Similarity and categorization

Ulrike Hahn

School of Psychology
Cardiff University

Michael Ramscar

School of Cognitive Science
University of Edinburgh

OXFORD
UNIVERSITY PRESS

4 *Categorization by simplicity: a minimum description length approach to unsupervised clustering*

Emmanuel M. Pothos and Nick Chater

There is a strong intuition that one important factor in determining psychological categories is that they should group similar items together—that categories should be seen as running along default lines in a psychological similarity space. To make this idea precise requires finding some 'objective' criterion for determining a 'good' classification, given a set of similarity data. This chapter provides such a criterion, based on an application of a simplicity principle, that can be viewed as a general criterion for cognition. To illustrate the approach, we address the specific illustrative problem of dividing a set of items into groups, on the basis of data consisting of pairwise similarities between the items. Clusters are defined as groups of items such that between-cluster similarity is minimized, while within-cluster similarity is maximized. A simplicity principle is used to assess the relative goodness of different clusterings on the same data set: a particular classification is good to the extent that it provides a short encoding of the similarity information. This criterion enables us to identify the optimal classification for a set of items. The utility of the present approach is illustrated by comparing the performance of well-known clustering algorithms on artificial data sets, and investigating clustering problems from the psychology literature.

Introduction

There is a powerful intuition that good, coherent categories group together things that are similar. What makes 'bird' a good category, we feel, is that birds share a great many properties (having feathers, laying eggs, and so on). What makes 'things that weigh a prime number of grams' a poor category is that objects that weigh a prime number of grams (perhaps my sofa, an ant, the milky way) have nothing in common at all, aside, of course, from weighing an even number of grams. Pushing this intuition to an extreme, one might suggest that similarity is the main determinant of conceptual structure, and that the concepts that we possess run along natural default lines in our internal similarity space. This line of thinking is embodied in, for example, early discussions of 'basic' categories (Rosch and Mervis 1975).

In recent years, there have been challenges to the centrality of similarity to categorization. There have been attempted demonstrations that other factors, such as background knowledge, are relevant (Rips 1989); and fears that there may be a vicious circularity in elucidating categorization in terms of similarity, when similarity itself seems to presuppose categorization (Hahn and Chater 1997). Moreover, Sloman *et al.* (Chapter 5, this volume) report several instances where one aspect of categorization performance, naming, is shown to dissociate from similarity judgements: that is, the perceived similarity of the objects Sloman *et al.* employed in their studies, could not predict the linguistic labelling of these objects. Likewise, Hampton (Chapter 2, this volume) points out that quite often the way 'similarity' is invoked in categorization may be vacuous because the notion of similarity is flexible enough to accommodate any set of experimental data (see also Hahn and Chater 1997). We agree with Hampton that at a broad level the notion of similarity imposes few computational constraints for models of categorization.

But let us leave aside these concerns, and assume, at minimum, that clustering items by similarity appears to be one important factor in determining what we view as good categories; and, furthermore, let us accept that the close relationship between similarity and categorization is worth exploring, even if the exact direction and nature of the dependence between the two is unclear.

With this in mind, a critical issue is: given a set of similarity data for a set of items, what is a natural way of partitioning those items into a 'good' classification? What principled method can we adopt? It might seem that there must be some standard answer to this problem in the statistical literature on clustering, but, as we shall discuss further below, the problem is instead avoided completely by many statistical methods (although the issue is addressed, using a different approach, and in a somewhat different setting, in important work by Gluck and Corter 1985). This raises the possibility that the idea of concepts falling along natural default lines in similarity space is doomed from the outset, because there is no agreed way of deciding what counts as a natural default line. Given a set of similarity data, perhaps, despite our intuitions, any classification is as good at capturing these similarities as any other. If this is the case, then the role of similarity as underpinning category structure would need to be completely rethought. On the contrary, though, if we could find such a criterion, this would provide a useful direction for seeking to test the idea that psychologically natural categories break up items according to clusters in similarity space, in a concrete way.

This chapter takes up this challenge. We propose that there is a preferred criterion for mapping similarity data into discrete sets of 'best' categories—based on a simplicity principle, that has been proposed as a general goal in many areas of cognition (Mach 1959; Leeuwenberg and Boselie 1988; Chater 1997, 1999). Roughly, the idea is that the cognitive system seeks to find the simplest explanation for the data it encounters—whatever the character of that data, and whatever the nature of the explanations that it postulates. Suppose that the data has the form of a matrix of similarities (which might be so, if similarity is psychologically basic, and easily computed—we note that this is, of course, very unclear, as we have mentioned above). Suppose further that the explanations for these data are simply different classes of items (intuitively, being part of the same category 'explains' why two items are similar—we shall see how this intuition can be made precise below). Then perhaps we can use a simplicity principle to determine which grouping is optimal—the best classification should correspond to the simplest explanation for the similarity data.

This chapter spells out one way in which this can work, for a particular representation of similarity data, and for a particular definition of the 'meaning' of a cluster. These specifics will be spelt out below. But the point of the exercise is, from the point of view of psychological theory, largely illustrative. We suggest that this kind of analysis may be relevant to how spontaneous categorization actually occurs in humans. Although our motivations are ultimately psychological, the goal of this chapter is to establish the technical viability of this kind of approach, rather than to report empirical data. We aim to put a new theoretical candidate into the ring. Another way of looking at the contribution of this chapter is as a piece of statistics—as providing a new method for building methods of clustering based on a simplicity principle. Since clustering is so important in the cognitive science, this work may be of fairly general interest, from this perspective, even to those who utterly reject the possibility that similarity may play any foundational role in the understanding of how categorization occurs.

The chapter is organized as follows. First we discuss the relationship between categorization and induction. We then provide an overview of existing clustering techniques and argue that these cannot be used to answer the question of what is a 'natural' classification for a set of items. We call a clustering 'natural' to the extent that there is evidence for it in the similarity structure of the objects classified. This discussion leads to the presentation of our own model, where a formal definition of classification goodness is suggested (that can be used to provide a quantitative measure for whether there is evidence for one classification as opposed to another). We next illustrate the measure with comparisons of different clustering algorithms on artificial data sets, and also by classifying some well-known data from the psychology literature.

Categorization and induction

There are several ways to make inferences from past experience to future events. Clustering, the partitioning of a set of instances into groups, is one of them, in the sense that the groups provide us with information about regularities in the instances. Clearly, if each new instance is completely unique, then clustering is useless. However, in most other cases, identifying an object as a member of a category will give us insights about several properties of the object, that we have not observed directly. For instance, if we are presented with a red, round object, of soft texture, and we infer from this information that the object is an apple, then we also know that this object is edible, has a characteristic taste, etc.

Inference from past experience is problematic, however, in the sense that there is no unique way to generalize from past experience to future events (Goodman 1954; Watanabe 1985). This problem carries over to clustering investigations: given a set of items, there is an intractably large number of possible divisions of the items into groups. The way this problem has been approached traditionally, is by defining a heuristic or criterion that would guide the classification of items; an example would be something along the lines of, 'cluster together the two most similar items in the set and proceed by combining these groups/items that are most similar to each other' (where similarity of two groups might be defined as the greatest similarity between any item in the first group and any item in the second one; this method is called single linkage and will be described more carefully in the following section).

Grouping by some function of similarity of the items, such as in the example above, is by far the most common approach (for alternatives, see for example, Medin *et al.* (1987), who suggest a classification system that attempts to extract rules describing different groups). With similarity as the guiding principle, the objective is to create groups of items so that more similar items end up in the same group. This is far from straightforward, however, because there is no unique way to assess similarity between items (Goodman 1954); thus, with different methods, one is likely to get different partitionings of the same data set, which is problematic if one is interested in making some inference on the basis of the partitioning.

So to summarize so far, clustering appears as an intuitive technique to organize a set of items in a way that inferences about new instances can be made. But, the usefulness of the method is reduced because different clustering methods will lead to different partitionings of the same data set, and there is no criterion to prefer one as opposed to the others. To reiterate some of the concerns raised by James Hampton in this volume, similarity independent of a particular model is not a very constraining grouping principle. So, if there are several possible classifications for the same data set, how could we decide among the best ones?

Some researchers have suggested that a way to get round the problem of evaluating different classifications on the same data set is by identifying the clustering methods that perform better on specific test data sets. Thus, the method of choice for a new data set will be determined by how similar it is to one of these test cases (Fraboni and Cooper 1989). In this work we aim for a more general solution to this problem. We will restrict ourselves to clustering on the basis of similarity information between items. This information could be in the form of vector distances, or confusabilities between items (Shepard 1987), or measures in terms of set theoretic properties of the items (Tversky 1977). Although we do not provide a solution to the problem of what is the most appropriate way to compute similarity between items (if indeed there is a solution to this problem), we will adopt a representation of similarity that is as general as possible, so that our model would be flexible with respect to how similarity is defined.

Our particular approach to formalizing simplicity is based on the minimum description length framework (MDL). But why is MDL an appropriate criterion for clustering? MDL is a particular formalization of the familiar notion of simplicity in inductive inference, according to which the simplest theory is the best (more accurately, the preferred theory is the one such that the description of the theory plus the description of the data in terms of the theory is least; Rissanen 1978, 1986a, 1986b; Wallace and Freeman 1987). William of Ockham (1285?–1349?) is credited to have first stated a principle of simplicity; Ockham's assertion was that 'entities are not to be multiplied beyond necessity' (*non sunt multiplicanda entia praeter necessitatem*), which has become to be understood as 'plurality should not be assumed without necessity' (see Bosch (1994) and Derkse (1993) for a discussion of the historical origins of the simplicity principle). Intuitively, the simplicity principle is useful, since in the absence of any information about the world our only strategy is to go for the 'simplest' (as stated above) hypothesis (see Barlow (1983) and Olshausen and Field (submitted) for an application in low-level vision). Formally, Vitanyi and Li (submitted, a) have shown the simplicity will identify both the most probable theory and the best theory for prediction (see also Bosch 1994). Furthermore, it has been suggested that many shortcomings of standard Bayesian models for generalization can be addressed by appeals to simplicity (Forster 1995; see Vitanyi and Li (submitted, a) for a demonstration of equivalences between simplicity and Bayesian methods, and Chater (1996) for an illustration of this equivalence in perception).

Thus, there is ample indication to suggest that simplicity is a good strategy. In subsequent sections we will describe a particular formalization of the simplicity principle (the minimum description length principle; Rissanen 1978) and show how this can be applied to classification.

Overview of clustering

In this section, we review some well-known methods of clustering; for more comprehensive reviews, see Krzanowski and Marriott (1995), Gordon (1994), or Everitt (1993). The objective of all methods is to identify a partitioning of the objects in a domain, such that the groupings will reflect regularities in the similarity structure between the items. The input to such procedures consists of a similarity matrix between the items (containing information about the similarity between two items, for all pairs of items); and the output, some representation of the cluster structure in the domain (see later). This differs from models of human categorization in an important way: the importance of similarity is taken for granted. In statistics, this makes perfect sense: we would only be interested in grouping things together if it would be meaningful to perceive some of these things as being more similar compared to others. In human classification, the situation is more complex. Identifying the similarity structure of a set of items as the starting point of categorization, provides us with a very tangible advantage. Namely, this enables us to specify the objective categorization quite easily: we select the classification for a set of items that best captures the similarity structure for these items. Granted, we have not yet discussed what 'bestness' is about (and, as will become apparent shortly, this is far from an easy problem). Nevertheless, this still goes a long way towards clarifying the problem of grouping, compared to the equivalent situation in human categorization where it is not even clear as to what categorization is about (for example, see the introduction of Chapter 5).

Back to clustering: the similarities between the items are typically derived on the basis of some metric distance, so that most methods assume that the similarities obey the metric axioms, i.e. minimality (the distance of any point to itself is zero, or the similarity of any item with itself is maximal), symmetry (similarity of item *a* with item *b* is the same as the similarity of item *b* with item *a*), and the triangle inequality (dissimilarity of items *a* and *b* plus dissimilarity of items *b* and *c* is at least as great as the similarity of items *a* and *c*). The metric axioms are clearly justified in a broad range of contexts; in the context of psychological investigations, it has been an issue of controversy whether similarity information derived empirically (for example, by asking participants in a psychology experiment to rate qualitatively the similarity between different objects) is consistent with the metric axioms (for an argument supporting this view, see Shepard (1987); for examples of violations of the metric axioms in psychological similarity judgements see Tversky (1977), and also Bowdle and Gentner (1997) for a more recent view of these issues).

The output of the clustering procedures can be, generally, of two kinds: either a hierarchy of groups such that the bottom level consists of individual items and the top level of an all-inclusive cluster, or one set of groups. In both types of cases, clusters are usually non-overlapping (for hierarchical approaches this applies to clusters at the same level; see Shepard and Arabie (1979) and Tenenbaum (1996) for exceptions).

The hierarchical clustering models (or agglomerative procedures) initially consider each item in a domain as a separate cluster. In each step, the two most similar items/clusters are combined together, until all items are included in the same cluster. Where such methods differ is mainly in the way the similarity between two clusters is computed. For example, in the single-link method, a much celebrated early approach (Sokal and Sneath 1963; Jardine and Sibson 1968; see Johnson (1967) for a more general discussion), the similarity between two clusters is defined as the greatest similarity between any item in the one cluster and any item in the other. Such a method will clearly lead to chain-like clusters, even if such groupings may not be appropriate for a data set (see Lance and Williams (1975) for criticisms of the single-link procedure, and Hartigan (1975) for a more general discussion and the plausibility of chain structures). However, an alternative approach, the complete-link method (Kuiper and Fisher 1975; Baker and Hubert 1976; Hubert and Baker 1977), where the *least* similarity between any point in one cluster and any point in the other will determine the similarity between the two clusters, is not without equivalent problems. For instance, complete-link methods are heavily biased to partition a set of items into clusters of more or less similar size.

The single- and complete-link methods are extremes in a continuum of methods, where similarity between two clusters depends on some function of the similarity between a point in the first cluster and another in the second. (see Hubert (1974) for an explication of this observation using graph theory). There are several other possibilities of procedures, the end result of which is a hierarchy of clusters for a set of items (see Norusis (1994) for an overview of methods that are common in practical clustering applications today).

The alternative class of clustering models involves algorithms that aim to provide a 'natural' classification for a set of items. That is, instead of producing a hierarchy of clusters, they would compute one classification (again, in most cases non-overlapping) that is supposed to be reflecting the optimal partitioning of the items, in the sense of making the regularities in the structure of the items as salient as possible. With hierarchical methods, such solutions are possible as well, if one provides a 'cut-off' criterion for the agglomeration procedure (that is a criterion to indicate when further combining of clusters is to be terminated; but note that such cut-off indicators are usually external to the actual algorithm).

The success of models in this second class depends on defining an appropriate criterion for how good a classification is. That is, there must be a measure to indicate when the optimal classification has been reached, and so terminate additional changes; MacQueen (1967) presents an early example (see also Banfield and Bassill 1977). Although the choice of such criteria is relatively unconstrained, researchers in this area have tried to come up with function of classification goodness that could be justified on a priori theoretical grounds. This reflects the general understanding that although hierarchical methods are more like tools to represent similarity information in a set of items in an elegant way, clustering procedures that construct a single classification attempt to address more directly the issue of what is the true group structure in a set of items (if indeed such a question is meaningful; Hartigan (1975) discusses the question of what is meant by a cluster or group).

For example, Wong (1993) relied on thermodynamics to derive an 'objective' function for determining classification goodness. In this context, objective is used to describe criteria that are meant to apply in as general conditions as possible; that is, criteria that are independent both of the type of investigation and the actual properties of the data set. Thermodynamics is an appropriate framework because it provides a formal description of the statistical properties of systems composed of an assembly of more elementary objects. More specifically, he defined cluster centres as the points that minimize a free-energy function (set up in a way to represent the clustering problem; note that at the lowest level the information about the items to be clustered that is manipulated relates to their similarity structure), and thus his procedure could provide an answer as to how many clusters can be postulated for a set of points. The actual method is an agglomerative one, so that a hierarchy of classifications is produced; but the emphasis is on the final answer—the number of clusters that optimally describe the domain—and so all hierarchical information is meant to be purely descriptive. A similar approach has been reported by Buhmann and Kuhnel (1993). Their calculation of cluster centres is similar to that of Wong (1993), in that it depends on minimizing a free-energy function. However, a novelty in their theoretical framework is that they also consider the trade-off between fit provided by the clusters (that is, how well the clusters describe the actual distribution of points) and complexity of cluster structure (the more clusters are postulated, the more complex the cluster structure will be, in the sense that it will be more difficult to decide where a new point belongs).

Generally speaking, as was the case with the hierarchical clustering procedures, methods such as the ones mentioned above will lead to similar but not identical results. Although there has been a major effort in defining 'objective' criteria for classification goodness, there are still several arbitrary features in approaches such as the one of Wong

(1993) and Buhmann and Kuhnel (1993), that preclude meaningful comparison between different classifications. For example, in the above approaches, the error—or distortion—terms have not been motivated in the same way that justified the use of free energies.

There is one important line of research that does directly address the question of how many categories it is appropriate to assign to a set of data, and this line arises from psychology, rather than statistics. Corter and Gluck (1992; see also Gluck and Corter 1985) developed a model of category utility. They argued that categorization is useful to the extent that it enables us to predict the features of instances (see also Anderson (1991) for a model of dynamic categorization based on a similar motivation). In particular, they defined category utility as

$$CU(c,F) = P(c)\sum_{k=1}^{m}[P(f_k|c)^2 - P(f_k)^2],$$

where f_k refer to features and c to a category, and showed that in a variety of situations, given a hierarchy of objects, their measure would identify correctly basic-level categories (Rosch and Mervis 1975). Category utility has been applied successfully in a number of categorization systems developed in the artificial intelligence literature (Fisher 1987, 1996; Gennari *et al.* 1989). The relationship between category utility and the present approach is an interesting area for future research. For now, we note simply that the two methods apply to different kinds of data—category utility requires that items consist of feature bundles; whereas the approach developed here assumes simply that pairwise similarity judgements between items are available, and makes no assumptions about how items are represented.

In sum, two points are evident: first, all methods of clustering in statistics start from similarity; this is considered the only meaningful motivation for doing classification in the first place. Secondly, in terms of discriminating between the possible methods, the situation is as bad as that in human categorization: there are several possible models, and it is not always clear as to what would be an effective strategy to compare them. Nevertheless, one shortcoming can be identified readily: none of the methods mentioned above provides us with any information regarding what could be an optimal classification for a set of items. Even the K-means clustering techniques required explicit information as to how many clusters need to be identified. This is the shortcoming that we address with the present work.

Clustering by simplicity

In this section we describe the general features of our approach; a more technical exposition can be found in Chater and Pothos (submitted). The general guiding principle is simplicity: classifications are good to the extent that they provide a parsimonious representation of the similarity structure of a set of items. More specifically, the version of simplicity we will adopt is the minimum description length principle (henceforth, MDL) (Rissanen 1978, 1986a, 1986b), according to which '. . . 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 and Rivest 1989, p. 227). As we will see, classifications can be thought of as theories on the regularities between a set of items, so that different 'theories' can be compared with respect to how simple a representation of a set of items they provide.

The description lengths in the MDL principle refer to how extensive the description of an object is in some (universal) programming language. That is, for an object x we ask how long is the shortest program to describe object x? For example, consider a sequence of natural numbers: 1, 2, 3, . . . 25. A simple program for this sequence would be 'start with 1 and increment in units of 1 until 25 is reached'. On the contrary, with a sequence of natural numbers presented randomly, it is unlikely that we can come up with a program that is shorter than simply listing the numbers one by one. In general, it will be the case that more regular objects will be described by shorter programs.

In order to apply the MDL formalism in clustering, we need to define what we mean by description lengths for clusters, similarity information, etc. Starting with similarity, at the outset we noted that we do not wish to restrict ourselves to any particular theory of similarity (Tversky 1977; Shepard 1987). The problem of defining a meaningful measure of similarity is one that is, to a large extent, separate from that of determining classification goodness; furthermore, it is unclear as to whether there can be a general representation of similarity (for an interesting suggestion see Chater and Hahn 1997). The way we avoid dealing with similarity definition issues, is by assuming a maximally general similarity representation, that would be compatible with any other more specific theory. In particular, we suggest that the similarity information between a set of items is in the form of the relative magnitude of pairwise inequalities: for items a, b, c, and d, we would have, for example, that similarity (a, b) is greater (or less) than similarity (c, d). Such a specification is suitable in the sense that it makes no assumptions at all about the properties of the data set. With respect to the metric axioms, for example, one can readily see how they could all be violated in a way consistent with the representation suggested.

With respect to applying the MDL principle to classification, the description length required to specify all the inequalities for a set of items can be computed easily by observing that for each inequality there are two possibilities: for the pairs (a, b) and (c, d) either the similarity between (a, b) is greater to that of (c, d), or it is less (in this work we have ignored ties; this is only a matter of simplifying methodology and all ideas can be easily extended to account for ties as well; note also, that with real-valued domains, ties would practically almost never arise). From information theory, (Cover and Thomas 1991), the codelength required to make a binary decision, that is a decision between only two possible outcomes, is one bit. Thus, with n objects (if we further assume minimality and symmetry, again for simplicity of exposition), we have $p = n \times (n - 1)/2$ pairwise relations between them, and so $p \times (p - 1)/2$ bits will be required to define the similarity structure of the objects in terms of inequalities between pairs of items.

Call the above quantity—the codelength required to describe all similarity information between a set of items—DL(raw), where DL stands for 'description length'. As mentioned before, a classification can be seen as a 'theory' that aims to flesh out regularities in this set of similarity relations. More specifically, we define clusters as collections of objects in a set, such that objects within a cluster are more similar than objects between clusters. Such a definition has been motivated as plausible in the human classification literature in the context of basic categories (Rosch and Mervis 1975), and is justified in the present work only in the sense that it captures, intuitively at least, all the specifications that are relevant for deciding whether a cluster is good or not good. The reason why we do not attempt a more formal justification of our choice for this definition of clusters is that, as will become apparent later, one is not needed, in the sense that our model does not depend on a particular cluster definition.

Saying that clusters correspond to objects such that between-cluster similarity is as low as possible, while within-cluster similarity is greater, can be used to specify many of the inequalities in the description of the similarity structure between the items. For example, suppose that we have a domain of four objects, *a*, *b*, *c*, and *d*, and we place objects (*a*, *b*) in one cluster, while objects (*c*, *d*) are in another. Then, from our definition of clusters, this entails the information that similarity (*a*, *b*) is greater than similarities (*a*, *c*), (*a*, *d*), (*b*, *c*), and (*b*, *d*), and similarly for similarity (*c*, *d*). Now, all these 'constraints' specified by the cluster may or may not be correct; however, generally speaking, a classification will be successful to the extent that it specifies many constraints that are correct.

So using a classification to describe the similarity structure of a set of items can reduce the description length required to specify all similarity information. Thus, DL(raw|clusters), the description length required for specifying the inequalities *with clusters*, will be less than DL(raw), depending on how many constraints the clusters specify. Since knowledge of each inequality is worth one bit, each constraint reduces DL(raw) by one bit as well.

As noted before, however, some of these constraints might be erroneous, so that the reduction in description lengths associated with a clustering configuration (call this compression, see later) will not be simply DL(raw) – DL(raw|clusters). Thus, one must consider the description length required for identifying the errors, which we call DL(errors). The way to do this is by considering the set of all constraints and deciding which subset of these constraints includes only all the erroneous ones (once an erroneous constraint has been found, the correct answer is known automatically, since for each pair of similarities there are only two possibilities; the required expression is obtained from standard combinatorics and is reported in Chater and Pothos (submitted)).

The final term that is needed before the MDL principle can be applied to the classification problem relates to the complexity of the cluster structure used. We start with a specification of the similarity relations only in terms of relative magnitudes between pairwise similarities, and 'clusters' is a construct that is postulated a posteriori; thus, we need to consider the savings provided by the constraints, only with regard to the additional costs required to describe the clusters. The description length to specify cluster configurations, DL(clusters), can be computed by considering all possible classifications on a set of items, and the number of bits needed to select the one actually used (see Chater and Pothos (submitted) for a derivation).

With the above terms, we can now apply the MDL principle: classification goodness is given by the difference

$$\text{compression} = \text{DL(raw)} - (\text{DL(raw|clusters)} + \text{DL(clusters)} + \text{DL(errors)}).$$

The better a classification, the more the constraints and the less the errors, so that the corresponding reduction in description lengths, compression, is likewise greater. With the above quantities, a numerical measure of classification goodness can be derived, so that different possibilities can be assessed quantitatively.

In what way is our formulation different from the others we mentioned? One could, for example, argue that in the same way there are arbitrary error terms in Wong's (1993) scheme, or the one by Buhmann and Kuhnel (1993), likewise in ours the particular way description lengths are computed will, no doubt, influence results so that the model will be biased to favour certain types of classification structures.

The difference lies in that although it is true that the actual specification of description lengths is somewhat flexible, the compression equation is maximally general and at the heart of the simplicity/MDL approach to generalization. For instance, another investigator might be able to come up with a better or more intuitive way to define clusters than ours, or a more efficient way to specify errors. Alternatively, for different data sets, different definitions of clusters might be more suitable (in terms of leading to greater compressions). However, despite all such differences, classification goodness measures calculated on the basis of the compression equation stated above are always directly comparable, *even for different methods*. This is because the complexity of description lengths in information theory is a notion that is both data and model independent; an object whose description is 40 bits long would be always simpler than an object whose description length is 50 bits long, regardless of the actual coding scheme that was used to compute these lengths (for the general theory of 'objective' measures of complexity in information theory, see Li and Vitanyi (1997); for application in psychology see Chater (1996), and in statistical inference, Juola *et al.* (1998)).

In the next sections, we illustrate these ideas by clustering artificial data sets with a set of well-known clustering procedures, as well as two novel clustering algorithms that optimize directly the compression measure of classification goodness in a local way. Although, as is typically the case, different procedures perform better with different data sets, in each case we can use the compression measure of classification goodness to identify the optimal clustering. Also, we classify some well-known data sets from the psychology literature, so as to further illustrate the uses of being able to identify the optimal clustering configuration on a set of items.

Artificial data sets

In this section, we present analyses of classification goodness on four artificial data sets. The data sets, shown in Figs 4.1–4.4, have all been constructed explicitly to reflect different data structures, while some points were also introduced by hand to make the final configuration slightly ambiguous. For example, although the Fig. 4.1 arrangement of points seems to be consistent with a two-cluster structure, the points in the middle cannot readily be assigned. Note that the similarity relations required as input to the classification goodness measure have been computed on the basis of the metric distances between the

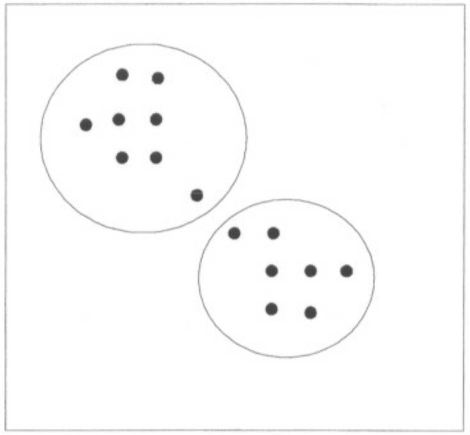

Fig. 4.1 Two overlapping clusters. The best cluster configuration resulted in a compression by about 2744 bits, almost 50% of the description length for the similarities without clusters (in this, and all other examples in this section, there are always 15 points; thus, there are 15 × 14/2 = 105 pairs of points, and hence 105 × 104/2 = 5460 pairs of similarities to be determined).

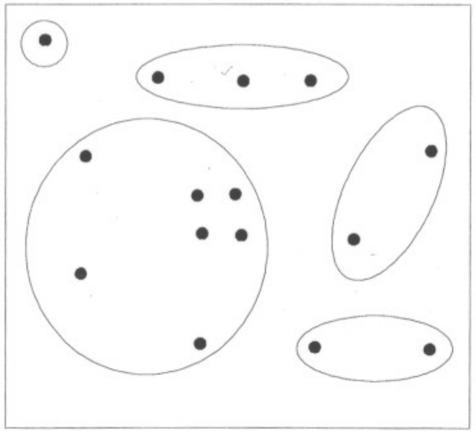

Fig. 4.2 Embedded category in random noise. Although most methods clustered the points corresponding to the embedded category together, the ambiguity in the random noise points resulted in poor overall solutions. The best compression achieved was only 1059 bits, almost a third of the compression possible in the straightforward two-clusters case.

points in the examples in Figs 4.1–4.4, but, as discussed before, the method does not require at all that the similarities between the objects are consistent with the metric axioms.

The clustering algorithms compared were the nearest neighbour, furthest neighbour,

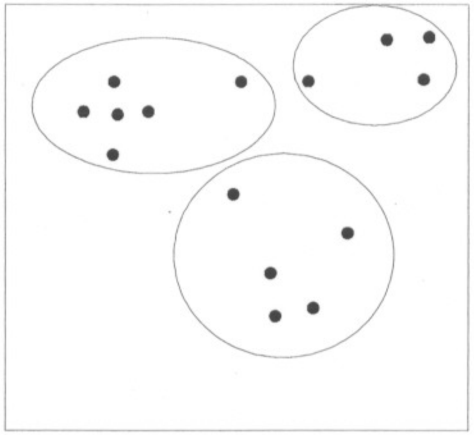

Fig. 4.3 A simple case of three clusters, with some noise. Although there was considerable structure, the best compression of 1987 bits again reflects the fact that some points were ambiguous, compared to the case in Fig. 4.1, where the structure was more transparent.

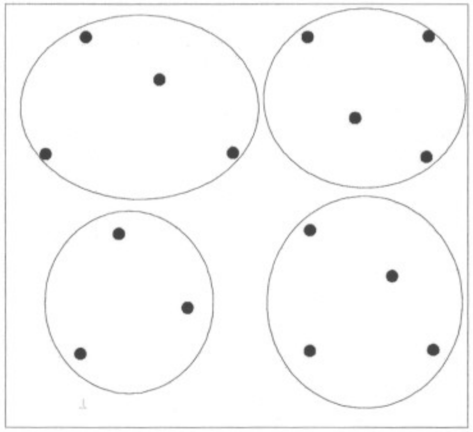

Fig. 4.4 Little obvious structure. We have constructed this data set so that no obvious structure would be readily discernible. Nevertheless, many algorithms succeeded in identifying nearly the optimal configuration (associated with a compression of 1276 bits).

between-groups linkage, within-groups linkage, centroid clustering, median clustering, and the Ward's method (methods described by Norusis 1994). Our choice of these methods simply reflects the fact that these are the most common procedures in clustering applications. Note that they are all agglomerative, hierarchical; that is, they initially consider all objects in a domain to be individual clusters, and in each step two objects or clusters are merged together. Since the simplicity criterion can be used to identify the optimal classification for a set of items (but not a hierarchy), for each of the above methods we identified the best classification for a specific method as that partitioning in the hierarchy that led to the greatest compression. For example, suppose we consider the nearest neighbour method; for a given data set, the nearest neighbour prediction for the best, or most natural, classification, will be that level in the hierarchy produced which leads to the greatest compression.

We also constructed two novel clustering agglomerative algorithms based on optimizing the compression criterion directly, in a local way. That is, as is the case generally with agglomerative algorithms, they start off with each item in a separate group. The two objects or clusters that are merged in each step, will be those such that the resulting classification is associated with a greater compression. One algorithm looks at simply how much compression is increased at each step (the 'absolute' method), while the other also takes into account other factors in preferring one merger as opposed to another (the 'relative' method; see Chater and Pothos (submitted) for details). Both these methods utilize local optimization algorithms in the sense that improvement over the current solution is assessed only with respect to possibilities in the next step. Thus, they do not guarantee to compute the best compression classification (but the problem of discovering the shortest code for an object is generally intractable).

For each data set, there is an 'objective' best classification, namely the one that is associated with least compression. Thus, the different methods can be evaluated (for these four data sets) according to how often they succeed in identifying the best compression solutions; the results are shown in Fig. 4.5.

Another straightforward dimension of comparison would be simply to ask what average compression was achieved by different methods for the four data sets. Such a calculation would be useful since a particular method might still be able to identify high compression classifications, even though they are not the optimal ones. Figure 4.6 shows these results, and the overall impression is the same as that from Fig. 4.5: namely, different methods (including our own) perform better in some situations, as opposed to others, and that overall no method can be credited with a clear advantage.

So what do these demonstrations show? One can argue that there is nothing remarkable in either the data sets, or that all methods perform, generally, at roughly similar levels.

The data sets may be simple, but clustering them has been far from trivial: this is reflected in the fact that we observed great between-method variability, in the number of best compression solutions discovered and the average compression for the four data sets, for each method. Thus, although the data sets represent classification structures that are highly intuitive to a human observer, none of the methods tested can reliably *predict* this classification structure. Moreover, the hierarchical methods we tested are even less constrained than ours, since they will provide not a single classification, but a family of possible partitionings. We think that the utility of our approach is that in each case we can evaluate different groupings according to how much compression they provide, and thus select the optimal one.

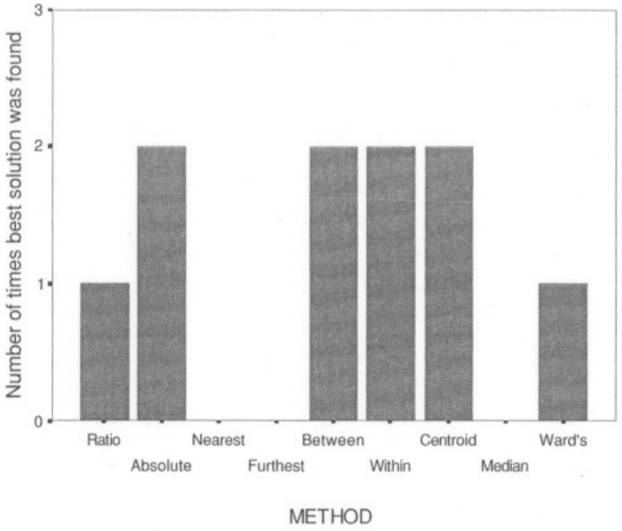

Fig. 4.5 Number of times each method discovered the best solution. Note that with the exception of the straightforward two-cluster case where most algorithms found the best solution (or a very similar one), in all the other cases on average about only two algorithms achieved the maximum compression.

Fig. 4.6 Average compression achieved by each of the algorithms tested across the four data sets we used. The total information content of each domain was 5460 bits.

Psychological data sets

In this section we present analyses on some well-known data from the psychology literature. These results are useful to the extent that the best classifications predicted by compression are consistent with those expected intuitively.

Miller and Nicely's data

Miller and Nicely (1955) were interested in investigating the phonological similarity structure between the consonants in the English alphabet. Thus, they had a set of participants make several same/different judgements on consonant pairs presented auditorially (they also used different filters and noise conditions; for our analysis, we selected the similarity information corresponding to least frequency distortion and maximum noise). They reasoned that more similar objects would be more often confused together, and so constructed a *confusability* matrix for the consonants, that is a matrix such that in each cell there was information about how many times the column consonant was confused with the row one (see Shepard (1980, 1987) for a general argument with respect to the psychological plausibility of such a procedure and associated implications; also, see Shepard (1962a) for a more general discussion about the nature of psychological relation).

Such a confusability matrix is equivalent to a dissimilarity matrix that is suitable for clustering procedures. Describing the similarity structure of a set of objects in such a way is more general than assuming a spatial representation (as we did in the examples with the artificial data sets), because the similarity relations in a dissimilarity matrix must obey only transitivity, and do not have to conform to the metric axioms (transitivity states that if similarity (a, b) is greater than similarity (b, c) and similarity (b, c) greater than similarity (d, e), then it must also be the case that similarity (a, b) is greater than similarity (d, e)). The similarity information required to compute compressions in our model is still more general, in that transitivity need not be adhered to either.

In practice, Miller and Nicely (1955) normalized the confusability data so that minimality and symmetry would be obeyed, and for ease of comparisons we will adopt their methodology as well (but, to reiterate, this is not required). Figure 4.7 shows the best compression grouping that we identified on the basis of confusability information.

For comparison, Fig. 4.8 shows a classification of the same consonants, now on the basis of a feature representation reported by Shillcock *et al.* (1992); consonants were coded along nine dimensions that were thought to summarize important characteristics of the language perception problem. The similarity of the two classifications can be used to examine whether the Shillcock *et al.* (1992) feature representation is indeed psychologically plausible (bearing in mind, however, that Miller and Nicely's data were collected under high noise conditions). The Rand index of classification similarity (Rand 1971; Fowlkes and Mallows 1983) for the clusterings in Figs 4.7 and 4.8 was 0.83, where 1 indicates identity, thus supporting Shillcock *et al.*'s representation.

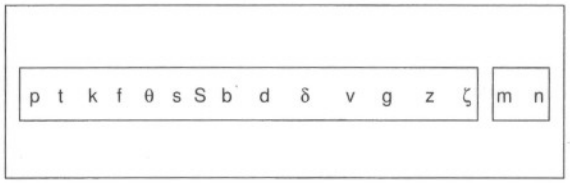

Fig. 4.7 Best cluster solution found for of Miller and Nicely's (1955) confusability matrices (signal-to-noise ratio, –12 db; frequency response of the sound-producing apparatus, 200-6500 cps; for more details see Miller and Nicely 1955). The compression achieved was about 2800 bits and the description length required for the similarity information without any clusters was 7140 bits.

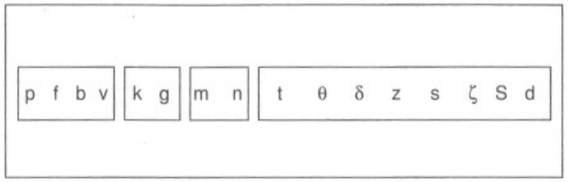

Fig. 4.8 The consonants of Fig. 4.7 classified according to a representation reported by Shillcock *et al.* (1992), whereby the different dimensions of the representation correspond to features relevant to the perception of linguistic objects (compression of the above grouping, about 2500 bits).

Ekman's data

Ekman (1954) reported a study very similar to that of Miller and Nicely (1955), but in the domain of colour. In particular, he asked participants to rate qualitatively the similarity between 14 colours, so that each participants made a similarity judgement for all possible pairs of colours. Furthermore, the participants' responses were averaged and transformed to a scale ranging from 0 to 1, and, as was the case with Miller and Nicely (1955), symmetry and minimality were imposed. Three main groups were identified (Fig. 4.9), a group whose members were primarily long wavelengths, another composed of short wavelengths, and finally a group of intermediate wavelengths. Note that the well-documented effect of the perceived similarity between extreme reds and extreme violets (thus suggesting a cyclic spatial structure for the wavelengths spanning the visible spectrum; see Shepard (1962b), but also Rodieck (1977) for a different view) can neither be seen in the similarity structure of our clustering solution, nor was (apparently) present in Ekman's results (confusability between 674#nm and 434#nm was only 0.16).

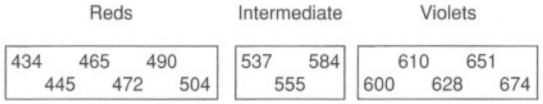

Fig. 4.9 Fourteen colours labelled according to their wavelength (nm) and classified on the basis of similarity ratings collected by Ekman (1954). We found the upper clustering to have the highest compression (1449 out of 4095).

Conclusions and future directions

This research addresses one very basic problem in classification: how can we decide which of the alternative classifications of a set of items is more suitable? Hierarchical classification techniques provide only an indirect answer to this question: the output of such procedures is a family of (hierarchically organized) groupings and it is up to the investigator to decide which one is the most appropriate. Other techniques exist, where a criterion is defined, allowing the derivational of one classification that is optimal relative to this criterion. However, it is generally the case that the criteria optimized in such procedures involve several arbitrary terms, that is terms that cannot be justified in a single theoretical framework. For instance, although the general theoretical motivation for the approach of Wong (1993) and Buhmann and Kuhnel (1993) is founded on thermodynamics and the concept of free energy, the way in which a cluster is defined is but one possibility of many equivalent ones compatible with their frameworks.

In our work, we argued that some version of the simplicity principle can provide the foundation for 'objective' classification. This is because clustering is a type of generalization inference, and simplicity has been suggested as a suitable guiding principle in generalization by several investigations in a variety of fields (Forster 1995; Vitanyi and Li submitted, a, submitted, b). The particular model we presented starts from a general, noncommitting representation of similarity between a set of objects, and interprets clusters as 'theories' about how these similarities are organized. Clusters were defined as sets of objects such that between-cluster similarities is lower than within-cluster similarities, for all pairs of objects. In this way, good classifications would reduce the description length required to specify the similarity structure between the objects, and, applying simplicity, the greater the overall compression, the better the classification.

'Objectiveness' arises in that although the exact simplicity values have been derived on the basis of our particular definition of clusters, and cost terms, the general equations for computing compression for a classification, on the basis of a classification, are model independent. That is, suppose that we have two classifications for a set of objects, A and B; the shortest possible description length for the objects with classification A is DL(A), and likewise for B, DL(B). Then, if DL(A) is less than DL(B) then it is the case that DL(A) is a better classification (according to simplicity) then DL(B), *even if the two description lengths were computed using different equations* (this approach is explicated in the theory of Kolmogorov complexity; see Li and Vitanyi (1997) for an introduction). Thus, our model avoids the problem of having a proliferation of clustering algorithms that cannot be compared with each other.

More generally, we hope that the idea that a simplicity principle for classification might be relevant to explain how people spontaneously group objects, and perhaps also to the nature of the categories that become fossilized into words of natural language. The direct psychological relevance of this work depends, of course, on using the framework developed here to explain empirical data—a project in which we are currently engaged.

Acknowledgements

This research was supported by ERSC grant reference R000222655 to Emmanuel Pothos and Nick Chater. We wish to thank Francis Marriott, Brian Ripley, and Ulrike Hahn for helpful comments and Steven Young for assistance with C programming.

References

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

Baker, F. B. and Hubert, L. J. (1976). A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering. *Journal of the American Statistical Association*, 71, 870–878.

Banfield, C. F. and Bassill, S. (1977). A transfer algorithm for bob-hierarchical classification. Algorithm 133. *Applied Statistics*, 26, 206–210.

Barlow, B. H. (1983). Understanding natural vision. In *Physical and biological processing of images*, J. O. Braddick and C. A. Sleight (ed.). Springer-Verlag, Berlin.

Bosch, A. P. M. van den (1994). Simplicity and prediction. Unpublished manuscript.

Bowdle, B. F. and Gentner, D. (1997). Informativity and asymmetry in comparisons. *Cognitive Psychology*, 34, 244–286.

Buhmann, J. and Kuhnel, H. (1993). Complexity optimized data clustering by competitive neural networks. *Neural Computation*, 5, 75–88.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, 103, 566–591.

Chater, N. (1997). Simplicity and the mind. *The Psychologist*, November, 495–498.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *Quarterly Journal of Experimental Psychology*, 52A, 273–302.

Chater, N. and Hahn, U. (1997). Representational distortion, similarity and the Universal Law of Generalization. In *Proceedings of SimCat 1997: An Interdisciplinary Workshop on Similarity and Categorization*, M. Ramscar, U. Hahn, E. Cambouropoulos, and H. Pain (ed.). Edinburgh University Press, Edinburgh.

Chater, N. and Pothos, E. M. Simplicity clusters. (submitted).

Corter, J. E. and Gluck, M. A. (1992). Explaining Basic Categories: Future Predictability and Information. *Psychological Bulletin*, 2, 291–303.

Cover, T. M.. and Thomas, J. A. (1991). *Elements of information theory*. John Wiley & Sons, New York.

Derkse, W. (1993). *On simplicity and elegance*. Eburon, Delft.

Earman, J. (1992). *Bayes or bust?* MIT Press, Cambridge, MA.

Ekman, G. (1954). Dimensions of color vision. *Journal of Psychology*, 38, 467–474.

Everitt, B. S. (1993). *Cluster analysis,* (3rd edn). Edward Arnold, London.

Fisher, D. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2, 139–172.

Fisher, D. (1996). Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Intelligence Research*, 4, 147–179.

Forster, M. R. (1995). Bayes and bust: simplicity as a problem for a probabilist's approach to confirmation. *British Journal of the Philosophy of Science*, 46, 399–424.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings (with comments and rejoinder). *Journal of the American Statistical Association*, 78, 553–584.

Fraboni, M. and Cooper, D. (1989). Six clustering algorithms applied to the WAIS-R: The problem of dissimilar cluster analysis. *Journal of Clinical Psychology*, 45, 932–935.

Gennari, J., Langley, P. and Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence*, 40, 11–62.

Gluck, M. A. and Corter, J. E. (1985). Information, uncertainty and the utility of categories. *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 283–287). Hillside, NJ: Erlbaum.

Goodman, N. (1954). *Fact, fiction and forecast*. Athlone Press, London.

Gordon, A. D. (1994). Identifying genuine clusters in a classification. *Computational Statistics and Data Analysis*, 18, 561–581.

Hahn, U. and Chater, N. (1997). Concepts and similarity. In *Knowledge, concepts and categories*, K. Lamberts and D. Shanks (ed.), (pp. 43–92). Psychology Press, Hove, England.

Hartigan, J. A. (1975). *Clustering algorithms*. Wiley, New York.

Hubert, L. (1974). Some applications of graph theory to clustering. *Psychometrika*, 39, 283–309.

Hubert, L. J. and Baker, F. B. (1977). Am empirical comparison of baseline models for goodness-of-fit in r-diameter hierarchical clustering. In *Classification and clustering*, J. Van Ryzin (ed.). Academic Press, New York.

Jardine, N. and Sibson, R. (1968). The construction of hierarchic and nonhierarchic classifications. *Computer Journal*, 11, 177–194.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32, 241–254.

Juola, P., Bailey, T. M., and Pothos, E. M. (1998). Theory–neutral domain regularity measurements. *Twentieth Annual Conference of the Cognitive Science Society*.

Krzanowski, W. J. and Marriott, F. H. C. (1995). *Multivariate analysis, Part 2: Classification, covariance structures and repeated measurements*. Arnold, London.

Kuiper, F. K. and Fisher, L. A. (1975). Monte Carlo comparison of six clustering procedures. *Biometrics*, 31, 777–783.

Lance, G. N. and Williams, W. T. A. (1975). A general theory of classificatory sorting strategies: I, Hierarchical systems. *Computer Journal*, 10, 271–277.

Leeuwenberg, E. and Boselie, F. (1988). Against the likelihood principle in visual form perception. *Psychological Review*, 95, 485–491.

Li, M. and Vitanyi, P. (1997). *An introduction to Kolmogorov complexity and its applications,* (2nd edn). Springer-Verlag, Berlin.

Mach, E. (1959). *The analysis of sensations and the relation of the physical to the psychical*. Dover Publications, New York. (Original work published 1886.)

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium in Mathematical Statistics and Probability*, pp. 281–297. University of California Press, Berkeley.

Medin, D. L., Wattenmaker, W. D., and Michalski, R. S. (1987). Constraints and preferences in inductive learning: an experimental study of human and machine performance. *Cognitive Science*, 11, 299–339.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, 27, 338–352.

Norusis, M. J. (1994). *SPSS Professional Statistics 6.1*, SPSS Inc., Chicago.

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

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson, London.

Quinlan, R. J. and 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.

Rips, L. J. (1989). Similarity, typicality, and categorization. In *Similarity and analogical reasoning*, S. Vosniadou and A. Ortony (ed.), pp. 21–59. Cambridge University Press, New York.

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

Rissanen, J. (1986a). Stochastic complexity and modeling. *Annals of Statistics*, 14, 1080–1100.

Rissanen, J. (1986b). Stochastic complexity and sufficient statistics. *Technical Report, IBM Research Laboratory*, San Jose.

Rodieck, R. W. (1977). Metric of color borders. *Science*, 197, 1195–1196.

Rosch, E. and Mervis, C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, 7, 573–605.

Shepard, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27, 125–140.

Shepard, R. N. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, 27, 219–246.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 210, 390–398.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237, 1317–1323.

Shepard, R. N. and Arabie, P. (1979). Additive clustering: representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86, 87–123.

Shillcock, R., Lindsey, G., Levy, J. and Chater, N. (1992). A phonologically motivated input representation for the modeling of auditory word perception in continuous speech. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Erlbaum, Hillsdale, NJ.

Sokal, R. R. and Sneath, P. H. A. (1963). *Principles of numerical taxonomy*. Freeman, San Francisco.

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

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

Vitanyi, P. and Li, M. (submitted, a). Minimum description length induction, Bayesianism, and Kolmogorov complexity.

Vitanyi, P. & Li, M. (submitted, b). On prediction by data compression.

Wallace, C. S. and Freeman, P. R. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B*, 49, 240–251.

Watanabe, S. (1985). Pattern recognition: human and mechanical. Wiley, New York.

Wong, Y. (1993). Clustering data by melting. *Neural Computation*, 5, 89–104.