available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Research Report

Comparison processes in category learning: From theory to behavior☆

Rubi Hammer[a,b,*], Aharon Bar-Hillel[c], Tomer Hertz[d], Daphna Weinshall[a,e], Shaul Hochstein[a,b]

[a]Interdisciplinary Center for Neural Computation, Edmond Safra Campus, Hebrew University, Jerusalem, 91904, Israel
[b]Neurobiology Department, Institute of Life Sciences, Hebrew University, Israel
[c]Intel Research, Israel
[d]Microsoft Research, Seattle, Washington, USA
[e]School of Computer Sciences and Engineering, Hebrew University, Israel

ARTICLE INFO

Article history:
Accepted 28 April 2008
Available online 13 May 2008

Keywords:
Category-learning
Categorization
Perceived similarity
Multidimensional scaling
Perceptron
Expectation-maximization

ABSTRACT

Recent studies stressed the importance of comparing exemplars both for improving performance by artificial classifiers as well as for explaining human category-learning strategies. In this report we provide a theoretical analysis for the usability of exemplar comparison for category-learning. We distinguish between two types of comparison — comparison of exemplars identified to belong to the same category vs. comparison of exemplars identified to belong to two different categories. Our analysis suggests that these two types of comparison differ both qualitatively and quantitatively. In particular, in most everyday life scenarios, comparison of same-class exemplars will be far more informative than comparison of different-class exemplars. We also present behavioral findings suggesting that these properties of the two types of comparison shape the category-learning strategies that people implement. The predisposition for use of one strategy in preference to the other often results in a significant gap between the actual information content provided, and the way this information is eventually employed. These findings may further suggest under which conditions the reported category-learning biases may be overcome.

1. Introduction

Acting adaptively in a complex and changing environment requires the ability to categorize objects and events, regardless of whether the agent is biological or artificial. Categorization can often be performed without supervision since objects can be represented as data points non-uniformly scattered in some multidimensional features space. In this case building categories can be reduced to grouping objects based on their relative proximity in the space (Duda et al., 2001). In particular, Shepard (1987) discussed this principle with respect to the mental representation of objects when referring to human cognition. The general rule that applies in this regard is that the smaller the distance between two objects in mental space, the greater their perceived similarity, and as a result the greater the probability that they will be grouped together

☆ This study was supported by grants from the Israel Science Foundation (ISF) and the US–Israel Binational Science Foundation (BSF), and an EU grant under the DIRAC integrated project IST-027787.

* *Corresponding author*. Interdisciplinary Center for Neural Computation, Edmond Safra Campus, Hebrew University, Jerusalem, 91904, Israel. Fax: +972 2 658 4985.
E-mail address: rubih@alice.nc.huji.ac.il (R. Hammer).

0006-8993/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2008.04.079

in the same category (Medin and Schaffer, 1978; Posner and Keele, 1968; Rosch et al., 1976). For the last three decades numerous studies have demonstrated that these ideas can serve as powerful tools both for designing unsupervised artificial classifiers (e.g. Fleming and Cottrell, 1990; Weber et al., 2000), as well as for explaining human behavior (e.g. Ashby et al., 1999; Pothos and Chater, 2002).

However, unsupervised categorization strategies do not always guarantee creation of proper categories. In fact, that when implemented in artificial classifiers they often fail to provide a satisfying solution, and in particular they fail to provide an explanation of human behavior and motivation. For this reason, many scholars claim that categorization often requires prior knowledge to determine the relevance or irrelevance of different object properties for the categorization task (Caramazza and Shelton, 1998; Keil, 1989; Murphy and Medin, 1985; Sloutsky, 2003; Smith et al., 1996; Tyler et al., 2000). Human conceptual knowledge is therefore not based only on the general similarity among objects, but rather on a more qualitative judgment of the features that are more relevant for categorizing targeted objects, and the degree of similarity among objects in these features (Diesendruck et al., 2003; Hammer and Diesendruck, 2005; Medin et al., 1993).

Fig. 1 illustrates a scenario in which unsupervised categorization may be expected to produce satisfactory results (Fig 1a), and a second scenario in which it may be expected to fail (Fig. 1b). Referring to the pictures in Fig. 1a, it may be expected that when asking a young child, who is not familiar with the presented animals, "Which of these animals is not of the same kind as the others?" he or she will probably identify animal 3 (a bird) as the odd one simply because it differs from the others, (elephants), in almost any possible aspect. In contrast, for the example illustrated in Fig. 1b, we may expect that when answering the same question, the naïve child will probably identify animal 2 (a colorful bird) as the odd one, simply because it dramatically differs from the others in color. But here the child would be wrong — animals 2, 3 and 4, are all ducks, while animal 1 is a seagull. In cases such as this, when irrelevant features overshadow more relevant ones, the result is often inappropriate categorization due to a lack of prior knowledge required for directing attention to the relevant features (such as the shape of the beak).

Thus, ignoring irrelevant, physically salient, features may require the mediation of directed attention. Early categorization models suggested that directed attention may affect the perceived similarity among objects and the mental representation of categories. That is, similarity judgment and categorization are context dependent and expected to be depend on directing attention to specific features according to their relevance in a specific context and not according to their physical salience (Nosofsky, 1986). This implies that the learner's prior knowledge drives his similarity judgment. Later category learning models (e.g. SUSTAIN — Supervised and Unsupervised STratified Adaptive Incremental Network; Love et al., 2004) assume that the learner's goals interact with objective factors such as the nature of the learning task and the structure of the world. These models also assume that prior knowledge, represented by the learner's goals, is an important factor in shaping our conceptual knowledge (see also Medin et al., 1993).

There is presently an ongoing debate concerning the nature of the prior knowledge required for generalizing a categorization rule, as well as the cognitive mechanism that enables obtaining this knowledge. One common belief is that category learning requires the direct learning of which features are most important within each specific object domain (Caramazza and Shelton, 1998; Keil, 1989; Murphy and Medin, 1985; Tyler et al., 2000). Others suggest that prior knowledge can be obtained without direct supervision, but categorization still requires being experienced with objects and their features. Such interaction with objects is expected to end up with rescaling features relevance for categorization (Sloutsky, 2003; Smith et al., 1996). More recently it was suggested that category learning should be seen as two different learning processes — classification and

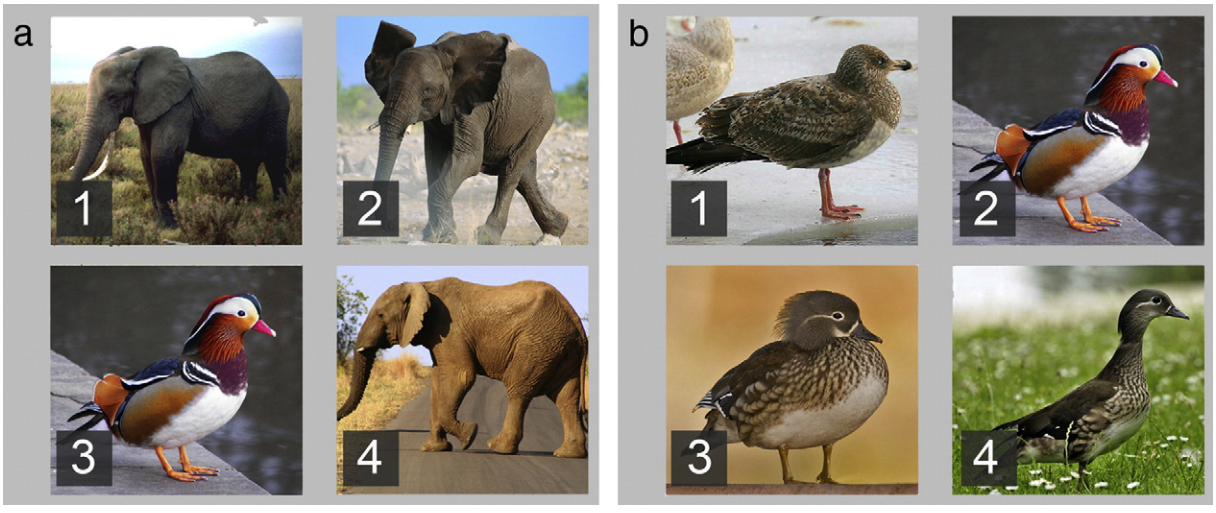


Fig. 1 – (a) Three elephants and a bird. The bird (#3) differs from the elephants in almost every aspect, making categorization easy and direct. (b) Three ducks and a seagull (#1) where naïve observers may erroneously think that #2 is the odd object due to its salient difference in colorfulness. When irrelevant features overshadow more relevant ones, inappropriate categorization can result due to the lack of prior knowledge directing attention to relevant features (such as beak shape).

inference (Erickson et al., 2005). While the first process encourages between category comparisons, the second encourages within category comparisons. These two processes may than highlight common and distinctive features that are expected to be more important for categorization. That is, learning from exemplars comparison may enable gaining useful knowledge on which features are more relevant for categorization.

In the current report we will examine *object comparison* as a cognitive mechanism enabling rapid learning of categorization principles driven by contextual constraints. Although similar ideas have already been presented (e.g. Markman and Gentner, 1993; Namy and Gentner, 2002), the theoretical limitations of learning by comparison have not been tested systematically until recently (Hammer et al., 2005, 2007, in press, submitted for publication). We now analyze further the theoretical attributes and limitation of category learning by comparison. We also provide additional analysis and interpretation of findings from human behavior studies, demonstrating the possible effect of these theoretical limitations on the strategies that people implement when learning new categories.

1.1. *"Birds of a feather flock together" — category learning by comparison*

Recent studies, in the fields of machine learning and human cognition, stressed the importance of object comparison for categorization. It was demonstrated that providing a clustering algorithm with only a small set of exemplars identified to be of the same class (denoted as a *chunklet* by Shental et al., 2004) is sufficient for improving categorization performance by artificial classifiers. This improvement is achieved mainly by re-evaluating the relevance of different object feature-dimensions to meet the constraints imposed by the presented categorical relations between the training examples (Bar-Hilel et al., 2003; Bilenko et al., 2004; Shental et al., 2004; Xing et al., 2002).

Similarly, when human subjects (including young children) are asked to compare a few exemplars identified to be from the same category, they are able to identify the features that are most important for categorization (Goldstone and Medin, 1994; Kurtz and Boukrina, 2004; Markman and Gentner, 1993; Namy and Gentner, 2002; Oakes and Ribar, 2005; Spalding and Ross, 1994). These studies demonstrate the importance of exemplar comparison for later similarity judgment and categorization. Nevertheless, these studies did not attempt to systematically assess the strengths and limitations of category learning by comparison. Moreover, they mainly focused on learning scenarios in which the learner was encouraged to look for similarities between compared exemplars. We will demonstrate here that learning by comparison is a complex process in which, for example, looking for differences may sometimes be more rewording than looking for similarities.

We now address the subject of category learning by comparison, presenting theoretical limitations of the comparison process and its usability by humans. We approach this question by discriminating between two types of comparison processes — comparison of exemplars identified to be from the same category (*Same-Class Exemplars*; also called *Positive Equivalence Constraints*), and comparison of exemplars identified to be from two different categories (*Different-Class Exemplars*; also called *Negative Equivalence Constraints*). We suggest that these comparison processes are complementary, and that they differ in their usability. Specifically we present evidence suggesting that learning by comparing same-class exemplars is expected to be quite independent of the guidance of an "expert supervisor". On the other hand, learning by comparing different-class exemplars requires further intervention in order to be both informative and effective. We propose that the usability of these comparison processes may be an essential element in shaping human conceptual knowledge.

As we have seen in Fig. 1b, categorizing by global similarity does not always provide proper categorization (e.g. discriminating between seagulls and ducks). In this case, comparing a few exemplars, for which the categorical relations are available, might be very useful for rescaling the importance of different feature-dimensions and thus reshaping the categorization rule. For example, it is sufficient to know that in Fig. 1b Animals 2 and 3 are of the same kind in order to conclude that the salient color dimension is not important for categorization (since two animals from the same category may differ dramatically in their color). Knowing that Animals 2 and 4 are from the same kind is even more informative, since now we can exclude both color and body weight from being relevant for categorization. As can be seen, same-class exemplars are quite useful for excluding salient *irrelevant* dimensions, but they are not necessarily sufficient for directly identifying *relevant*, less salient, dimensions. Yet when informed that Animals 1 and 3 are *not* of the same kind, (although highly similar in their color and global shape), we can conclude that finer differences in other features, such as differences in head shape, are more relevant for categorizing these animals correctly. As we can see, knowing the categorical relation between a few exemplars can be quite useful even when the objects' names (categories labels) are not used.

We further claim that category learning by comparison is embedded in many everyday life scenarios, as well as in many experimental category learning tasks. For example, when a parent points two animals in the presence of his child and says, "You see, these are both ducks", the child can conclude that the two are of the same kind. When the parent points two animals in the presence of his child and says, "You see, this one is a duck and that one is a seagull" the child can conclude that the two are from different kinds. This way the child can learn about features that are common to the same category members or features that are important for discriminating members of different categories. Comparison process can take place also when no labels are presented — each time a learner makes a same/different decision and receive a feedback for this decision, he can retrieve the actual categorical relation from the provided feedback. Furthermore, category learning by comparison does not necessarily require any direct supervision — simple observation on the way objects in the world behave and interact may provide clues for their categorical relation. These observations may enable the learning of a categorization rule, which can be later generalized to other objects.

In the next section we discuss in detail the differences between learning from same-class and different-class exemplars. We will start with an illustration of the qualitative difference in the usability of same-class exemplars and different-class exemplars for category learning, and we will suggest that this may require two different processes in order to optimize the learning from both comparison types. We will also provide a

quantitative analysis showing that same-class exemplars are more often useful for category learning than different-class exemplars; (a more detailed and formal analysis is provided in the Appendixes).

1.2. *Underlying differences between comparing Same-Class and Different-Class Exemplars*

We present theoretical limitations of category learning by comparison. In particular we demonstrate that, although the comparison of either same-class or different-class exemplars can be used for category learning, these two processes differ in their usability as follows:

1. As the distance (in a multidimensional feature space) between the compared exemplars increases, the comparison process becomes more informative for same-class exemplars and less informative for different-class exemplars. We measure the distance between exemplars in two ways: (a) The number of dimensions in which two compared exemplars are significantly separated (simplifying the representation to binary dimensions and using the L1, city block, metric). This indicates the number of relevant or irrelevant dimensions. In Appendix A we suggest a means for quantifying the information content of same-class exemplars comparison and different-class exemplars comparison when referring to the object space as to a hypercube. We show that as the number of irrelevant dimensions increases, comparing same-class exemplars will become more informative, while comparing different-class exemplars will become less informative. (b) The Euclidean distance between exemplars. This measure may be used to estimate the information provided by the comparison process regarding a specific dimension. Instead of referring to dimensional relevance as a dichotomy, the Euclidean distance between a pair of exemplars on each specific dimension can provide a finer indication for its information content regarding the possible relevance of this dimension. Furthermore, the Euclidean distance between the compared exemplars can provide an indication for the information content of the compared exemplar pair when there are dimensions which cannot be considered as independent for the categorization process. In Appendix B we use the relatively simple case of a binary linear classifier as a means for demonstrating the relation between the Euclidean distance between the compared exemplars, and the information content of the comparison process. We show that as the Euclidean distance between the compared exemplars increases, comparing same-class exemplars will be more informative, while comparing different-class exemplars will be less informative.
2. The property of belonging to the same-class is transitive but the property of belonging to different-classes is not. This differentiates the usability of same-class and different-class exemplars for the purpose of packing together objects into clusters. In Appendix C we quantify the contribution of transitivity by quantifying the information content of same-class exemplars and different-class exemplars in graph partitioning. We suggest that transitivity makes same-class exemplars comparison much more efficient then different-class exemplars comparison even when not considering any metrical assumptions about the structure of the object space.

Taken together, these differences lead us to expect that same-class exemplars will be much more informative for category learning than different-class exemplars. Nevertheless, we describe next the possible contributions of different-class exemplars for category learning.

1.2.1. *Relation between inter-exemplar distances and their usability*

As mentioned above, as the distance between compared exemplars increases, the information provided by the comparison of same-class exemplars also increases, while the information available from different-class exemplars decreases. The information content of a comparison between paired exemplars may be quantified as the reduction in volume of the version space, i.e. the space of hypotheses consistent with the constraints that have been previously seen in the learning process. Since determining a general measure and analyzing the structure of version space is difficult, we focus here on some relatively simple cases.

Fig. 2 provides some intuition for the above statement, using simplified scenarios analogous to those presented in Fig. 1. Fig. 2a represents a condition reminiscent of Fig. 1a, where the two categories are well separated in all dimensions. In this case unsupervised classifiers are expected to categorize the data points correctly as illustrated in Fig. 2c, which shows classification by two simple models: Perceptron – represented by blue dashed line and Gaussian Mixture model – represented by red dashed ellipses. Here, unsupervised classifiers will be successful simply because the true categorical assignment corresponds well with the overall distance between objects in the multidimensional feature object space. Fig. 2b represents a condition reminiscent of Fig. 1b, with a smaller distance in the relevant dimension of head shape than in the within-category distance in the irrelevant dimension (color). Here, our unsupervised classifiers are expected to fail, as illustrated in Fig. 2d — instead of categorizing the data points as ducks and seagulls, the data points will be categorized according to their color, which is not relevant for categorizing the targeted animals.

Comparison of paired exemplars might be quite useful in conditions where the global distance (dissimilarity) between objects is not a good predictor for their proper categorical assignment. Fig. 2e illustrates a condition in which our classifiers are provided with an indication (pink arrow) that two very close objects are in fact not from the same category. The comparison of these different-class exemplars will be quite informative for updating our Perceptron — this piece of information acts as a constraint (*Negative Equivalence Constraint*), forcing the classifier to relocate and change the orientation of the category borderline between the data points. Since the distance between the different-class exemplars in the relevant dimension is larger than the distance in the irrelevant dimension, the borderline is now updated to be orthogonal to the relevant dimension as required. Since the distance between the different-class exemplars in the relevant dimension is still quite small, this comparison makes it possible to localize the borderline quite accurately. This example shows how different-class exemplars might be quite useful in highlighting relevant dimensions that would have been otherwise disregarded.

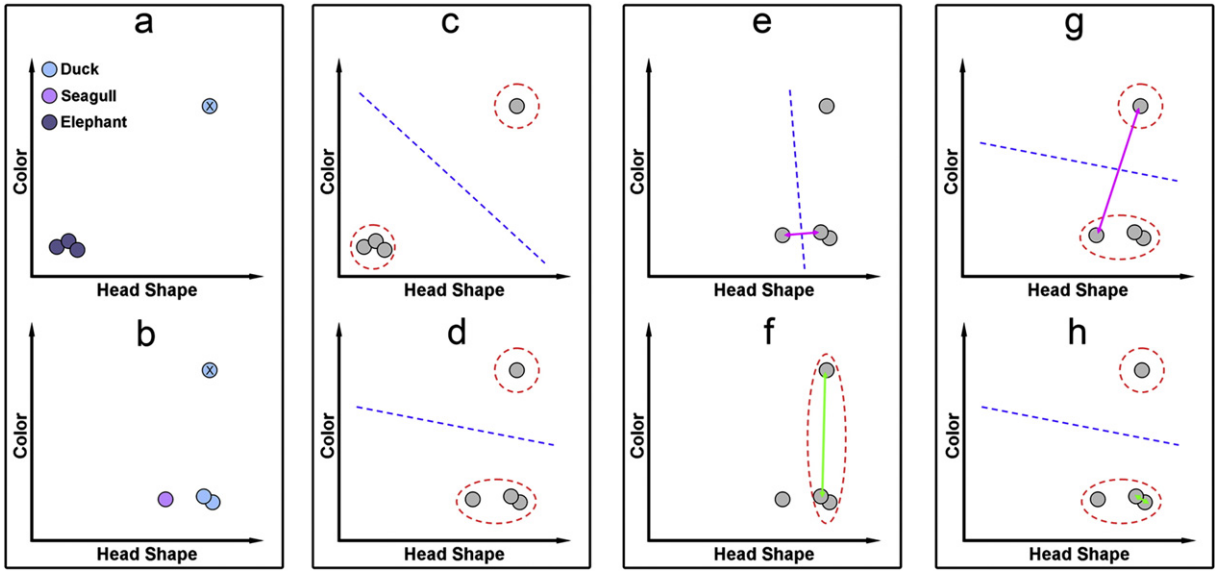


Fig. 2 – Illustration of expected performance of common classifiers using a simplified representation for animal categories in accordance with Fig. 1. (a, b) Representation of the actual categories as presented in Fig. 1: light blue circles represent ducks (the colored duck marked with a cross), light purple represents the seagull; dark purple the elephants. (c, d) Representation of expected categorization executed by unsupervised Perceptron (dashed blue line) and Expectation-Maximization classifier (dashed red ellipses) in the two different scenarios illustrated in panel a, and panel b, respectively. Examples for the contribution of informative different-class exemplars, indicated by pink arrow (e), and same-class exemplars, indicated by green arrow (f), in improving performance. Examples for poorly informative different-class exemplars (g) and same-class exemplars (h).

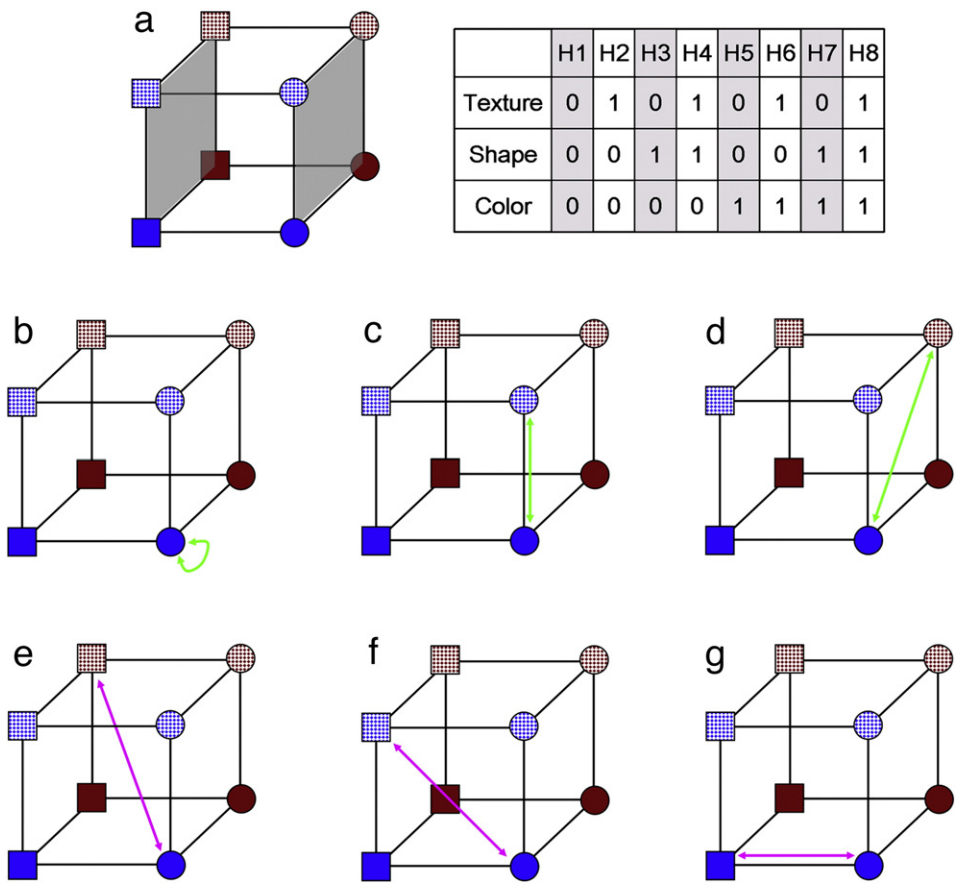
Same-class exemplars are useful for category learning in a different way. Fig. 2f shows a condition in which the same-class exemplars indicates that two objects, which differ dramatically in their color, can still be from the same category. This constraint (*Positive Equivalence Constraint*) suggests that the color dimension is not relevant for categorization, despite the fact that objects can be easily separated along this dimension. As the number of irrelevant dimensions increases, paired same-class exemplars can provide more information by indicating that two exemplars, which are distant in more than one dimension, are nevertheless from the same category. If our classifier represents categories by calculating a Gaussian Mixture, providing it with such same-class exemplars can also assist it in updating the center of the category from which these two exemplars are taken, by simply calculating the mean coordinates of the specified set of exemplars (Hammer et al., 2007; Shental et al., 2004).

Same-class exemplars and different-class exemplars are not always informative. In Fig. 2g we show an example of poorly-informative different-class exemplars: Knowing that two distant objects are not from the same category is not useful for improving performance as compared to unsupervised categorization (Fig. 2d). Such different-class exemplars are not informative in two respects: First, since the distance between the different-class exemplars, in the relevant dimension, is smaller than it is in the irrelevant dimension, the orientation of the borderline will not be updated sufficiently. In this case the large distance in the irrelevant dimension overshadows the smaller distance in the relevant one. Second, since the overall Euclidean distance between the two exemplars is quite large, there is a large uncertainty concerning the actual location of the borderline. Furthermore, when the number of categories is larger than two, there is a larger probability that a third, hidden, category may be present between distant different-class exemplars.

Fig. 2h illustrates poorly-informative same-class exemplars. Being informed that two nearby objects are from the same category does not provide much information since it does not capture the possible variance permitted within this category, or within which directions such variance is permitted. Such cases are conceptually similar to a case when we are informed that an object is from the same category as itself.

The illustrations in Fig. 2 suggest that same-class exemplars indeed differ from different-class exemplars in the way they can be used. But the two types of comparison also differ quantitatively in their information content, so that same-class exemplars will be more often informative for learning. This idea is illustrated in Fig. 3. Fig. 3a demonstrates a case of exemplars defined by 3 dimensions (color, texture and shape), with the single relevant dimension for categorization being shape. Thus the two target categories differ in their values only in this dimension — one category is constrained to the subspace (surface) of square shapes, while the other is constrained to the circle subspace. The two other dimensions cannot be taken into account as relevant for categorization since the within category variation is as large as the between category variation in these dimensions.

Fig. 3b provides an example of poorly informative "same-class exemplars" — the indication that an object is in the same



	H1	H2	H3	H4	H5	H6	H7	H8
Texture	0	1	0	1	0	1	0	1
Shape	0	0	1	1	0	0	1	1
Color	0	0	0	0	1	1	1	1

Fig. 3 – (a) A simple example of three dimensional space with binary dimensions — color (red vs. blue), shape (square vs. circle), and texture (full vs. doted). The gray surfaces indicate the two target categories (differing only in shape). The table on the right illustrates the hypotheses space (eight possible hypotheses) — all possible combinations of relevant dimensions (marked as "1") and irrelevant dimensions (marked as "0"). (b–d) Same-class exemplars differing in 0, 1 or 2 dimensions, going from low to high information content. (e–g) Different-class exemplars differing in 3, 2 or 1 dimension, going from low to high information content.

class as itself. Of course this indication changes nothing about what we know of the relevance or irrelevance for categorization of the three given dimensions. It excludes none of the hypotheses described in the Hypotheses table, $\log_2 8/8=0$ bit. Similarly, two different same-class exemplars that are "sufficiently close" can be treated in the same way (leaving the question of what determines "sufficiently close" open). The same-class exemplar pair in Fig. 3c provides 1 bit of information since it indicates that two objects from the same category may differ in color. It excludes all the hypotheses in which color is relevant (H5, H6, H7 H8), $-\log_2 4/8=1$ bit. The same-class exemplar pair in Fig. 3d provides 2 bits of information since it indicate that two objects from the same category may differ both in color and texture. It excludes all the hypotheses in which color, texture or both, are relevant (H2, H4, H5, H6, H7 and H8), $-\log_2 2/8=2$ bits. The latter same-class exemplar pair reduces the uncertainty in partitioning the object space quite dramatically since it leaves only the possibility that the shape dimension is the relevant one — despite the fact that the constraint only provides indirect evidence for this claim.

Fig. 3e provides an example for poorly informative different-class exemplars — it indicates that two objects that differ in all three dimensions are not from the same category. This indication changes little about what we know of the relevance or irrelevance of the three dimensions, excluding only the possibility that non of the dimensions is relevant (H1), $\log_2 7/8=0.19$ bit. Although the different-class exemplar pair in Fig. 3f differs in only two dimensions, it still provides little information — we may guess that shape is the relevant dimension for categorization, but we may just as well guess that the texture dimension is the relevant one. In fact, when not knowing the number of possible relevant dimensions, these different-class exemplars do not even exclude the possibility that the color dimension is relevant as well (excluding only H1 and H5), $\log_2 6/8=0.41$ bit. Note that these two examples of different-class exemplar pairs demonstrate that there are more possibilities of poorly informative different-class exemplar pairs than of poorly informative same-class exemplar pairs. Finally, the different-class exemplar pair in Fig. 3g exemplifies informative different-class exemplar pairings, directly indicating that shape is relevant — since this is the only dimension differentiating the two exemplars identified to be from two different categories (excluding H1, H2, H5 and H6), $\log_2 4/8=1$ bit.

In Appendix A we provide a formal proof for the above statement showing that informative different-class exemplars, and poorly informative same-class exemplars, are both relatively rare. Furthermore, the information content of a typical same-class exemplar pair increases as the number of dimensions in which they differ increases (using a city block metric for estimating the information content in reducing the hypotheses space). The information content of different-class exemplar pairs will behave in the opposite way. As a result, when the total number of dimensions increases, particularly with respect to the number of relevant dimensions, so does the information provided by a randomly selected same-class exemplar pair, while the information provided by a randomly selected different-class exemplar pair decreases.

In Appendix B we describe a method for estimating the information content of same-class and different-class exemplar pairs as a function of the Euclidean distance between the paired exemplars. This enables a non-discrete computation of the information content of same-class and different-class exemplar pairs. It also provides a measure for estimating the information content of same-class and different-class exemplar pairs when dimensions cannot be considered as independent, such as in information-integration category learning tasks (Ashby and Ell, 2001). Note that this analysis is based on the Euclidean distance, and in its current formulation it only shows qualitative differences between same-class vs. different-class exemplars. Specifically, it shows that as the distance between paired same-class exemplars increases, their information content increases as well, while for different-class exemplar pairs the relation between information and distance is reversed.

Together, the two analysis provided Appendix A (L1 metric) and Appendix B (L2 metric) demonstrate a clear difference for the usability of same-class vs. the usability of different-class exemplars comparison in the context of the two metrics known to be most relevant for explaining human similarity judgment (Shepard, 1987) and categorization strategies (Ashby and Ell, 2001).

1.2.2. *Graph partitioning and transitivity*

The advantage of same-class exemplars over different-class exemplars does not result only from the different relationship between distance and information for the two types of comparison processes. In this section we show that using same-class exemplars is much more beneficial for packing together objects into clusters than the use of different-class exemplars, in a more general case, even when generalization according to distance in specified dimensions is not relevant. This property of same-class exemplars can be helpful in summing together pieces of information when constructing an internal representation of newly-learned categories. For example, transitivity may facilitate formation of a category prototype by averaging common relevant features of packed-together objects. Similarly, a small subset of packed-together objects can be used as a set of exemplars representing the category.

Whenever the number of categories is larger than two, same-class exemplars will be much more useful for graph partitioning. This results mainly from the fact that the property of belonging to the same-class is transitive, but belonging to different-classes is not: For example, being informed that objects A and B are from the same category and that B and C are from the same category is sufficient for concluding that all three objects are from the same category. On the other hand, knowing that D and E are from different categories, and that E and F are from different categories, does not tell us much about the categorical relation between D and F. As the number of objects and categories increases, the contribution of transitivity in "packing objects" into categories also increases. Note that the above statement concerning transitivity of same-class exemplars is true only when assuming that each object belongs only to one category. In real life scenarios this is not always the case. For example, an animal can be classified both as a dog and as a mammal. In this way, a dog will not be perceived to be from the same category as a cat in one comparison context (dogs vs. cats), but it can be perceived to be from the same category as another cat in another comparison context (e.g. mammals vs. reptiles). For the purposes of our discussion here, we assume that the comparison context is identical for all referred objects and that there is no overlapping between categories.

Fig. 4 provides an illustration for the differences between same and different-class exemplars in graph partitioning. To keep the illustration simple, we provide three labeled points (although the formal theoretical basis provided in Appendix C is for the harder, more general, case where no labels are provided. In that case same-class exemplar pairs have an advantage also when the number of categories is only two). As can be seen, indicating a small random set of same-class pairs is sufficient to significantly reduce graph uncertainty. This is not the case for different-class pairs. The hypotheses space even in such simple example is quite large — all the combinations for coloring the six gray points using three colors which is $3^6=729$ different possibilities. Assigning the right color to all the gray points is easy when provided with same-class pairs, and due to transitivity it does not require direct connection to either one of the labeled points. Assigning the right color to the gray points is hard when provided only with different-class pairs. In this case, in order to know the color of a point it needs to be directly connected to *all* the labeled points from the categories to which it is not related.

In order to provide a formal basis for the theoretical difference between different-class and same-class exemplars for graph partitioning, we analyze the problem in two ways. First, in Appendix C.1, we show a qualitative difference between the two comparison processes — clustering with different-class exemplars is related to the problem of finding the maximal cut in a graph, which is known to be very hard (NP-complete). In contrast, clustering with same-class exemplars is related to the analogous problem of finding the minimal cut in a graph, for which efficient polynomial algorithms are known.

Secondly, in Appendix C.2, we define the notion of information for both types of comparison processes, and obtain a lower bound on the difference in information content between same-class exemplars and different-class exemplars. This provides a quantitative measure for comparing the usability of the two comparison process. Specifically, the information content of different-class exemplars is inversely related to the number of different graph colorings for the graph defined by the negative constraints. Computing this number is very hard (again, an NP-hard problem), with no known approximations (Khanna et al., 2000). More importantly, for random graphs it is known that the number of solutions tends to be very large whenever there is a solution to

the coloring problem. In contrast, the number of colorings for a graph defined by same-class exemplars is rather small due to transitivity. Thus, the difference in information content between same-class exemplars and different-class exemplars is typically very large.

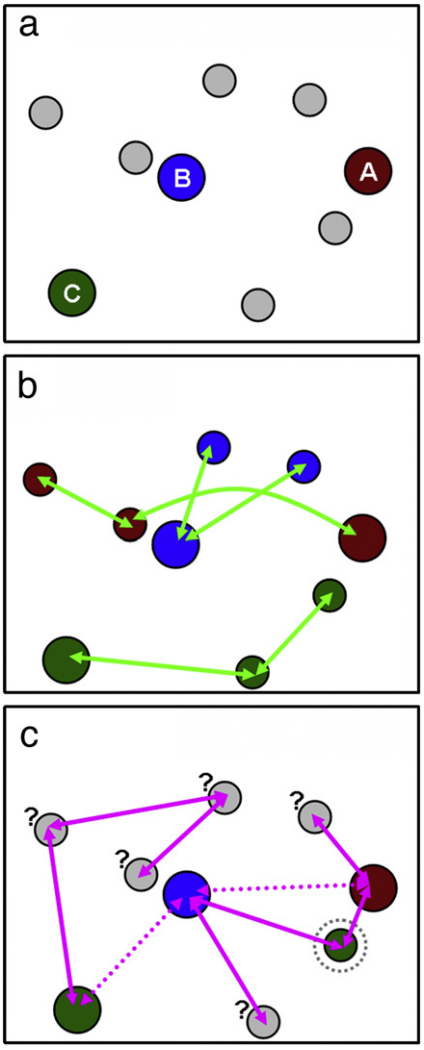


Fig. 4 – A simple illustration for the advantage of same-class exemplar pairs over different-class pairs in graph partitioning. Note that spatial proximity is not a relevant factor in this illustration (a) A graph in need of partitioning. The three labeled (colored) points A, B, C are data points each representing an exemplar from a different category. The categorical identities of the gray data points are not given. The task is to color correctly the gray points while using the minimal number of constraints, thus reducing the task complexity. (b) A small set of six same-class indications (green arrows) is sufficient for correctly coloring the graph. Note that due to transitivity, there are many more combinations that enable the retrieval of this result with six same-class indications. (c) A small set of seven different-class indications (pink arrows), in addition to the different-class indications provided by the labels (dotted pink arrows) are not sufficient to significantly reduce the uncertainty in the graph. In this case, when only a few indications are provided, we can decisively color only points connected directly to labeled points from all the other categories (e.g. the green point in the dotted circle).

1.3. *Research hypothesis*

We argued and demonstrated above that same-class exemplar pairs qualitatively differ from different-class exemplar pairs in their usability for category learning. Furthermore, in most natural scenarios same-class exemplar pairs are expected to be significantly more informative than different-class exemplar pairs. For this reason, we expect that although both comparison types can contribute to category learning, they may in fact involve different learning mechanisms.

We hypothesize that people use same-class and different-class exemplars in different ways, and with different proficiency levels. In particular, since same-class exemplars are expected to be more often informative in everyday life scenarios, people will develop superior abilities for using same-class exemplars than for using different-class exemplars, even in cases where the objective amount of information provided is identical. In order to test this hypothesis, we conducted an experiment in which we measured the usability of a small set (three pairs) of same-class vs. different-class exemplars in a simple rule-based category learning task. In these category learning tasks, the categorization rule could be learned by either same-class or different-class exemplars (see Experimental procedures). Using novel stimuli also enabled us to exclude the intervention of other factors such as prior domain specific knowledge and features saliency effects. This task also simulates best the case of categorization according to the L1 metric analyzed above.

Furthermore, we manipulated the "*Level of Supervision*" — the amount of intervention and instructions provided by the experimenter during the task, and tested its effect on participants' performance. We discriminate between three conditions: In the first condition same-class and different-class exemplars were *randomly* selected. This condition is expected to simulate the more common everyday life scenario when no expert supervisor provides the learner with carefully selected exemplar pairs that maximize the information. According to the above theoretical analysis, we expected that performance with a small random set of same-class exemplars will be significantly better than performance with a small random set of different-class exemplars. This is to be expected because in the random *Same-Class Exemplar* condition participants are objectively provided with more information than in the random *Different-Class Exemplar* condition. In fact, there is high probability that a random set of three same-class exemplar pairs is sufficient for providing all the information needed for perfect performance in the performed categorization task (see Experimental procedures and Appendix A for details).

In the second condition, the sets of pairs of same-class and different-class exemplars were selected deliberately so that they contained all the information needed for learning the categorization rule. This was done by using same-class exemplar pairs as the one described in Fig. 3c, and different-class exemplar pairs as the one described in Fig. 3g. Here we expected that most participants will again perform quite well when same-class exemplars are introduced, but many will fail to use even these *Informative* different-class exemplars since they are less trained with the proper strategy for using this comparison type.

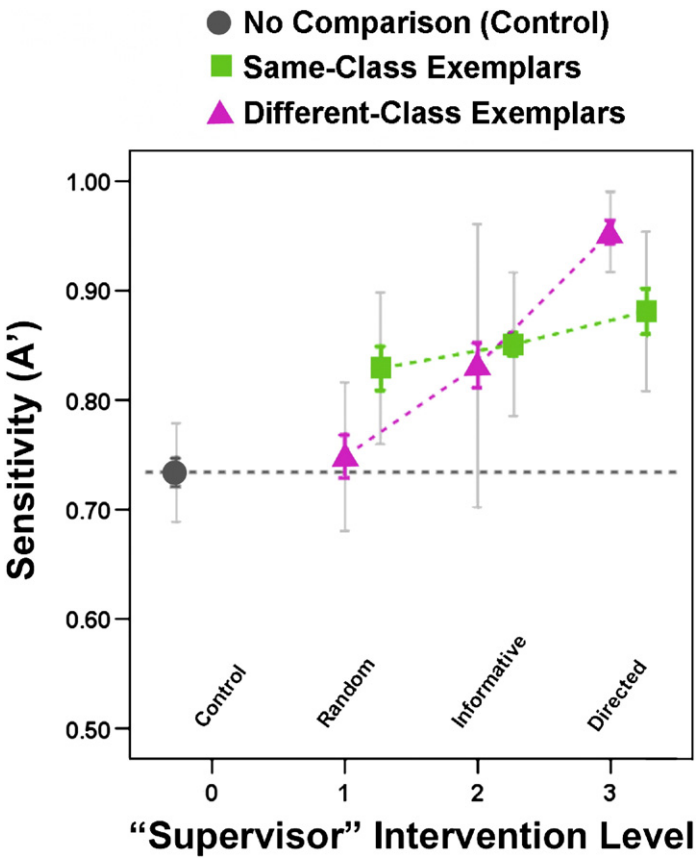


Fig. 5 – Participant sensitivity (A') as a function of *Comparison Type* and *Level of Supervision*. Dark gray (circle) data-point and dashed line — participant mean performance in the *Control*, no comparison, condition. Green (square) data-points — mean performance in the learning from *Same-Class-Exemplar* condition. Pink (triangle) data-points — mean performance in the learning from *Different-Class-Exemplar* condition. Thick colored error bars represent standard errors. Light-gray error bars represent standard deviations; (note that the number of participants for the *Random* and *Informative plus Directions* conditions was 12, while for the *Informative* condition it was 40).

In the third condition participants were not only provided with similarly, and sufficiently, *Informative* same-class and different-class exemplar pairs as in Condition 2, but at the beginning of the experiment, the experimenter also directed the participants, coaching them about the strategy for learning from the provided same-class or the different-class exemplar pairs. The directions were simple and straight forward: we asked participants to look for all the features in which the paired exemplars are different, as well as those in which they are similar. We further directed participants to integrate the information provided by the three different pairs of exemplars used for the rule learning. We expected that such *Directions* will be significantly more beneficial for improving performance in the different-class exemplars conditions, since most participants are expected to be already quite experienced in the use of same-class exemplars from their everyday life experience. In order to eliminate any possible advantage in these *Same-Class Exemplar* conditions, with respect to their correspondent *Different-Class Exemplar* conditions, we also avoided the use of transitive same-class exemplar pairs.

As a baseline, we tested a group of participants in similar categorization tasks but with no supervision at all (participants were not trained with same-class nor different-class exemplar pairs).

2. Behavioral results

Our main hypothesis was that there will be no significant effect of the *Level of Supervision* on performance in the *Same-*

Table 1 – Performance mean and standard deviation (SD) in the different conditions

Comparison type/ level of supervision	Random	Informative	Informative plus directions
Same-Class Exemplar	M=0.83 SD=0.07 *n*=12	M=0.85 SD=007 *n*=40	M=0.88 SD=0.07 *n*=12
Different-Class Exemplar	M=0.75 SD=0.07 *n*=12	M=0.83 SD=0.13 *n*=40	M=0.95 SD=0.04 *n*=12
Control	M=0.73 SD=0.05 *n*=12		

Class Exemplar condition, but that there will be such an effect in the *Different-Class Exemplar* condition. We measured participant ability to learn the new categories by using the non-parametric sensitivity measure A′ (Grier, 1971), calculated from participants' Hits (correctly identifying a test exemplar as a member of the target category) and False-Alarms (incorrectly identifying a test exemplar as a member of the target category).

A battery of independent sample *t*-tests show that categorization performance in all conditions of learning by comparison was significantly better than performance in the *Control*, no comparison condition ($p<0.02$ for all learning conditions), except for the *Random Different-Class Exemplar* condition $t(22)=0.62$, $p=0.54$ (see Fig. 5 and Table 1). This result suggests that comparison is useful, as long as the compared exemplars differ in an informative manner. Nevertheless, the following analysis shows that the process of learning by comparison is more complex.

In order to evaluate the effect of *Level of Supervision* on sensitivity, we first calculated the nonparametric Spearman correlation between level of supervision (ordinal scale: 1 – *Random* exemplar pairs; 2 – *Informative* exemplar pairs; 3 – *Informative* exemplar pairs *plus Directions*) and A′ score, for each experimental condition separately. We found no significant correlation between *Level of Supervision* and A′ in the *Same-Class Exemplar* condition, $p(64)=0.19$, $p=0.14$, but the correlation between *Level of Supervision* and A′ in the *Different-Class Exemplar* condition was highly significant, $p(64)=0.60$, $p<0.0001$. This result supports our hypothesis that learning from different-class exemplars is a process highly dependant on the learning conditions while learning from same-class exemplars is a more robust process, less affected by the learning conditions.

A two-way ANOVA with A′ as the dependent variable, level of supervision (*Random*, *Informative*, and *Informative plus Directions*), and comparison type (*Same-Class Exemplars* vs. *Different-Class Exemplars*) as between-subject factors, revealed no main effect of comparison type, $F(2, 122)=0.62$, but a significant effect of level of supervision, $F(2, 122)=12.41$, $p<0.0001$, $\eta_p^2=0.17$. Importantly, there was a significant interaction between comparison type and level of supervision, $F(2, 122)=4.42$, $p<0.02$, $\eta_p^2=0.07$. Post-Hoc independent sample *t*-tests on the effect of comparison type within each level of supervision showed that in the *Random* exemplars condition A′ score was significantly higher when participants were trained with *Random* same-class exemplars (M=0.83; SD=0.07) than when they were trained with *Random* different-class exemplars (M=0.75; SD=0.07), $t(22)=2.87$, $p<0.01$. There was no such comparison type effect in the *Informative* exemplars condition, $t(78)=0.85$, $p=0.40$. Surprisingly, performance in the *Informative plus Directions* condition was better when participants were trained with different-class exemplars (M=0.95; SD=0.04) than when they were trained with same-class exemplars (M=0.88; SD=0.07), $t(22)=-3.08$, $p<0.005$.

Moreover, one-way ANOVAs showed a significant effect for level of supervision on the A′ score only in the *Different-Class Exemplar* condition $F(2, 61)=10.98$, $p<0.001$, but not in the *Same-Class Exemplar* condition $F(2, 61)=1.81$, $p=0.17$. Post-Hoc Scheffe tests showed that in the *Different-Class Exemplar* condition, performance with *Random* exemplars (M=0.75; SD=0.07) and with *Informative* exemplars (M=0.83; SD=0.13) were both significantly lower than performance in the *Informative plus Directions* condition (M=0.95; SD=0.04) ($p<0.005$ in both cases). There was no significant difference in the A′ score between *Random* exemplars and *Informative* exemplars, $p=0.07$. Nevertheless this difference was significant when using the Post-Hoc LSD test, $p<0.05$.

Same-Class and *Different-Class Exemplar* conditions did not only differ in the mean level of performance, but they also significantly differed in their variances. Levene's test for homogeneity of variance showed a significant effect for *Level of Supervision* on performance variance only in the *Different-Class Exemplar* condition $F(2, 61)=6.72$, $p<0.005$, but not in the *Same-Class Exemplar* condition $F(2, 61)=0.27$, $p=0.76$. Further tests showed that there was no significant difference in variance between the *Random Exemplar* conditions, $F(1, 22)=0.01$, $p=0.94$, but there was such an effect for the *Informative Exemplar* conditions when the variance in the *Informative Different-Class Exemplar* condition (SD=0.13) was significantly higher than in the *Informative Same-Class Exemplar* condition (SD=0.066), $F(1, 78)=13.94$, $p<0.001$. In the *Informative plus Directions Same-Class Exemplar* (SD=0.07) and *Different-Class Exemplar* (SD=0.04) conditions the pattern was reverse $F(1, 22)=5.04$, $p<0.05$ (see also the standard deviations in Fig. 5).

This latter analysis suggests that when presented with same class exemplars, participants show quite similar performance levels and they are quite capable of learning the categorization rule even with little intervention by a supervisor. On the other hand, when presented with *Informative* different-class exemplars, participants demonstrate a wide range of abilities. When participants are also provided with directions specifying the learning strategy, this has no effect on performance variance in the *Informative* (*plus Directions*) *Same-Class Exemplar* condition, but it narrows down the variance in the *Informative* (*plus Directions*) *Different-Class Exemplar* condition.

3. Discussion

3.1. Summary

We reported here in detail computational properties of category learning by comparison and their effect on human performance. We suggested that the process of learning by comparison should be treated as two separate processes: Learning from same-class exemplar comparison vs. learning from different-class exemplar comparison. These processes differ qualitatively from each other: As the distance between same-class exemplars increases, the information content of their comparison also increases, while as the distance between different-class exemplars increases, their information content is reduced. Moreover, same-class exemplar comparison is transitive but different-class exemplar comparison is not. We further showed that the two learning processes also differ quantitatively, so that learning from same-class exemplars is expected to be more often informative than learning from different-class exemplars. In this respect, we suggest that though the two learning processes seem to be useful for the same goal, they in fact differ and should be treated as complementary. This may require two different strategies (algorithms) for maximizing performance when using each of the comparison types. We further suggested that the two processes may not evolve in a similar way in humans.

We propose that these differential properties of same-class and different-class exemplars may affect their usability, even

under conditions in which the two types of comparison can be used quite efficiently and to the same extent. In order to test this, we designed an experiment in which we tested people's abilities in executing (separately) the two learning processes in a rule-based category learning task — in which both comparison types can be similarly effective. We further tested the usability of same-class vs. different-class exemplar comparisons under different levels of supervision — i.e. the amount of intervention by a supervisor, (the experimenter), in the selection of the provided exemplars for the learning phase and the directions provided to the learner (participant).

We expected that learning from same-class exemplars would be less affected by such intervention than learning from different-class exemplars for two reasons: First, since even the comparison of arbitrary same-class exemplars is expected to be quite informative, a small set of randomly selected same-class exemplars is likely to provide most of the information needed for a specified categorization task. This is why we did not expect a significant difference between the *Random* vs. the *Informative Same-Class Exemplar* conditions. Since the comparison of *Random* different-class exemplars is not expected to be sufficiently informative, we expected that providing participants with *Informative* different-class exemplars might enhance their performance. Secondly, the availability of highly informative same-class exemplars in everyday life, together with poor availability of informative different-class exemplars, could drive people to adopt an appropriate strategy for learning from same-class exemplars, but not from different-class exemplars. This is why we expected that regarding performance in the *Informative Different-Class Exemplar plus Directions* condition, the directions that were provided would be much more effective in improving performance compared to the *Directed Same-Class Exemplar* condition. The reported findings strongly support our hypotheses.

3.2. *Implications*

The reported theoretical analysis, together with the behavioral findings, suggests that the differentiating properties of information building blocks may have a dramatic effect in shaping the most fundamental cognitive abilities of humans, and probably of other living species. Furthermore, the current findings suggest a means for predicting human category-learning limitations in different conditions, and perhaps possible ways for overcoming such limitations.

As we have shown, comparing same-class exemplar pairs is expected to be always quite useful, but even in a simplified category learning task they are not sufficient for perfecting performance. Same-class exemplar comparison is imperfect even when the objective conditions enable perfect performance. In reference for this statement, we recently executed a computer-simulation for testing similar conditions to those tested in the behavioral study reported here (Hammer et al., 2007). This simulation showed that a constrained EM (Expectation-Maximization) algorithm always performs almost perfectly in a categorization task, when it is trained with few paired same-class exemplars — even when these are randomly selected. The algorithm performance level when trained with few same-class exemplar pairs was $A' > 0.9$ in all tests, (and $A' = 1$ in most of them), suggesting that the objective amount of information needed for perfect performance was available. However, our human participants failed in achieving this level of performance when trained with same-class exemplars, although they were quite capable of achieving similar performance levels in the *Informative plus Directions Different-Class Exemplar* condition.

On the other hand, comparing different-class exemplar pairs is not expected to be useful in most everyday life scenarios. This statistical fact clearly emerges from the theoretical analysis provided here. This analysis suggests that in the absence of an expert supervisor that knowingly selects informative different-class exemplar pairs and "feeds" them to the learner, executing a "learning from different-class exemplars" process is expected to be of little value. Furthermore, even when informative different-class exemplars are selected by an "expert supervisor" (as was the case in the *Informative Different-Class Exemplar* condition) participants' performances differed dramatically, suggesting that different people execute different learning strategies when faced with informative different-class exemplars; (for further analysis of the distribution pattern in such conditions, see Hammer et al., in press). We suggest that the later results from the former — i.e. the fact that informative different-class exemplars are rare results in a poor proficiency level, observed in many participants, for executing correctly the process of "learning from different-class exemplars".

Nevertheless, comparing informative different-class exemplar pairs can sometimes be quite rewarding: When executed correctly, as occurred in the *Informative plus Directions Different-Class Exemplar* condition, this comparison process enables superior performance than that achieved by executing same-class exemplar comparison. We suggest that this difference between same-class and different-class exemplar comparisons is an outcome of the fact that same-class exemplar comparison directly indicates only which dimensions are irrelevant, while different-class exemplar comparison may directly indicate the relevant ones. Since eventually generalization of a categorization rule requires identification of the relevant dimensions, even perfectly identifying all the irrelevant dimensions will not be sufficient for perfecting performance. As demonstrated in Fig. 1b, knowing that two distinct birds are from the same category is at most sufficient for knowing that the salient features that distinguish the two are not important for determining if two birds are from the same kind or not. Nevertheless it is not sufficient for identifying the relevant features for perfecting bird categorization.

In fact, the superiority of the constrained EM algorithm, when tested with same-class exemplars, emerges from its ability to execute Principle Component Analysis (PCA) as a first step. This provides the algorithm with a representation for all the possible relevant dimensions in which there is informative variance. Later, by executing the "learning from same-class exemplars" step, the algorithm updates its covariance matrix to fit the constraints imposed by the same-class exemplars. This could be described as if the EM algorithm "identifies" the irrelevant dimensions which then enable it to identify also the complementary set of relevant dimensions. Humans do not, and apparently cannot, behave similarly: We are limited and driven by the physical properties of objects in the world and their representation in our perceptual systems. The salience of object features interact with our requirements and it may happen that physically salient features will overwhelm our judgment by overshadowing less salient, but potentially more

relevant, features — i.e. features that better predict the behavior or core properties of a perceived object.

It might be possible that people execute a kind of PCA — that is, they can learn without any supervision the variability within the different feature-dimensions in the world. But even if they do, it seems that such learning in humans is still limited by the prior biases of our perceptual system. One can further claim that such perceptual biases, if they exist, are also driven by evolutionary constraints shaping our perceptual system, so that features that were more important for our survival have become perceptually more salient. But since feature-importance seems to be context dependent, such an evolutionary mechanism is likely to be quite limiting. Furthermore, it is not unlikely that changes in our everyday life demands exceed evolutionary changes in our perceptual system, perhaps forcing other brain areas to become involved in a creative way with facing these increased demands. Learning by comparison, and more specifically learning by comparing different-class exemplars, might be such a "higher" learning process. These processes becomes most valuable whenever there is little correspondence between the overall similarity between objects and their categorical identity — situations in which unsupervised categorization will fail to satisfy our needs.

Comparing different-class exemplars can be highly useful when the compared exemplars are selected carefully, but even then it is not always sufficient for perfecting performance (even when adult university students are tested). Nevertheless, as demonstrated, the process of category-learning from informative different-class exemplars can be triggered in adults by using simple directions (but see Hammer et al., submitted for publication for relevant findings with children). Also, the constrained EM algorithm we tested frequently failed when trained with informative different-class exemplars, suggesting that its architecture is not the appropriate one for using this source of information. Here, the flexibility of the human brain prevailed, being able to rapidly adopt an optimized strategy, as was demonstrated in the *Informative plus Directions Different-Class Exemplar* condition.

The current report suggests that the conditions for which supervised category-learning is executed should be considered with care, and that simple changes in the selection of the information building blocks, as well as the strategies implemented for using them, are extremely important in some conditions but not in others. Specifically we provide the intuition for the conditions in which different-class exemplar comparison will be of great value. We also suggest that when an expert supervisor is unavailable, it might be much more efficient to limit the learning process to learning by comparing same-class exemplars. These principles do not contradict previous category learning models (e.g. Anderson, 1991; Goldstone and Medin, 1994; Love et al., 2004; Nosofsky, 1986), and in fact we think they can be (and should be) naturally integrated into these models.

4. Experimental procedures

4.1. *Subjects*

104 students (mean age 24 ± 3.2, 60 female and 44 male) from the Institute of Life Sciences at the Hebrew University of Jerusalem

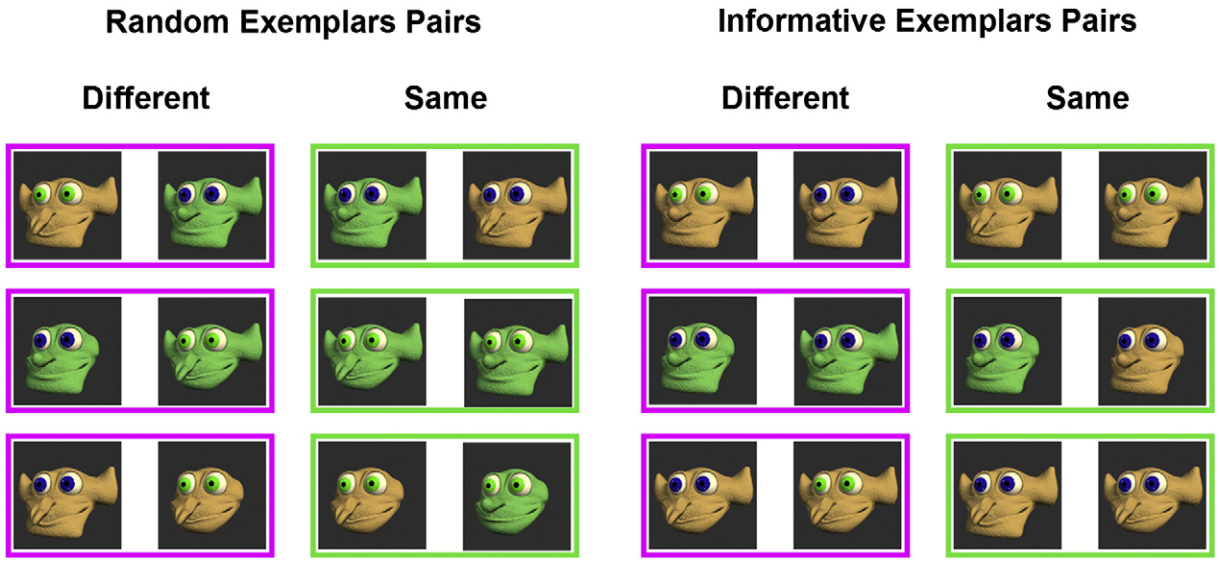


Fig. 6 – Examples of stimuli used in the experiment during the learning phase. On the left — triads of paired stimuli used in the *Random Different-Class Exemplar* (pink frames) and *Same-Class Exemplar* (green frames) conditions. On the right — triads of paired stimuli used in the *Informative Different-Class Exemplar* (pink frames) and *Same-Class Exemplar* (green frames) conditions. For all four conditions, the examples shown have eye color and ear shape as relevant dimensions. The random set of different-class exemplar pairs is not very helpful for identifying the relevant dimensions since each pair differs in more than one dimension, not all of which are relevant. The random set of same-class exemplars is quite useful for identifying the relevant dimensions since each pair of exemplars differs in more than one irrelevant dimension. The informative different-class exemplar pairs were selected so as to ensure that each pair would specify only one relevant dimension and the triad of pairs would specify all the relevant dimensions. The informative same-class exemplar pairs were selected so as to ensure that each pair would specify only one irrelevant dimension and the triad of pairs would specify all the irrelevant dimensions.

participated in the experiment. We obtained written consent from participants. Participants were randomly assigned to the different experimental conditions: 12 performed both the *Control* and the two *Random* conditions in a within-subject design (order of the *Same-* vs. *Different-Class Exemplar* conditions was counterbalanced and the *Control* condition was always first); 80 participants performed the *Informative* conditions in a between-subject design (40 performed the *Same-Class Exemplar* condition and 40 the *Different-Class Exemplar* condition), and 12 participants performed the *Informative plus Directions* condition in a within-subject design (with counterbalanced order).

4.2. *Materials*

Computer-generated pictures of "alien creature faces" were used as novel stimuli, as shown in Fig. 6. Each face was characterized by a unique combination of 5 potentially task-relevant dimensions: shape of chin, nose and ears, and color of skin and eyes. We designed 10 different sets of 32 stimuli each, such that for each set, all combinations of the 5 binary dimensions were presented in each of the 10 experimental trials. All sets were used in each experimental condition. Two or three (of the 5 possible) dimensions were selected as relevant for category definition on each trial, so that same-class exemplars had to have the same features (values) for all relevant dimensions and different-class exemplars had to differ in at least one of these.

4.3. *Procedure*

For each experimental trial, the task was to decide which stimuli (creatures) belong to the same category as a given standard. Thus, participants needed to learn by comparing either same-class exemplars, or different-class exemplars, which are the relevant dimensions for the current category-learning trial. Later, in the test phase, participants performed the categorization task and compared the trial standard stimulus with the other test stimuli solely on the basis of these dimensions.

Participants were told, (while performing a warm-up trial), that during each experimental trial they would have to learn which of the 32 "alien creatures" (test stimuli) belongs to the same tribe as the one identified as "chief" (a standard representing the target category). They were instructed that each trial in the experiment would be independent of the other trials and would necessitate learning a new way of categorizing the aliens into tribes. Participants were not informed that for each trial 2 or 3 dimensions were chosen as trial-relevant. In general, we did not give participants specific instructions which would clarify the optimal categorization strategy or the structure of the categories; rather, participants were simply told that during each trial they would have to use the clues provided for identifying members of the chief's tribe. Participants were also instructed that they would have limited time to respond, and that they should perform the task not only accurately, but also as quickly as possible.

On each trial, 3 pairs of either same-class or different-class exemplars appeared simultaneously for 20 s in order to allow participants to compare each exemplar pair and integrate the information provided by the set of pairs. For the *Same-Class Exemplar* condition, participants were instructed that the two creatures that are presented together within a single green frame are necessarily of the same kind. In the *Different-Class Exemplar* condition, participants were instructed that two creatures that are presented together in a pink frame are necessarily of different kinds. There was no learning phase in the *Control* condition. After 20 s the learning phase was terminated and the test phase began. In the test phase, participants were given 50 s to select (by drag-and-drop) from the array of 32 stimuli presented simultaneously on the screen, those that he or she thought belong to the standard's category. The trial was then terminated and the next experimental trial began. All together, participants performed 10 category learning tasks in each experimental condition.

4.3.1. *Control and Random conditions*

Participants in the *Random*, lowest *Level of Supervision*, were tested on three experimental conditions: In the first, they categorized stimuli without the learning phase (*Control* condition). This condition was needed to assess the contribution of learning by comparison that was tested in the other experimental conditions. In the second and third experimental conditions, participants were provided with *Randomly*-generated *Same-Class* or *Different-Class Exemplars*. These randomly generated pairs were consistent with the task-assigned categories, but no attempt was made to control the information they provided as a set (i.e., their selection was random). In a sense, these *Random* conditions were designed to represent expected real-world scenarios in which the classifier is provided with haphazard indications of the categorical relations between stimuli and not those that are necessarily most useful for good categorization (see Fig. 6-left for examples).

Note that for reasons mentioned in the Introduction, in the *Random Same-Class Exemplar* condition the information provided by three randomly selected pairs almost always sufficed for identifying the task-relevant dimensions. This was not the case for the *Random Different-Class Exemplar* condition, where the information provided was almost as poor as in the *Control* condition.

4.3.2. *The Informative conditions*

Participants in the *Informative*, intermediate *Level of Supervision* conditions, performed categorization tasks in which both same-class and different-class exemplars were deliberately selected so as to provide all the information needed for perfect performance. The goal here was to test participants' inherent proficiencies in the comparison of same-class vs. different-class exemplars, when both types are similarly informative.

4.3.3. *The Informative plus Directions conditions*

The procedure in the *Informative plus Directions*, high *Level of Supervision* condition was identical to that of the *Informative* condition with exactly the same sets of same-class and different-class exemplars presented in the learning phase. The only difference between these two *Level of Supervision* conditions was in the instructions provided during the warm-up trial of each condition. Participants in the *Informative plus Directions* condition were encouraged to compare the paired exemplars and to derive a generalized categorization rule from their similarities and differences. The directions were simple

and straightforward, and all participants easily learned the category-learning strategy. In particular, before performing the *Informative plus Directions Same-Class Exemplar* condition, participants were informed that they should exclude the dimension discriminating between each two paired exemplars, since this dimension was necessarily irrelevant for the categorization task, and reserve judgment about the rest of the dimensions, for which the paired exemplars had identical features, since they may or may not be relevant. Before performing the *Informative plus Directions Different-Class Exemplars* condition, participants were informed that they should take into account the dimension discriminating between each two paired exemplars because, as the only differentiating dimension it must be relevant for the categorization task.

Appendixes

Appendix A. Dimension reduction in a hypercube space (L1 Metric)

We compare the contribution of *Same-Class Exemplar* pairs vs. *Different-Class Exemplar* pairs for cases in which 1— a dimension is either relevant or irrelevant, 2— the dimensions are independent, and 3— dimensions are pseudo-binary — i.e. the number of values in each dimension can be greater than two, but we assume that for each dimension there is at most one orthogonal border line decisively separating between categories. We show that the number of objects in each category is not an important factor for reaching our conclusion. Under these assumptions we can refer to the object space as to a hypercube.

For the case of a *Same-Class Exemplar* pair, its information content (in bits) can be obtained from the number of irrelevant dimensions it specifies. This will be the number of dimensions in which the within-category variation, presented by the *Same-Class Exemplar* pair, is similar to the observed between-category variation. A *Same-Class Exemplar* pair will be poorly informative (0 bits) only when it is composed of two nearby objects from the same category (objects that are both taken from the region of the same vertex of the hypercube; see text Fig. 3b).

In the case of *Different-Class Exemplar* pairs, relevant dimensions can be *directly* identified as the dimensions in which two compared objects differ significantly. Nevertheless, as the number of dimensions in which they differ increases, it becomes less clear which of these dimensions are relevant for discriminating between the categories. As a result, *Different-Class Exemplar* pairs can at most provide 1 bit. This will happen only when the compared objects differ in only one dimension.

The following analysis provides an assessment of the amount of information provided by a small set of *Same-Class Exemplar* pairs vs. a small set of *Different-Class Exemplar* pairs, as a function of the total number of dimensions in the feature space, and the number of dimensions relevant for the categorization task.

Let:

- c the number of categories.
- d the number of relevant dimensions; assuming binary dimensions, $c=2^d$.
- D the total number of dimensions.
- n the number of objects in the neighborhood of each vertex.
- N the number of objects in each category, $N=2^{D-d}n$.

It follows that,

1. Total number of *Same-Class Exemplar* (*SCE*) pairs (with #bits=0, 1, 2… $D-d$):

$$SCE=\frac{N_c(N-1)}{2}=\frac{2^{D-d}2^dn\left(2^{D-d}n-1\right)}{2}=\frac{2^Dn\left(2^{D-d}n-1\right)}{2}$$

$$=\frac{2^{2D-d}n^2-2^Dn}{2}$$

2. The number of poorly informative *Same-Class Exemplar* (*poorSCE*) pairs (with #bits=0):

$$poorSCE=\frac{2^Dn(n-1)}{2}$$

3. The number of highly *informative Same-Class Exemplar* (*infSCE*) pairs (with #bits≥1):

$$\inf\ SCE=\frac{2^Dn\left(2^{D-d}n-1\right)}{2}-\frac{2^Dn(n-1)}{2}=\frac{2^Dn^2\left(2^{D-d}-1\right)}{2}$$

Therefore the ratio between the number of highly informative and the number of poorly informative Same-Class Exemplar pairs is:

$$\frac{2^Dn^2\left(s^{D-d}-1\right)}{2^Dn(n-1)}>2^{D-d}-1\geq 1 \text{ whenever } D>d$$

i.e., whenever there is at least one irrelevant dimension, most of the *Same-Class Exemplar* pairs will be highly informative. Furthermore, as the number of irrelevant dimension increases, a larger portion of *Same-Class Exemplar pairs* will be highly informative.

4. Total number of *Different-Class Exemplar* (*DCE*) pairs (#bits≤1):

$$DCE=\frac{N^2c(c-1)}{2}=\frac{2^{2(D-d)}2^dn^2\left(2^d-1\right)}{2}=\frac{2^{2D-d}n^2\left(2^d-1\right)}{2}$$

$$=\frac{2^{2D}n^2-2^{2D-d}n^2}{2}$$

5. The number of highly informative *Different-Class Exemplar* (*infDCE*) pairs (with #bits=1):

$$\inf\ DCE=\frac{n^2cd}{2}=\frac{n^22^dd}{2}$$

6. The number of poorly informative *Different-Class Exemplar* (*poorDCE*) pairs (with #bits<1):

$$poorDCE=\frac{2^{2D}n^2-2^{2D-d}n^2}{2}-\frac{n^22^dd}{2}$$

$$=\frac{n^22^d\left(2^{2D-d}-2^{2D-2d}-d\right)}{2}$$

Therefore the ratio between the number of highly informative and the number of poorly informative Different-Class Exemplar pairs is:

$$\frac{n^2 2^d d}{n^2 2^d \left(2^{2D-d} - 2^{2D-2d} - d\right)} = \frac{d}{\left(2^{2D-d} - 2^{2D-2d} - d\right)} < 1 \text{ whenever } D > d$$

i.e., the majority of *Different-Class Exemplar* pairs are poorly informative. Furthermore, as the total number of dimensions increases, a larger portion of *Different-Class Exemplar* pairs will be poorly informative.

The properties of hypercubes suggest that as the dimensionality of the hypercube space increases, the vast majority of diagonals are expected to be from an order higher than 2 (Bowen, 1982). For that, the above calculation suggests that the information content for most *Same-Class Exemplar* pairs is expected to be more than 1 bit. For *Different-Class Exemplars*, in similar conditions, most pairs are expected to provide much less than 1 bit of information. The difference in the information content between the two types of comparison processes increases as the total number of dimensions increases, and particularly as this number increases with respect to the number of relevant dimensions.

To conclude, in the context of a category learning task in which we dichotomically discriminate relevant dimensions from irrelevant ones, the information content of a haphazard small set of *Same-Class Exemplar* pairs will be significantly larger than the information content of a small set of *Different-Class Exemplar* pairs.

Appendix B. Similarity and information content of same-class and different-class constrained pairs for a binary perceptron (L2 Metric)

The intuition says that for a *Same-Class Exemplas* constraint, the information value increase with the Euclidean distance between the constrained points, while for *Different-Class Exemplar* constraints the opposite thing happens. A natural measurement for the information value of a constraint is the reduction it incurs to the volume of the 'version space', i.e. the space of all possible hypotheses. An informative constraint is not consistent with a large portion of the version space, and hence it leaves a small number of possible hypotheses.

However, placing a measure and analyzing the structure of the version space is relatively hard in all but the simplest cases. We therefore focus here on the relatively simple case of a binary linear classifier passing through the origin, as the one used in certain versions of the Perceptron or linear SVM. This classifier is characterized by a weight vector $w \in R^d$, and given an input pattern $x \in R^d$ its output is a binary label $y \in \{-1, 1\}$ given by the formula

$$y = sign(w \cdot x)$$

The advantage of using this simple classifier is that the hypotheses space (the set of all possible w parameters) has a relatively simple structure. It can be equated with the unit sphere $W = \{w : \|w\| = 1\}$, and it has a natural prior measure, i.e. the uniform measure over the sphere.

A *Same-Class Exemplar* constraint $(x_1, x_2, 1)$ demands that

$$(w \cdot x_1 > 0 \;\&\; w \cdot x_2 > 0) \text{ or } (w \cdot x_1 < 0 \;\&\; w \cdot x_2 < 0).$$

This can be summarized by demanding that $(w \cdot x_1)(w \cdot x_2) > 0$.

Similarly, a *Different-Class* constraint demands that $(w \cdot x_1 > 0 \;\&\; w \cdot x_2 < 0)$ *or* $(w \cdot x_1 < 0 \;\&\; w \cdot x_2 > 0)$ and equivalently $(w \cdot x_1)(w \cdot x_2) < 0$.

Looking at this characterization, we can see that a *Different-Class Exemplar* constraint $(x_1, x_2, -1)$ is equivalent to the *Same-Class Exemplar* constraint $(x_1, -x_2, 1)$ since

$$(w \cdot x_1)(w \cdot x_2) > 0 \quad \text{iff} \quad (w \cdot x_1)(w \cdot (-x_2)) < 0$$

This means that for such a binary classifier every *Same-Class Exemplar* constraint can be converted to a *Different-Class Exemplar* constraint and vice versa, so there is no inherent difference between the information values of the two constraint types. However, the two types of constraints radically differ with respect to the way they are affected by the similarity between the constrained data points.

The main difference between the two types of constraints is sketched in Fig. 7. In these sketches we assume for simplicity that the classifier operates in R^2, i.e. each data instance is described using two measurements only. In Fig. 7a, a typical linear classifier

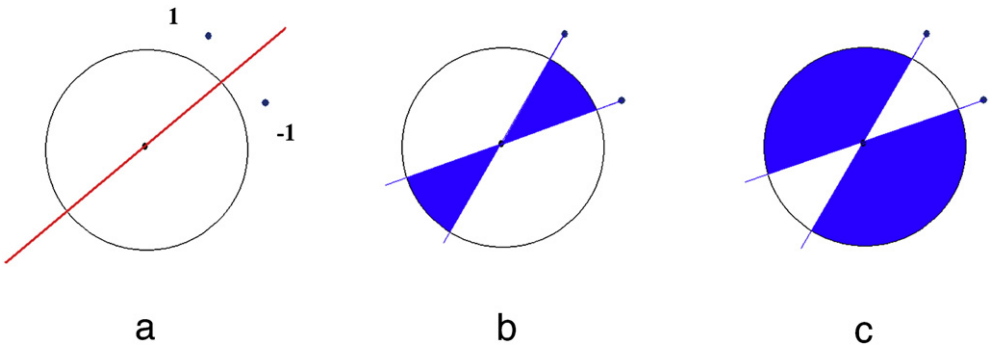


Fig. 7 – Illustration for the main difference between *Same-Class* and *Different-Class* constraints on a two-dimensional space.

is plotted: the classifier is characterized by a single borderline passing through the origin, which separates between data point labeled by 1 and those labeled by −1. a separator in this example can be characterized by a single number — the separator's direction in [0, 2π]. Notice that such a separator only discriminates between two points if their direction is different. Hence a natural measurement for the similarity between two points is the cosine of the angle between them, expressed by

$$\cos\ (\theta(x_1, x_2)) = \frac{x_1 \cdot x_2}{\| x_1 \| \cdot \| x_2 \|}$$

When this angle is small, the two points are similar and the cosine will be close to 1. When they are far from each other and create an angle of 90 degrees, the similarity as measured by the cosine will be 0. Finally, when two points are antipodes, their cosine-based similarity will be −1.

In Fig. 7b we assume that a *Same-Class Exemplar* constraint is given between the two drawn data points. Since these points are known to be together (i.e. both should be on the same side of the linear separator), they disallow all the possible separators passing through the indicated blue (dark) sector. Hence the fraction of the version space disallowed by such a constraint is the angle between the two points, divided by the measure of all the version space, i.e.

$$p = \frac{1}{\pi}\ \arccos\left(\frac{x_1 \cdot x_2}{\| x_1 \| \cdot \| x_2 \|}\right)$$

It can hence be seen that the closer the points, the higher the cosine between them (i.e. their similarity), but the angle between them (the arc-cosine) tends toward zero and so is their information content p.

In Fig. 7c we see the effect a *Different-Class Exemplar* constraint has on the version space. The two points in this figure are assumed to be from different classes, and again the area of disallowed separators is marked in blue (dark area). This time the disallowed area is proportional to the complementary angle and the expression for the information content is

$$p = \frac{1}{\pi}\ \arccos\left(\frac{-x_1 \cdot x_2}{\| x_1 \| \cdot \| x_2 \|}\right)$$

This time points with high similarity (high cosine) give negative values in the arc-cosine argument, leading to large excluded angles and hence high information content.

The points illustrated in the Fig. 7 for two dimensional instances can be proved and extended to arbitrary high instance dimension under mild assumptions, including mainly a uniform prior over the space of allowed separators.

Appendix C. Graphs partitioning and transitivity

This Appendix provides a formal basis for the computational difference between *Same-Class Exemplar* pairs and *Different-Class Exemplar* pairs in graph partitioning. This difference, unlike those of Appendix 1 and 2, is not limited by any metrical assumptions.

Notation. We represent data points in a graph G $G=\{V, E\}$, where the set of nodes V of size N corresponds to the data points, and the set of edges E of size M corresponds to the given paired exemplars, either *Same-Class* or *Different-Class* (but not both). The task is to divide the data-points into K classes.

Appendix C.1. The complexity of satisfying Same-Class vs. Different-Class pairs

Assume $K=2$, and the task is therefore to partition the data into two clusters. Each partition is represented by C — the set of all edges from E which connect nodes assigned to different clusters; the set C is called the cut of graph G. Each cut is assigned a cost — the number of edges in C.

Appendix C.1.1. Enforcing Same-Class Exemplar constraints is manageable. Given *Same-Class Exemplar* pairs, we seek a partition which violates as few edges as possible, representing *Same-Class Exemplar* constraints. Finding this partition is equivalent to finding the minimal cut in the above graph. There are known efficient algorithms to solve this problem. Thus, in the complexity hierarchy of computer science, this problem is considered tractable.

Appendix C.1.2. Enforcing Different-Class Exemplar constraints is hard. Given *Different-Class Exemplar* pairs, we seek a partition which violates as many edges as possible, representing *Different-Class Exemplar* constraints. Finding this partition is equivalent to finding the maximal cut in the graph defined above. There are no known efficient algorithms to solve this problem (although approximate solutions can be produced by using meta-heuristic search methods). Therefore, in the complexity hierarchy of computer science, this problem is almost certainly intractable.

Appendix C.2. The information content of Same-Class vs. Different-Class pairs

We define the information of a set of paired exemplars E to be the difference between the entropy H of all the partitions of the set of nodes V to K clusters, and the entropy H_G of all such partitions consistent with E. Assuming that each allowed partition is assigned equal probability, the entropy H_G is equal to the log of the number of allowed partitions. We are interested in the difference between the information of *Same-Class* and *Different-Class* constraints, namely in

$$I = \left(H - H_G^+\right) - \left(H - H_G^-\right) = H_G^- - H_G^+ = \log \frac{\#_G^-}{\#_G^+}$$

where the entropy superscript + or − denotes, respectively, whether the set of constraints is *Same-Class* or *Different-Class*, $\#_G^-$ denotes the number of partitions consistent with E if the constraints are *Different-Class*, and $\#_G^+$ is similarly defined if the constraints are *Same-Class*.

N_C denotes the number of connected components of G. In particular, if the graph G has no loops, $N_C=N-M$. For a general graph with N_C connected components, we note that each connected component in G has at least one legal coloring (by assumption). Here N_C^l denotes the number of connected components with l or more elements $N_C=N_C^1\geq N_C^2\geq\ldots\geq N_C^K$. The following can now be shown:

The information gain of Same-Class over Different-Class pairs satisfies

$$I \geq \sum_{l=2}^{K} N_C^l \ \log\ (K-l+1)$$

We can therefore state that, if $N \gg N_C$, the information content of *Same-Class* constraints is exponentially larger than *Different-Class* constraints; (for the detailed calculation and further comments regarding this analysis, see Hammer et al., 2007).

REFERENCES

Anderson, J.R., 1991. The adaptive nature of human categorization. Psychol. Rev. 98 (3), 409–429.

Ashby, F.G., Ell, S.W., 2001. The neurobiology of human category learning. Trends Cogn. Sci. 5, 204–210.

Ashby, F., Queller, S., Berretty, P.M., 1999. On the dominance of unidimensional rules in unsupervised categorization. Percept. Psychophys. 61, 1178–1199.

Bar-Hilel, A., Hertz, T., Shental, N., Weinshall, D., 2003. Learning distance functions using equivalence relations. The 20th International Conference on Machine Learning, pp. 11–18.

Bilenko, M., Basu, S., Mooney, R.J., 2004. Integrating constraints and metric learning in semi-supervised clustering. Proceedings of the 21st International Conference on Machine Learning, pp. 81–88.

Bowen, J.P., 1982. Hypercubes. Pract. Comput. 5 (4), 97–99.

Caramazza, A., Shelton, J.R., 1998. Domain-specific knowledge systems in the brain: the animate–inanimate distinction. J. Cogn. Neurosci. 10 (1), 1–34.

Diesendruck, G., Hammer, R., Catz, O., 2003. Mapping the similarity space of children and adults' artifact categories. Cogn. Dev. 118, 217–231.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. John Wiley and Sons Inc.

Erickson, J.E., Chin-Parker, S., Ross, B.H., 2005. Inference and classification learning of abstract coherent categories. J. Exper. Psychol., Learn., Mem., Cogn. 31 (1), 86–99.

Fleming, M., Cottrell, G., 1990. Categorization of faces using unsupervised feature extraction. Proc. IEEE Int. Joint Conf. Neural Networks, vol. 2, pp. 65–70.

Goldstone, R.L., Medin, D.L., 1994. Time course of comparison. J. Exper. Psychol., Learn., Mem., Cogn. 20, 29–50.

Grier, J.B., 1971. Nonparametric indexes for sensitivity and bias: computing formulas. Psychol. Bull. 75, 424–429.

Hammer, R., Diesendruck, G., 2005. The role of feature distinctiveness in children and adults' artifact categorization. Psychol. Sci. 16 (2), 137–144.

Hammer, R., Hertz, T., Hochstein, S., Weinshall, D., 2005. Category learning from equivalence constraints. Proceedings of the 27th Annual Conference of the Cognitive Science Society.

Hammer, R., Hertz, T., Hochstein, S., Weinshall, D., 2007. Classification with positive and negative equivalence constraints: theory, computation and human experiments. In: Mele, F., Ramella, G., Santillo, S., Ventriglia, F. (Eds.), Brain, Vision, and Artificial Intelligence: Second International Symposium, BVAI 2007. Lecture Notes in Computer Science. Springer-Verlag Press, Berlin Heidelberg, pp. 264–276.

Hammer, R., Diesendruck, G., Weinshall, D., Hochstein, S., (submitted for publication). The Development of Category Learning Strategies: What Makes the Difference?

Hammer, R., Hertz, T., Hochstein, S., and Weinshall, D., (in press). Category learning from equivalence constraints. Cognitive Processing.

Keil, F.C., 1989. Concepts, kinds, and cognitive development. MIT Press, Cambridge, MA.

Khanna, S., Linial, N., Safra, S., 2000. On the hardness of approximating the chromatic number. Combinatorica 1 (3), 393–415.

Kurtz, K.J., Boukrina, O., 2004. Learning relational categories by comparison of paired examples. Proceedings of the 26th Annual Conference of the Cognitive Science Society.

Love, B.C., Medin, D.L., Gureckis, T.M., 2004. SUSTAIN: a network model of category learning. Psychol. Rev. 111, 111309–111332.

Markman, A.B., Gentner, D., 1993. Structural alignment during similarity comparisons. Cogn. Psychol. 25, 431–467.

Medin, D.L., Schaffer, M.M., 1978. Context theory of classification learning. Psychol. Rev. 85, 207–238.

Medin, D.L., Goldstone, R.L., Gentner, D., 1993. Respects for similarity. Psychol. Rev. 100 (2), 254–278.

Murphy, G., Medin, D.L., 1985. The role of theories in conceptual coherence. Psychol. Rev. 92, 289–316.

Namy, L.L., Gentner, D., 2002. Making a silk purse out of two sow's ears: young children's use of comparison in category learning. J. Exp. Psychol. Gen. 131, 5–15.

Nosofsky, R.M., 1986. Attention, similarity, and the identification-categorization relationship. J. Exp. Psychol. Gen. 115, 39–57.

Oakes, L.M., Ribar, R.J., 2005. A comparison of infants' categorization in paired and successive presentation familiarization tasks. Infancy 7, 85–98.

Posner, M., Keele, S., 1968. On the genesis of abstract ideas J. Exp. Psychol. 77, 353–363.

Pothos, E.M., Chater, N., 2002. A simplicity principle in unsupervised human categorization. Cogn. Sci. 26, 303–343.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyes-Braem, P., 1976. Basic objects in natural categories. Cogn. Psychol. 8, 382–439.

Shental, N., Bar-Hillel, A., Hertz, T., Weinshall, D., 2004. Computing Gaussian mixture models with EM using equivalence constraints. Proceedings of Neural Information Processing Systems, NIPS 2004.

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

Sloutsky, V.M., 2003. The role of similarity in the development of categorization. Trends Cogn. Sci. 7, 246–251.

Smith, L.B., Jones, S.S., Landau, B., 1996. Naming in young children: a dumb attentional mechanism? Cognition 60, 143–171.

Spalding, T.L., Ross, B.H., 1994. Comparison-based learning: effects of comparing instances during category learning. J. Exper. Psychol., Learn., Mem., Cogn. 20 (6), 1251–1263.

Tyler, L.K., Moss, H.E., Durrant-Peatfield, M.R., Levy, J.P., 2000. Conceptual structure and the structure of concepts: a distributed account of category-specific deficits. Brain Lang. 75, 195–231.

Weber, M., Welling, M., Perona, P., 2000. Unsupervised learning of models for recognition. European Conference on Computer Vision, Dublin, Ireland, pp. 18–32.

Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S., 2002. Distance metric learning with application to clustering with side-information. Advances in Neural Information Processing Systems, vol. 15. The MIT Press.