Rational Categories

Emmanuel M. Pothos (pothos@psy.ox.ac.uk)
University of Oxford; Department of Experimental Psychology
South Parks Road, Oxford OX1 3UD, UK

Nick Chater (nick.chater@warwick.ac.uk)
University of Warwick; Department of Psychology
Coventry CV4 7AL, UK

Abstract

We adopt the interpretation of rationality according to which an organism's behavior is rational if it is optimally adapted to its environment (Anderson, 1990, 1991a, 1991b). Rationality, according to this view, often implies mechanisms that are as informationally efficient as possible. We interpret the problem of basic-level categorization (Rosch & Mervis, 1975) as one of data compression within an information theory framework, to define a framework whereby the best classification on a set of items is the one that maximally compresses the description of the similarity structure of these items. This framework is then used to examine whether participants in two experiments classified meaningless items in a way that reflected such a compression bias. In addition to the implications for human basic-level categorization, an objective criterion is established for assessing the relative merits of alternative clustering solutions on the same domain.

Introduction

A fundamental problem for any living creature trying to survive or for any statistician faced with noisy data, is to identify how much (if any) useful structure exists in a noisy input. Classification aims to identify groups of individuals so that within category similarity is greater than between category similarity. We have interpreted the classification problem in information theory terms and used the minimum description length principle (MDL, Rissanen, 1978) to specify a framework whereby partitioning a set of objects into groups is favored to the extent that there is a data compression advantage. The MDL principle states that "...the best theory to infer from a set of data is the one which minimizes the length of the theory and the length of the data when encoded using the theory as a predictor for the data." (Quinlan & Rivest, 1989). Lengths of objects refer to a binary string description of these objects in some arbitrary programming language; the binary string can be seen a series of binary questions that would be needed to specify an object, so that smaller lengths correspond to simpler objects. Applying the MDL principle to categorization requires specifying a way to code for the similarity structure in a set of items (which would give us a measure of the information content of the domain) and also defining the meaning of categories. The latter must include the information cost (that is the extra code that would be needed) for specifying a classification of the items in categories, and a description of how this classification may reduce the description length of the items.

Quinlan and Rivest (1989) have reported a decision tree method based on MDL, whereby different exemplars are partitioned into disjoint groups, with tree structure costs defined in terms of the information required by an imaginary computational procedure creating the tree. Our approach is similar to theirs in spirit, but we have tried to define clustering-configuration complexity costs and error costs in more general information-theoretic terms, so that the link with psychological processes would be more direct.

Rationality and information gain in psychological models

The traditional views of rationality originate from Aristotle, whereby rationality was understood to be fundamentally associated with the capacity to reason according to the rules of (classical) logic. This view has survived to the present day (e.g., Brown, 1989), despite overwhelming evidence that humans very often fall prey to an alarmingly large number of logical fallacies, and also despite the discovery of many other logical systems, so that (a priori at least) there would be no reason why classical logic should provide the normative model for thought (see, e.g., Braine et. al., 1995, for an example of a reasoning model based on classical logic and Evans et al., 1991, for a more general review; Chater & Oaksford, 1993, consider the suitability of traditional deductive models of reasoning in more general terms).

Recently another view of rationality has been suggested, formulated independently of an adherence to any particular rule system. Anderson (Anderson, 1990, 1991a, 1991b; see also Stich, 1990) has argued that an organism is rational to the extent that it is optimally adapted to the environment. An account of rationality along such lines requires to specify the nature of the problems an organism is being faced with in its struggle for survival, so that its behavior can be applauded as rational (or dismissed) to the extent that these problems are being overcome or not. A problem common to all living creatures is that of identifying the underlying regularities in a noisy input; that is the process through which in a set of

instances their common characteristics are identified, while irrelevant individual differences are discarded.

For instance, in the domain of low level perception, Barlow (1983) has proposed that in the absence of any a priori expectations on the statistics of the world, the only means we have in understanding our environment is encoding it with as little redundancy as possible (that is minimizing description length). Redundancy in a set of instances refers precisely to their common features which would not be needed in describing these instances as members of the set. In other words, the point of identifying such features is to simplify the representation of the objects encoded. For instance, Olshausen and Field (1996): Minimizing redundancy in any representation leads to statistical independence so that the relevant entities will be easier to cope with (as this would enable the segregation of the perceptual input into objects that have, generally, little in common, so that they can be processed separately; see also Barlow, 1974).

Turning to higher level cognitive functions, accounts of performance on various reasoning tasks have recently been proposed that are based on information theory (Chater & Oaksford, submitted; Oaksford & Chater, 1994), suggesting that data compression might have a central role across a wide range of cognitive mechanisms.

The above discussion illustrates that rationality in many contexts implies a process of minimizing description lengths, or compression, as suggested by the MDL principle[1]. Consistent with the above, we will present a "rational" model of low level categorization, in the sense that categorization will be seen as a process of extracting as much redundancy as possible from a given domain. Such a model is further suggestive of a coherent view of human rationality/cognition through compression.

Basic Categories

Grouping together a set of items under the same category label necessarily involves a trade-off between the simplicity of the category structure, and how informative categories are relative to their members (Komatsu, 1992). If we want the members of a category to share as many features as possible (so that category instances will be more homogeneous), then we must use many categories; in such a case, labeling a new instance as a member of a category will be very informative, because there will instantly be many features one can assume for this instance. On the other hand, this view taken to the extreme would readily recommend having a different category for each instance, so that the obvious objective of categorization–summaring information about a group of instances–would seem to lost. The competing pressures between few general categories, that would likewise assume few properties for their members, and many specific categories, which, however, will be less useful, seem to imply that there may be an optimal level in the generality/specificity trade-off.

These intuitions have been captured by Rosch and Mervis's (1975) concept of "basic categories." In their formulation, "cue validity" was defined as the conditional probability that an object is in a category, given that it has some cue (or attribute) associated with the category. On the assumption that "in the domains of both man-made and biological objects, there occur information-rich bundles of attributes that form natural discontinuities" they defined basic categories as the categories "...for which the cue validity of attributes within categories is maximized" (Rosch & Mervis, 1975). That is basic categories are categories whose elements share the most attributes among themselves and as few attributes as possible with members of other categories. Thus, according to Rosch and Mervis, in our hierarchy of concepts, where higher level concepts are more general, there exists a level that is optimally suited for describing our environment. Indeed empirical support for such a claim is abundant in the psychology literature. For instance, Murphy and Smith (1982) reported evidence that basic level concepts are easier to learn than either their subordinates or superordinates, Rosch et al. (1976) showed that subjects would agree the most about which attributes are possessed by the members of basic level categories, etc. (see Corter & Gluck, 1992, for an overview of basic categories research).

Our own formulation is aimed towards providing a criterion to identify basic categories, in a way consistent with Rosch and Mervis's intuition that these categories ought to be optimally descriptive of the domain, neither too specific nor too general.

Classification by MDL

The problem of identifying the set of groups that best capture the statistical structure of the domain can be redescribed as follows. Suppose we are interested in describing a set of ordered relations among n objects (so that we have $n*(n-1)/2$ relations), of the form $a<b$, $a<d$, $b<c$ etc.)[2]. If we were to describe each relation individually then we would require $n*(n-1)/2$ bits (ignoring ties, for each pair of elements, say for a and b, we have two, a priori equally probable, possibilities: either a is greater or b is greater; since we have only two available choices each decision costs one bit).

If there are regularities in the domain, then it may be possible to describe it more efficiently by partitioning the objects into groups, or clusters. Such a framework enables us to employ the MDL principle: The question of whether objects fall into clusters becomes a question of whether a classification of these objects would require less code length

[1] Moreover, processes that minimize description lengths can be equivalent to Bayesian inference, which has a number of justifications other than the ones alluded to here; see Chater, 1996.

[2] A description of the similarity structure among a set of items in this way is non-parametric; this is a desirable feature for any model specified over internal representations, since the scales in such cases are usually entirely arbitrary.

(i.e. less information), compared to the default choice of simply specifying all the inequalities individually.

Creating a cluster is equivalent to saying that all the distances within clusters are less than all the distances between clusters so that the total number of relations that needs to be described is reduced. In particular, since we noted above that knowledge of the relation between any two distances costs one bit, if s constraints are introduced by a set of clusters, then the information gain (or compression) is s bits. However, in general no cluster will be perfect, so that some of these constraints will be wrong, and also the particular way clusters divide up the data set needs to be described.

Therefore, in order to determine the extent to which a cluster configuration is advantageous, information gain needs to be balanced against the cost of specifying the errors in the constraints and the cost of describing the cluster configuration. If there are e errors in the constraints then the number of bits required to identify them will be $\lg(s+1)+\lg({}^{s}C_{e})$, as ${}^{s}C_{e}$ is the number of ways in which we can select e objects out of s.

To code for the clusters, we first need to specify the number and sizes of clusters (on the assumption that there are no empty clusters) and also the particular assignment of objects to clusters. Thus, we need to take into account all the possible cluster structures, and for each one of them all the possible ways we can assign objects to clusters. Assuming that each such possibility is equally probable, if there are D of them, then this code will be log2D bits long, where D is given by $\sum_{v=0}^{n}(-1)^{v}\frac{(n-v)^{r}}{(n-v)!\,v!}$ (Feller, 1970; n is the number of nodes and r is the number of items; for a more extensive discussion see Chater & Pothos, manuscript).

Summing up, the compression associated with transmitting a cluster solution instead of all relations individually would be (all relations) { (all relations) (constraints - costs) } or (constraints-costs) so that we have: Compression = Constraints Costs, where Costs are given by $\left(\log(s+1)+\log({}^{s}C_{e})\right)+\sum_{v=0}^{n}(-1)^{v}\frac{(n-v)^{r}}{(n-v)!\,v!}$.

The above framework allows one to examine the "goodness" of cluster configurations partitioning a domain of objects in different ways: A particular classification is more successful compared to an alternative one to the extent that it leads to a greater compression. Figures 1 – 4 show four data sets and the optimal cluster configurations derived by clustering algorithms directly optimizing a compression criterion (Chater & Pothos, manuscript). In the first three cases the data points are partitioned in the way that seems to reflect most faithfully the structure of the domains, while in the fourth case the low information gain associated with the final configuration suggests what is intuitively obvious, namely that there is very little structure in this data set.

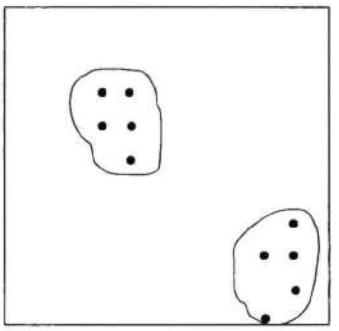

Figure 1: Two clusters. Compression: 491 bits , out of a total unprocessed information content of the domain of 990 bits.

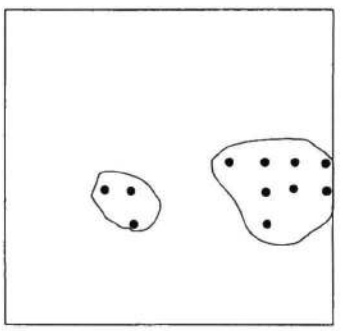

Figure 2: One small cluster and a big one: Compression: 678 bits out of 1485 bits.

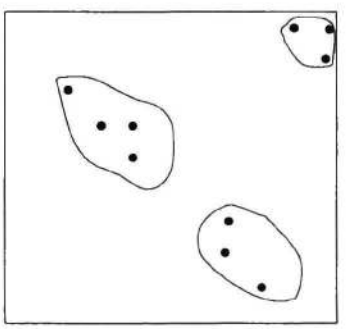

Figure 3: Three clusters. Compression: 383 bits out of 990. Smaller compression to before indicates that a two cluster domain is a more redundant one.

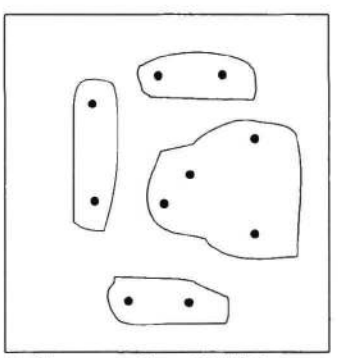

Figure 4: Compression: 181 bits out of 990.

Comparisons with human results

The above data sets were used to construct simple stimuli, varying along two physical dimensions. In this section we consider how the optimal clustering solutions suggested by our framework compare with the way human participants would classify these stimuli. Note that we assume that the physical dimensions of the stimuli directly map onto the dimensions of some internal representation of these items (e.g., see Shepard, 1987; Shin & Nosofsky, 1992); this would appear a reasonable approximation since the stimuli were simple enough to preclude the possibility of participants selectively encoding them over some subspace of the relevant dimensions.

Materials

Stars were constructed so that their inner diameters corresponded to the vertical dimension of the data sets in the previous section, while their outer diameters corresponded to the horizontal dimension. Thus, each data point specified an inner and outer star and these were blended together, to emphasize the impression of an individual entity, as shown in Figure 5. The outer diameter on each star could very from 108mm to 198mm, while the inner diameter from 10mm to 100mm. All stars were printed on A4 sheets of paper in black and ink.

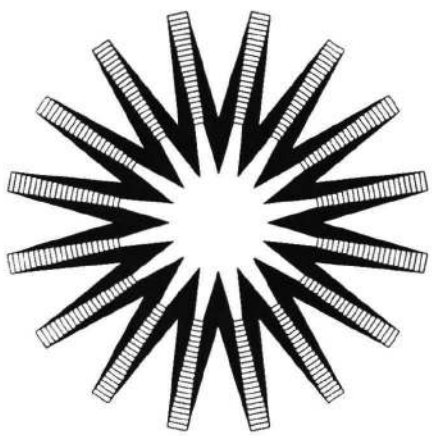

Figure 5: An example of the stimuli constructed from the data sets presented above.

Procedure

Participants received in succession (order randomized for each person) sets of stars corresponding to the data sets in Figures 1 – 4 and were simply asked to partition the stars in groups. In the first condition, they were only told to divide the items in a way that seemed "natural and intuitive" and that there was no limit to how many groups they could use, but that they should not use more than what would think is necessary. In the second condition, we presented the classification problem in a realistic context (stars were meant to be mailed to customers) and provided extensive instructions (and visual aids) that made the particular aspects of performance we were interested in as obvious as possible. This procedure has been motivated from the finding that quite often a pragmatic context in a reasoning problem improves performance (Cheng & Holyoak, 1985).

We tested 28 individuals in each conditions, in a between-subjects design. The experiment lasted for approximately 10 minutes.

Results

We were interested in identifying information-theoretic parameters that would be predictive of participants' classification results. The general hypothesis is that basic-level categorization processes aim, to a certain extent, to represent a domain with as little redundancy as possible. Therefore, insofar that there is structure in a domain, classification would aim to flesh out this structure as much as possible, and, alternatively, how readily structure is perceived in a domain should be a function of the degree to which this domain can be compressed.

We first tested the hypothesis that some particular cluster structure was preferred, against the null hypothesis that they were all equally likely. We therefore calculated for each data set the number of different solutions participants produced, to derive an expected chance frequency for each solution through the ratio (total number of solutions) / (number of distinct solutions). Note that this is an extremely conservative measure of chance frequencies as the number of unique solutions produced was much less than the total possible number of solutions.

A single sample chi-square test was then used to examine the deviation of the frequency of any one given solution from what would be expected by chance. In condition 1 (simple instructions), in data sets 1, and 2 the partitions that were significantly more frequent than chance were the best compression ones (for the two data sets respectively: χ^2 (1) = 28.3, 127.1; in all cases $p<.001$). In condition 2, that was the case for data sets 1, 2, and 3 (all p-values less than .01). There seemed to be no preference for any particular classification for the fourth data set in both cases. The frequencies with which the best compression solutions were produced for each pattern is shown in Figure 6.

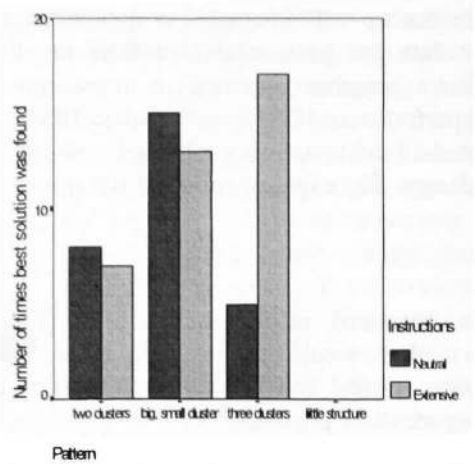

Figure 6: Number of times the "best compression" solution was found in condition 1 (neutral instructions) and condition 2 (extensive instructions).

We then looked at the extent to which the total compression possible with a data set was in any way predictive of the difficulty involved in partitioning the data set into different groups. We reasoned that data sets which were associated with a high possible compression should be easier to partition, which would lead to less between participant variability. Figure 7 shows the number of different solutions that were produced with frequencies greater than one[3] for the four data sets, in each condition. With the simple instructions, for the first two data sets, only three solutions were produced more often than once, while with the other data sets, participants would be a lot less likely to agree on the optimal way to partition the domains. With the extensive instructions these results were replicated, and also little variability was observed with the third data set as well.

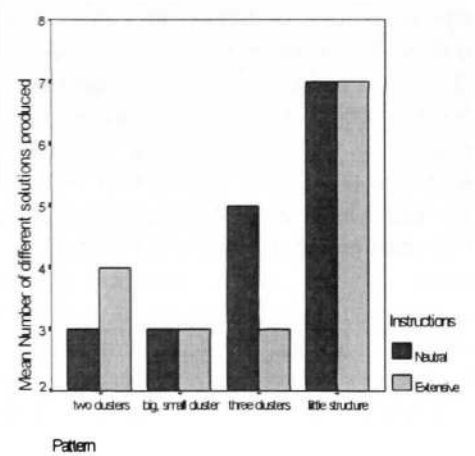

Figure 7: Variability in participants' solutions for the conditions with neutral and extensive instructions.

[3] Looking simply at how many different solutions were produced with each data set is not very useful, since there is considerable noise; this is to be expected with small samples considering the unconstrained nature of the task.

Thus, both our main expectations were confirmed: Participants classified the items in each of the data set in way so that the best compression classifications were indeed more likely to be produced. Also, the extent to which a pattern could be compressed was a measure of how "difficult" the classification problem was, since, in general, there was more (between-subject) solution variability in the cases of lower (possible) maximum compression. Instructions also helped participants "solve" the classification problem in data set 3, as is manifest both by a decrease variability in unique solutions produced, and an associated increase in the number of times the best compression solution was identified. Even with the simple instructions, however, in data set 3, participants' responses reflect some selective bias to the best compression solution[4].

Conclusions and future directions

We have presented a model of basic-level categorization (Rosch and Mervis, 1975) based on the MDL principle. Although we have not specified the algorithmic details in this work (see Chater and Pothos, manuscript), the model will partition a set of items in such a way so that the (information-theoretic) description of the similarity structure of these items will be maximally compressed.

The main predictions of our model were that people classifying meaningless items would generally prefer the "best compression" classifications and also that the difficulty of a classification situation would depend on how much compression is (theoretically) possible in each case. These predictions were generally confirmed in two experiments, where the variable manipulated was thoroughness of instructions. In the case of extensive instructions, performance was improved in one of the data sets.

The major confounding of the present work is that the representation of the items used was based on an experimenter-defined description of the items. We argued that with simple, meaningless stimuli this is not a problem; nevertheless, there is always the possibility of different individuals encoding the similarity structure in different ways. Thus, in additional work we aim to directly apply our framework on representations of the items derived directly from confusability experiments (Shepard, 1987).

Another aim is to extent the present framework to describe hierarchical category structures, where the constraints of low level categories would be preserved if they offer an information-theoretic advantage. This would have the advantage of enabling us to describe human categorization

[4] Rand similarities (Rand, 1971; Milligan & Cooper, 1986) of all solutions produced, to the best compression solution and the most popular solution, were compared and it was found that a given solution was more likely to be more similar to the best compression one than to the most popular one.

more generally, without restricting ourselves to basic categories only.

Acknowledgments

The first author was supported by the UK Medical Research Council (reference number: G78/ 4804), the Bodossaki foundation, St. Peter's College, Oxford, and the A. S. Onasis foundation (reference: Group S-076/1996-97).

References

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991a). Is human cognition adaptive? *Behavioral and Brain Sciences, 14,* 471-517.

Anderson, J. R. (1991b). The adaptive nature of human categorization. *Psychological Review, 98,* 409-429.

Barlow, B. H. (1983). Understanding natural vision. In J. O. Braddick & C. A. Sleight (Eds.), *Physical and Biological Processing of Images.* Berlin: Springer-Verlag.

Barlow, B. H. (1974). Inductive inference, coding, perception, and language. *Perception, 3,* 123-134.

Braine, M. D. S., O'Brien, D. P., Noveck, I. A., Samuels, M. C., Lea, B. L., Fisch, S. M., Yang Y. (1995). Predicting Intermediate and Multiple Conclusions in Propositional Logic Inference Problems: Further Evidence for a Mental Logic. *Journal of Experimental Psychology: General, 124,* 263-292.

Brown, H. I. (1989). *Rationality.* London: Routledge. Chater, N. (1996). Reconciling Simplicity and Likelihood Principles in Perceptual Organization. *Psychological Review, 103,* 566-591.

Chater, N., & Oaksford, M. (1993). Logicism, mental models and everyday reasoning. *Mind and Language, 8,* 72-89.

Chater, N. & Oaksford, M. (submitted). A Rational Analysis of Syllogistic Reasoning.

Chater, N., & Pothos, E. M. (manuscript). Non-parametric clustering by minimum description length.

Cheng, P. W., Holyoak, K. J. (1985). Pragmatic Reasoning Schemas. *Cognitive Psychology, 17,* 391-416.

Corter, J. E. & Gluck, M. A. (1992).Explaining basic categories: Feature predictability and information. *Psychological Bulletin, 2,* 291-303.

Dirlam, D. K. (1972). Most Efficient Chunk Sizes. *Cognitive Psychology, 3,* 355-359.

Evans, St B. T. J. (1991). Theories of Human Reasoning: The Fragmented State of the Art. *Theory & Psychology, 1,* 83-105.

Evans, St B. T. J., Newstead, S. E., Byrne, R. J. M. (1991). *Human Reasoning, The Psychology of Deduction.* Lawrence Erlbaum Associates, Hove.

Feller, W. (1970). *An introduction to probability theory and its applications.* New York, Wiley.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical neurons. *J. Opt. Soc. Am. A., 4,* 2379-2394.

Komatsu, L. K. (1992), Recent Views of Conceptual Structure. *Psychological Bulletin, 112,* 500-526.

Milligan, G. W. & Cooper, M. C. (1986). A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. *Multivariate Behavioral Research, 21,* 441-458.

Murphy, G. L. & Smith, E. E. (1983). Basic-level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior, 21,* 1-20.

Oaksford, M. & Chater, N. (1994). A Rational Analysis of the Selection Task as Optimal Data Selection. *Psychological Review, 101,* 608-631.

Olshausen, B. A. & Field, D. J. (submitted). Learning efficient codes for natural images, The roles of sparseness, overcompleteness, and statistical independence.

Quinlan, R. J. & Rivest, R. L. (1989). Inferring Decision Trees Using the Minimum Description Length Principle. *Information and Computation, 80,* 227-248.

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association, 66,* 846-850.

Redington, F. M., Chater, N. & Finch, S. (1993). Distributional information and the acquisition of linguistic categories, A statistical approach. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 848-853). Hillsdale, NJ, Erlbaum.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14,* 465-471.

Rosch, E. & Mervis, B. C. (1975). Family Resemblances, Studies in the Internal Structure of Categories. *Cognitive Psychology, 7,* 573-605.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyles-Brian, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382-439.

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science, 237,* 1317-1323.

Stich, S. (1990). *The fragmentation of reason.* Cambridge, MA, MIT Press.

Tenenbaum, J. B. (1996). Learning the structure of similarity. In *Advances in Neural Information Processing Systems 8.* San Mateo, CA, Morgan Kaufman.