Modeling Human Transfer Learning with the Hierarchical Dirichlet Process

Kevin R. Canini
Computer Science Division
University of California
Berkeley, CA 94720
kevin@cs.berkeley.edu

Thomas L. Griffiths
Department of Psychology
University of California
Berkeley, CA 94720
tom_griffiths@berkeley.edu

Abstract

Transfer learning can be described as the distillation of abstract knowledge from one learning domain or task and the reuse of that knowledge in a related domain or task. In categorization settings, transfer learning is the modification by past learning experience of prior expectations about what categories are more likely to exist in the world. While transfer learning is an important and active research topic in machine learning, there have been few studies of transfer learning in human categorization. We propose an explanation for transfer learning effects in human categorization, implementing a nonparametric Bayesian statistical model – the hierarchical Dirichlet process – to make theoretical and empirical evaluations of its effectiveness in explaining these effects.

1 Introduction

When forming hypotheses and drawing inferences, people abstract and reuse the knowledge gained from one domain in other, related domains. This behavior, called *transfer learning*, is central to people's ability to quickly adapt to new situations given a limited amount of data. Despite the importance of transfer learning as part of an explanation for how people learn new concepts, most studies of human category learning have focused on settings where people learn independent categories, with little opportunity for learning of one category to influence learning of another.

Recent work on rational models of human category learning has begun to take an approach that can support a novel kind of transfer learning. In these models, categories are represented using a set of clusters of objects. Learning a system of categories involves learning which objects cluster together, and which clusters belong to each category. These models allow for the possibility that categories share clusters of objects, providing a way for transfer learning to take place: the clusters induced by learning one category can be used to inform learning about other categories. For example, if learning the category of "cats" leads to the formation of a cluster corresponding to striped cats, one might be faster to learn the category of "striped objects", having the expectation that if one of the members of the cluster belongs to the new category, the other members will also.

We focus on the hierarchical Dirichlet process (HDP). Here, each category is a different "group" of observations, to use the terminology from [1]. The HDP provides a way to generate categories that share clusters, with the tendency towards sharing being regulated by the parameters of the model. As α increases, the number of clusters used to represent each category increases. As γ increases, the amount of sharing of these clusters between categories decreases. The model can be used to estimate the density associated with each category by inferring the values of these parameters and the assignment of stimuli to clusters and clusters to categories that make the observed data most probable. This can be done using standard Markov chain Monte Carlo algorithms developed for the HDP. As a consequence, this model allows us to explore how the capacity to share clusters between

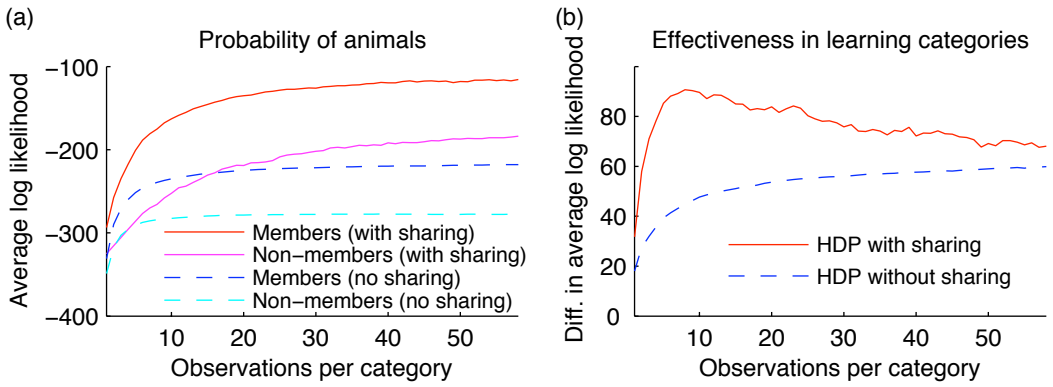

Figure 1: Results of the simulated learning experiment with animal data. (a) Traces of the average log-likelihood of category members and non-members for the HDP model with sharing and the HDP model without sharing. (b) The difference between the average log-likelihood of category members and the average log-likelihood of non-members for the two models. For all plots, results are averaged across 20 random observation sequences.

categories influences category learning by comparing the performance of models that allow sharing (with γ free to vary) and models that do not (with $\gamma \to \infty$).

2 Transfer learning with real-world categories

People engage in transfer learning not just in laboratory settings with contrived stimuli, but also when learning natural categories of real-world objects, such as different kinds of animals. We performed simulated learning experiment in which we evaluated the degree to which the type of transfer learning that the HDP uses is helpful in learning categories of animals. The data used in this simulation is a matrix of binary feature values for 129 animals created by human participants [2]. Our approach is to build a set of semantic, psychologically relevant categories from the human-generated features and compare the performance in learning these categories for an HDP model that allows clusters to be shared among categories and an HDP model that does not.

We measured the performance of each learning algorithm on each trial by first computing the average log probability of the unobserved members of each category and the average log probability of the non-members of each category. These two quantities were then averaged across the 19 categories. The resulting four curves are shown in Figure 1 (a). We also plotted the difference between these two curves for each model in Figure 1 (b), giving a measure of the accuracy of the two learners in estimating the categories over time.

The results show that allowing the HDP to share clusters of animals among categories improves learning performance significantly. The difference between the average log-likelihood of the category members and the average log-likelihood of the non-category members is much greater for the HDP model with cluster-sharing, especially during the early stages of learning. This coincides with the ability of human learners to generalize quickly from small amounts of data, as opposed to many machine learning algorithms which require much more data to perform well.

3 Experiment with human learners

We conducted an experiment in which subjects sequentially learned multiple category systems over the same set of stimuli, creating an opportunity for domain knowledge to be transferred from one category system to another. When learners engage in transfer learning in a categorization environment, they are not simply learning what objects belong to each category; they are also learning how to learn about categories in general. In the hierarchical Bayesian framework, this is realized by placing a distribution over the categories in which they are probabilistically dependent, so that knowledge about one category affects the resulting distribution over the others.

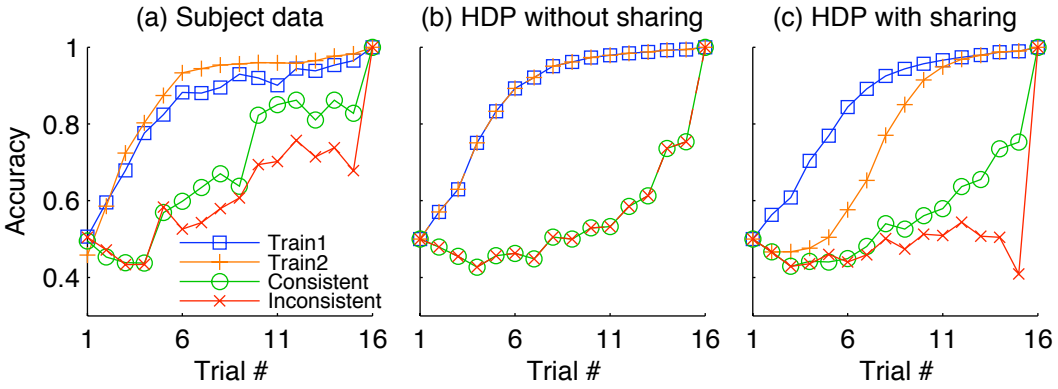

Figure 2: Experimental results. Each plot shows the accuracy in categorizing the remaining unlocked stimuli for each of the 16 trials of each session. (a) The data collected from the participants. (b) The predictions made by the HDP model without sharing. Since category systems are modeled independently, there are no transfer effects. (c) The predictions made by the HDP model with sharing. The model provides an explanation for the difference between the two test conditions.

The experiment was comprised of 3 sessions completed in sequence. In each session, participants learned to partition a set of stimuli into two categories. The stimuli represented objects which have 4 binary-valued features, and the same $2^4 = 16$ objects were used in each session. The category members were chosen as follows. In the first session, the stimuli were split into two categories based on their feature value on a single, randomly-chosen dimension. In the second session, their category membership was based on their feature value on a different randomly-chosen dimension. In the third session, category membership was determined using an exclusive-or (XOR) rule. The two dimensions involved in the XOR rule depended on which of two conditions the subject was assigned to. In the *consistent* condition, the XOR rule used the same two dimensions that were used in the first two sessions. In the *inconsistent* condition, it used the remaining two dimensions.

The stimuli used in the experiment were adopted from a previous categorization experiment [3] and consisted of geometric figures with 5 binary-valued dimensions: size (small or large), color (blue or purple), border (yellow or white), texture (smooth or dotted), and diagonal cross (present or absent). The dimensions were found to be independent and equally salient by a multidimensional scaling of pairwise similarity ratings [3]. Each subject was exposed to only 16 of the stimuli, selected by eliminating those having a randomly-chosen feature value.

The results of the experiment are shown in Figure 2 (a). There was a statistically significant difference between the observed accuracies in the consistent and inconsistent test conditions. Participants were more accurate when learning an exclusive-or (XOR) rule based on the two dimensions that they used in the training sessions than when learning and XOR rule based on the other two dimensions. We fit an HDP model with cluster-sharing and one without sharing to the results. The performance of the two learning algorithms is shown in Figure 2 (b) and (c). The HDP without sharing treats categories as completely independent; therefore, the two test conditions are equivalent under this model. The HDP model with sharing shows an increased learning rate for the consistent test condition as compared to the inconsistent test condition, providing a possible explanation for the difference found in human performance.

References

[1] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. Technical Report 653, Department of Statistics, University of California, Berkeley, 2004.

[2] Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods, Instruments, and Computers*, 37:547–559, 2005.

[3] Yasuaki Sakamoto and Bradley C. Love. Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, 133(4):534–553, 2004.