A Framework for Concept Formation[1]

J. Daniel Easterlin
Pat Langley

Irvine Computational Intelligence Project
Department of Information and Computer Science
University of California, Irvine 92717

INTRODUCTION

Our approach to concept formation differs from the traditional view. In this paper, we outline an alternative view of concept learning, and argue that the goals of the learner play a central role in this process. We propose that goals act to determine significant features of the world, and that without such goals as a basis, concept formation is a semantically empty data summarization task. We begin by examining the components of the concept formation process. After laying this foundation, we review previous approaches to concept formation in these terms, rejecting two of the assumptions upon which this work has been based – the presence of a tutor and the "all–or–none" character of concepts. This leads us to propose an alternative model of the concept formation process, in which goals and prototypes figure prominently.

THE COMPONENTS OF CONCEPT FORMATION

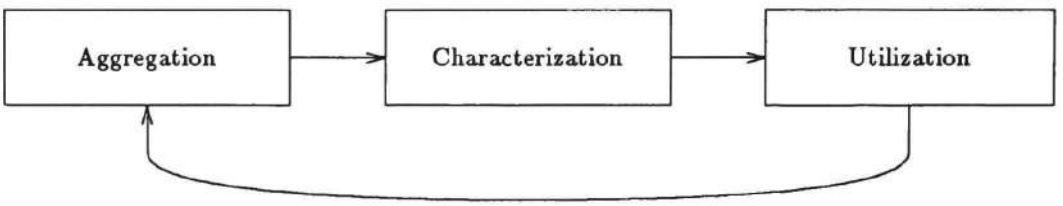

Figure 1. The Components of Concept Formation.

Previous research in machine learning suggests that the process of concept formation can be divided into three distinct components. The first of these – aggregation – involves grouping instances of the concept into collections. The second component – characterization – involves generating some description of the instances in the aggregate. The final subprocess – utilization – involves making use of the resulting description. Let us examine each of these components in more detail.

Aggregation

Aggregation is a process of collection, in which objects or instances of some concept (possibly still to be learned) are grouped together into a set. Aggregation is not a process of description, but involves collecting entities into an aggregate, from which a description or characterization can subsequently be formed. In the task of learning from examples as studied by machine learning researchers, the aggregation process is made trivial (Hunt, Marin & Stone, 1966). The tutor provides explicit aggregation of the examples into sets of positive and negative instances, and some characterization of the positive instances is generated. In contrast, in the task of learning search heuristics, aggregation must be performed by the learning system itself (Mitchell, Utgoff & Banerji 1983). Instances that led to the successful solution of a problem are aggregated as positive instances of the responsible rule's use, while instances that led away from the solution path are aggregated as negative instances. Thus, aggregates are generated by the learner on the basis of performance, rather than relying on a tutor, as in the case of learning from examples. The utility of discussing aggregation as a distinct process in concept formation is that it focuses attention on what constitutes significance for the system. As a result, one begins to question the plausibility of existing aggregation techniques, and to explore alternative methods.

Characterization

Characterization is the process that is usually discussed in machine learning under the name "concept learning". It involves constructing a description for an aggregate of entities, based on individual descriptions of each entity. This may occur either incrementally or non–incrementally, depending on whether instances are presented simultaneously or one at a time. Researchers in machine learning have proposed a number of computational methods for characterization, and these constitute their contributions to concept formation.

[1] This research was supported in part by the IBM Corporation, and in part by a gift from Hughes Aircraft Company.

Utilization

The utilization process integrates the characterization or concept description with the performance element. In the case of recognition, it contains a matching process for identifying positive instances of the description that were constructed during the characterization process. Following recognition, some action may be taken or a metric may be applied to test the adequacy of the description in recognizing the instance. For the most part, concept descriptions are used to recognize positive instances of the concept when they occur in the problem domain. Since most AI learning research has assumed "all–or–none" concepts, the recognition process has typically involved a "complete matching" mechanism, in which all conditions must be satisfied before instances of the concept are recognized.

TRADITIONAL APPROACHES TO CONCEPT FORMATION

We are interested in concept formation as it occurs in complex, reactive environments that are similar to the real–world. In this section, we review the components of the concept learning process, and find that such environments lead one to reject some important assumptions upon which earlier machine learning work has been based.

Tutors and Aggregation

Previous research on concept learning has assumed careful guidance by a tutor, despite the intuition that humans learn most of their concepts through experience with the world. For example, children clearly learn concepts such as "dog" and "chair" before they know the words for these concepts. In the traditional approach to learning concepts from examples, a tutor trivializes the aggregation problem, by providing positive and negative instances of the concept to be learned. In contrast, learners in the real world must aggregate instances in some other manner.

All–or–None Concepts and Characterization

According to the classical view described by Smith and Medin (1981), a concept is defined by necessary and sufficient conditions, and machine learning researchers have used a similar notion of concepts. In other words, for an object to be recognized as a positive instance of some concept, it must satisfy all of the conditions specified in the concept description, and additional features have no effect. However, most of our everyday concepts are "fuzzy", with exemplars being better or worse, rather than instances or non-instances (Rosch & Mervis, 1975). For instance, a robin is a better instance of the concept "bird" than a penguin, and some chairs are better than others (e.g., ones that are missing a leg). Such concepts cannot be described in terms of necessary and sufficient conditions, so some other representation is required.

Smith and Medin (1981) have outlined two alternatives to the all–or–none framework. The "prototype" (or "probabilistic") approach, first proposed by Rosch and Mervis (1975), assumes that there exists an abstract representation of each concept, and that instances are judged to be better or worse examples depending on the *degree* to which they match this representation. Smith and Medin also discuss the *exemplar* approach, in which concepts are represented not as abstract structures, but as disjunctions of many specific instances. Both approaches have their advantages, and evidence exists for both theoretical frameworks. In this paper, we will focus on the prototype–probabilistic approach for a simple reason – this approach is computationally much more tractable.

The most common AI methods for characterization are generalization and discrimination. Upon closer examination, we find that these methods encounter serious difficulty when applied to "fuzzy" concepts. The problem is that both methods rely on a strong distinction between positive and negative instances: generalization finds structures held in common among positive instances, and discrimination finds differences between positive and negative instances.

Complete Matching and Utilization

Recognition involves determining the most appropriate concept to describe the current data. This task is considerably simplified by the assumption that concepts are defined by a set of necessary and sufficient conditions. However, we have already argued that real–world concepts cannot be defined in this manner. Thus, complete matching of object to characterization must be rejected.

AN ALTERNATIVE APPROACH TO CONCEPT FORMATION

In the previous section, we argued that real–world concepts are not learned from a tutor, and that they cannot be described in terms of necessary and sufficient conditions. If we hope to account for the process of concept formation, this forces us to propose new techniques for aggregating sets of instances, for characterizing the resulting aggregates, and for recognizing the best concept for a given situation.

Goals and Aggregation

By rejecting the traditional assumptions of tutor-provided instances, we must find some other solution to the problem of aggregation. We believe that the *goals* of the learner play a major role in this process, and presume that at each point in time, the agent has one or more active goals (possibly organized as a hierarchy of goals and subgoals). In describing their means–ends analysis theory of human problem solving, Newell and Simon (1972) distinguish between three types of goals. Although each of these goal types can be used to direct the aggregation process, the most obvious examples involve apply–operator goals, in which one wants to apply an operator to some object or state. For instance, suppose the agent is tired, and decides to apply the operator *sit–down*.† This operator requires some object upon which to sit, and the agent will scan its immediate environment for a likely candidate. The important point is that by applying its operator to candidate objects, the agent will discover that some objects produce better results than others. These will be good instances of "sittable" objects, while others (such as chairs with wobbly legs, or with a tack on their seat) will be poor instances. In any case, objects to which the operator has been applied that more or less satisfy the goal are passed on to the characterization process.

Our attention to the importance of goals arises primarily from recognizing aggregation as a distinct process demanding a supporting basis. As this basis, goals play a dual role – they identify which objects should be grouped together for input to the characterization mechanism, and they provide a test indicating the degree to which the desired state has been achieved. Thus, they tie objects and operators to experience by indicating their relative value in satisfying goals. Certain objects and operators are rated higher than others, since the application of particular operators to particular objects manifests properties of those objects that are instrumental in satisfying the posted goals, while others are not. Objects are thus rated higher to the degree they manifest functional properties in achieving goals. This provides the feedback necessary to identify some combinations of objects and operators in experience as more significant to the learner than others.

In summary, we believe that the learner's goals direct the aggregation process. Furthermore, objects and operators are grouped together according to the degree to which their interaction satisfies goals. It is through this interaction between operators and objects that objects manifest properties which are functional in satisfying goals. Thus, not only do goals identify significant objects and operators, but they further suggest the existence of significant functional properties within an object.

The Representation of Concepts

Traditional approaches to characterization assume the all–or–none nature of concepts, which simply does not hold for many everyday object concepts. As a result, we must find another solution to the characterization problem. Our approach must be able to represent "fuzzy" concepts, and to incrementally modify these descriptions in response to new instances of the concept. In real–world concepts, some features and relations are more important than others. Thus, our representation must include some measure of each feature's *criteriality*. We specify this in terms of a weight ranging from zero to one, with zero denoting low importance and one indicating high importance. Of course, these numbers have little meaning detached from the utilization process. In our framework, conditions (features or relations) with high weights contribute more to the overall degree of match than conditions with low weights. Moreover, conditions with very high weights (near one) must be matched for a reasonable overall match to result. As a result, the notion of all–or–none concepts emerges as a special case of this scheme, in which all conditions have weights of one.

Since we are concerned with object concepts, we believe that structures similar to Binford's (1971) generalized cylinders will prove adequate. This representation has the advantage of combining structural relations between the components of an object with numeric features of those components. This is an important characteristic, since the real–world has both structural and numeric aspects. For instance, a prototypical chair might be represented with the components of four legs, a seat, and a back arranged in particular spatial relations to each other. In addition, each component would be described by numeric features, such as length, diameter, and orientation (normalized for the overall size of the object). In addition, the use of numeric features leads to a novel interpretation of the weights on each feature. With each numeric feature, one can associate a *mean* value of the positive instances that have been observed, and a *standard deviation* of those values. High standard deviations imply that a wide range of values of the feature are satisfactory, while low standard deviations imply that only a narrow range of values for the feature is acceptable. Thus, one might use the *inverse* of the standard deviation for a feature as its associated weight. This would give low criteriality to features with widely varying values, and high criteriality to features with nearly constant values. For instance, the legs of a chair are nearly always half the length of the entire chair's

† Obviously, the action *sit-down* is not primitive in any sense; it is a high–level operator (or macro–operator) that must be acquired from experience.

height. Thus, this feature would have a low standard deviation and be highly criterial, giving it an important role in judgements of prototypicality.

However, recall that we are assuming the characterization process receives more than prototypical instances as input. Rather, it is given the degree to which each object satisfies the agent's goals. We would like our learning mechanism to use this information in creating the concept description. In order to do this, we require more than a feature's mean value; we require a function relating features and operators to the "goodness" of an object in satisfying goals. We propose constructing this function by considering operators in addition to physical features as relevant to the object's goodness and including them as a special type of feature in the concept description. By regressing goodness against the values of all feature types, we derive a relation between important physical features of the object, those operators that operate on the features, and values of goal satisfaction. Thus, we have a description of the object that expresses the object's functional properties as they relate to goal satisfaction and the object's physical characteristics for use in recognition. The representation is an equation providing both a means for predicting values of goal satisfaction and a measure of the goodness of fit for such predictions. Returning to our interest in the criteriality of features for predicting goal satisfaction, the percentage of variance in goal satisfaction accounted for by an individual feature can be taken as a measure of that feature's criteriality in the concept definition.

The Characterization Process

Let us now turn to the mechanism of characterization, by which the learner goes from instances of some concept to a description of that concept. Within the current framework, we are assuming that the aggregation process has determined which object and operators should be incorporated into the concept description, and that aggregation also provides the degree to which the object and operators satisfy the relevant goal. The task of characterization is to modify the existing description to better predict the "goodness" of the current object. We also assume that instances are processed incrementally, since the agent generally interacts with one object at a time (or a few at most). Thus, each instance leads to only minor modifications in the concept description. Before a concept description can be altered, it must first be created, and issues arise about the nature of such initial descriptions. Since early descriptions are based on a single instance, one might make each feature very criterial by having a weight of one. Through experience, as additional instances are observed and variation among feature values (including operators) occurs, constraints on the feature values become looser.

Once a stable description has been formed, the feature values of new instances are used to modify the regression coefficients associated with each feature. By retaining the number of instances that have been observed so far, one can easily compute a revised equation that includes the new feature value. This accommodates gradual changes in the concept description over time. For example, if the learner began to see chairs with longer legs, his coefficients for the "length of leg" features would slowly be revised. Thus, this method can respond to changing environments, unlike most traditional approaches to concept learning.

However, if the agent encounters an object with feature values that fall far outside previous experience, this is an occasion to generate a disjunctive version of the current concept. For instance, if one sees a chair in which the legs are substantially longer than expected (such as a baby's high–chair), then it is natural to distinguish this from other chairs that more closely match one's expectations. Such variants are stored near to the initial concept, but are characterized independently of the original version. Note that this implies the order of presentation is relevant to learning. If gradual changes in feature values are observed, a single concept will be learned; however, if instances with extreme values are alternated, disjunctive concepts will be acquired instead.

Goal-Indexed Partial Matching and Utilization

Traditional AI approaches to concept learning assume that complete matching can be used for recognition. However, in rejecting the notion of necessary and sufficient conditions, we are inevitably led to replace this with some form of partial matching mechanism. Hayes–Roth (1978) has argued that partial matching is computationally expensive, and the best known algorithm is exponential in the general case. Therefore, we would like to take advantage of constraints to make the task manageable.

Recall that we are assuming different weights on the various conditions composing the concept description. To a certain extent, we can constrain the partial matching process by attempting to match more criterial conditions (those with higher weights) first, and leaving less criterial features and relations until later. This leads to a best–first search through the space of partial matches, and is much more attractive than an exhaustive version. Like all heuristic search approaches, the method is not guaranteed to find the optimal solution (in this case the best partial match), but it will nearly always find a satisfactory one with considerably less effort.

However, recall also that the agent must choose between hundreds and thousands of competing concepts, and it is unlikely that the above method will suffice. Fortunately, in this framework concepts are created because their instances have been instrumental in achieving the learner's goals. Thus, it is natural to organize concepts around the goals they help satisfy. If we index concepts by the goals with which they are associated, then the agent can use its currently active goals as probes to retrieve potentially relevant concepts. As a result, the partial match is constrained to those concepts likely to aid in achieving the current goal, presumably a few instead of thousands.

Since it is central to the recognition process, we should say a little more about the partial matching mechanism. Given the description of some object and the characterization of some concept, the matcher returns a mapping between the two structures, along with the degree to which the match was successful. If the match was high, then the agent can infer that the object will prove ideal for satisfying the goal under which its concept was indexed. If the match is only fair, then it may still want to use the object, provided no better objects are found in the immediate vicinity. Furthermore, since information about which operators to apply is included in the object concept, guidelines for instrumental use of the object derive not from additional problem solving, but directly from the concept's content. Thus, inferences regarding the object's functionality co-occur with recognition.

Goals ⇒ Aggregation
Incremental weighting ⇒ Characterization
Goal–indexed partial matching ⇒ Utilization

Figure 2. An alternative approach to concept formation.

CONCLUSION

In the preceding pages, we identified three components of the concept learning process – aggregation, characterization, and utilization – and found that earlier work relied on two assumptions that made each of the tasks manageable. The first involved the presence of a tutor, who made the aggregation problem trivial by providing positive and negative instances of the concept to be learned. The second involved the notion that concepts are all–or–none in nature, so that they can be described by a set of necessary and sufficient conditions. Since we were concerned with concept formation in real–world settings, we rejected these two assumptions. However, this forced us to propose new methods for dealing with the three components of concept formation. In response, we proposed an alternative framework in which goals were used to aggregate experience. In this approach, goals are also used to index and retrieve potentially relevant concepts, reducing the task of partial matching against prototypes to reasonable proportions. Finally, we proposed a method for incrementally characterizing concepts, based on weighting numeric features and operators in terms of their observed variance. This gave us both physical criteria for recognizing future instances of the concept and information about the functional properties of objects. Taken together, we believe that these methods constitute a viable alternative to traditional approaches to concept formation, and in our future work we plan to instantiate the framework as a running system, and to test its learning abilities in a complex, reactive environment.

REFERENCES

Binford, T.O. (1971) Visual perception by computer. In *Proceedings IEEE Conference on Systems Science and Cybernetics.*

Hayes–Roth, F. (1978) The role of partial and best matches in knowledge systems. In D.A. Waterman & F. Hayes–Roth (Eds.) *Pattern–directed inference systems*, NY: Academic Press, 557–576.

Hunt, E.B., Marin, J., & Stone, P. (1966) *Experiments in induction.* NY: Academic Press.

Mitchell, T.M., Utgoff, P.E., & Banerji, R. (1983) Learning by experimentation: Acquiring and refining problem solving heuristics. In R.S. Michalski, J.G. Carbonell, & T.M. Mitchell (Eds.) *Machine learning*, Palo Alto, CA: Tioga Publishing Company, 163–190.

Newell, A., & Simon, H. (1972) *Human problem solving.* Englewood Cliffs, NJ: Prentice–Hall.

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

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