Modeling the Effect of Category Use on Learning and Representation

Kenneth J. Kurtz (kjk@northwestern.edu)
John T. Cochener (cochener@howard.psych.nwu.edu)
Douglas L. Medin (medin@northwestern.edu)
Department of Psychology, Northwestern University, 2029 Sheridan Rd
Evanston, IL 60208-2710 USA

The formation of categories is based on learning to perform various cognitive tasks and not just categorization per se (see Yamauchi & Markman, 1998). Most models of categorization are committed to explaining the learning, representation, and use of categories solely in terms of taxonomic classification. Recent evidence shows that category structure can be derived from and organized to support the use of examples for particular tasks or goals (e.g., Ross, 1997; Medin, Lynch, Coley, & Atran, 1997).

The ORACLE model (One Representation Across Channels of LEarning) addresses concept formation from the perspective of learning how to represent examples in order to support performance on naturally occurring cognitive tasks. This "you are what you eat" approach stresses that concepts emerge not so much in order to represent the world as it is, but to represent the world relative to the learner's needs and demands. ORACLE is based on two key ideas: 1) input examples are re-represented through error-driven learning to improve task performance (Rumelhart, Hinton, & Williams, 1986); and 2) conceptual organization emerges from numerous iterations of parameter optimization on multiple interwoven processing tasks.

As in a traditional multi-layer network, inputs tend to become represented more closely in the constructed multidimensional space of a hidden layer to the extent they share the same teaching signal at the output layer. ORACLE has two further design principles: 1) different channels of learning lead to different sets of outputs, but share a single set of hidden units; and 2) an anchor channel is dedicated to auto-associative learning (reproducing input information at output). The number and nature of the channels depends upon the tasks or goals the inputs participate in for the learner. Classification is the function being approximated, but the output classes correspond to relevant uses, linguistic labels, or implications; not to established categories. ORACLE predicts that internal representations will tend to emerge that emphasize elements of the input which are most useful for performance across the channels of learning. Therefore, we attempted to simulate Ross' (1997, E1) category use effect.

Participants in Ross' study learned to diagnose fictitious diseases and to select the right treatment. Learning trials were: guess the disease, feedback, guess the treatment, feedback. Each patient consisted of one symptom perfectly predictive for both tasks, one symptom perfectly predictive for diagnosis only, and one non-predictive symptom. After study, participants tested on disease classification performed better on features relevant to both disease and treatment (96%) than on features relevant only to disease (80%). Learning to treat the diseases influenced diagnosis.

The ORACLE architecture consisted of an input layer and a hidden layer of three hidden units (n=3) with projections along three channels of learning to the task outputs (disease, treatment, and auto-association). The layers were feedforward and fully connected. A patient was presented to ORACLE by activating three input units in a 3x12 array based on the twelve possible symptoms and the three possible presentation positions.

Each epoch of standard back-propagation training consisted of randomly ordered presentation of the sixteen patients in all possible symptom orders at a low learning rate. To simulate the two types of training trials, an alternating scheme was used in which half the trials set targets on the disease outputs and half on the treatment outputs. Auto-associative targets stayed set.

The category use effect does not appear initially, but tends to emerge as training proceeds. After 10,000 epochs, 11 of 13 ORACLE runs performed better on features relevant to both tasks. Mean activation of the correct disease output was 96% for the double-relevant and 88% for single-relevance features. ORACLE produces the category use effect by learning anchored representations that support diagnosis and treatment.

References

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32,* 49-96.

Rumelhart, D.E., Hinton, G.E., & Williams, J.R. (1986). Learning internal representations by error propagation. In D.E. Rumelhart & J.L. McClelland (Eds). *Parallel Distributed Processing, Vol. I.* Cambridge, MA: MIT Press

Yamauchi, Y., & Markman, A.B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39,* 124-148.

Ross, B.H. (1997). The use of categories affects classification. *Journal of Memory and Language, 37,* 240-267.