Title
Inferring knowledge of properties from judgments of similarity and argument strength

Permalink
https://escholarship.org/uc/item/1cj0n0nd

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 26(26)

ISSN
1069-7977

Author
Stromsten, Sean

Publication Date
2004

Peer reviewed

Inferring knowledge of properties from judgments of similarity and argument strength

Sean Stromsten (sean_s@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139-4307 USA

Abstract

Psychological similarity has been invoked to explain many phenomena, including judgments of the strength of inductive arguments (Osherson et al., 1990). The present work follows the suggestion of Tenenbaum and Griffiths (2001) that judgments of similarity and judgments of argument strength cohere because they are essentially judgments of the same kind, which consult the same knowledge of properties of objects or classes. I work backward from people's judgments of argument strength and similarity to the knowledge of properties–specifically, knowledge of probable property extensions–that might explain the coherence among those judgments. I show that the knowledge inferred can be used to predict other such judgments. I then examine this knowledge for structural properties such as taxonomic organization.

Induction, or generalization from examples, is a central cognitive capacity in need of two kinds of explanation: (1) What representations and processes underly induction? (2) Why do we have those representations, and carry out those processes? That is, to the degree that they work, what relation to right reason explains their success? I focus here on the second question, and with respect to just one much-studied inductive task, category-based induction.

To illustrate this task, consider the following inductive argument (after Osherson, et al. , 1990)

Chimpanzees require biotin for hemoglobin synthesis.
Gorillas require biotin for hemoglobin synthesis.

Mammals require biotin for hemoglobin synthesis. (1)

Horizontal lines separate conclusions from their premises. The premises assert facts about categories of objects, and the conclusions do not (in general) follow deductively.

Osherson et al. collected extensive judgments of the strength of such arguments–that is, the subjective probability of the conclusions, given the premises. The arguments contained various mixtures of ten species of mammals in the premises, but all conclusions were about either 'horses' or 'all mammals' (the set of all mammals is approximated,

in all models discussed here, by the set of ten mammals used in the arguments)[1].

In order to study argument strength, rather than particular knowledge of predicates, the premises and conclusion assert so-called 'blank' predicates of species, about which experimental participants will not have direct knowledge. The biological sound of the predicates, and the fact that they are asserted to be true of all members of one or more species, are clues that they are biological properties. The intention, then, is that participants have no choice but to fall back on categorical biological knowledge.

Osherson et al. propose the *similarity-coverage* model, which predicts the judged strength of these arguments as a function of judgments of pairwise similarity among the species of animals in them. The strength $g(X, Y)$ of a conclusion, according to this model, is a weighted sum of (1) the similarity of the premise categories X to the conclusion category Y, and (2) the degree to which the diversity of the premise categories 'covers' the lowest superordinate category S including both the premise categories and the conclusion category:

$$g(X, Y) =$$
$$\alpha \max_i sim(X_i, Y) +$$
$$(1 - \alpha) \sum_j \max_i sim(X_i, S_j).$$

---

[1]In what follows, in addition to the 81 judgments studied by Osherson et al. , I use data on 28 additional judgments, collected by Sanjana and Tenenbaum (2003). They designed these additional generalization judgments to demonstrate effects which their Bayesian model could explain, but which the Osherson et al. model could not. Again, 'horse' was the only species in the conclusions. The innovation was repeated examples of the same species, which required a cover story that makes such examples reasonable. Participants observed a set of example animals–individual animals–with a particular disease, and were then asked to judge the probability that horses could get the disease. Trusting that participants assume that disease susceptibility is a species property, I aggregate these data with the Osherson et al. data.

Osherson et al. test their model against a number of robust qualitative patterns in the way the plausibilities people assign to such arguments relate to the similarities of the categories used. A few examples of these patterns will illustrate the utility of the similarity and coverage terms. The argument

Chimpanzees require biotin for hemoglobin synthesis.
___
Gorillas require biotin for hemoglobin synthesis.   (2)

is stronger than the argument

Chimpanzees require biotin for hemoglobin synthesis.
___
Dolphins require biotin for hemoglobin synthesis.   (3)

because gorillas are more like chimpanzees than dolphins are. The argument

Chimpanzees require biotin for hemoglobin synthesis.
Dolphins require biotin for hemoglobin synthesis.
___
Mammals require biotin for hemoglobin synthesis. (4)

is stronger than argument (1), which may be explained by the greater 'coverage' of the set of mammals by 'chimpanzees and dolphins' than by 'chimpanzees and gorillas'.

It may strike the reader that these intuitions require more than purely psychological, *ad hoc* explanations, for surely they are *correct*. If so, they require normative (Bayesian) explanation. This point has been addressed by several authors, beginning with Heit (1998).

There are a number of other reasons for dissatisfaction with an explanation of judgments of argument strength in terms of judgments of similarity, having nothing to do with the degree of predictive success of the similarity-coverage model. The most obvious, perhaps, is that similarity and argument strength are judgments of equal status, equally in need of explanation. Another objection is that the judged similarity of x to y is not a stable, context-free property of the pair (Tversky, 1977). If judgments of similarity must be computed on-the-fly, as judgments of the strength of arguments presumably are, then whatever knowledge is consulted when computing similarities could be consulted when computing argument strengths, without computing similarity as an intermediary. This is, in essence, the kind of explanation proposed in the Bayesian models of Sanjana and Tenenbaum (2003) and Kemp and Tenenbaum (2003). For purposes of direct comparison, they predicted argument strengths from similarities, just as Osherson, et al. did, but did so by way of inferring taxonomic knowledge presumed to underly both similarity and argument strength judgments.

Bayesian generalization

Before discussing the details of particular proposals, I will briefly review the notion of category-based induction as Bayesian generalization, as formulated by Tenenbaum and colleagues. We assume that:

- The premise categories are random samples from the set c of categories having the target 'blank' property.

- Prior to receipt of any examples, the generalizer has a hypothesis space H, where each hypothesis $h \in H$ is a possible extension for the target property. The generalizer also has a probability distribution over H, which represents the prior degree of belief that each candidate is the extension of the target property. This prior distribution may itself be sensitive to (conditional on) other information, for instance, about the *kind* of property being generalized.

The probability that a category y is a member of the set c, given a set of n examples $\mathbf{x}$ drawn at random from c, can be found by summing over hypotheses:

$$P(y \in c | \mathbf{x} \sim c, \xi) = \sum_h P(y \in c | c = h) P(c = h | \mathbf{x} \sim c, \xi).$$

Here $\mathbf{x} \sim c$ means that the examples $\mathbf{x}$ are random draws from c, and ξ represents background information. The first term is 1 if $y \in h$, and 0 otherwise. The second term can be re-written in an enlightening form by Bayes rule:

$$P(y \in c | \mathbf{x} \sim c, \xi) = \frac{\sum_{h \ni y} P(\mathbf{x} \sim c | c = h) P(c = h | \xi)}{\sum_{h'} P(\mathbf{x} \sim c | c = h') P(c = h' | \xi)}.$$

The terms $P(\mathbf{x} \sim c | c = h)$ represent the probability of seeing just the examples $\mathbf{x}$ in n draws from h. Assuming that items in h are drawn with equal probability, then the probability of drawing any particular item in a single draw is $1/|h|$. Then $P(\mathbf{x} \sim c | c = h)$ is $|h|^{-n}$, if h contains all the examples in $\mathbf{x}$, and zero otherwise. The likelihood term $P(\mathbf{x} \sim c | c = h)$ depends only on the examples and the contents of h, so we see now that ξ represents information we may have, prior to seeing the examples, about the probability of the various possible extensions. In what follows, I suppress this term to make the notation simpler.

Further abbreviating $P(c = h)$ to $P(h)$, we can re-write the above as

$$P(y \in c | \mathbf{x} \sim c) = \frac{\sum_{h \supset y \cup \mathbf{x}} |h|^{-n} P(h)}{\sum_{h' \supset \mathbf{x}} |h'|^{-n} P(h')}. \quad (1)$$

Note that the sum in the denominator can be broken into two sums: one is the same as that in the numerator, and the other is over those hypotheses that contain the $\mathbf{x}$ but *not* y. Generalization, then, depends on two weighted sums: one over the properties common to both $\mathbf{x}$ and y, and another over those distinctive to $\mathbf{x}$. Each summand is weighted by both its prior plausibility and its likelihood or 'fit' to the examples.

The two terms have different jobs to do. The fit for extension h–that is, $|h|^{-n}$–gives an advantage to smaller extensions, which is exponential in the number of examples. Without a likelihood term sensitive to the number of examples, we miss an important phenomenon: given that examples are consistent with two extensions, increasing the number of examples ought to shift weight to the more specific extension. For instance, suppose our prior gives high weight to the classes 'mammal' and 'rodent'. Then, given 'mouse' as an example of a species with property P, either class is quite plausible. But adding the further examples 'gerbil' and 'hamster' ought, intuitively, to give a strong advantage to 'rodent', because the selection of three rodents from the larger class is highly coincidental. The likelihood term captures this focusing effect.

Without prior preferences for some extensions over others, the likelihood or 'fit' term will always favor the extension consisting of just the examples, and will have no preference among larger extensions of the same size. For example, given 'mouse' and 'gerbil' as examples of species with some property, generalization to 'turtle' will be just as strong as that to 'hamster'. A prior favoring the natural class 'rodents' over 'rodents minus hamsters, plus turtles' prevents this bizarre behavior.

Similarity as a function of generalization probabilities Tenenbaum and Griffiths (2001) have argued that the similarity of x to y is a function of the probability of generalizing from x to y, or vice-versa, or both. This move gives the infamously slippery notion of similarity some solid footing on the ground of reason, because generalization has a normative foundation in Bayesian statistics. They also show how this view rationalizes earlier work on formalizing similarity and generalization.

For present purposes, we need not delve deeply into the question of just how generalization proba-

bilities determine similarities. I assume, as Osherson et al. do, that similarity is symmetrical, and, further, that it has this particularly simple form:

$$sim(x, y) \equiv \qquad (2)$$
$$\frac{P(y \in c | x \sim c) + P(x \in c | y \sim c)}{2}.$$

Intuitively, this definition says that two items are similar to the degree that one is likely to have a property that the other exemplifies.

Previous work on Bayesian modeling of category-based induction

Various restrictions on the form of the prior could be entertained. For instance, each species might correspond to a location in a low-dimensional Euclidean 'psychological space', with higher priors assigned to sets contained by convex or connected regions. The restricted families of priors investigated by Tenenbaum and colleagues are based on binary trees, with species at the leaves. The sets with highest priors are those corresponding to single subtrees, but some probability is assigned to sets picked out by multiple subtrees. Sanjana and Tenenbaum use a generic method for assigning probabilities to disjunctions of a basis set of hypotheses (in this case, single subtrees), while Kemp and Tenenbaum define a simple 'mutation' process that can generate arbitrary hypotheses, but assigns lower probability to those that require many mutations, or mutations over short branches.

The proponents of these tree-based priors stress that taxonomic trees are not just another restricted family of priors; they are also an independently-motivated *theory* of the domain. People around the world seem to organize creatures into 'folk taxonomies' (Atran, 1995), and the geneology of species does, indeed, form a tree. This kind of theory may be applicable in domains besides biology: even artifact kinds are often the result of a process of copying and modifying earlier designs.

One obvious way to compare various proposed families of priors is to compare predictive accuracies: fit the parameters (for instance, the locations of the points in a metric-space model, or the topology and branch lengths of a tree) to subsets of the judgments and see how well each model predicts the rest.

Rather than competing with previous models on data fit, I take a complementary, 'empirical Bayes' approach (see, for instance, Gelman, et al. , 1995): I place *no* constraints on the form of the prior, find priors that do a good job predicting the data, and then examine those priors for structural properties.

This strategy has an obvious pitfall: an unrestricted search for a prior that makes the data probable may over-fit the accidental properties of the training data, especially, as in this case, when there are many more parameters than data points. Before examining the prior for interesting structural properties, therefore, I demonstrate that the model is not over-fitting so badly as to be uninformative.

Computing a prior from judgments

For any given hypothesis space and prior, Bayesian generalization yields point estimates for a set of similarities and/or argument strengths. To accommodate noisy human data, I take these point estimates to be central tendencies.

In what follows, I refer to the model's prediction of the ith judgment, given a prior, θ, as $j_i^m(\theta)$ (this is given by either equation 1 or equation 2, above). The actual human judgment I denote j_i^h. A simple noise model that respects the constraint that both generalization probabilities and similarities must be between 0 and 1 assumes that

$$log\left(\frac{j_i^h}{1 - j_i^h}\right) \sim N\left(log\left(\frac{j_i^h}{1 - j_i^h}\right), \sigma^2\right).$$

In words, we apply a transform to each model prediction that may (conveniently) take on any real value, and assume that the similarly-transformed human judgment is normally distributed around this transformed prediction.

A bit of work (omitted here) reveals that the log-likelihood (up to an additive constant) of a set of judgments **j** is

$$P(\mathbf{j}|\theta) = \qquad\qquad\qquad (3)$$
$$\sum_i \log\left(\frac{j_i^h}{1 - j_i^h} + \frac{1 - j_i^h}{j_i^h} + 2\right) +$$
$$\frac{1}{2\sigma^2}\left(\log\left(\frac{j_i^h}{1 - j_i^h}\right) - \log\left(\frac{j_i^m(\theta)}{1 - j_i^m(\theta)}\right)\right)^2.$$

The log likelihood of a set of judgments has a complicated but readily-computed gradient with respect to the prior, involving only the second term in equation 3, which can therefore be optimized by off-the-shelf techniques. I used the method of conjugate gradients, stopping whenever several iterations produced less than a set increase in the log likelihood of the training data. The model was parameterized by 'soft-max' parameters z, where the prior probability of extension i is given by $\theta_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$. On each run, the z were randomly initialized such that the θ were nearly uniform.

proportion used in training		correlations of model and data on remaining	
args	sims	arguments	similarities
0	1	.50±.026	n. a.
1	0	n. a.	.88±.006
0.5	0.5	.61±.029	.77±.033
0.9	0.9	.80±.026	.72±.104
0	0.5	.29±.046	.60±.064
0.5	0	.54±.030	.67±.043
0	0.9	.41±.033	.79±.084
0.9	0	.67±.038	.82±.018

Table 1: Predictions of held-out data given various training data. All rows show averages of ten runs, with associated standard errors.

Predicting held-out judgments

Remarkably, this rather lavishly parameterized model does a reasonable job of predicting randomly held-out judgments when fit to the rest.

Tuned to the judgments of argument strength, the model's predictions of pair-wise similarity agree strongly with the actual judgments, approaching a correlation of 0.9. A number of experiments, using various proportions of each kind of judgment as training data, are reported in table 1.

This model does relatively poorly on the task that has been the focus of the previous work–predicting the argument strengths, given the similarities. A possible explanation for the deficit relative to the other published fits is that the assumptions about the form of the prior made explicitly by using a tree with mutations (and perhaps implicitly in the similarity-coverage model) are essentially correct, in which case opening up the space of priors, as I have done, can only reduce predictive accuracy. As further evidence of over-fitting, early stopping would usually have yielded better predictions, although I could not find a single stopping rule that consistently did so.

Given these results, we can expect that the priors converged to will reflect both the underlying structure of people's knowledge and, to some degree, peculiarities of the data set fit by the over-parameterized model. In the next section, I examine the priors converged on for taxonomic structure.

The 'shape' of the prior

For the purpose of examining the structure of the prior that best explains the data, I focus on results obtained by optimizing the prior over the entire set of judgments.

If we examine the hypotheses with highest priors, certain patterns can be found by eye or statistical

test. Table 2 lists the 10 sets with the highest average prior probability in a typical optimization run.

If the most probable sets are those corresponding to sub-trees of a taxonomic tree, then we should expect that most pairs of such sets will obey taxonomic constraints: either one will contain the other, or they will be disjoint. There are a suspiciously large number of these containment relations among the top-ranked sets–randomly generated collections of sets have as many containment relations between pairs as the top-ranked 100 sets only about 40 out of 1000 times. There is an even more extreme number of disjoint pairs–exceeded not even once in 1000 random sets. Forcing the random sets to match the top-ranked 100 in number of members makes no difference to these results.

However, there are also quite a few partially overlapping sets, which is not what we would expect from a single, strictly-observed tree. The overlap is notably non-arbitrary, however. For instance, the sets 'chimp, gorilla, mouse, squirrel', 'chimp, gorilla, dolphin, seal', and 'mouse, squirrel, dolphin, seal' are composed of just the three pair 'dolphin, seal', 'chimp, gorilla', and 'mouse, squirrel' ('Mouse, squirrel' is not shown here, but ranked 14th in this solution. 'Horse, cow', another pair one might expect, is not far behind.).

What this might point to is a 'mutation' process, as suggested by Kemp and Tenenbaum (2003). While there are sets above that could only be explained by mutations, if a single tree is assumed, they seem to be restricted to cases where the mutations could occur over relatively long branches; members of the very short subtrees, such as 'dolphin, seal', seem to be present or absent in tandem, as predicted by the mutation process.

Another possibility is that the prior reflects uncertainty over *several* taxonomies. Uncertainty about just which taxonomy to consult may be of two kinds: uncertainty about which taxonomy is *correct*; and uncertainty about which taxonomy is *relevant* to the property under consideration. The first is a commonplace of probabilistic modeling, and quite intuitively understandable, in this case. If I perform bottom-up, agglomerative clustering by eye, using the two-dimensional multidimensional scaling solution in figure 1, I come up with the tree topology used in both the Sanjana and Tenenbaum and the Kemp and Tenenbaum papers. But only the lowest-level clusterings are obvious. Is the 'seal, dolphin' cluster closer to the 'gorilla, chimp' cluster than the 'mouse, squirrel' cluster is? It is hard to tell.

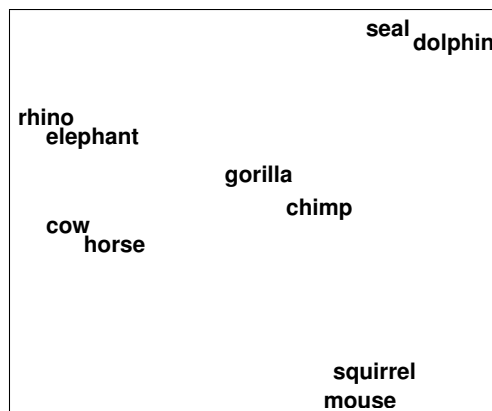The second kind of uncertainty is about which of several trees is relevant. Even if some properties are

Figure 1: A two-dimensional MDS solution for the similarities of the ten mammals (Euclidean metric, variance accounted for = .81)

distributed according to a particular tree/mutation process, others are likely not to be. This is true even if we restrict attention to biological properties of the kind that are likely to be universal across a species (and which therefore are sensible fodder for the kinds of judgments we consider here). 'Deep' biological properties, such as having a certain organ or metabolic process, are quite likely to respect the 'tree of life'–that representing the genealogy of species. The distribution of other species properties, such as what and how members eat, may be quite random with respect to this tree, but might still respect a different tree.

How might people come to have these priors?

I proceeded above with no constraints on the form of the prior over possible extensions of a new predicate. People or machines asked to make these judgments, however, have no such luxury. They must assume that the extension of the new predicate is systematically related to some known predicate or predicates (and, more generally, that predicates are likely to have systematically related extensions), or have no basis for generalization.

In addition to positing coherence among new properties and old ones, real learners must learn from the kind of data available in the real world. Similarity-like data may sometimes be available, but they are not necessary; people can observe objects and their properties–for instance, that cows, horses, elephants and rhinos all eat grass. Lists of such properties are standard fodder for machine-learning methods, including agglomerative clustering or more sophisticated tree-finding methods. Several strate-

rank	horse	cow	chimp	gorilla	mouse	squirrel	dolphin	seal	elephant	rhino
1	horse	cow	chimp	gorilla	mouse	squirrel	dolphin	seal	elephant	rhino
2							dolphin	seal		
3			chimp	gorilla	mouse	squirrel				
4					mouse	squirrel	dolphin			
5			chimp	gorilla			dolphin	seal		
6	horse	cow		gorilla		squirrel			elephant	rhino
7					mouse	squirrel	dolphin	seal		
8	horse	cow	chimp	gorilla	mouse	squirrel			elephant	rhino
9			chimp	gorilla						
10	horse	cow								rhino

Table 2: The 10 sets with the highest prior probability, on a single optimization over all judgments. There are many instances of nesting, but they are not strictly compatible with any single taxonomic tree.

gies of tree-learning from such data have been applied to a number of standard machine-learning datasets in Kemp et al. (2003).

Summary and discussion

I have suggested a novel technique of general utility for fitting a Bayesian model to a set of judgments. I applied this technique to a large collection of human judgments. Without imposing a taxonomic form on the prior, the prior of a Bayesian model optimized to fit human judgments nevertheless shows significant conformity to taxonomic constraints. It seems that either participants have a bias, in the domain of animals, toward priors that respect the taxonomic constraints, or the raw facts about mammals have this structure (which would, in turn, *justify* a taxonomic bias).

The technique is not limited to the case of a structureless prior over a small set of possible extensions. Any prior that has tractable derivatives with respect to its parameters could be so optimized. In the case of a larger number of categories, whose power set is too large for enumeration, an approximate gradient could be computed using a sample from the current estimate of the prior.

A principled alternative to using held-out data to check models, and to using null-distribution hypothesis tests to look for structure in the prior, is Bayesian model comparison: compare the marginal likelihoods of various structures. For most interesting structure classes, the sums or integrals involved are intractable, but they can be approximated by Markov Chain Monte Carlo or other methods.

Acknowledgments

References

Atran, S. (1995). Classifying nature across cultures. In Smith, E. E. and Osherson, D. N., eds., *An Introduction to Cognitive Science*, volume 3. MIT.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In Oaksford, M. and Chater, N., eds., *Rational Models of Cognition*, 248-274. Oxford.

Kemp, C. and Tenenbaum, J. B. (2003). Theory-based induction. In *Proceedings of the Twenty-fifth Annual Meeting of the Cognitive Science Society*.

Kemp, C., Griffiths, T. L., Stromsten, S., and Tenenbaum, J. B. (2003). Semi-supervised learning with trees. *Advances in Neural Information Processing 16*.

Osherson, D., Smith, E. E., Wilkie, O., López, A., and Shafir, E. (1990). Category-based induction. *Psychological Review, 97(2)*, 185-200.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Sanjana, N. and Tenenbaum, J. (2003). Bayesian models of inductive generalization. In Becker, S., Thrun., S., and Obermayer, K., eds., *Advances in Neural Information Processing Systems 15*. MIT.

Tenenbaum, J. B. and Griffiths, T. L. (2001). Generalization, similarity, and Bayesian Inference. *Behavioral and Brain Sciences, 24*, 629-641.