

CHAPTER 6

Theory-Guided Concept Formation

DOUG FISHER

MICHAEL PAZZANI

1. Introduction

Previous chapters in this volume concentrated on empirical or inductive approaches to concept formation. These models assume that observations are independent; the only link among them is implicit in the common language used for the intrinsic properties of observations, such as size and shape. Empirical concept formation is guided by similarities and differences among observations as expressed in this language.

However, psychological findings suggest that humans do not regard observations as independent, but as interacting in complex ways. The structural representations surveyed by Fisher and Pazzani (Chapter 1, this volume) and by Thompson and Langley (this volume) begin to address these issues, but they do not capture all the subtleties that are likely to confront a learner. For example, relationships among features may not be an explicit part of the environmental input to a system. Thus, a novice card player learning about straights (hands in which the cards can be placed in ascending order by rank) relies on internalized background knowledge about successive rank values (Vere, 1978). In addition, a learner must also be attuned to relationships across observations, not simply within observations. For example, (`successor(Ace, King)`) is a relation that is internal to a card hand, whereas the expression `stronger(hand1, hand2)` defines a relation between card hands.

In sum, observations often cannot be treated in isolation. Rather, a rich set or 'theory' of interconnections link observations. These connections may vary in the granularity of relationships: features within or across observations can be connected, as can whole observations.

In addition, these relationships must often be inferred from a learner's background knowledge — they need not be an explicit part of the environmental input. This chapter briefly reviews the psychological literature that motivates this view, and examines computational models that have been proposed to deal with it.

2. Form and Function

Recall that studies by Medin, Wattenmaker, and Hampson (1986), Ahn and Medin (1989), and others found a strong unidimensional bias on the part of subjects during sorting. Fisher and Pazzani (Chapter 1, this volume) suggested extensions to these studies intended to encourage family resemblance clustering, but the experimenters (Medin et al., 1986; Ahn, 1990) were able to do this by suggesting the relevance of additional 'functional' features, together with theories about how intrinsic features and function were related. For example, they discussed 'hammering' prior to having subjects sort objects with features that would be relevant to this task: those with a majority of features like graspable and flat head. An explanation for the emergence of family resemblance sorting in this case is that by introducing the hammering descriptor, subjects were able to retain their bias for unidimensional concepts (i.e., those that one could use to hammer); this behavior was consistent with family resemblance sorting when attention was limited to the perceivable descriptors.

These studies suggest the strong role that *function* plays in human category development (e.g., Nelson, 1977). In particular, some evidence suggests that many categories arise because constituent members share similar functionality (e.g., a class of objects are used for hammering; a class of objects can be rolled), rather than shared *form* (flat head, spherical). Of course, similarity in structure is still a valuable heuristic guide in categorizing observations (Ross & Spalding, this volume), but at least in some cases this similarity is not the reason such classes arose originally. The importance of functionality becomes even more pronounced in categories for which there appears to be no structural similarity at all. In part, these correspond to what Barsalou (1983) defines as *ad hoc* categories, which arise spontaneously based on (roughly) shared function (e.g., things that will sell at a garage sale). Such categories tend to be transient and generated on demand,

in part because the lack of structural similarity makes categorization with respect to them difficult.

Some research on unsupervised machine learning has also recognized the importance of the form/function distinction (Nordhausen, 1986). In fact, an important rationale for distinguishing clustering and characterization in unsupervised systems is suggested by psychological findings that functional knowledge guides the search for clusterings, and structural information guides the search for characterizations. One example of 'functional clustering' is provided by GLAUBER (Langley, Zytkow, Simon, & Bradshaw, 1986), which is typically viewed as a model of scientific discovery. In this system, observations about molecular structures are represented by intrinsic features (e.g., taste), as well as by reaction relations among molecules. For instance, the interaction of different substances is represented by events like

(reacts inputs {HCl NaOH} outputs {NaCl}),
(reacts inputs {HCl KOH} outputs {KCl}), and
(reacts inputs {HNO₃ KOH} outputs {KNO₃}).

Upon seeing these and similar reactions, GLAUBER forms the clusters {HCl, HNO₃}, {NaOH, KOH}, and {NaCl, KCl, KNO₃}. Intuitively, these clusters represent the chemical classes of acids, alkalis, and salts, respectively. The system then uses these clusters to define the general relation (reacts inputs {acids alkalis} outputs {salts}).

GLAUBER has a number of limitations as a model of functional clustering, but it illustrates the promise of this paradigm: clusters of observations (e.g., molecular structures) bias the way that higher-level processes (e.g., chemical reactions) are characterized. Thompson and Langley (this volume), Reich and Fenves (this volume), Scott and Markovitch (this volume), and others (Fisher, 1986; Schlimmer, 1989; Handa, 1990) view this process as reformulating the description language that biases learning (Subramanian, 1989) and problem solving.

Much of the existing research on concept formation assumes that surface similarity indicates functional similarity and vice versa. However, we have seen that some psychologists hypothesize more than a heuristic connection between form and function. Rather, the assumption is that a 'theory' of a domain — a set of relations — serves to link the observable features with the function that these objects serve. We now turn to models that assume background knowledge and that require the learner to infer linkages among the observables.

3. Exploiting Background Knowledge

A landmark study by Chi, Feltovich, and Glaser (1981) required subjects to sort physics problems into categories of their own design. This unsupervised task revealed significant differences between the criteria that 'experts' (i.e., graduate students in physics) and 'novices' (i.e., undergraduates) used to create categories. Novices tended strongly to sort based on 'surface' features — those that were referred to in the problem statement (e.g., reference to an 'inclined plane' or 'friction'), while experts formed categories of problems that required the application of similar solution strategies (e.g., application of Newton's second law). In addition, experts generally required more time to produce an initial sort than novices. These findings suggest that expert subjects make an initial 'qualitative analysis' of a problem prior to classification. This preliminary analysis is responsible for inferring 'derived' (intermediate, deep) features from the surface features given in the problem statement.

The use of underlying principles or derived features to guide categorization by human subjects has been noted and qualified in many studies (Medin, 1989; Wisniewski & Medin, this volume). In addition to the Chi et al. studies, Faries and Reiser (1988) found that goal-directed, problem-solving situations accentuated reliance on inferred features. Importantly, inference of derived features need not be static and limited to a single preprocessing stage prior to classification. Rather, inference may interact continuously with classification (Seifert, 1989). Finally, despite the importance of inferred features, it would be a mistake to conclude that surface features are not useful. In fact, there are many cases when such features are exploited, even by experts (Ross, 1987; Ross & Spalding, this volume). Notably, humans appear to exploit surface features when they are well correlated with the derived features that are deemed most useful. For example, if one is calculating the time required to travel between two points, then the mention of 'boats' (and travel by water) versus 'cars' (and travel by land) indicates the relevance of a water current. To the extent that currents affect time, rate, and distance calculations, 'boats' will be exploited to classify the problem relative to past problems that were complicated by currents. Thus, as we have noted, surface features often provide a good heuristic guide to more fundamental, theory-based similarities, though this need not always be the case (e.g., travel by boat on a lake may not involve currents at all).

The importance of coupling inference and classification has not been lost on machine learning researchers. One paradigm that illustrates inference through background knowledge is *explanation-based learning* (Mitchell, Keller, & Kedar-Cabelli, 1986; DeJong & Mooney, 1986). In supervised settings, a learner searches a domain theory of inference rules in an attempt to link (i.e., explain) how an observation's surface features indicate membership in a teacher-provided target category. Recent research has explored the identification of derived features (i.e., a so-called boundary of operationality) that should be inferred and exploited during classification (Braverman & Russell, 1988; Yoo & Fisher, this volume). In addition, Mooney (this volume) shows that the explanation-based paradigm need not be limited to supervised scenarios, just as empirical learning includes supervised and unsupervised representatives. Rather, objects that participate in similar ways across explanatory structures provide an opportunity for 'functional clustering' of the type described in the previous section — the difference is that relationships between observations are inferred from background knowledge rather than being an explicit part of the input as in GLAUBER.

Mooney (this volume) also points out that some explanation-based methods are easily adapted to facilitate attribute prediction. That is, an explanatory network may support inferences about many aspects of an observation, not simply its membership in a target category. For example, consider a voter who wants to categorize politicians based on their views along selected issues. A domain theory may contain very general inference rules such as

If a senator and the president are of the same party,
and the president is not a lame duck,
Then predict the senator's view equals that of the president,

or relatively specific rules like

If a senator is from the midwest,
Then predict that the senator supports farm aid.

In any case, a domain theory need not be directed at predictions along a single dimension, nor does this limitation apply to explanatory structures derived from such a theory.

To some extent, attribute prediction is also suggested by Stepp and Michalski (1986) and Mogensen (1987) in their nonincremental, conceptual clustering system, CLUSTER/CA. They suggest inference as a preprocessing step that fills in unknown values of observations prior to

clustering. CLUSTER/CA also includes a second, novel form of background knowledge known as a *goal-dependency network*, which links an agent's goals with observable features. For example, when a voter categorizes politicians, certain goals may play a greater role in clustering than others. If a high-level goal is to Limit-Soviet-Expansionism, then the voter's 'goal-dependency network' may indicate that issues such as MX-missile and Contra-aid should play a role in category construction, and a politician's views on issues such as Welfare-cuts should not. Thus, an agent's current goals will lead certain features to be used in clustering while others are not. This type of theory-driven attention promises to speed an agent's convergence on useful features for concept learning (Seifert, 1989), relative to purely empirical attention mechanisms surveyed by Fisher and Pazzani (Chapter 1, this volume).

In addition to explanation-based strategies, case-based approaches represent a second major theory-driven paradigm (e.g., Kolodner, 1987; Hammond, 1987). These models posit that many problems are best solved by appealing to specific problem-solving experiences (i.e., cases) of the past. An exemplar of this approach is a recent system by Shinn (1989) that incrementally 'clusters' problem-solving episodes (e.g., menu plans) into an abstraction hierarchy (somewhat like an EPAM discrimination net), labeling arcs of the tree by surface features (e.g., a lunch meal) and goal features (e.g., low-salt). New problems trigger the retrieval of similar past problems that are then modified to fit the particular constraints of the new situation. More generally, case-based approaches are concerned with a relatively tight coupling between inferring and classification (Owens, 1990; Schank et al., 1990), like that noted in humans (Seifert, 1989). Many systems in this paradigm can also be cast as concept formation systems, just as some concept formation systems can be viewed as case-based systems (Langley, 1989).

A number of authors (Braverman & Wilensky, 1990; Porter, Bareiss, & Holte, 1990; Fisher, Yoo, & Yang, 1990) point to a merger of the two primary theory-driven paradigms: explanation-based and case-based learning. Briefly, case-based systems typically have some inductive capability that renders them less sensitive to imperfections in a domain theory than explanation-based systems (see Mitchell, Keller, & Kedar-Cabelli, 1986), while a domain theory, if available, provides a flexible avenue for problem solving when available cases do not provide sufficient experience on which to draw. An influential system in this regard is Pazzani's (1987) OCCAM, which incrementally clusters 'events'

(e.g., international sanctions) using a UNIMEM-like strategy (Lebowitz, 1987; Fisher & Pazzani, Chapter 1, this volume). It is worth describing OCCAM in some detail, since it illustrates several ways in which theory-driven and inductive mechanisms can interact within a concept formation system.

4. A Case Study in Theory-Driven Concept Formation

The OCCAM system clusters 'events' in an attempt to facilitate accurate predictions about future situations. For example, the system might be called upon to categorize international-sanction incidents like the following:

In 1983, Australia refused to sell uranium to France, unless France ceased nuclear testing in the South Pacific. South Africa sold France uranium at a premium and France continued nuclear testing.

In 1961, the Soviet Union refused to sell grain to Albania if Albania did not rescind economic ties with China. China sold Albania wheat at a below market price and Albania continued the ties with China.

The objective in this domain is to facilitate the system's ability to accurately predict the outcome (i.e., successful or unsuccessful) of new sanction events.

Without background knowledge OCCAM uses UNIMEM's approach to clustering, which is based on similarity over the surface features. Briefly, this strategy tends to group observations that share many features (e.g., selling country, buying country, selling price, decade of sale) and segregate observations that share few features. These clusters are characterized by conjoining the features common to *all* examples in the cluster.¹ Notice that clusters formed in this manner may include successful and unsuccessful incidents; as a consequence, the outcome feature is dropped from the cluster's characterization, and the cluster will not be useful for purposes of predicting outcome.

Thus, consistent with psychological findings, reliance on surface similarity can be quite limiting in OCCAM. In response, the system incorporates explanation-based methods to increase the rate and accuracy

1. Early versions of UNIMEM did not allow exceptions to the 'predictable' features that make up a category's characterization; concepts were strictly conjunctive. Later versions of UNIMEM overcame this limitation (Lebowitz, 1987), but OCCAM is based on earlier versions.

of learning by excluding irrelevant features that can lead to categories that hinder prediction of outcome (Seifert, 1989; Stepp & Michalski, 1986). Using background knowledge, OCCAM explains how the incident's surface features led to the success or failure of the sanction. In the first example above, South Africa provided a product (i.e., uranium) for economic reasons while China provided grain for political reasons. Thus, explanations for these events differ in some important ways. In general, each example with a distinct explanation creates a new cluster that is characterized by the features deemed relevant by explanation-based learning. As training proceeds, sanction incidents that succeed or fail for the same reason are grouped together. An alternative way of viewing this process is that instead of creating clusters of examples which are similar to each other in a space defined by the surface features, the clusters are created in an explanation space whose attributes are defined by the inference rules used to explain the examples (Mooney, this volume; Medin, 1989).²

However, explanation-based capabilities do not alleviate all problems. For example, a characterization that covers the two examples given above would not include the price of the commodity sold, since it is not present in the second example, even though the price is a critical contributor to success in the first example. Because categories are arranged hierarchically, this discriminating feature of the explanation structure is not used to initially guide classification at the most general levels of the hierarchy. More extreme examples of this effect occur when radically different explanations are clustered together by virtue of an identical outcome. This is undesirable because the different explanations basically define a disjunctive concept, but they are stored as a simple conjunction of one term (i.e., shared outcome).

In sum, OCCAM's conjunctive representational bias sometimes promotes clusters with irrelevant surface features, and overgeneralizes explanation structures, thus eliminating relevant features from its descriptions. Collectively, this can degrade predictive accuracy. We designed an experiment to analyze the extent of these tendencies in more detail. In each condition, training involved a series of 15 actual sanction incidents, and testing compared OCCAM's predictions to those of a Rand Corporation expert on a set of hypothetical incidents.

2. In the UNIMEM-like 'similarity' computation performed by OCCAM, a feature shared by a new event and cluster contributes a 1 to the match and a 0 otherwise. This occurs whether the features are surface or derived.

One condition was run in which all irrelevant surface features were deleted from the training examples and no domain theory was used. This removed any possible effect that irrelevant features could have during learning. Intuitively, we might expect that this would result in the same accuracy as that obtained by OCCAM with explanation-based learning, but accuracy was not as good. Thus, as Ahn (1990) found in experiments with human subjects, background knowledge does more than identify relevant features; it biases the way that relevant features are integrated into cohesive categories.

A second condition implemented a particular strategy of combining 'form' and 'function' in concept formation. In particular, OCCAM was run using background knowledge for clustering; thus explanation structures biased the formation of categories, in that derived features were included in the similarity computation that guided clustering. However, a 'knowledge-free' characterization strategy used all (relevant and irrelevant) surface features. Despite the biasing effect of background knowledge during clustering, the use of irrelevant features to characterize classes led to inaccurate classification of test cases. This highlights the variable contribution of surface similarity in guiding classification; in the sanctions domain it had little heuristic value.

Finally, when background knowledge was used to bias clustering (as above), and characterization occurred over only the relevant surface features, accuracy results were identical to those obtained in the standard OCCAM approach, in which features derived from background knowledge also participate in characterization. This last experimental condition highlights the dual role of derived features: they are valuable guides both in the formation of categories and in the characterization of these categories. Their presence in characterizations triggers inferencing during classification in OCCAM and related systems (e.g., Yoo & Fisher, this volume), and apparently in humans as well (Seifert, 1989).

5. Concluding Remarks

We have briefly surveyed psychological findings that motivate theory-driven approaches to concept formation, and examined some computational mechanisms that address this task. One important insight is that observations are often not independent. Rather, observations may interact, and these interactions influence clustering and characterization just

as intrinsic properties do. Often, form and function interact in subtle ways, each suggesting alternative categorizations.

For example, a traditional zoological partitioning of animals into mammals, birds, reptiles, amphibians, and fish is largely dictated by similarity over intrinsic properties, but interrelationships among animals (e.g., that lions eat zebras) dictate an ecological categorization that includes herbivores, carnivores, and omnivores. In fact, one often wants these orthogonal partitionings to coexist in a knowledge base, thus relating to Fisher and Pazzani's (Chapter 1, this volume) concerns with clumping and 'dag' organizations. Similar concerns from their discussions take on added complexity when form and function interact.

Finally, relations among observations are often not explicit in the environmental input. Instead, one must fill in the gaps from internal background knowledge. The explanation-based and case-based paradigms provide some guidance on how inference, categorization, and learning interact, though considerable research remains to be done before the field realizes a robust coupling of these processes within a single model.

Acknowledgements

The first author was supported by Grant NCC 2-645 from NASA Ames Research Center, and the second author was supported by Grant IRI-8908260 from the National Science Foundation. Discussions with Woo-Kyoung Ahn helped elucidate some issues, but she should not be held responsible for our interpretations of various data.

References

- Ahn, W., & Medin, D. L. (1989). A two-stage categorization model of family resemblance sorting. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 315-322). Ann Arbor, MI: Lawrence Erlbaum.
- Barsalou, L. W. (1983). Ad hoc categories. *Memory and Cognition*, *11*, 211-227.
- Braverman, M. S., & Wilensky, R. (1990). Toward a unification of case-based reasoning and explanation-based learning. *Working Notes of the AAAI Spring Symposium on Case-Based Reasoning* (pp. 80-84). Palo Alto, CA: AAAI Press.

- Chi, M., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.
- Faries, J. M., & Reiser, B. J. (1988). Access and use of previous solutions in a problem-solving situation. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 433–439). Montreal, Quebec: Lawrence Erlbaum.
- Fisher, D. H. (1986). A proposed method of conceptual clustering for structured and decomposable objects. In T. M. Mitchell, J. G. Carbonell, & R. S. Michalski (Eds.), *Machine learning: A guide to current research*. Boston, MA: Kluwer.
- Fisher, D. H., & Chan, P. K. (1990). Statistical guidance in symbolic learning. *Annals of Mathematics and Artificial Intelligence*, 2, 135–148.
- Fisher, D., Yoo, J., & Yang, H. (1990). Case-based and abstraction-based reasoning. *Working Notes of the AAAI Spring Symposium on Case-Based Reasoning* (pp. 7–11). Palo Alto, CA: AAAI Press.
- Hammond, K. (1987). *Case-based planning: An integrated theory of planning, learning, and memory*. San Diego, CA: Academic Press.
- Handa, K. (1990). CFIX: Concept formation by interaction of related objects. *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*. Nagoya, Japan.
- Kolodner, J. L. (1983). Reconstructive memory: A computer model. *Cognitive Science*, 7, 281–328.
- Kolodner, J. L. (1987). Extending problem solver capabilities through case-based reasoning. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 167–178). Irvine, CA: Morgan Kaufmann.
- Langley, P. (1989). Toward a unified science of machine learning. *Machine Learning*, 3, 253–259.
- Langley, P., Zytkow, J. M., Simon, H. A., & Bradshaw, G. L. (1986). The search for regularity: Four aspects of scientific discovery. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2). San Mateo, CA: Morgan Kaufmann.
- Lebowitz, M. (1987). Experiments with incremental concept formation: UNIMEM. *Machine Learning*, 2, 103–138.

- Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, 44, 1469-1481.
- Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, 19, 242-279.
- Michalski, R. S., & Stepp, R. E. (1983). Learning from observation: Conceptual clustering. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach*. San Mateo, CA: Morgan Kaufmann.
- Mogensen, B. (1987). *Goal-oriented conceptual clustering: The classification attribute approach* (Tech. Rep. No. UILU-ENG-87-2257). Urbana: University of Illinois, Department of Computer Science.
- Nelson, K. (1977). Some evidence for the cognitive primacy of categorization and its functional basis. *Merrill-Palmer Quarterly of Behavior and Development*, 19, 21-39.
- Nordhausen, B. (1986). Conceptual clustering using relational information. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 508-512). Philadelphia, PA: Morgan Kaufmann.
- Owens, C. (1990). Functional criteria for indices and labels. *Working Notes of the AAAI Spring Symposium on Case-Based Reasoning* (pp. 64-68). Palo Alto, CA: AAAI Press.
- Pazzani, M. (1987). Inducing causal and social theories: A prerequisite for explanation-based learning. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 230-241). Irvine, CA: Morgan Kaufmann.
- Porter, B., Bareiss, R., & Holte, R. (1990). Concept learning and heuristic classification in weak-theory domains. *Artificial Intelligence*, 45, 229-263.
- Quinlan, J. R. (1990). Learning logical definitions from relations. *Machine Learning*, 5, 239-266.
- Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 629-639.
- Schank, R. C. (1982). *Dynamic memory*. Cambridge: Cambridge University Press.

- Schank, R., Brand, M., Burke, R., Domeshek, E., Edelson, D., Ferguson, W., Freed, M., Jona, M., Krulwich, B., Ohmaye, E., Osgood, R., & Pryor, L. (1990). Towards a general content theory of indices. *Working Notes of the AAAI Spring Symposium on Case-Based Reasoning* (pp. 36–40). Palo Alto, CA: AAAI Press.
- Schlimmer, J. C. (1989). Refining representations to improve problem solving quality. *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 457–460). Ithaca, NY: Morgan Kaufmann.
- Seifert, C. M. (1989). A retrieval model using feature selection. *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 52–54). Ithaca, NY: Morgan Kaufmann.
- Shinn, H. (1989). *A unified approach to analogical reasoning*. Doctoral dissertation, School of Computer Science, Georgia Institute of Technology, Atlanta.
- Shrager, J. (1987). Theory change via view application in instructionless learning. *Machine Learning*, 2, 247–276.
- Stepp, R. E., & Michalski, R. S. (1986). Conceptual clustering: Inventing goal-oriented classifications of structured objects. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2). San Mateo, CA: Morgan Kaufmann.
- Stepp, R. E. (1987). Concepts in conceptual clustering. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (pp. 211–213). Milan, Italy: Morgan Kaufmann.
- Subramanian, D. (1989). Representational issues in machine learning. *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 426–429). Ithaca, NY: Morgan Kaufmann.
- Vere, S. A. (1978). Inductive learning of relational productions. In D. A. Waterman & F. Hayes-Roth (Eds.), *Pattern-directed inference systems*. New York: Academic Press.
- Wisniewski, E. (1989). Learning from examples: The effect of different conceptual roles. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 980–986). Ann Arbor, MI: Lawrence Erlbaum.