Cognitive Models of Test-Item Effects in Human Category Learning

Xiaojin Zhu, Bryan R. Gibson, Kwang-Sung Jun, Timothy T. Rogers, Joseph Harrison, Chuck Kalish

Departments of Computer Sciences, Psychology, and Educational Psychology
University of Wisconsin-Madison
Madison, WI 53706 USA

Abstract

Imagine two identical people receive exactly the same training on how to classify certain objects. Perhaps surprisingly, we show that one can then manipulate them into classifying some test items in opposite ways, simply depending on what other test items they are asked to classify (without label feedback). We call this the Test-Item Effect, which can be induced by the order or the distribution of test items. We formulate the Test-Item Effect as online semi-supervised learning, and extend three standard human category learning models to explain it.

The Test-Item Effects in Human Category Learning

A computer can hold a trained classifier fixed during testing.
A human cannot.

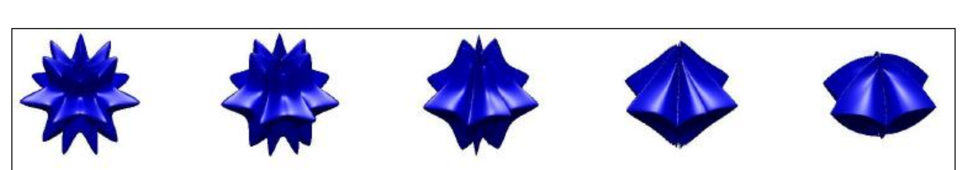
Test-Item Effect: Unlabeled test items change the classifier in human's mind. Two otherwise identical people A, B receiving exactly the same training data can be made to disagree on certain test items x, simply by manipulating what other test data they are asked to classify, *without label feedback*.

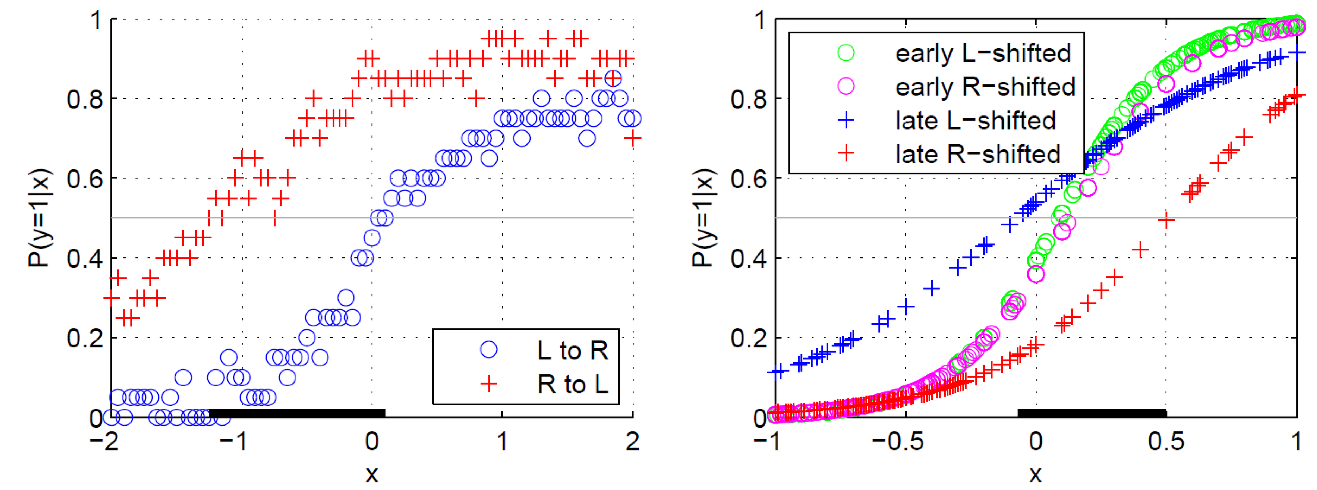
Test-Item Effect 1: Order of test items
40 subjects, 1D feature space, 10 labeled items {(x=-2, y=0), (2,1)} *5
Two conditions, 20 subjects each:
- L to R: test item −2,−1.95,−1.9, . . . , 2
- R to L: reverse order.

Results: Subjects in the "L to R" condition tend to classify more test items as y = 0, and vice versa. For test items in [−1.2, 0.1], a majority-vote among subjects will classify them in opposite ways in these two conditions.

Test-Item Effect 2: Distribution of test items
22 subjects, same feature space, 20 labeled items {(-1,0), (1,1)}*10
Test items drawn from two-component GMM. Two conditions:
- L shifted: GMM means at -1.43 and 0.57
- R shifted: GMM means at -0.57 and 1.43

Results: early (in first 50 test items) decision boundaries the same; late (after 700 test items) boundaries shifted according to condition





The Test-Item Effect due to order (Left) and distribution (Right)

Test-Item Effects as Online Semi-Supervised Learning

The key is to update classifier upon unlabeled data. Standard human category learning models in psychology (equivalent to supervised learning models) cannot explain test-item effects. We propose and compare three online semi-supervised extensions:

Semi-Supervised Exemplar Model
= self-training Nadaraya-Watson kernel estimator; extends the generalized context model (Nosofsky, 1986)

Algorithm 1 Semi-Supervised Exemplar Model
- **Parameter:** kernel bandwidth h
- **for** $n = 1, 2, \ldots$ **do**
  - Receive x_n, predict its label by thresholding $r(x_n) = \sum_{i=1}^{n-1} \frac{K(\frac{x_n - x_i}{h})}{\sum_{j=1}^{n-1} K(\frac{x_n - x_j}{h})} \hat{y}_i$ at 0.5
  - Receive y_n (may be unlabeled), update model:
  - **if** y_n is unlabeled **then**
    - $\hat{y}_n = r(x_n)$
  - **else**
    - $\hat{y}_n = y_n$
  - **end if**
- **end for**

Semi-Supervised Prototype Model
= incremental EM on GMM (Neal & Hinton, 1998), but without revisiting old items; extends (Posner & Keele, 1968)

Algorithm 2 Semi-Supervised Prototype Model
- **Parameter:** Prior encoded in ϕ
- Initialize $\theta^{(0)}$ from ϕ (see M-step below)
- **for** $n = 1, 2, \ldots$ **do**
  - Receive x_n, classify by $q(y) = P(y|x_n, \theta^{(n-1)})$
  - Receive y_n (may be unlabeled), update model
  - **E-step:**
  - **if** y_n is unlabeled **then**
    - $\phi = \phi + \mathbb{E}_q[\tilde{\phi}(x_n, y)]$
  - **else**
    - $\phi = \phi + \tilde{\phi}(x_n, y_n)$
  - **end if**
  - **M-step**: Let $\phi = (n_0, s_0, ss_0, n_1, s_1, ss_1)$. Compute $\theta^{(n)}$ as follows: $\alpha = \frac{n_1}{n_0+n_1}$, $\mu_0 = \frac{s_0}{n_0}$, $\sigma_0^2 = \frac{ss_0}{n_0} - \left(\frac{s_0}{n_0}\right)^2$, $\mu_1 = \frac{s_1}{n_1}$, $\sigma_1^2 = \frac{ss_1}{n_1} - \left(\frac{s_1}{n_1}\right)^2$
- **end for**

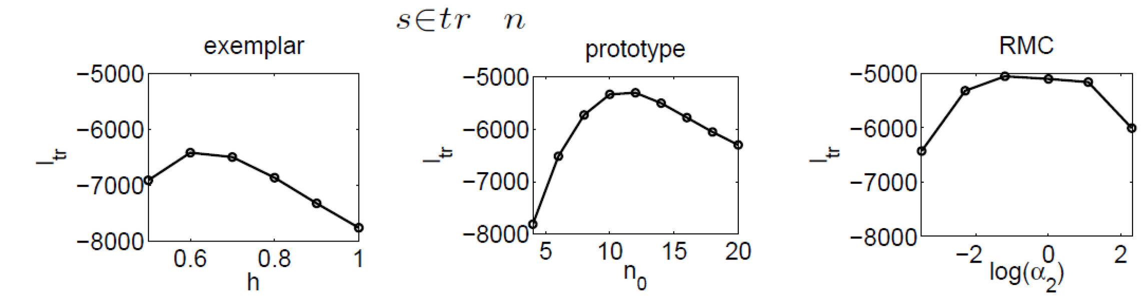
Semi-Supervised Rational Model of Categorization
= Dirichlet Process Mixture Model with marginalization over y; extends (Anderson, 1990)

Algorithm 3 Semi-Supervised Rational Model of Categorization
- **Parameters:** $\alpha_2, \mu_0, \kappa_0, \alpha_0, \beta_0, \alpha_1, \beta_1$
- Initialize m empty particles; y_0 =unlabeled
- **for** $n = 1, 2, \ldots$ **do**
  - Receive y_{n-1} (may be unlabeled) and x_n
  - Re-sample m particles
  - Predict y_n with new particles
- **end for**

Which Model Fits Humans Better?

Parameter tuning: divide subjects into "training" and "test" groups. Maximize training group human prediction likelihood.

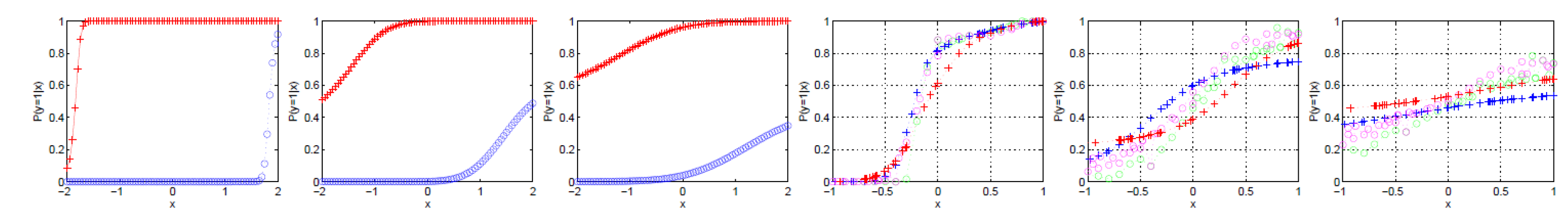
$$\ell_{tr}(\theta) \equiv \sum_{s \in tr} \sum_{n} \log P(h_n^{[s]} \mid x_{1:n}^{[s]}, y_{1:n-1}^{[s]}, \theta)$$



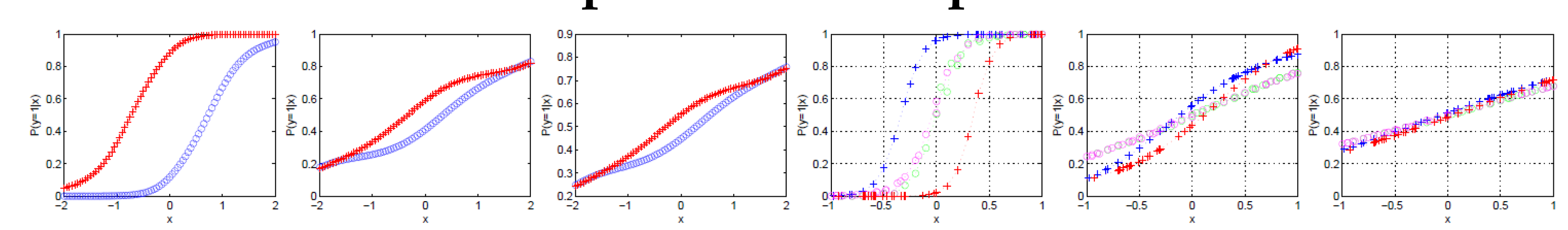
The learned parameters and the test group log likelihood:

	exemplar	prototype	RMC
$\hat{\theta}$	$h = 0.6$	$n_0 = 12$	$\alpha_2 = 0.3$
$\ell_{te}(\hat{\theta})$	-3727	-2460	**-2169**

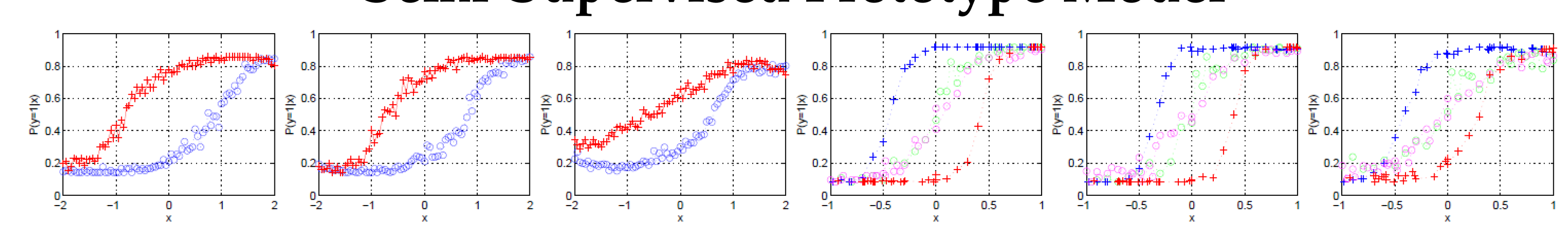
Model behavior under different parameters:



Semi-Supervised Exemplar Model



Semi-Supervised Prototype Model

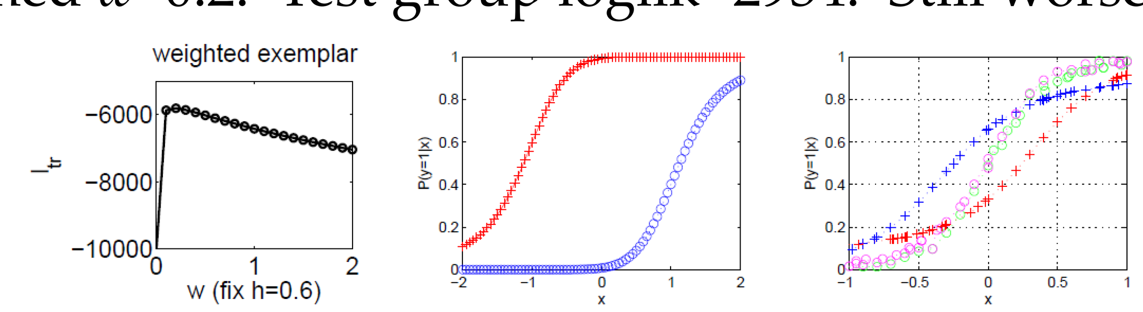


Semi-Supervised Rational Model of Categorization

Observations:
1. All models exhibits test-item effects;
2. *Semi-supervised RMC has the best fit*
3. Semi-supervised exemplar model is particularly poor
   - What if we down-weight unlabeled items?

$$r(x) = \sum_{i=1}^{n} \frac{w_i K(\frac{x - x_i}{h})}{\sum_{j=1}^{n} w_i K(\frac{x - x_j}{h})} y_i$$

   - Learned w=0.2. Test group loglik -2934. Still worse.



This work is supported in part by AFOSR FA9550-09-1-0313, NSF IIS-0916038, IIS- 0953219, DLS/DRM-0745423, and the Wisconsin Alumni Research Foundation.