Prediction Performance as a Function of the Representation Language in Concept Formation Systems*

Mirsad Hadzikadic

Department of Computer Science
University of North Carolina
Charlotte, NC 28223
mirsad@unccvax.uncc.edu

Abstract

Existing concept formation systems employ diverse representation formalisms, ranging from logical to probabilistic, to describe acquired concepts. Those systems are usually evaluated in terms of their prediction performance and/or psychological validity. The evaluation studies, however, fail to take into account the underlying concept representation as one of the parameters that influence the system performance. So, whatever the outcome, the performance is bound to be interpreted as 'representation-specific.' This paper evaluates the performance of INC2, an incremental concept formation system, relative to the language used for representing concepts. The study includes the whole continuum, from logical to probabilistic representation. The results demonstrate the correctness of our assumption that performance does depend on the chosen concept representation language.

Introduction

Concepts lie at the core of human thought, perception, speech, and action. Consequently, the issue of *concept formation* represents an important research problem of interest to researchers from diverse disciplines, including psychology, philosophy, linguistics, and artificial intelligence. The section on concept formation partially summarizes past work in the above disciplines.

One of the far-reaching decisions to be made by every investigator/system designer is the language(s) for representing concepts and instances. The representation language defines not only how *easily* a concept can be learned, but, more importantly, *what* kind of concept can be acquired. Also, it seems plausible that the same representation cannot be equally well suited for different tasks in different application domains under different circumstances. Therefore, the goal of this paper is to *evaluate the relationship between performance and representation language in concept formation systems.* The 'Concept Representation' section provides a brief overview of different representation formalisms, while the following section explains the specifics of two evaluation methods, i.e., prediction accuracy and psychological validity.

The experimental tool used in this process is INC2 (Hadzikadic and Elia, 1991; Hadzikadic and Yun, 1989), an incremental, similarity-based concept formation system. The INC2's architecture, briefly explained in the 'Representation Continuum' section, allows us to easily modify its representation language both statically and dynamically in order to understand a potential correlation between performance and representation.

The remaining sections of the paper summarize the results of our analysis with respect to both prediction performance and psychological evaluation.

Concept Formation

Concept formation refers to the incremental process of constructing a hierarchy of concept descriptions (categories) which characterize objects in a given domain. A system which can accomplish this task can be used both as an aid in organizing and summarizing complex data and as a retrieval system which can predict properties of previously unseen objects. Such a system will be useful in domains where knowledge is incomplete or classifications and/or human experts do not exist.

Most existing *concept formation* systems use hill-climbing methods to find suboptimal clusterings of objects to be characterized. Six existing systems which share all of the above features are COBWEB (Fisher, 1987), CLASSIT (Gennari, Langley, and Fisher, 1989), UNIMEM (Lebowitz, 1987), CYRUS (Kolodner, 1984), WITT (Hanson and Bauer, 1989), and INC2 (Hadzikadic and Elia, 1991).

Researchers from disciplines other than computer science, e.g., psychology, philosophy, and linguistics, have been very active in this area as well. For example, Wittgenstein's research (1953) is associated with the ideas of *family resemblance.* Family resemblance introduces the idea that members of a category may be related to one another without all members having any properties in common that define that category.

Brown (1958) begins the study of what will later

*This work was supported by the grants from the College of Engineering and the Office of Academic Affairs, UNCC.

become known as basic-level categories. *Basic-level* categorization places the cognitively basic categories in the 'middle' of a general-to-specific hierarchy. Generalization and specialization, then, proceed upward and downward, respectively, from the basic level.

Finally, Rosch and her collaborators (1976) suggest that thought in general is organized in terms of *prototypes* ('best' examples) and basic-level structures. Their work establishes research paradigms in cognitive psychology for demonstrating family resemblance and basic-level categorization.

Concept Representation

The system that established the field of conceptual clustering, CLUSTER/2 (Michalski and Stepp, 1983) used a logic-based representation to represent both instances and concepts. The concepts were represented as conjunctions of necessary and sufficient features (logic expressions). The membership in a class was defined as all or none, depending on whether the instances possessed the required features or not.

In contrast, many researchers (as indicated in the previous section) have suggested that some instances are better examples of the concept than others, and that instances of the concept are distributed all over the space defined by the concept features. The best example (prototype) is the center of that space, with 'good' examples gravitating toward the center, while the 'bad' ones lie at the concept's periphery. Clearly, a logic-based representation, in its original form, cannot capture such distributional information. *Probabilistic* concept representations (Smith and Medin, 1981), however, handle this problem easily by associating a probability (weight) with each feature of a concept definition. This weight is usually implemented as the conditional probability $p(f \mid C)$ of the feature f's presence, given category C. In literature, it is often referred to as *category validity* of the feature. The retrieval and prediction, using probabilistic concepts, are usually based on the comparison between the sum of the feature weights and a given threshold (Smith and Medin, 1981). Both COBWEB and INC2 systems are based on a hierarchical probabilistic representation of concepts, where the hierarchical structure eliminates the weakness of simple probabilistic representations, namely their inability to capture non-linear correlations among features.

Probabilistic representations are more general than the logic-based ones in a sense that the former can simulate the latter by dropping all features with the category probability of less than 1.0. In addition, it is easy to imagine a continuum of probabilistic representations which differ in the value of their feature drop threshold. The drop threshold will range from 0.0 (initial probabilistic representation) to 1.0 (logic representation).

Performance Tasks

The choice of the drop threshold (and ultimately the representation) may influence the performance of the system. Prediction and psychological validity are the two performance tasks most frequently used in concept formation systems.

Prediction refers to the process of drawing inferences in regard to the category membership of previously unseen instances. It assumes existence of two key components: (1) a set of concepts known to the system, and (2) a domain-independent heuristic which indicates the likelihood of each concept being the target category. Concept formation systems usually rely on heuristics developed in psychology to guide the classification process. For example, INC2 utilizes the *contrast model* (Tversky, 1977) to compute the similarity between two objects/concepts and *family resemblance* (Wittgenstein, 1953) to decide whether to place an object into the category or not. On the other hand, COBWEB makes use of *category utility* (Gluck and Corter, 1985) to find the optimal clustering at each level of the hierarchy.

Psychological validity, on the other hand, emphasizes the importance of psychological findings (human subject studies) and measures the extent of their overlap with the results of the concept-formation systems. These findings include *typicality, basic level categories, and intra- and inter-category similarity.* More often than not, concept formation systems rely on their heuristic evaluation function (category validity in COBWEB; contrast model and family resemblance in INC2) to demonstrate 'human-like' performance as a side effect.

Representation Continuum

The experimental tool used in this evaluation study is INC2, an incremental concept formation system which builds a hierarchy of concept descriptions. The leaves of the hierarchy are objects (singleton concepts). The root of the hierarchy has associated with it a description which is a summary of the descriptions of all objects seen to date by the system.

In addition to features and hierarchical pointers, each concept description contains an estimate of its cohesiveness, given in the form of *family resemblance* (Wittgenstein, 1953). Family resemblance is defined as the average similarity between all possible pairs of objects in a given category. The similarity function used by INC2 represents a variation of the contrast model (Tversky, 1977), which defines the similarity between an object and a category as a linear combination of both common and distinctive features. As a result, INC2 implements a hill-climbing strategy which encourages advancement toward the maximal improvement of the hierarchy as measured by the increase in the family resemblance of the host concept.

INC2 uses a probabilistic representation to store concept descriptions. A description of each concept C is defined as a set of features f (attribute-value pairs). Each feature has a conditional probability $p(f \mid C)$ associated with it. Thus, representing the color feature of red apples would take the form (*color red* 0.75). The 0.75 means that members of this category are red 75% of the time. Since members of a given concept may reside in distinct portions of the hierarchy, the adopted representation formalism is referred to as a *distributed probabilistic concept hierarchy.*

The only threshold introduced in INC2 is a drop threshold. This threshold allows for concept descriptions to be either probabilistic or logical. It can be set anywhere between 0.0 and 1.0, and means that any feature with the conditional probability below this threshold should be dropped[1] from the concept description. The value of 1.0 for this threshold would yield a logical concept description. It is easy to imagine systems with different values for the drop threshold, e.g., 0.75 (each instance should have at least 3/4 of the features in common with other instances of the category), or 0.5 (at least 1/2 common features).

The drop threshold is static in nature, i.e., the same value is used at every level of the hierarchy and for all instances, no matter what their time of arrival or path of incorporation happens to be. However, the nature of classification calls for a dynamically adjusted threshold rather than a fixed one. For example, all features are important at the top level of the hierarchy, no matter how low their probabilities might be, due to the diversity of objects in the domain as well as the potential noise in object descriptions. Therefore, the drop threshold should be set close to 0.0. At the lower levels of the hierarchy, however, certain patterns have been detected, resulting in high conditional probabilities for 'relevant' features and low probabilities for the ones not significantly present in those patterns. Since all categories at the lower levels have few members, all the features found in their descriptions will have relatively high conditional probabilities. To avoid the interference of irrelevant features with the retrieval process, the drop threshold should be set close to 1.0. The intermediate categories will, then, require the drop threshold somewhere between 0.0 and 1.0, depending on the level of the hierarchy (the lower the level, the higher the drop threshold).

In order to accommodate this type of reasoning, INC2 relies on family resemblance to provide an estimate of the drop threshold value. Family resemblance is naturally set close to 0.0 at the root (summarizing the whole universe) and to 1.0 at the leaves. Consequently, INC2 automatically sets the drop threshold to the value of the family resemblance of the parent category during both classification and retrieval. That value increases with the object traversing the hierarchy downward. INC2, therefore, performs a context-sensitive classification/retrieval due to its adaptive behavior that changes from level to level of the hierarchy. In that process, INC2 uses different representations to describe objects/categories at different levels of the hierarchy, possibly moving from the probabilistic representation (drop threshold = 0.0) at the top level to the logical one (drop threshold = 1.0) at the leaves.

The idea of a dynamically adjusted drop threshold, coupled with the fact that features are only dropped temporarily (until the changing environment will have brought them back into the foreground of the system's attention), effectively emulates the idea of tracking *concept drift* (i.e., adapting to concepts that change over time) as advanced by Schlimmer and Granger (1986).

Prediction Performance Evaluation

At this point, the reader should have a sufficient understanding of INC2's representation formalism to appreciate the context in which the probabilistic-vs-logical-representation experiment has been carried out. We will briefly describe, next, the domain of clinical audiology in which the experiment took place, and then the experiment itself.

The audiology domain consists of 200 cases, 58 features, and 24 ideal categories[2]. The distribution of cases across the categories varies from one to 48 per category. Half of the categories are represented by only one or two cases. Such a distribution certainly makes learning almost impossible for those categories that are under-represented. The cases include noise in the form of incorrect and/or missing features. On average, each case has only 11 features with known values.

The probabilistic-vs-logical-representation experiment involved four different sizes of the training set (20, 50, 100, and 150) and six different values for the drop threshold (variable, 0.0, 0.25, 0.5, 0.75, and 1.0). The size of the test set was kept constant at all times (45 objects -- 22.5% of the total object set). Figure 1 summarizes the percentage of correct responses, averaged over five runs with randomly chosen objects, for all of the above cases.

[1] This happens only temporarily since new object acquisitions may bring that feature back into the concept description.

[2] Provided by Prof. Jergen from the Baylor College of Medicine and Bruce Porter of the University of Texas at Austin.

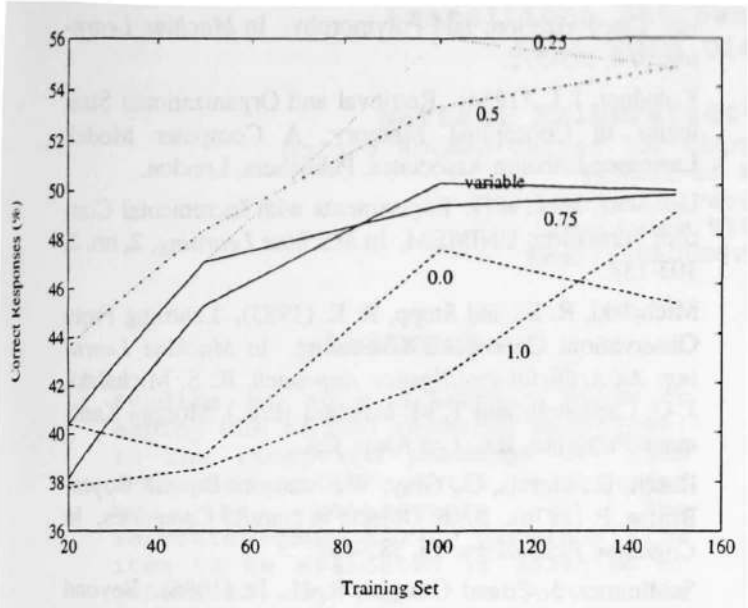

Figure 1: Prediction performance for both variable and fixed drop thresholds.

In the case of the set with a low number (20) of input objects, the variable drop threshold was outperformed by all the fixed-version values except for 1.0. The picture, however, changed for larger sets (e.g., 50, 100, and 150). The variable threshold clearly outperformed both probabilistic and logical representations, while scoring comparably to the 0.75 case. At the same time, the probabilistic representation consistently demonstrated better performance than the logical one, though not decisively so.

Unexpected results, however, came from the strong performance of the 0.25 and 0.5 cases, which clearly proved to be the best choice in our experiments. The 0.5 performed better than the 0.25 in the experiments with a low number of training objects (actually, even the 0.75 case was as good as the 0.25 under those conditions), while the 0.25 demonstrated its strength in the cases with a large number of input objects. These results seemed to indicate that neither storing all features nor 'forgetting' those that do not hold for all instances of the concept maximizes the performance of the system or provides a clear advantage over one another.

In addition, the results demonstrated the need for 'forgetting' those features that were irrelevant for the category membership. It remained unclear, however, how to 'recognize' them. Forgetting the features that do not hold for at least a half of the concept instances proved to be beneficial for the low number of training instances. An increased number of training objects provided some new evidence about the importance of certain features, and the drop threshold had to be lowered in order to improve the system performance. This evidence is in line with the reasoning behind the variable drop threshold, which adopts higher values for the nodes closer to the leaves (summarizing but a few input cases) and lower values for the nodes closer to the root (those that accumulate higher levels of experience).

Psychological Evaluation

In addition to its prediction performance, INC2 has been evaluated in terms of the psychological validity of its results. There are three issues of special interest here: typicality, basic level categories, and intra-category similarity vs. inter-category dissimilarity.

Due to the uneven distribution of instances, two classes *(cochlear age* and *cochlear unknown)* accounted for 70% of all retrievals. In order to evaluate the quality of retrieved objects in this domain, we decided to closely examine the objects from one of those classes, *cochlear age.* First, we calculated the average similarity of each object with all other members of the category. The similarity ranged from 0.0 to 0.527. The objects with the similarity greater than or equal to 0.5 were considered to be 'good' examples of the category. Then, we reviewed the list of often-retrieved objects and noticed that over 60% of them were among the examples regarded as 'good.' This finding was consistent with the prototype theory.

In addition, we reviewed all objects retrieved at least once, and for each such object calculated its average similarity. As expected, the frequency of retrieval was roughly proportional to the average similarity of the object. Consequently, we can conclude that the INC2-generated hierarchies demonstrate typicality effects similar to those generated by human subjects.

Due to the strategy adopted in its concept formation algorithm (place an object into the category if it increases the family resemblance of the category), INC2 always incorporates the object at its basic level. While traversing the hierarchy, and before it will have reached the basic level, the object encounters more and more familiar objects and categories, i.e., the ones it has more features in common with than with any previously encountered object/category. That will stop at the basic level, however, since the remaining objects/categories will begin having more and more differing features due to their increased specialization within the hierarchy. It is important to notice that objects may have their basic level at different levels of the hierarchy (depending on the order of objects and local context), thus leading to the notion of a *distributed basic level.*

Finally, the issue of intra-category similarity vs. inter-category dissimilarity is addressed implicitly in INC2,

again through its algorithm. Namely, the system will place an object into the category which maximizes the increase in the category's family resemblance (compactness). Consequently, the category that receives the object will pull its instances somewhat closer to its imaginative center, thus positioning itself away from other 'gravitation points' in the instance/category space. This process will automatically reduce the force (similarity) between the category and the surrounding concepts.

Summary

This paper has evaluated the relationship between performance and adopted category/object representation. We varied the representation from probabilistic to logical, and compared their corresponding performance on the prediction task. An alternative approach, variable representation, was evaluated as well. It was characterized by the constant switching among different representation schemas according to the value of the compactness of the categories stored at different levels of the hierarchy. The variable-threshold approach worked consistently better than either the probabilistic or the logical representation. It did not, however, match the success of the fixed, middle-of-the-road-valued drop threshold.

This last observation represents our research agenda. We will continue to search for the ways to automatically set the optimal value for the variable drop threshold. In addition, we will extensively evaluate the system in terms of the cost/accuracy trade-off as it moves from probabilistic to logical representation.

References

Brown, R. (1958). How Shall a Thing be Called? *Psychological Review,* **65,** 14-21.

Fisher, D. H. (1987). Knowledge Acquisition Via Incremental Conceptual Clustering. In *Machine Learning,* **2,** 2, 139-172.

Gennari, J. H., Langley, P., and Fisher, D. H. (1989). Models of Incremental Concept Formation. In *Artificial Intelligence,* **4,** 1-3, 11-61.

Gluck, M. A. and Corter, J. E. (1985). Information, Uncertainty, and the Utility of Categories. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society,* 283-287, Irvine, CA, Lawrence Erlbaum.

Hadzikadic, M. and Yun, D. Y. Y. (1989). Concept Formation by Incremental Conceptual Clustering. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence,* 831-836, Detroit, MI.

Hadzikadic, M. and Elia, P. (1991). Context-Sensitive, Distributed, Variable-Representation Category Formation. *Proceedings of the Thirteenth Annual Meeting of the Cognitive Science Society,* 269-274, Chicago Illinois.

Hanson, S. J. and Bauer, M. (1989). Conceptual Clustering, Categorization, and Polymorphy. In *Machine Learning,* **3,** 4, 343-372.

Kolodner, J. L. (1984). Retrieval and Organizational Strategies in Conceptual Memory: A Computer Model. Lawrence Erlbaum Associates, Publishers, London.

Lebowitz, M. (1987). Experiments with Incremental Concept Formation: UNIMEM. In *Machine Learning,* **2,** no. 2, 103-138.

Michalski, R. S., and Stepp, R. E. (1983). Learning From Observation: Conceptual Clustering. In *Machine Learning: An Artificial Intelligence Approach,* R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (Eds.), Morgan Kaufmann Publishes, Inc., Los Altos, CA.

Rosch, E., Mervis, C., Gray, W., Johnson, D., and Boyes-Braem, P. (1976). Basic Objects in Natural Categories. In *Cognitive Psychology,* **18,** 382-439.

Schlimmer, J. C. and Granger, R. H., Jr. (1986). Beyond Incremental Processing: Tracking Concept Drift. *Proceedings of the Fifth National Conference on Artificial Intelligence,* 502-507, Philadelphia, PA.

Smith, E. E. and Medin, D. L. (1981). Categories and Concepts. Harvard University Press, Cambridge, MA.

Tversky, A. (1977). Features of Similarity. *Psychological review,* **84,** 327-352

Wittgenstein, L. (1953). Philosophical Investigations. MacMillan, New York.