicated, but is based directly on Gluck and Corter's simple result. It was modified to allow some attributes to be ignored when learning categories. The term $P(a_{A_j})$ differentiates goal attributes from others. It is 1 if A_j is a goal attribute and 0 otherwise. The term N_I is the number of instances observed, $matchfnc$ is the procedure used for matching, and I is a particular instance that is matched. The match function may be any metric that assigns a degree of match between a particular instance and each disjoint category. Simple possibilities are an arithmetic or geometric average.

3.2 Model 2: Learning overlapping categories simultaneously

The major difference between this model and the last is that this model attempts to pull as much predictive structure out of the instances as it can. It allows simultaneous multiple classification. During prediction, an instance is multiply classified and predictive information is combined from the recognized categories to make a prediction. As an illustration of the idea, imagine that each of several guards has a limited view of the outer wall of a fortress, and that they each report to a

commander, who collects the separate pieces of information. The commander can compose the different fragments to permit informed predictions.

Learning in this model consists of adding new categories, modifying the descriptions of categories, and modifying parameters of the composition function. The system can concurrently learn new categories and learn how to combine the information provided by the categories. If the size of a pet fish is not well predicted from information about pets and fish, either a new pet-fish category would be added or parameters of the composition function would be altered to handle this exception.

More specifically, the model assumes that there are several *non-disjoint* sets of two or more *disjoint* categories. The simplest case is when each set has two possibilities. For example, one category may split plants into those that live underwater and those that do not. Another may split plants into those that are large and those that are small. One category from each the non-disjoint sets could be accessed by an instance. Both learning and prediction use Equation 2 to classify instances into each set in the same manner as in Model 1.

Composition is achieved by assuming that the non-disjoint sets and their activated members are attributes and values of instances that can be categorized by another set of disjoint categories. For each attribute, Equation 2 and the accessed categories are used to select a category from which to generate predictions.

3.3 Empirical illustration

The above models were compared to COBWEB using a cross-classification domain. The domain had twelve binary attributes grouped into four clusters of three mutually predictive attributes. In each cluster, each attribute's values were consistently paired with values on the other two attributes.

Testing was divided into 5 runs for each model. For each run, the domain of 16 instances was randomly ordered and split into a 12 instance training set and a 4 instance test set. After each instance in the training set was presented, the model's ability to make correct predictions about the test instances was determined.

The averaged results demonstrate the predicted benefit of overlapping categories. Both of the overlapping category models achieved a maximum score of approximately 70% correct after all training instances had been seen (67.5% for the first model and 72.5% for the second). COBWEB, on the other hand, achieved only 48.5% correct. One interesting result is that COBWEB only required one instance to achieve a prediction accuracy of $50\% \pm 2\%$. In contrast, the overlapping category models required four and three instances respectively. This suggests that the improved accuracy might be associated with slower initial learning.

4 Discussion

Disjoint categories divide the world at its joints (Rosch, 1978). The different levels of a taxonomy divide the world into either coarse or fine pieces. In these terms, overlapping categories allow the system to identify and consolidate pieces that recur often. Instead of repeatedly defining categories that extract the same piece, that piece is learned once and made generally available.

4.1 Theory-based concept learning

The above statement implies that general predictive rules could be learned and used in a system that permits overlapping categories. These general rules may constitute a theory that can then influence later learning. A simple way this might happen is if previous learning had established a

set of interpredictive values, and future learning assumed that the set of values remained interpredictive. When learning in a new domain, past learning can transfer, possibly permitting faster learning.

Murphy and Medin (1985) argue that general theories do not simply coexist with categories; they interact in many ways. Most importantly, a background theory can provide an explanatory principle for category membership. Although the current representation for theories includes no relationships aside from cooccurrence, conceptual coherence can be simulated with overlapping categories.

There may, for example, already be categories that predict weight from height, and that predict height from gender. Now the learner must learn to partition individuals based on their weight, but they are not given either the weight or the height. All of the information necessary to reason that women are generally lighter than men is available. The background theory permits the identification of important attributes, such as gender, and provides a weak theory to explain why the lighter category contains mostly women.

To fully develop the use of overlapping categories as theories, a more complex representation is needed, including causal and other relations between attributes.

4.2 Conceptual combination

Several overlapping categories can be learned and can be recombined into novel configurations about which the system will have inferential commitments. Someone might, for example, expect a toy elephant to be gray because all elephants they knew were gray and toys have no characteristic color. This general notion is almost identical to the intuition behind conceptual combination (Osherson & Smith, 1982). People are assumed to have separate meaning representations for various nouns and adjectives that can be combined. The major goal of the conceptual combination research is to determine how inferences from separate representations are combined, especially when they are competing. That is also a major issue for models of overlapping categories. The first model presented above makes no commitments about how it might combine two or more categories

The second model, however, adopts a relatively novel perspective on conceptual combination. It does assume a simple multiplicative combination rule, but unlike existing models of combination (Osherson & Smith, 1982), it can modify how the combination is achieved as it learns. So in contrast to earlier approaches, there is not a simple fixed mechanism, but rather a more flexible adaptive method.

4.3 Summary

Although taxonomies have long been the standard model for human categories, they do not always provide the most economical storage nor the best predictions. In particular, in cross classification domains, structures that permit overlapping categories are superior. Two models were presented that learned overlapping category structure, and they were demonstrated to be superior to a method that relies on non-overlapping structure.

5 Acknowledgements

I would like to thank Dorrit Billman for her advice, and I would like to thank Nancy Smith Martin and Tom Hinrichs for assistance with earlier drafts of this paper.

References

Anderson, J. R. (1988). The place of cognitive architectures in a rational analysis. In *Proceedings of Tenth Annual Conference of the Cognitive Science Society*, Montreal, Canada.

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

Brachman, R. J., Fikes, R. E., and Levesque, H. J. (1983). Krypton: A functional approach to knowledge acquistion. *IEEE Computer*, 16:67–73.

Bruner, J. S., Goodnow, J. J., and Austin, G. A. (1956). *A Study of Thinking.* Wiley, New York.

Fisher, D. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning.*

Gluck, M. and Corter, J. (1985). Information, uncertainty, and the utility of categories. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, Irvine, CA.

Gluck, M. A. and Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, 27:166–195.

Martin, J. D. (1989). Reducing redundant learning. In *Proceedings of Sixth International Workshop on Machine Learning*, Ithaca, NY.

McClelland, J. L. and Rumelhart, D. E. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol I.* MIT Press, Cambridge, MA.

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

Minsky, M. (1975). A framework for representing knowledge. In Winston, P. H., editor, *The Psychology of Computer Vision.* McGraw-Hill, New York.

Murphy, G. L. and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92:289–316.

Osherson, D. N. and Smith, E. E. (1982). Gradedness and conceptual combination. *Cognition*, 12:299–318.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.

Touretsky, D. S. (1986). *The Mathematics of Inheritance Systems.* Pitman Publishing, London.