Inducing Rule-like Learning in Connectionist Architectures

Eric Yeh, Matthew Crossley
SRI International

Summary
Models of rule-based learning typically come in the form of symbolic programs. Since the human brain is connectionist in nature, these models cannot be considered as implementation-level explanations. However, as typical connectionist networks learn slowly and do not generalize this learning well to novel circumstances, they do not display the signatures of rule-based learning. We show that using information theoretic regularization, artificial networks behave in a manner consistent with human rule-based category learning. The resulting networks can also be used to generate excellent explanations on a proxy human decision-making transfer learning task.

Experimental Setup
1. Perceptual task: Classify 30x30 sine-wave gratings into two categories, determined by hyperplane through two stimulus dimensions, Orientation and Frequency.

Example gratings classified into categories A and B along hyperplane separating frequency and orientation of gratings.

2. Train network to regress orientation and frequency from the gratings image, then add a Softmax to classify the category.

\[
P(W_{ij}) = \frac{|W_{ij}|}{\sum_{kl} |W_{kl}|}
\]

\[
C_{info}(W) = -\sum_{i,j} P(W_{ij}) \log P(W_{ij})
\]

W = Weights, C_{info} = Information Theoretic Regularization penalty

Rule-Based Classification Experiments
Three perceptual experiments tested mean accuracies on validation against learning rate (amount of training data). Rates compared on two data-dictated classification conditions in humans: Rule-Based (RB) and Information Integration (II). Experiments 2 and 3 featured a distribution change, marked by Pre/Post, with network weights transferred from the Pre to Post network.

1. Rule-based classification learning is fast

Condition: RB  
Condition: II

Means for both conditions are shifted in the Post stage.

The rule-based network is able to match human learning rate curves for RB and II conditions.

2. Rule-based classification generalized to novel stimuli

Condition: RB  
Condition: II

As with human studies, RB condition demonstrates transfer effects, while II does not.

3. Rule-based classification adapts to label swaps

Condition: RB  
Condition: II

Network is able to recover more quickly in RB condition than II.

Classification Transfer Experiments
Test “explanations” learned by declarative perceptual network by transferring simple rules for classification to a simpler network.

1. Use bottleneck perceptual network to “teach” a one-nearest-neighbor classifier to determine if a digit image is even or odd. Used a row-wise variant of information theoretic regularization on the Softmax classification layer.

2. “Rules” are exemplars and label. One-nearest-neighbor is a proxy for simple human decision making. Exemplars drawn from each bottleneck node, using simple average of top 10 activating images.

\[
C_{rowinfo}(W) = -\sum_{i,j} P(W_{ij}) \log P(W_{ij})
\]

C_{rowinfo} = Row-wise Information Regularization penalty applied at bottleneck and Softmax.

3. Exemplars from rule-like learning performed best. They were also less confusable than those drawn from other methods, measured on a sum of inter-exemplar similarities.

3. Exemplars from rule-like learning performed best. They were also less confusable than those drawn from other methods, measured on a sum of inter-exemplar similarities.

For this experiment, a 10 node bottleneck was used to produce 10 exemplar and label “rules.”

4. Exemplars from network using row-wise information regularization were more diverse than those using L2.

Future Work
- Use a more realistic online learning model, making it better match human experiments.
- Does rule-like learning entail information entropy-like synaptic connectivity?

Acknowledgments
Supported by SRI International Information Computing Sciences internal research and development.