UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Extending the Locally Bayesian Learning Model to Exemplar-Based Categorization with Continuous Features

Permalink

https://escholarship.org/uc/item/57r329zj

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Chang, Yu-Wei
Aytac, Sinem
Mendoza Gonzalez, Cindy G
et al.

Publication Date

2024

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Extending the Locally Bayesian Learning Model to Exemplar-Based Categorization with Continuous Features

Yu-Wei Chang (ychang01@syr.edu)
Sinem Aytaç (saytac@syr.edu)
Cindy Mendoza (cgmendoz@syr.edu)
Michael L. Kalish (mlkalish@syr.edu)
Daniel Corral (dcorral@syr.edu)
Department of Psychology, Syracuse University,
430 Huntington Hall, Syracuse, NY 13244 USA

Abstract

The Locally Bayesian Learning (LBL) approach bridges the gap between optimal Bayesian learning and suboptimal performance that arises from human behavior. Although this learning model has considerable potential, it has been underdeveloped and has remained in its original form for several decades. In this paper, we extend the original LBL model to an exemplar approach, which we refer to as the exemplar-LBL model. Two notable features of this extension are that (a) the model can take continuous features as inputs and (b) can conduct exemplar-based categorization. We report various simulations, which show that the model can generate numerous important predictions about category learning. Additionally, we introduce the *extra-learning hypothesis*, which can account for how classification and observation training can produce differential learning. Our results showcase scenarios under which classification training is superior to observation training and other instances in which the opposite occurs.

Keywords: Locally Bayesian Learning; Categorization; Classification Training; Observational Training; Exemplar-based Learning.

Introduction

The Bayesian framework has been applied to different areas of human cognition, which has led to important insights (e.g., Kersten et al., 2004; Lee, 2006; Tenenbaum & Griffiths, 2001; Xu & Tenenbaum, 2007). These models hold the potential to enhance our understanding of human cognition by addressing how learners update their beliefs when they encounter new data during problem solving (Chater et al., 2010; Griffiths et al., 2008; Perfors et al., 2011). However, there is a gap between the optimal performance predicted by Bayesian models and the sub-rational or suboptimal behavior that humans engage in.

Kruschke (2006) argued that computational constraints explain why people cannot fully use Bayes' rule during inferential decision making. To account for this gap, Kruschke proposed a Locally Bayesian Learning (LBL) approach. The basics of the LBL are that a cognitive task can be separated into different modules, in which each of them represents a psychological process. For example, in a category learning task, people learn about which parts of an item they should attend to (e.g., shape, color, size) and learn to associate these elements to the corresponding category labels. According to the LBL model, learning within modules is purely Bayesian, but the communication between modules may not conform with the Bayesian framework. Instead, an approximate message is passed from module to module, which makes its behavior as a whole non-Bayesian.

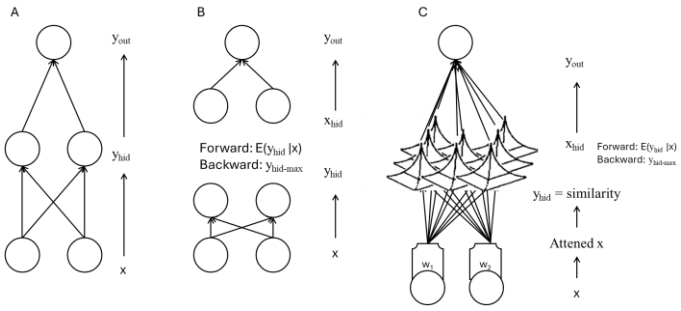


Figure 1: Diagram of (A) Globally Bayesian Learning (GBL), (B) Locally Bayesian Learning (LBL), and (C) exemplar-Locally Bayesian Learning (exemplar-LBL). Taken and edited from Sanborna and Silva (2013).

Globally Bayesian Learning (GBL) and Locally Bayesian Learning (LBL) Models

Kruschke (2006) constructed Globally Bayesian Learning (GBL) and LBL in a three-layer neural network and compared the two approaches to show the properties of LBL. The illustrative models were applied to a simple associative learning task that first mapped two features to two attentionally filtered features and then mapped attentionally filtered inputs to the outcome. The activation of a node was the sum over the weighted incoming connections and was rescaled by a sigmoid function.

$$y_{hid} = \text{sig}(W_{hid} \cdot x_{hid}) \quad (1)$$

Where W_{hid} was a 2×2 hidden weight matrix and x_{hid} was a 2×1 vector of inputs. $W_{hid} \cdot x_{hid}$ was the dot product. Because the hidden nodes, as the attended features, represented the input features, Kruschke (2006) constrained the corresponding weights (i.e., from 1st input node to 1st hidden node and from 2nd input node to 2nd hidden node) to be excitatory, with values of either 4 or 6, and the noncorresponding weights to be inhibitory, with values of either 0 or -4. Any particular set of weights is a hypothesis, which makes 16 hypotheses (i.e., 2^4) for the hidden weights.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

Similarly, the activation of the output layer was computed from x_{out}.

$$y_{out} = sig(W_{out} \cdot x_{out}) \quad (2)$$

Where W_{out} was a 1×2 hidden weight vector and x_{out} was a 2×1 vector representing the attended features. For simplicity, each output weight was allowed to take the values of 5, 0, or -5. As a result, there were 9 (i.e., 3^2) combinations and thus hypotheses of outcome weights. The activation of the single output node, y_{out}, represented the probability of selecting one outcome category over the other exclusive category.

In the GBL approach, all layers were treated as one integrated mapping. The lower layer output was passed to the next layer as the input in a probabilistically distributed form. As a result, the global hypothesis space was the cross of W_{hid} and W_{out}, yielding 144 (i.e., $2^4 \times 3^2$) hypotheses. When provided with a target output *t*, the probability distribution of the hypotheses was updated according to Bayes' theorem.

On the other hand, the LBL model splits the layers into two modules: (a)_the mapping from input to hidden (i.e., the "lower" module) and (b) the mapping from hidden to outcome (i.e., the "upper" module). Each module updated the representation with Bayes' rule but was not provided with the entire state of the probability distribution represented in the other module. The upper module only received the mean output value from the lower module as the input. Also, when a target output *t* was provided, the weights in the upper module were first updated, then only the value of $y_{hid\text{-}max}$ that maximizes the probability of the target was passed to the lower module. The lower module was blind to *t* and only updated its weights according to $y_{hid\text{-}max}$. As a result, the LBL model had in total 25 (i.e., 2^4+3^2) hypotheses.

Merits of Locally Bayesian Learning (LBL)

In this illustrative example, the GBL model had to conduct Bayesian learning on 144 hypotheses while the LBL model only had 25. The comparison is more extreme if the number of input nodes or output nodes is increased.

The large hypothesis space of the GBL model not only makes the model fitting computationally demanding and intractable, but also raises questions as to whether people can work on updating their beliefs about that many hypotheses, which can occur either overtly or covertly. Thus, it is more plausible to assume separate cognitive modules processing in the Bayesian approach themselves with simple communication between them, rather than an aggregated global Bayesian module with overwhelming computations.

A more significant merit of the LBL approach is that it can produce order effects. Kruschke (2006) demonstrated that LBL can predict (a) highlighting (assigning ambiguous stimuli as belonging to the less frequent of two categories, rather than to the more common category; Medin & Edelson, 1988), (b) blocking (the previously learned stimuli preventing the learning on latter stimuli; Kamin, 1968), and (c) unovershadowing (the release from overshadowing; Larkin et al., 1998). In contrast, the GBL approach cannot readily produce these phenomena if they assume the equal representation of instances, regardless of their order (e.g., Anderson, 1990; Dayan, et al, 2000).

Although the LBL approach can account for people's near-optimal Bayesian learning (Kruschke, 2006; Sanborn & Silva, 2013), it has typically been discussed at an abstract, theoretical level or with only simple tasks. The underdevelopment of LBL models was induced by the limitation that previous LBL models only take discrete values as input and works on simple associative learning tasks. In light of this issue, our goal here is to extend the LBL model to predict performance on tasks that involve relatively complex cognitive processing, which allows for continuous inputs. We decided to focus on category learning, not only because categorization is foundational to many practices of cognition, but also following the suggestions from Kruschke (2006).

In this study, we revised the LBL model, with the exemplar-based approach, to extend it to a category-learning task. Simulation results are reported to show the properties and predictions of different scenarios with our extension of the model. Moreover, we demonstrate that the new model can differentiate observational and classification training in learning simple categories. As described in the results sections, the implementation illustrates the merits of the LBL models in modeling a trial-order-related algorithm and predicting how expectations (priors) affect learning. The implementation results provide insights into when classification training leads to better learning than observation training, as well as when observation training leads to better learning than classification training.

Model

We extended the LBL model to account for exemplar-based categorization, which we will refer to as the exemplar-LBL. This model has three layers that are separated into two modules. The lower module maps the input to the exemplars and the upper module maps the exemplars to the category label response.

The input remains the same as the features. The weights from the inputs to exemplars represent the selective attention given to each feature dimension. As a result, the weights from the same input node to the exemplars have the same values. The weights can take real positive numbers, but for simplicity, we only consider these values being 0, 0.2, 0.4, 0.6, 0.8, or 1, where weight = 0 indicates ignoring the corresponding feature. We also do not constrain the sum of these attention weights to be 1, which allows the features to be both attended to or both be ignored. The prior on the combination of the weights is set to be uniform.

Based on the attention weights, the activation of each exemplar node represents the similarity between the input item and the exemplar. The similarity between two items is computed as the approach in the Generalized Context Model (GCM; Nosofsky, 1986).

$$d_{ij} = \left[\sum_{m=1}^{M} w_m \left|x_{im} - x_{jm}\right|^r\right]^{1/r} \quad (3)$$

The computation of the distance between item *i* and exemplar *j* is shown in equation 1. *M* is the total feature dimensions in the psychological space. w_m denotes the attention weight on feature dimension *m*. x_{im} denotes the value of item *i* on dimension *m* and x_{jm} denotes the value of exemplar *j* on the same dimension; *r* determines what distance calculation is used. We set r to be one, indicating the city-block distance calculation method.

$$S_{ij} = e^{-cd_{ij}{}^p} \quad (4)$$

Equation 4 shows the formula for computing the similarity between item *i* and exemplar *j*. The value for *c* is the sensitivity parameter reflecting the rate at which similarity declines with distance. The value *p* determines the shape of the function relating similarity to distance, and for simplicity, is set to 1.

From the activation of the exemplars, the upper module in the exemplar-LBL maps the exemplars to the category outcome. Following Kruschke (2006), the weights between the exemplar nodes and the output node are allowed to take the values of 5, 0, or -5. The prior on the weights is assigned a probability proportional to the density under a pseudo-Gaussian distribution with $M = 0$ and $SD = 5$, which favors the weight combinations that have more zeros. For the outcome, we consider a two-exclusive-categories (i.e., Category A and Category B), so there is only one output node where a value closer to 1 indicates a higher probability of classifying the item into Category A.

Basically, the model is similar to an attention learning covering map model (ALCOVE model; Kruschke, 1992), but with a locally Bayesian updating approach. The difference, at the computational level (Marr, 1982), is assuming layers only pass partial information during error-driven learning. Although other means might also be feasible, in this study, the computational theory is implemented with the algorithm (i.e., the algorithm level; Marr, 1982) that learning is toward to the maximum a posteriori (MAP).

Results

In this section, we report the simulation results of the exemplar-LBL model to demonstrate the properties and the predictions of different scenarios with the model.

Simulated Prediction

Considering the simplest settings, the model could be asked to categorize items that consist of two-features into one of two categories. As a result, there would be two input nodes, and one output node that indicates the probability of classifying the item into Category A. The value on both features can take any real number but for illustrative purposes, we separated both features into 7 levels. The scales were standardized when fed into the model.

We can test on different category structures including uni-dimensional rules, conjunctive rules, XOR rules, or other non-linear separable structures. However, the first two structures were too easy; the model almost achieved ceiling performance, even just after one 8-exemplar training block. On the other hand, the nonlinear boundary categories required more trials to learn and became more computationally demanding. As a result, we opted to simulate the XOR rule, which were learnable and able to show the learning curve with different model settings and scenarios.

In the following simulation results, each block of training contains these eight items in the following order: (1,1,A), (2,2,A), (6,6,A), (7,7,A), (1,7,B), (2,6,B), (6,2,B), and (7,1,B).

A model prediction was said to be better if it had a higher average accuracy, which for convenience, was computed as (the probability of responding ‘A’ on ‘A’ exemplars) + (1 - the probability of saying ‘A’ on ‘B‘ exemplars) divided by the total number or exemplars. Since there were only two category labels, chance performance was .5.

Exemplar-LBL on Learning XOR First, the simulation results showed that the larger the training trials, the better the performance. After one block of training, the average accuracy was only .503. As shown in Figure 2, the prediction showed a unidimensional structure. However, after four blocks, the accuracy was increased to .744, and the model was able to demonstrate the XOR pattern. Adding more training trials further increased performance. Specifically, after eight blocks of training, accuracy reached .839. This result shows that in the model, learning accumulates in the expected manner.

Second, the exemplar-LBL model predicts order effects and also represents trial-by-trial learning, just as in the original LBL model (Kruschke, 2006). When presenting all the A exemplar trials prior to the B exemplar trials, even with the same total number of training trials as the 4-block condition, accuracy was greatly reduced and dropped to .504. Furthermore, as shown in Figure 2, the model did not learn the XOR in this setting.

The reason for the model’s inability to learn can be explained by online trial-by-trial learning. After learning the batch of the Category A exemplars, the model classified all items into the A category. However, in the first few trials in the Category B batch, the probabilities of classifying an item as being from Category A decreased significantly. Critically, this decrease extended to all items (i.e., both Category A and B items). After this decrease, there were no trials in the Category B batch that could guide the learning of Category A resulting in the model’s inability to differentiate the Category A and B exemplars.

Furthermore, by changing the training targets from binary (1 & 0) to probability, we modeled the probabilistic category learning. For example, we set the model to learn all A exemplars to be A with 90 % probability and all B exemplars to be A with 10% probability (i.e., the (0.9 & 0.1)). The performance accuracy decreased to .693. When we incorporated more uncertainty and set the model to learn (0.75 & 0.25), performance dropped further to .622.

The exemplar-LBL model can also take different priors on both the hypotheses between the input to hidden layer and the hidden to output layer. The difference in prior is a way to model people’s prior knowledge, bias, or preferences on the

task. We first tested the model's sensitivity to changes in the priors on the weights between hidden to output layer by increasing or decreasing the concentration of the prior on specific regions of hypothesis space.

For example, a belief could be that items sharing similar features should be in the same category and items with both features very different, such as the pairs (1,1)-(7,7) and (1,7)-(7,1), should be in different categories. We implemented this prior by assigning higher likelihood to the hypotheses that align with this belief. The accuracy of the 4-block-XOR was .650, which was lower than the accuracy without manipulation on the priors (.744). On the other hand, when implementing the prior that items with dissimilar features should be in the same category, which aligned with the XOR structure, the performance after four training blocks was higher than the one with the original prior and increased to .799.

A similar pattern was observed when manipulating the priors on the input to hidden layer attention strengths. Performance slightly increased to .825 when the model held a strong prior that both dimensions should be equally and moderately or highly attended to. With a prior that both dimensions should not be equally attended to, which was a conflict prior to the XOR structure, performance trivially decreased to .740. Due to space limitations, additional manipulations of the priors are not reported here. However, across different attempts, the model showed that the more extreme the prior, the larger the change (improvement or deterioration) that can be expected in the model's performance compared to a flat prior.

Parameter Recovery

Next, we conducted parameter recovery on the exemplar-LBL model. The model was fitted to the data that it generated, which produced the recovered parameters results. Parameter recovery was evaluated by examining the correlation between the generated parameters and the recovered parameters. A higher correlation shows that the model is better able to discover the true parameters of the data. Here, we focused on the *c* parameters, which affected the calculation of similarity. The *c* parameters were sampled from a Gamma distribution of values 2 and 1, so extreme values were less likely to be selected, and the mode was 1. We used the optim() function in R to conduct the optimization.

The results of the correlations between the 20 pairs of the generated parameters and the recovered parameters are shown in Figure 3. Although the correlation was significant (r = .54, p = .01), visually, it is apparent that the model struggled to recover the larger *c* values.

An Illustrative Implementation

In this section, we demonstrate an application of exemplar-LBL model to generate predictions on when classification training would lead to better learning than observational training.

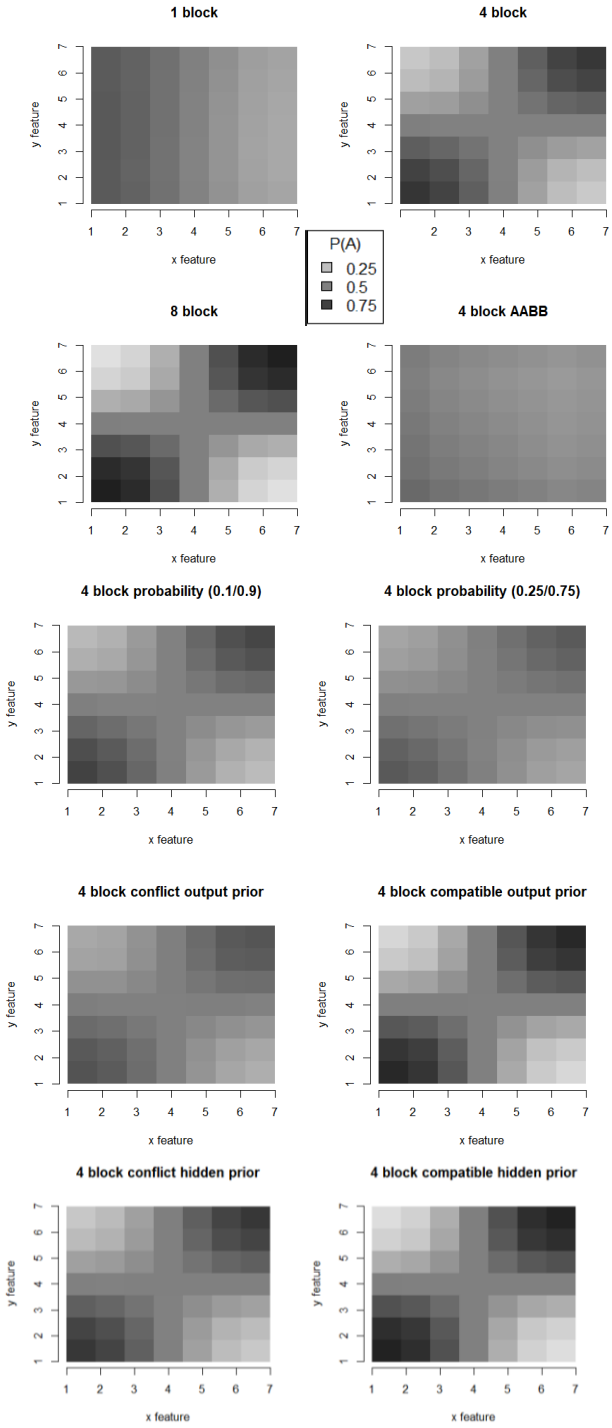


Figure 2: Visualized category structure predicted by the exemplar-LBL model. The darker the cell, the higher the probability of making a classify it into Category A judgment. For illustrative purposes, the category spaces were separated into 7x7 matrix.

In classification training, participants are presented with some stimuli and are asked to make a classification judgment by selecting the correct category label. Next, participants are presented *corrective feedback*, wherein they are shown the

Table 1: Model predictions.

Model \ Instance		(1,1)	(2,2)	(6,6)	(7,7)	(1,7)	(2,6)	(6,2)	(7,1)	Accuracy
Perfect XOR		1	1	1	1	0	0	0	0	
		Observation								
1-block		.65	.61	.37	.36	.64	.61	.37	.35	.503
4-block		.79	.69	.70	.78	.22	.29	.29	.21	.744
8-block		.88	.80	.80	.88	.12	.20	.20	.12	.839
all A then B		.59	.52	.40	.43	.51	.47	.44	.48	.504
conflict output prior		.67	.63	.63	.67	.34	.37	.37	.32	.650
compatible output prior		.84	.76	.76	.84	.16	.25	.24	.16	.799
conflict hidden prior		.78	.69	.70	.78	.22	.30	.30	.22	.740
compatible hidden prior		.87	.78	.78	.87	.13	.22	.22	.13	.825
probability form	0.1/0.9	.74	.65	.65	.73	.27	.34	.34	.27	.693
	0.25/0.75	.65	.59	.59	.65	.35	.40	.40	.35	.622
		Classification								
1-block		.63	.59	.39	.38	.62	.59	.39	.38	.502
4-block		.72	.62	.63	.69	.31	.36	.37	.29	.666
8-block		.82	.73	.73	.82	.18	.27	.27	.18	.775
all A then B		.60	.55	.39	.41	.50	.46	.47	.50	.504
conflict output prior		.67	.61	.62	.64	.36	.38	.39	.32	.636
compatible output prior		.85	.76	.76	.85	.15	.24	.24	.15	.804
conflict hidden prior		.75	.67	.67	.74	.26	.34	.35	.25	.703
compatible hidden prior		.73	.68	.67	.74	.27	.35	.36	.25	.698
probability form	0.1/0.9	.73	.63	.64	.72	.28	.34	.34	.28	.685
	0.25/0.75	.66	.62	.62	.66	.33	.38	.38	.34	.641

Note. Each row was a prediction on the four items after training. The numbers show the probabilities of classifying that item into Category A. The actual rule was the XOR structure that (1,1), (2,2), (6,6), and (7,7) were items from Category A and the other exemplars were items from Category B. If not specified, there were four training blocks. The accuracy was the average probability of correctly categorizing the items.

correct response as well as whether their response was correct. Thus, participants can learn the categories by trial and error. In contrast, in an observational learning task, the category label is presented with the stimulus, and participants are only asked to study and learn it.

Previous studies have shown inconsistent advantages of one form of training over the other. (e.g., Ashby et al., 2002; Hsu & Griffiths, 2010; Levering & Kurtz, 2015). Here, we implement a simple algorithm to differentiate the classification and observational learning: participants learn from their own responses when asked to classify the item in a classification task. As a result, participants have an extra learning opportunity on a classification trial compared with an observation trial.

This approach demonstrates the power of the exemplar-LBL model, because one major characteristic of the extra-learning hypothesis is that initial learning opportunity always occurs prior to feedback. Since the LBL approach, as described, before is sensitive to the training trial-order, it is better able to assess the potential of the extra-learning hypothesis in explaining learning patterns.

Observation Versus Classification in exemplar-LBL In exemplar-LBL, when asked to classify the training items, the model generates the responses by making explicit predictions about the current input and weights. Next, the generated response and the cues are treated as a separate training trial that goes into the model and updates the weights. To eliminate randomness in model predictions, the responses are represented as probabilities of responding 'A' (between 0 and 1) rather than being assigned a Bernoulli-generated label of either 0 or 1. Another benefit of this approach is that it considers different degrees of belief, numerically. We set up the parameter *e* (for extra learning), indicating the probability that the model learns from its own response.

Observation Versus Classification Results. Built on the extra-learning hypothesis, the exemplar-LBL model differentiates classification learning from observational learning by providing an extra learning opportunity on its own classification response. Table 1 shows these simulation results. The model performed better with classification training than with observational training when (a) the priors matched the to-be-learned rules (see *compatible prior* rows in Table 1) or (b) the training was noisy or probabilistic (see *probability form* rows in Table 1). Conceptually, classification training outperforms observational training when the benefit of having the additional self-learning opportunities outweighs the risk of forming incorrect representation of the categories, and vice versa.

Discussion

In this study, our goal was to extend the LBL model to be able to handle continuous stimulus dimensions rather than discrete stimulus features. Based on this goal, we revised the LBL to be an exemplar-based model that has two modules. In the lower module, the model conducts Bayesian updating on the selective attention weight on each feature dimension. With the attention weights, the model computes the similarity between the item and the exemplar as the activation on the corresponding exemplar node in the hidden layer. The upper

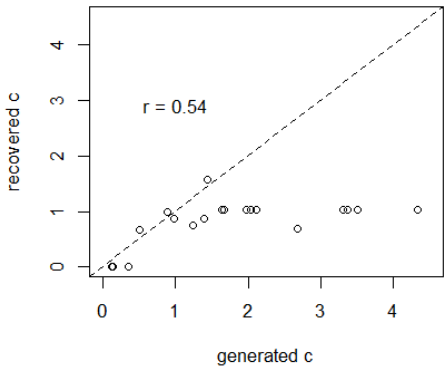


Figure 3: The parameter recovery results.

module also conducts Bayesian learning on its own. The associative weights between hidden and the output layer connect the exemplars to the category labels and thus direct the assignment of the current item based on its relative similarities to the exemplars.

We evaluated this exemplar-LBL model by conducting simulations on category learning tasks with different settings. The results showed that the model captured learning performance and the trial-order effect at the individual level.

One potential application on these results is to model learning-during-test (LDT) processes. As Nosofsky and Hu (2022) mentioned, the LDT process takes part in how people generalize their learning to unseen items. Since LDT on each trial should only affect subsequent judgments, the model for LDT should account for order effects, which is a critical strength of the LBL (Kruschke, 2006) and the exemplar-LBL models. The extra-learning hypothesis on classification training also corresponds to LDT.

Besides, the merits of the LBL models being able to predict trial-order effects, compared to some other associative models (e.g., Rescorla & Wagner, 1972; Shanks, 1992), the LBL model emphasizes the importance of including attention shifting in categorization models (also see Kruschke, 2009).

For our secondary question, we studied how classification and observational training could differ. We proposed the extra-learning hypothesis, which posits that people learn from their own response, even before receiving corrective feedback. By implementing this mechanism into the exemplar-LBL model, we showed that classification training outperformed observation training when the probability of responding correctly is high and when the learning opportunities are scarce.

One possible explanation for the difference between classification and observation training is the different levels of commitment to a given hypothesis. Under the classification task, participants might sharpen and enlarge the likelihoods of their hypotheses to make it easier to find the maximum likelihood that a given hypothesis is true, which is then used to generate the classification response. As a result, participants who receive classification training might be more likely to commit to a smaller set of hypotheses, while participants who receive observational training may engage in more explorative hypothesis testing. If the participants have prior beliefs of the category structure that align with the task assignment, they might therefore learn better with classification training as they commit to the true structure sooner. On the other hand, if the starting priors are farther away from the true structure in the task, participants should learn better with observational training, as they should be less committed to their hypotheses and more open to alternative hypotheses. The extra-learning algorithm implements the distinct levels of commitment to the corresponding hypotheses by assuming that learning can occur from one's own responses during classification training.

However, slight adjustments in the parameter values or model settings can result in notable changes in the outcomes, causing a shift in which training mode (i.e., classification or observation) yields better learning. This result might therefore shed light on why the results in the literature have been inconsistent, wherein some studies have shown benefits of classification over observation training (Ashby et al., 2002), whereas others have shown the opposite pattern (e.g., Levering & Kurtz, 2015).

Nevertheless, the success of differentiating classification and observational training indicates the potential of the LBL models on predicting the spacing effect (e.g., Cepeda et al., 2006; Donovan & Radosevich, 1999) and the benefits of interleaving (Taylor & Rohrer, 2010) in the category learning or other psychological fields.

For future directions, a major next step is to fit the model to behavioral data. It is also important to compare the current exemplar-LBL model to the GBL model as well as other non-Bayesian models. It is also interesting to study if the partially selected message passing between modules can induce rapid performance changes and thus relieve the disadvantage of the neural-network-based models on predicting a steep, stage-shape learning curve on an individual level.

Conclusion

In this study, we extended the LBL model to include exemplar-based categorization and to deal with continuous stimulus dimensions. This exemplar-LBL model is similar to ALCOVE as both incorporate the connectionist network with item features as the input, exemplars as the hidden, and the attention strengths as the input-to-hidden associations. But apart from it, the current model assumes the communication of summarized but not full information between layers. The simulation results showed that this exemplar-LBL model had desirable properties when fit to category learning data and is a promising candidate to explain trial-order effects. Moreover, with the hypothesis that classification training provides extra learning opportunities to learn from the classification responses, the exemplar-LBL model can explain task differences between classification and observational training. We note that the results from the extra-learning hypothesis that we introduce can shed light on the inconsistent results between classification and observation training in the category-learning literature.

References

Anderson, J. R. (2013). The adaptive character of thought. Psychology Press.

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. Memory & cognition, 30(5), 666-677.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological bulletin, 132(3), 354.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811-823.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. Nature neuroscience, 3(11), 1218-1223.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. Journal of Applied Psychology, 84(5), 795.

Hsu, A., & Griffiths, T. (2010). Effects of generative and discriminative learning on use of category variability. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 32, No. 32).

Griffiths, T. L., Kemp, C., & B Tenenbaum, J. (2008). Bayesian models of cognition.

Kamin, L.J. (1968). "Attention-like" processes in classical conditioning. In M.R. Jones (Ed.). Miami Symposium on the Prediction of Behavior, 1967: Aversive Stimulation. Coral Gables, Florida: University of Miami Press (Pages 9-31).

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol.*, *55*, 271-304.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. Psychological review, 99(1), 22.

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. Psychological review, 113(4), 677.

Larkin, M. J., Aitken, M. R., & Dickinson, A. (1998). Retrospective revaluation of causal judgments under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1331.

Lee, M. D. (2006). A hierarchical bayesian model of human decision-making on an optimal stopping problem. *Cognitive science*, *30*(3), 1-26.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. Memory & cognition, 43(2), 266-282.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. San Francisco: W. H. Freeman

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *117*(1), 68.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. Journal of experimental psychology: General, 115(1), 39.

Nosofsky, R. M., & Hu, M. (2022). Generalization in Distant Regions of a Rule-Described Category Space: a Mixed Exemplar and Logical-Rule-Based Account. Computational Brain & Behavior, 5(4), 435-466.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302-321.

Rescorla, R. A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. Classical conditioning, Current research and theory, 2, 64-69.

Sanborn, A. N., & Silva, R. (2013). Constraining bridges between levels of analysis: A computational justification for locally Bayesian learning. *Journal of Mathematical Psychology*, *57*(3-4), 94-106.

Shanks, D. R. (1992). Connectionist accounts of the inverse base-rate effect in categorization. Connection Science, 4(1), 3-18.

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. Applied cognitive psychology, 24(6), 837-848.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, *24*(4), 629-640.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological review*, *114*(2), 245.