Class Representation and Image Retrieval with Non-Metric Distances*

David W. Jacobs
NEC Research Institute
4 Independence Way
Princeton, NJ 08540, USA
dwj@research.nj.nec.com

Daphna Weinshall†
Inst. of Comp. Science
The Hebrew University
91904 Jerusalem, Israel
daphna@cs.huji.ac.il

Yoram Gdalyahu
Inst. of Comp. Science
The Hebrew University
91904 Jerusalem, Israel
yoram@cs.huji.ac.il

Abstract

One of the key problems in appearance-based vision is understanding how to use a set of labeled images to classify new images. Classification systems that can model human performance, or that use robust image matching methods, often make use of similarity judgments that are non-metric; but when the triangle inequality is not obeyed, most existing pattern recognition techniques are not applicable. We note that exemplar-based (or nearest-neighbor) methods can be applied naturally when using a wide class of non-metric similarity functions. The key issue, however, is to find methods for choosing good representatives of a class that accurately characterize it. We show that existing condensing techniques for finding class representatives are ill-suited to deal with non-metric dataspaces.

We then focus on developing techniques for solving this problem, emphasizing two points: First, we show that the distance between two images is not a good measure of how well one image can represent another in non-metric spaces. Instead, we use the vector correlation between the distances from each image to other previously seen images. Second, we show that in non-metric spaces, boundary points are less significant for capturing the structure of a class than they are in Euclidean spaces. We suggest that atypical points may be more important in describing classes. We demonstrate the importance of these ideas to learning that generalizes from experience by improving performance using both synthetic and real images.

*This paper is based on "Condensing Image Databases when Retrieval is based on Non-Metric Distances", by D. Jacobs, D. Weinshall and Y. Gdalyahu, which appeared in the 6th IEEE International Conference on Computer Vision, January 1998, and on "Supervised Learning in Non-Metric Spaces", by D. Weinshall, D. Jacobs and Y. Gdalyahu, 1998, NIPS, (forthcoming).

†Vision research at the Hebrew University is partially supported by DARPA through ARL Contract DAAL01-97-0101. This research was done while DW was on sabbatical at NEC Research Institute.

1 Introduction

The availability of large image databases makes it possible to classify a new image by querying the database. To identify the class of a new object or scene from an image, we need to build up some representation of classes using previously seen images. Two approaches to this problem have dominated computer vision, pattern recognition and cognitive science. In one *parametric* approach, some compact representation is generated that attempts to capture what is essential to each class. This might be a prototype of the class, a list of essential features, a parameterized model, or a set of geometric surfaces that delineate the set of images belonging to the class, encoded in a neural net. In a second, exemplar-based approach, some or all previously seen examples are stored and compared directly to new images.

Almost all work following these approaches has assumed that images may be thought of as vectors that can be compared for similarity using the Euclidean distance. However, this may not be a valid assumption. A number of recent approaches in computer vision compare images using measures of similarity that are not Euclidean, and in fact not even metric, in that they do not obey the triangle inequality. This can occur because the triangle inequality is difficult to enforce in complex matching algorithms that are statistically robust. Moreover, much research in psychology suggests that human similarity judgments are also not metric. This raises serious questions about the extent to which existing work on classification can be applied using complex models of similarity, either in computer vision or as cognitive models.

The problem is illustrated in Fig. 1, showing the different decision boundaries that arise when using three non-metric distance functions, in comparison with the metric Manhattan and Euclidean distances.

Non-metric distances turn up in many application domains, such as string (DNA) matching, collaborative filtering (where customers are matched with stored "prototypical" customers), and retrieval from image data bases. Fig. 7 shows one example, the output of an algorithm for judging the similarity of the silhouettes of different objects [9]. Given a series of labeled pictures of common objects (cars and cows, Fig. 7a,b), we may wish to identify new silhouettes (Fig. 7d) based on their similarity to the previously seen ones. [1] reviews many such algorithms, and discusses why their use will typically lead to non-metric distances.

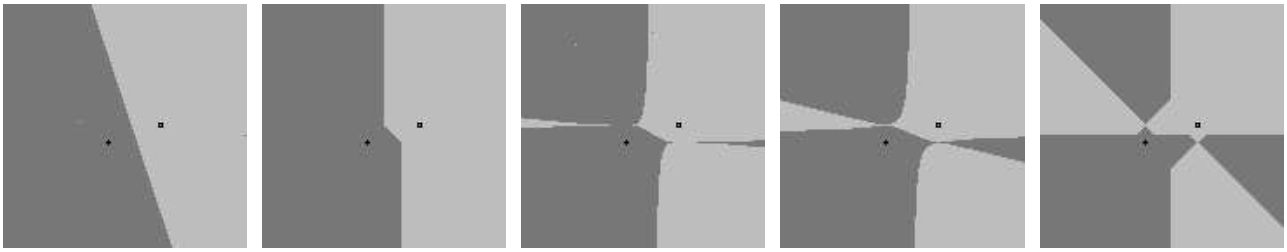


Figure 1: The Voronoi diagram for two points using, from left to right, p-distances with $p = 2$ (Euclidean distance), $p = 1$ (the Manhattan distance, which is still metric), the non-metric distances arising from $p = .5$, $p = .2$, and the min (1-median) distance. The p-distance between 2 points (x_1, y_1) and (x_2, y_2) is: $(|x_1 - x_2|^p + |y_1 - y_2|^p)^{\frac{1}{p}}$; the min distance is $min(|x_1 - x_2|, |y_1 - y_2|)$. The min distance in 2-D is illustrative of the behavior of the other median distances in higher dimensions. The region of the plane closer to the one point is shown in dark gray, and closer to the other in light gray.

In this paper we make two contributions to this problem. First, we show that existing classification methods can indeed encounter considerable difficulties when applied to non-metric similarity functions. Then we show how exemplar-based approaches can be modified so that they are appropriate for a range of different non-metric similarity functions, including some commonly used ones.

More specifically, we first make the minimalistic assumption that higher-order representations of images (e.g., as a set of features) are *not* given; we are only given a (possibly complex) comparison algorithm which takes as input any two raw images and returns the similarity between them. Thus the choice of class representation needs to be image-based, and classification can only be based on nearest neighbors. We then consider an approach to nearest neighbor classification in which representative images are selected from each class. A new image is then classified according to its nearest neighbor among these representatives.

Existing methods select representatives so that every member of the class has a representative nearby. That is, one image is chosen as a stand-in for others that are close to it. We show that this technique is often ill-suited to non-metric spaces. Rather, we show that it is preferable to pick stand-ins for an image such that the image and its stand-ins have similar distances to other, previously seen images. We demonstrate analytically some of the conditions under which this strategy is preferable, and show experimentally that it is effective.

The rest of this paper is organized as follows: In Section 2 we discuss why non-metric distances are often the method of choice in applications, and why most existing work on supervised

classification is ill-suited to handle these cases. In Section 3 we describe our approach to the organization (or condensing) of a non-metric image database. In Section 4 we describe a concrete condensing algorithm, and compare it to other algorithms using both simulations and real data. Finally, in Section 5 we discuss the possibility of relaxing the assumptions in this paper and building parametric supervised learning algorithms suited to specific non-metric distances.

2 Some Challenges Raised by Non-metric Distances

2.1 Why non-metric distances are important

Our work is motivated by work on image similarity that suggests that practical and psychologically valid measures of similarity are non-metric. Note that some authors use the phrase *non-metric* to imply a qualitative, or rank ordering of distances. We use *non-metric* in the standard mathematical sense. A distance function, $D(I_1, I_2)$, between pairs of images, is metric when:

1. $D(I_1, I_2) \geq 0$
2. $D(I_1, I_2) = 0$ if and only if $I_1 = I_2$.
3. $D(I_1, I_2) = D(I_2, I_1)$ (symmetry).
4. $D(I_1, I_3) \leq D(I_1, I_2) + D(I_2, I_3)$ (the triangle inequality).

We are interested in spaces in which the last condition fails. Failure of symmetry is also of interest, but this is beyond the scope of the present paper.

Distance functions that are robust to outliers or to extremely noisy data will typically violate the triangle inequality. One group of such functions is the family of image comparison methods that match subsets of the images and ignore the most dissimilar parts, see [1, 9, 2, 20]. As one example, Huttenlocher et al.[17, 18] perform recognition and motion tracking by comparing point sets using the Hausdorff distance. They consider only a fixed fraction of the points for which this distance is minimized. By not considering the most dissimilar parts of the images, these methods become both robust to image points that are outliers, and non-metric. We call these non-metric methods median distances. A k-median distance between two vectors (or images, represented as

vectors) $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is defined as:

$$d(\mathbf{x}, \mathbf{y}) = k\text{-}median\{|x_1 - y_1|, \ldots, |x_n - y_n|\}$$

where the k-median operator returns the k-th value of the ordered difference vector. Related robust techniques, such as M-estimation which identify outliers and weigh them less heavily, also lead to non-metric distances [13].

Also, the use of non-metric L^p distances, with $0 < p < 1$ has been suggested for robust image matching [7]. An L^p distance (or *p-distance*) between two vectors $\mathbf{x}, \mathbf{y}$, is defined as:

$$d(\mathbf{x}, \mathbf{y}) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} \tag{1}$$

Thus, $p = 2$ gives the Euclidean distance, and $p = 1$ gives the Manhattan distance. These distances are non-metric for $p < 1$ (see Royden [23]). As shown in Fig. 2, they are less affected by extreme differences than the Euclidean distance, and can therefore be more robust to outliers.

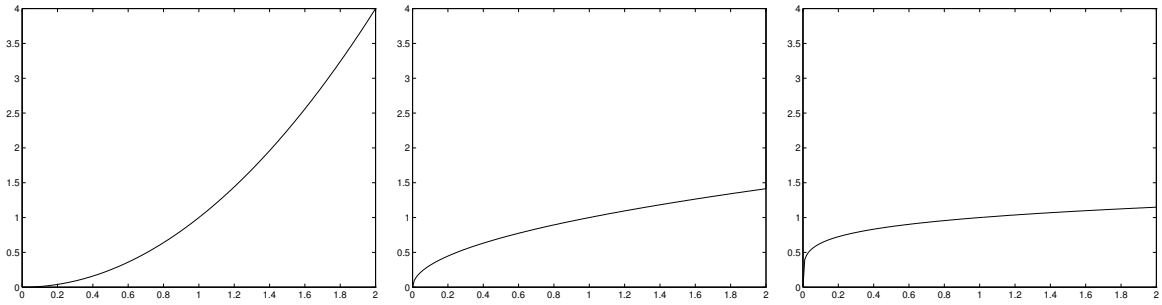


Figure 2: Graphs of x^2 (left), $x^{0.5}$ (middle) and $x^{0.2}$ (right). We can see that when properly normalized, for p-distances less than 1 the cost function rises quickly at first, than more slowly, so that extreme values do not dominate total cost.

A second reason for non-metric distances to arise is that image distances may be the output of a complex algorithm, which has no obvious way of ensuring that the triangle inequality holds. Jain et al. [19] , for example, perform character recognition and other image comparisons using a deformable template matching scheme that yields distances that are not symmetric and do not obey the triangle inequality. Related elastic matching methods have been widely used (e.g., [1, 9, 15, 26, 30, 25]) in ways that do not appear to lead to metric distances. In fact, Basri et al.[1] show that contradictions can exist between forcing such methods to obey the triangle inequality and other goals that are desirable in deformable template matching.

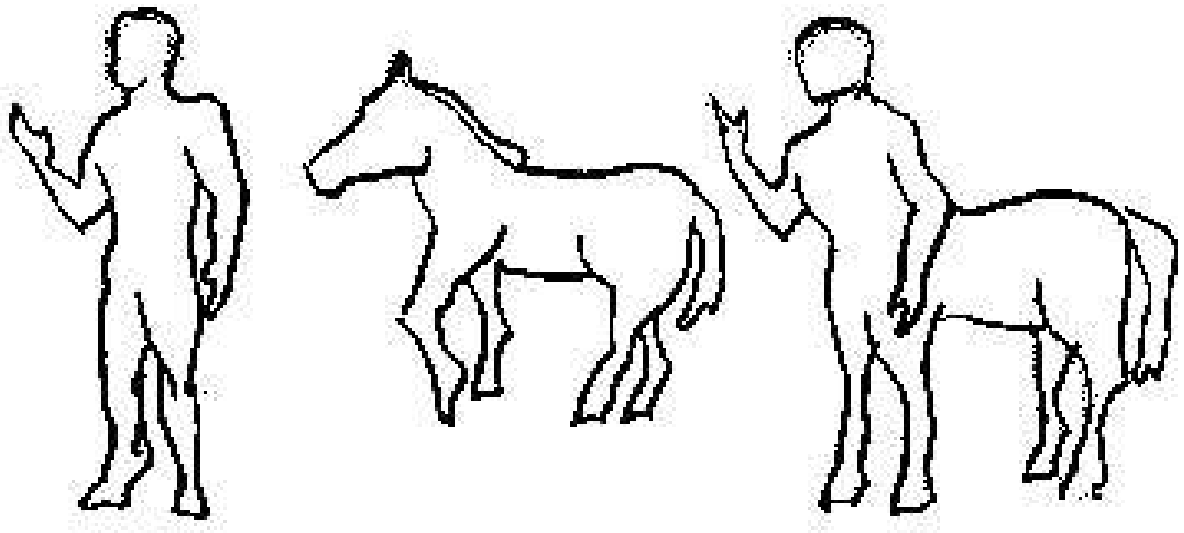


Figure 3: Judgements of visual similarity between these three images may violate the triangle inequality.

Finally, we are interested in image comparison methods that model human vision. This may also be desirable in many applications. However, there is much work in psychology that suggests that human similarity judgments are non-metric. Most notably, Tversky et al. (e.g., [27]) showed in a series of studies that human similarity judgments often violate metric axioms: in particular, the judgment need not be symmetric (one would say "North Korea is similar to Red China", but not "Red China is similar to North Korea"), and the triangle inequality rarely holds (transitivity should follow from the triangle inequality, but the similarity between Jamaica and Cuba and between Cuba and Russia does not imply similarity between Jamaica and Russia). Figure 3 provides a visual analog of this example. Many observers will find that the centaur is quite similar to the person and to the horse. However the person and the horse are quite different from each other. For observers who determine that the dissimilarity between the horse and person is more than twice as great as the dissimilarity between either shape and the centaur, there is a violation of the triangle inequality. Intuitively, this occurs because when comparing two images we focus on the portions that are very similar, and are willing to pay less attention to regions of great dissimilarity. Consequently, any computer vision system that attempts to faithfully reflect human judgments of similarity is apt to devise non-metric image distance functions (see also [24] for discussion of this issue).

2.2 Why existing methods are not suitable to deal with non-metric distances

In this paper we specifically focus on the problem of supervised classification in non-metric spaces. Despite the importance of non-metric distance functions, most work on supervised classification has considered the case in which the training dataset consists of points in Euclidean space, with the L^2 norm. For example, in linear discriminant analysis and in many multi-layer neural network models, hyperplanes are found that separate labeled points as well as possible. Similarly, the mean point of a set is used in methods such as k-means clustering, and minimal linear spanning bases are used in methods such as Radial Basis Functions [22] or PCA methods. As we will discuss in Section 5, such approaches would require significant modification to be applied to classification in non-metric spaces.

We should note that, in contrast to supervised learning, there has been a good deal of work on clustering, or *unsupervised* learning that is applicable to non-metric spaces. An early example of such work, which stresses the importance of non-metric distances, is [21]. Brand [4] has proposed clustering based on E-M in non-metric spaces. In addition, clustering methods based on graph partition or physical models (e.g., [3, 16]) are suitable for non-metric clustering. These works do not directly address the issue of using these clusters for classification of new data (though see [21] for some comments).

Throughout most of this paper we assume only that we can compute distances between images. We do not have a mapping from images to vectors in some vector space. In this case, it is simply not possible to construct parametric descriptions of classes, such as linear separators. Only exemplar based methods need not face these problems, because they represent classes using elements of the training set directly, rather than attempting to construct new objects in the space. In particular, we consider nearest neighbor methods which classify a new item as belonging to the same class as the item in the training set to which it is closest.

One can compute nearest neighbors in a non-metric space with the brute force approach, computing the distance to a test image from all elements in the training set. Recently, [11] have shown improvements over nearest neighbor methods by representing each object by its distance to all others, and then applying support vector machines to the resulting vectors. However, these approaches can result in impractical space and time complexity, because they requires storing all

previously seen images, and require explicit computation of many distances. In particular time complexity may be high because many similarity functions used in computer vision are complex, and require considerable computation. Many techniques have therefore been devised to speed up nearest neighbor classification. Some of these methods build tree structures that rely on the points lying in a Euclidean [8] or at least in a metric space [29].

A more heuristic approach to decrease the time and space complexity of nearest neighbor classification - the use of *condensing* algorithms - has been proposed for metric spaces but is potentially applicable in non-metric spaces as well. These algorithms select subsets of the original training classes that are as small as possible, and that will correctly classify all the remainder of the training set, which is then discarded. For example, Hart [14] proposed an iterative algorithm that initializes a representative set, and then continues to add elements from the training set that cannot be correctly classified by the representative set, until all remaining training elements can be correctly classified. Gowda & Krishna [10] propose an algorithm like Hart's, but which explicitly attempts to add first to the representative set the points nearest the boundary. Fukunaga & Mantock [12] propose an iterative algorithm, that attempts to optimize a measure of how well distances to the representative set approximate distances to the full training set.

Dasarathy [6] more explicitly searches for the smallest representative set that will correctly classify the training set. We will discuss this algorithm in more detail, and show experiments with it in Section 4. Dasarathy notes that once an element is added to the representative set, it is guaranteed that other elements of the same class will be correctly classified if they are closer to this representative element than to any element of any other class. A greedy algorithm is used, in which the representative set is augmented at each step by the training element that guarantees correct classification of the largest number of elements not yet guaranteed to be correctly classified. This basic step is followed by subsequent steps that refine the representative set, but in practice these are not found to be important.

Condensing algorithms do not explicitly rely on distances that obey the triangle inequality. However, we will discuss in the next section two significant problems that arise when applying existing algorithms to non-metric space, and discuss basic ways of resolving these problems.

3 Our Approach

Our goal is to develop condensing methods for selecting a subset of the training set, and then classifying a new image according to its nearest neighbor among this subset. The goal is to find a subset of the training set which minimizes errors in the classification of new datapoints: a representative subset of the training data whose nearest distance to most new data items approximates well the nearest distance to all the training set. Thus we emphasize that the representative set maintains the same generalization function as the whole dataset.

3.1 Basic intuition: looking beyond distance

In designing a condensing method, one needs to understand *when is one image a good substitute for another?* Our answer to this question is what most distinguishes our approach from previous work on nearest neighbors in metric spaces. So we begin by considering the answer to this question that is implicit in previous work. In what follows, let i_1 and i_2 denote two arbitrary elements of the training set, and let j denote a new image that we must classify. Let $d(i, j)$ denote the distance between two element, i and j.

In a metric space, the triangle inequality guarantees that the distance between i_1 and i_2 bounds the difference between $d(i_1, j)$ and $d(i_2, j)$, that is, $|d(i_1, j) - d(i_2, j)| \leq d(i_1, i_2)$. Thus when $d(i_1, i_2)$ is small, we know that one of these images can substitute for the other. However, in a non-metric space, the value of $d(i_1, i_2)$ need not provide us with any information about the relationship between $d(i_1, j)$ and $d(i_2, j)$. Our first observation is that we must use more information than just the distance between two datapoints to decide whether the presence of one in the training set makes the other superfluous.

More specifically, what we really need to know is when two images will have similar distances to other images, yet unseen. We define the **redundancy** $R(i_1, i_2)$ of two images i_1, i_2 as the probability that $|d(i_1, j) - d(i_2, j)|$ is small for some arbitrary image j. In this section we make general points which are independent of the way we measure the redundancy of two images. (In the next section we use specific functional forms to measure redundancy.) This idea is quite simple, but it requires a different outlook than that taken in previous metric-space condensing algorithms.

Existing condensing methods focus on choosing representative points from the boundaries

between classes. Boundary points are especially important to retain as representatives in metric spaces because they influence the decision boundary between two classes in the region that is near the classes. Our second main intuition is that it is important to retain atypical examples in the training set rather than just boundary points. An "atypical" image is any image that is dissimilar from most other images in the class, and especially from the already chosen representatives of the class. Atypical points can be important to retain as representatives because points that are far from the other representatives can have a big effect on the decision boundary between the two classes, especially (but not only) in non-metric spaces.

3.2 A 2-D domain

Let us illustrate these points in a simple 2-D domain. Thus, in this section, all "images" are 2-D vectors. In this domain we will use the non-metric 1-median distance, i.e., the min distance. The min distance may not be a good robust estimator; our goal is only to demonstrate ideas that apply to other median distances in higher dimensions. The relevant geometrical structure is the Voronoi diagram: a division of the plane into regions in which the points all have the same nearest neighbor.

Our first example is Fig. 1, showing that non-metric distances (Fig. 1, three right-most pictures) can produce much more complex Voronoi diagrams than do metric distances (Fig. 1, two left-most pictures). Fig. 4 further illustrates the complex structures that classes can take with non-metric distances.

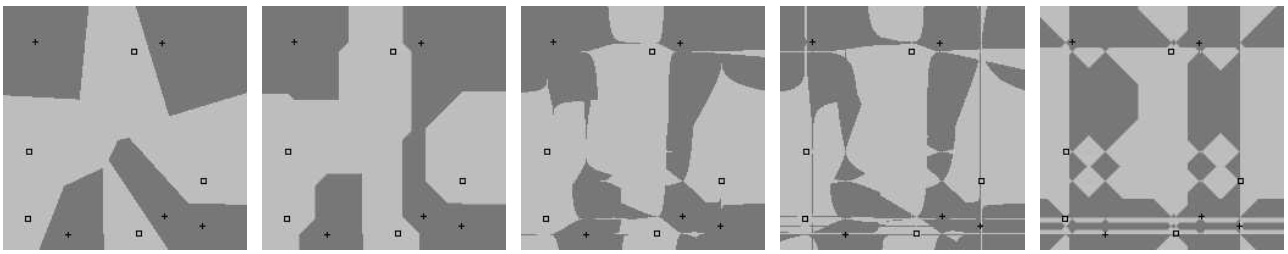


Figure 4: The Voronoi diagram for two sets of points, each containing 5 points (shown as squares and crosses). The distance functions used are, from left to right, p-distances with $p = 2$ (Euclidean distance), $p = 1$ (the Manhattan distance), $p = 0.5$, $p = 0.2$, and the median distance. The region of points closer to the class of crosses is shown in black, and the region closer to the class of squares is shown in white.

Next, Fig. 5 shows a simple example illustrating the potential value in looking beyond just the distance between images to measure their interchangeability as representatives of a class. In the top

right we show the Voronoi diagram, using the median distance, for two clusters of points (i.e., each point in the plane is labeled according to the cluster to which its nearest neighbor belongs). One cluster, P, consists of four points (labeled p_1, p_2, p_3, p_4 in the upper left, otherwise shown as black squares) all close together both according to the median distance, and to Euclidean distance. The second cluster, Q, (labeled $q_1, ...q_5$ in the upper left, otherwise shown as black crosses), all have the same x coordinate, and so all are separated by zero distance when using the median distance; however, the points are divided into two subgroups, q_1 and q_2 on top, and $q_3,\ q_4,\ q_5$ on the bottom.

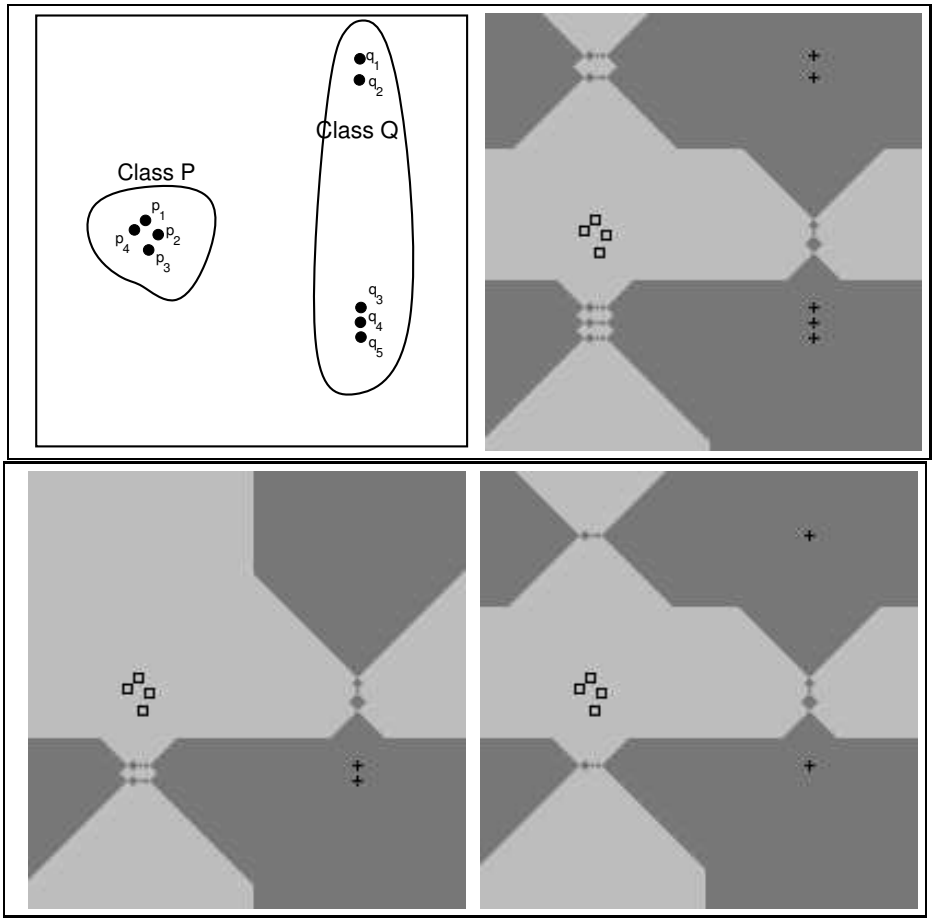


Figure 5: Top: Two clusters of labeled points (left) and their Voronoi diagram (right) using the median (min) distance. Bottom: The Voronoi diagram resulting when only q_3 and q_4 represent class Q (left) and when q_2 and q_3 are chosen as representatives (right).

To characterize Q well, it is important to pick representatives from both the bottom group and the top group. To illustrate this, the bottom left-hand figure shows the Voronoi diagram produced when we represent Q using q_3 and q_4, while the bottom right figure shows the Voronoi diagram resulting when we choose q_2 and q_3. Clearly the latter choice produces a Voronoi diagram much

more faithful to the true one.

Existing condensing algorithms cannot handle this situation. First, every element of Q will correctly classify all the other elements; no point will be preferable in this regard. Second, q_3, q_4 and q_5 are nearest to class P, and would be judged boundary points. So although cluster Q is really split into two distinct groups, this is not directly reflected in their distances to each other, or in their distances to the boundary of cluster P. However, one can detect the two subgroups of cluster Q in a natural way by looking at the pairwise redundancy of elements of Q; for example, although $d(q_1, q_5) = 0$, $R(q_1, q_5)$ is small because the distances from q_1 and q_5 to the elements of P are quite different. Noticing this difference between q_1, q_5 is important if we want to do more than classify already seen images well, while keeping the decision boundary similar using only representative subsets.

In summary, by considering a 2-D domain we first see that robust, non-metric distances lead to rather odd, non-intuitive decision boundaries between clusters. Second, we see that for a median distance, the extent to which two images have similar distances to other images can be a good predictor of their interchangeability as class representatives, much better than just the distance between them. Third, we see that boundary points play a much weaker role in determining the important parts of the decision boundary in non-metric spaces than they did in metric spaces. We wish to emphasize that while these results are illustrated in a 2-D domain, it should be clear that they extend to higher dimensions as well.

3.3 Correlation vs. distance

In deciding how well one image can substitute for another as a class representative, we are primarily interested in how similar their distances to new images will be. In other words, we are interested in estimating the redundancy of the two images i_1, i_2: the probability that $|d(i_1, j) - d(i_2, j)|$ is small for some arbitrary image j. We compare two methods of estimating redundancy: (1) the distance between the two images; and (2) the correlation between their *distance vectors*: $\{d(i_1, k) | \forall k \in T, k \neq i_1, i_2\}$ and $\{d(i_2, k) | \forall k \in T, k \neq i_1, i_2\}$, where T is the training set. In Section 3.3.1 we show that when the Euclidean distance is used for classification there is no need to look at correlation, pairwise image distance alone gives an excellent indication of how well one image may substitute for another. In Section 3.3.2 we show that, when a robust distance such as the

median is used, vector correlation gives a reliable estimate of redundancy while pairwise similarity can fail miserably.

3.3.1 Euclidean space: distance predicts redundancy

In Appendix A we show that in the Euclidean space, proximity of points implies large redundancy. More specifically, let P_1, P_2, and P_3 denote three training points in the Euclidean space. Assume $d(P_1, P_2) < d(P_1, P3)$. Let Q denote a new, test point, and let d_i denote the distance from Q to P_i. In the appendix we show that for most distributions of test data Q, with probability greater than $\frac{1}{2}$ we will have $|d_1 - d_2| < |d_1 - d_3|$. We also describe the unlikely form the density on Q would have to take for d_3 to provide a better approximation of d_1 than does d_2. Thus if we use image proximity to predict redundancy, we will only be fooled by special distributions of test cases. (But we are not likely to be fooled if the training sample is taken from the same odd distribution as the test data.)

Relying on a specific estimator of redundancy, such as the correlation of distance vectors, to predict interchangeability, instead of relying on distances, can be either disadvantageous or advantageous: The disadvantage is that similarities of distances to the training set can be noisy. The training set may accidentally contain a preponderance of atypical images, leading to inaccurate predictions. At the same time, as long as the test set is sampled from a symmetric distribution in space, the distance between two points perfectly predicts their "true" similarity no matter how "bad" the sampling of the training set is. The possible advantage of using redundancy in Euclidean spaces is that if the training and test sets both come from the same non-uniform and biased distribution, we will learn this bias from the training set, and use it to better predict performance on the test set.

3.3.2 Median: distance does not predict redundancy

We now consider the case of median distances. For this case, we show that if one image pair is separated by a smaller distance than another, this does not necessarily mean that their redundancy is higher. More specifically, we will show an example where for the median distance, unlike the Euclidean case, the distance between two images is *not* monotonically related to their similarity to new images.

We will now describe an example where this condition does not hold for the median distance.

Let all the coordinates of P_1 differ from those of P_2 by the same small amount ϵ. Let the first half plus one of the coordinates of P_3 be identical to P_1, and the rest be different by the same very large amount σ. The median distance between P_1 and P_2 is ϵ, while the median distance between P_1 and P_3 is 0. Thus, with the median distance, P_3 is closer to P_1 and P_2 is farther than P_1. However, for any point in the space, its distances to P_1 and P_2 cannot differ by more than ϵ. On the other hand, a point can have similar distances to P_1 and P_3 only in two cases. First, if its most similar coordinates to these objects, the ones used to determine the median distance, are the first ones (which P_1 and P_3 share). Second, for special configurations in which it is $\frac{\sigma}{2}$ away from each point in some of its remaining coordinates (other special cases are possible too, for which the new point is at least $\frac{\sigma}{2}$ from both points). Thus $|d_1 - d_2|$ is always small, while $|d_1 - d_3|$ is small only for a tiny fraction of possible test points. This shows that the distance between two points can in fact be an arbitrarily poor predictor of future distance similarity. The closest pair of points need not be the pair with the most similar distances to new points.

At the same time, for the median distance, clearly distance vector correlation is a good predictor of future distance similarity when the images in the test set come from the same distribution as the images in the training set. Moreover, since $R(P_1, P_3) \approx 0$ even if only a small part of the data space is sampled (as long as the sample includes enough points that share more than half their coordinates with P_3 and fewer than half with P_1), the distance vector redundancy will distinguish P_1 and P_3 in spite of the fact that $d(P_1, P_3) = 0$, even when the distribution of the training set is different from the distribution of the test set. Thus, in principle, the distance vector correlation can predict future distances to completely new datapoints, coming from previously un-sampled areas of the dataspace.

3.3.3 Discussion

We have shown that when a Euclidean distance is used, there is no need to look at past distance redundancy. Distance alone is an excellent indication of how well one image may substitute for another. But when a robust distance, such as the median, is used, the picture changes completely. In this case we can obtain more reliable results by focusing instead on past distance redundancy.

4 Comparison of algorithms

We now draw on these insights to produce concrete methods for representing classes in non-metric spaces, for nearest neighbor classification. In the following we will first describe four different condensing algorithms (Section 4.1). In Section 4.2 we will describe a series of simulations comparing the different algorithms under various conditions. In Section 4.3 we will describe a comparison of the different algorithms given a real data set of classes of silhouettes.

4.1 Description of algorithms

We compared the following four algorithms: the first two algorithms, random selection and boundary detection, represent old condensing ideas; the last two algorithms, atypical selection and correlation cover, apply new ideas discussed above for class representation in non-metric spaces. Each algorithm selects a representative set of examples $\mathcal{S}$ of size Q_c.

In describing these algorithms, we use the word *cover* as follows: Let $\mathcal{P}$ denote a class of images. For $p_1, p_2 \in \mathcal{P}$, p_1 *covers* p_2 if and only if $d(p_1, p_2) < d(p_2, q), \forall q \notin \mathcal{P}$. That is, choosing p_1 as a representative guarantees correct classification of p_2.

Random selection: for every class $\mathcal{C}$, compute $\mathcal{S}$ by: randomly (but without repetitions) choose Q_c examples from $\mathcal{C}$.

Every algorithm for the selection of class representation should be able to perform better than this simplest brute-force algorithm.

Boundary detection: for every class $\mathcal{C}$, compute $\mathcal{S}$ as follows:

1. compute an approximation to the minimal cover of size $\leq Q_c$ using a greedy algorithm
2. until size of $\mathcal{S}$ is Q_c, add boundary points which are furthest from $\mathcal{S}$

This algorithm is described in detail in appendix B.1.

Part 1 of this algorithm resembles the first iteration of Dasarathy's algorithm [6]; subsequent iterations were not found by [6] to significantly affect the results. While capturing the essence of most condensing algorithms which look for class boundaries, our implementation

ignores important differences which address the issue of computational efficiency, since such differences are not relevant for our purpose here.

Atypical selection: for every class $\mathcal{C}$, compute $\mathcal{S}$ as follows:

1. compute an approximation to the minimal cover of size $\leq Q_c$ using a greedy algorithm
2. until size of $\mathcal{S}$ is Q_c, add atypical points (not necessarily on the boundary) which are furthest from $\mathcal{S}$

This algorithm is described in detail in appendix B.2.

This algorithm tests our second observation, that class boundaries fail to capture important class structure, and that "atypical" points - which are far from the representative set - should also be included in the class representation.

Selection based on correlation: our most important computational observation is that in non-metric spaces looking only at the distances between datapoints is not sufficient; we argued that some comparison of the distances to other points should be a better measure of "similarity" between datapoints. The following implementation tests this idea.

We compare datapoints by simply measuring how well their vectors of distances are correlated. More specifically, given two datapoints X, Y and their corresponding distance vectors $\mathbf{x}, \mathbf{y} \in \mathcal{R}^n$, where $\mathbf{x}$ is the vector of distances from X to all the other training points and $\mathbf{y}$ is the vector of distances from Y to all the other training points, we measure the correlation between the datapoints using the statistical correlation coefficient between $\mathbf{x}, \mathbf{y}$ (in future work, we will investigate other measures of correlation):

$$corr(X, Y) = corr(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} - \mu_x}{\sigma_x} \cdot \frac{\mathbf{y} - \mu_y}{\sigma_y}$$

Above μ_x, μ_y denote the mean of $\mathbf{x}, \mathbf{y}$ respectively, and σ_x, σ_y denote the standard deviation of $\mathbf{x}, \mathbf{y}$ respectively.

We explored two ways of using correlation between datapoints. Version 1 below is significantly less efficient than the previous algorithms, since it requires the computation of correlation between any two datapoints. Version 2 only requires the computation of correlation between any two datapoints within the same class. Since both versions performed

roughly the same in all our tests on both simulated and real data, version 2 is our preferred algorithm; in subsequent discussions only results with version 2 are described.

Version 1:

Repeat the atypical selection algorithm as described above in B.2, but whenever the distance between two datapoints is used - use instead the correlation between them.

Version 2:

For every class $\mathcal{C}$, compute $\mathcal{S}$ as follows: using a greedy algorithm, choose Q_c points that maximize V_X - a combined measure of distance cover and correlation at each datapoint X.

Define V_X as $V_X = (N_X + 1) \cdot \frac{1-Corr_X}{2}$, where: $Corr_X$ is the maximal correlation of X with points in $\mathcal{S}$, and N_X is the number of not-yet-covered points which get covered by X.

This algorithm is described in detail in appendix B.3.

Note that this algorithm chooses representatives by combining N_X, which measures how well each point covers previously uncovered points in terms of distances, with $1 - Corr_X$, which indicates how atypical (measured by how uncorrelated) the point is relative to previously chosen points.

4.2 Simulations

To compare the four algorithms described above, we simulated data representing various conditions:

1. For simplicity, each datapoint was chosen to be a vector in $\mathcal{R}^7$ or $\mathcal{R}^{25}$. 30 clusters were randomly chosen, each with 30 datapoints.

2. To study how the structure of the data affect the performance of each algorithm, we simulated 3 cases:

   (a) **"Vanilla":** the points in each class form a small cluster in the data vector space. Specifically, the center of each cluster is chosen randomly in space, and the class members are chosen from a spherical normal distribution spread around the chosen center.

(b) **One outlier:** the points in each class cluster around a prototype, but many class members vary widely in one dimension (which may be different for the different class members). Specifically, the center of each cluster is chosen randomly in space, and the class members are chosen from a spherical normal distribution; for about half the points, however, one coordinate (randomly chosen) takes an arbitrary value totally different from the center value.

(c) **Irrelevant features:** the points in each class cluster around a prototype, but many class members vary widely in a small number of dimensions (less than half, and fixed for the different class members). Specifically, the center of each cluster is chosen randomly in space, and the class members are chosen from one of 2 normal distributions spread around the chosen center: half the points are chosen from a spherical normal distribution, and half the points are chosen from an elongated elliptical normal distribution.

In case (a) above robust distances are not really required, whereas in cases (b),(c) robust methods are needed to guarantee good performance.

3. We simulated 4 distance functions: Euclidean (L^2), $L^{0.5}$, $L^{0.2}$, and median. Note that the middle two are non-metric but bounded, i.e., the violation of the triangle inequality is bounded by a constant scaling factor, while the median distance can arbitrarily violate the triangle inequality (see Appendix C).

During the simulations, 1000 test datapoints were randomly chosen from a uniform distribution in $\mathcal{R}^7$ and $\mathcal{R}^{25}$. Each test datapoint was classified based on: (1) all the data, (2) the representatives computed by each of the four algorithms described in Section 4.1. For each algorithm, the test is successful if the two methods (classification based on all the data and based on the chosen representatives) give the same results. Thus for each algorithm we attach a score of percent correct: the percentage of test datapoints that scored success in this simulation block. We repeated each block 20 times, to gather statistics on the variability in the percent correct value (mean and standard deviation).

Fig. 6 summarizes representative results of some of our simulations. Note that our test data comes from a different distribution than the training set. By using uniformly distributed test data,

we estimate the volume of the difference in the voronoi diagrams produced by the complete set and the representative subset.

4.3 Test with Real data

To test our method with real images, we used the local curve matching algorithm described in [9]. This curve matching algorithm is designed to compare curves which may be quite different, and return the distance between them. The steps of the algorithm include: (1) the automatic extraction of feature points with relatively high curvature, (2) feature matching using dynamic programming and efficient alignment; the final distance is the median of the inter-feature distances. Thus this algorithm is non-metric, due to both the noisy selection of features and the final median step.

In our test we used 2 classes, with 30 images each. The first set included images of 2 similar cars from different points of view. The second set included images of a cow from different points of view, including occluded images. Two of the original images are shown in Fig. 7d. The contours from the cow class, which were automatically extracted, are shown in Fig. 7a, and the contours from the car class are shown in Fig. 7b. 30 images were used as test images; the contours extracted from these images are shown in Fig. 7c. These are also cow contours: some were obtained from different viewpoints of the same cow, and some from the same viewpoints with more occlusion.

We used the four algorithms described in Section 4.1 to compute representative sets of 5 and 7 contours for each class. We then compared the classification of the test data based on these representatives, to the classification obtained using all the data. Results (percent correct scores) are given in Fig. 8.

4.4 Discussion of results

In Fig. 6b,c the correlation-based algorithm performs significantly better than any other algorithm given any of the three non-metric distance functions; given the Euclidean distance, its performance is similar (or slightly, but significantly, better) as compared to the boundary and atypicals methods. Surprisingly, the boundary method performs **significantly worse** than a random selection of representatives with the median distance and the $L^{0.2}$ distance.

In Fig. 6a, which represents the "vanilla" case where the data lacks any "interesting" structure and where a class is just a clump of entities which are truly close to each other and well separated

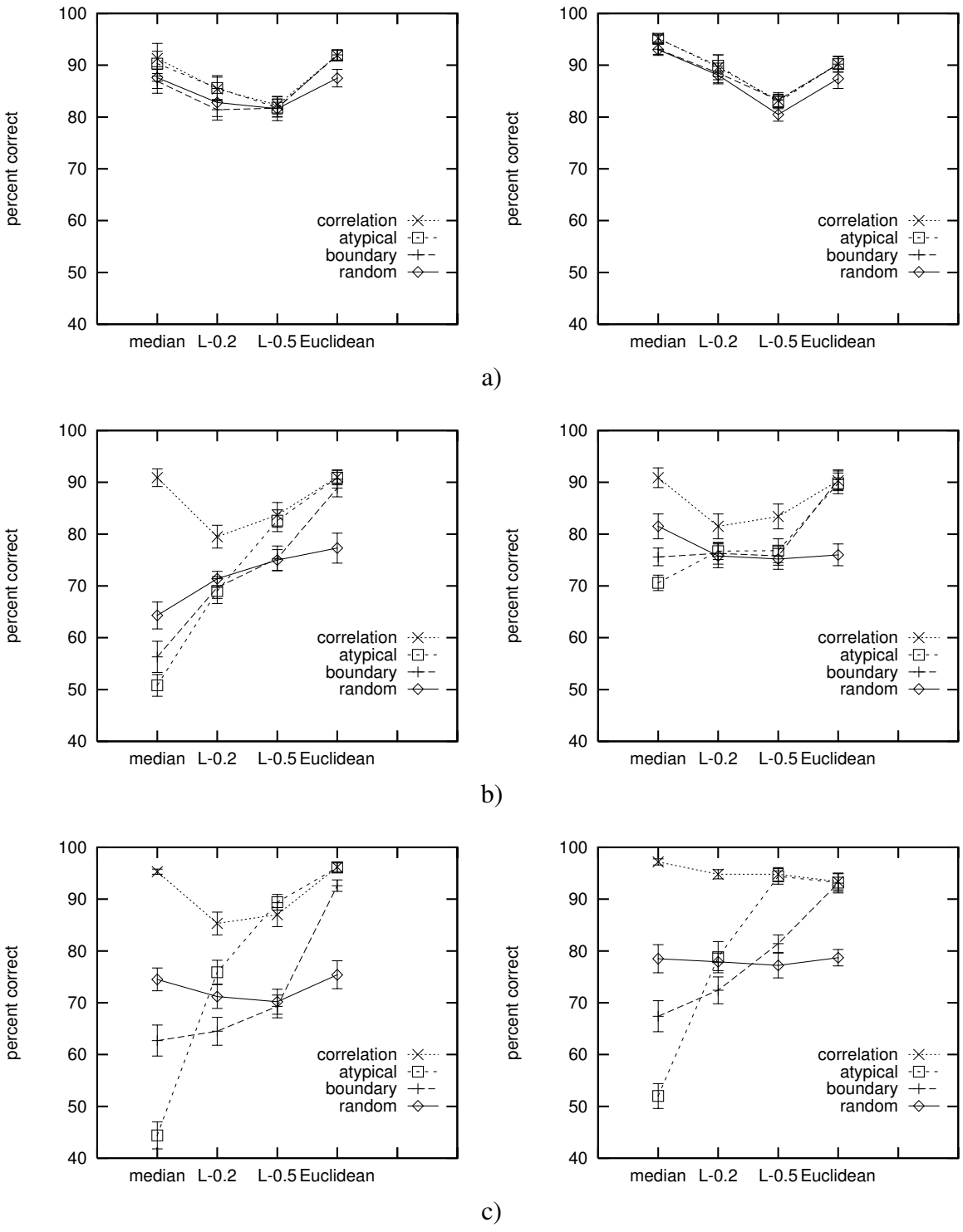
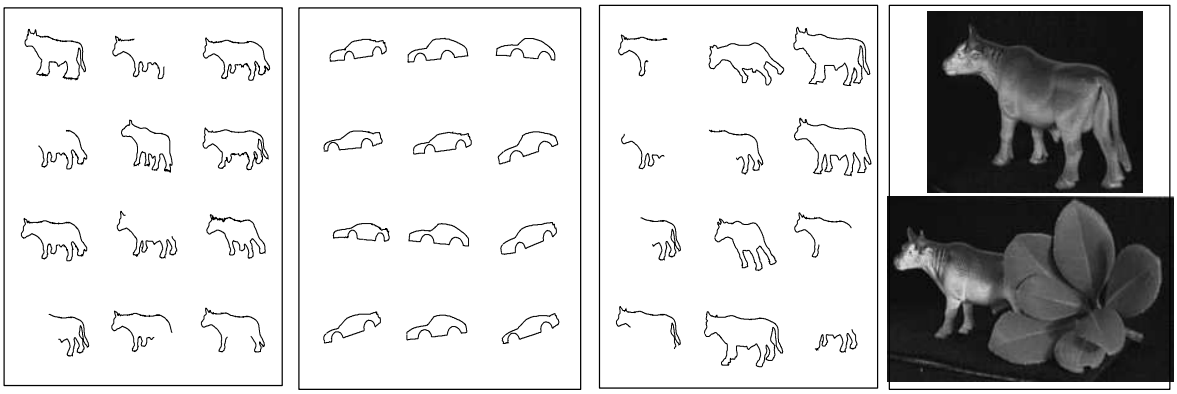


Figure 6: Simulation results, with data chosen from $\mathcal{R}^7$ (left column) and $\mathcal{R}^{25}$ (right column). Plotted are the values of percent correct scores, as well as error bars giving the standard deviation calculated over 20 repetitions. Each graph contains 4 plots, giving the percent correct score for each of the four algorithms described in Section 4.1: random (selection), boundary (detection), atypical (selection), and (selection based on) correlation. Each plot gives 4 values, for each of the different distance functions used here: median, $L^{0.2}$, $L^{0.5}$ and L^2. (a) Data is chosen from a normal distribution. (b) Data is chosen from a normal distribution, where in half the datapoints one coordinate significantly differs from the distribution's center. (c) Data is chosen from 2 concentric normal distributions, one spherical and one elongated elliptical.



a) b) c) d)

Figure 7: Real data used to test the four algorithms: a) 12 examples from the first class of 30 cow contours. b) 12 examples from the second class of 30 car contours. c) 12 examples from the set of 30 test contours. d) 2 examples of the real images from which the contours in a) were obtained.

from the other classes, all the methods are comparable in performance. In this case, one would not actually need to make use of a robust, non-metric distance function. Occasionally, especially with the Euclidean distance, the random selection performs significantly worse than the other methods, although its score is not much lower (and the difference may not be worth the additional effort).

Finally, the results with the real data as shown in Fig. 8 also confirm this picture: The correlation method performs much better than any other method. The boundary method does only slightly (but probably not significantly) better than random.

Also note that the atypical selection algorithm for which we show results here computes atypicals based on distance, and does not perform noticeably better than the other distance-based algorithm. As mentioned above, although we do not show results here, the same algorithm, computing covers and atypicals using correlation, performs approximately as well as the other correlation-based algorithm shown here.

4.5 Summary of results

Both the simulations and the real data clearly demonstrate an advantage to our method over existing methods in the classification of data in non-metric dataspaces. Almost as importantly, in metric spaces (4th column in Fig. 6a-c) or when the classes lack any "interesting" structure (Fig. 6a), our method is not worse than any existing method. Thus it should be generally used to guarantee good

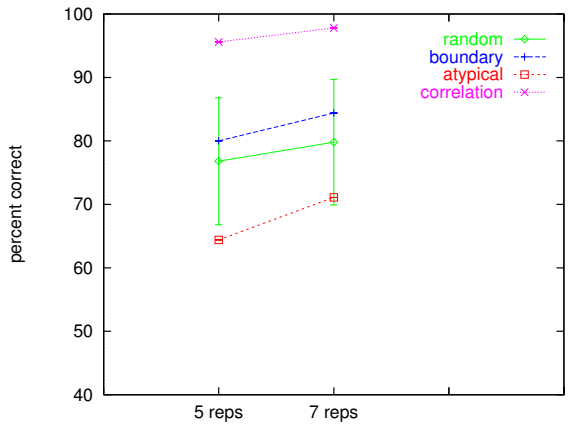


Figure 8: Results with real data: each graph contains 4 plots, giving the percent correct score for each of the four algorithms described in Section 4.1: random (selection) where standard deviation of performance is also plotted, boundary (detection), atypical (selection), and (selection based on) correlation. The number of representatives chosen by the algorithm was limited to 5 (first column) and 7 (second column).

performance when the nature of the data and the distance function is not known a priori.

Note that although the random method sometimes performs as well as the other methods, it does not provide any criteria as to how many points should be selected for the representative set. This is a rather important hidden advantage of the three principled methods over the random method, especially the correlation method which uses a threshold to determine how many points should be selected.

5 Extension: Parametric methods

Up to now, we have considered non-parametric methods of supervised learning, in particular nearest neighbor classification. This is naturally suited to learning in non-metric spaces, and can be used when our distance function is a black box. However, parametric methods have been very successful in Euclidean spaces. In this section we consider the extension of parametric methods to data described by non-metric distances. We start by relaxing our assumption that the distance function is a black box, and consider specific non-metric distances.

Parametric methods typically represent objects as vectors in a high-dimensional space, and represent classes and the boundaries between them in this space using geometric constructions or probability distributions with a limited number of parameters. Some of the simplest and most

popular parametric methods look for a hyperplane that best separates two classes. Up to now we have considered our distance function as a black box, so that the construction of objects like hyperplanes simply wasn't possible. But even if we considered distances defined over a feature vector space, such as the robust median distance, Fig. 1 tells us that linear discriminants and hyperplane separators are still inappropriate.

A linear discriminant between two classes in the Euclidean space is effective when each class can be represented by a prototypical example. This follows the simple observation that using a linear discriminant is equivalent to performing nearest neighbor classification with respect to two points (i.e., the two prototypical examples representing each class), where the separating hyperplane is perpendicular to the line connecting these two points and located midway between them. Thus the most obvious generalization of linear discrimination to non-metric space is via the use of prototypical examples and nearest neighbor classification (or networks of Radial Basis Functions).

More specifically, we propose that the natural generalization is to seek a prototype instance of each class and perform nearest neighbor classification using these prototypes, and using the distance function that is appropriate to the space. Optimal prototypes are defined as points in the vector space such that nearest neighbor classification according to the given distance produces correct classification of the training set. Unfortunately, finding the optimal prototypes for most distance functions, including the median, is prohibitively computationally demanding.

A simpler approach is to represent each class by its generalized "centroid". In the Euclidean space, the centroid (or mean) is the point $\bar{q}$ whose sum of squared distances to all the class members $\{q_i\}_{i=1}^n$ - $\left(\sum_{i=1}^n d(\bar{q}, q_i)^2\right)^{\frac{1}{2}}$ - is minimized. Suppose now that our data come from a vector space where the correct distance is the L^p distance from (1). Using the natural extension of the above definition, we should represent each class by the point $\bar{q}$ whose sum of distances to all the class members

$$\left(\sum_{i=1}^{n} d(\bar{q}, q_i)^p\right)^{\frac{1}{p}}$$

is minimal. It is now possible to show that for $p < 1$ (the non-metric cases), the exact value of every feature of the representative point $\bar{q}$ must have already appeared in at least one element in the class[1] (see [7] for discussion of closely related results).

[1]The proof goes as follows: since x^p is monotonic, the minimum of the sum, E, is obtained for the same $\bar{q}$ as the

Moreover, the value of these features can be determined separately with complexity $O(n^2)$ or less (less for $p = 1$, where the median feature obtains the minimum), and total complexity of $O(dn^2)$ given d features. $\bar{q}$ is therefore determined by a mixture of up to d exemplars, where d is the dimension of the vector space. Thus there are efficient algorithms for finding the "centroid" element of a class, even using certain non-metric distances. Moreover, the point which replaces the centroid in L^p spaces for $p < 1$ contains values from at most d datapoints, which means that the representation is "sparse" - a desirable property for data compression.

We will illustrate these results with a concrete example using the corel database, a large commercial database of images pre-labeled by different categories (such as "lions"), where non-metric distance functions have proven effective in determining the similarity of images [5]. The corel database is very large, making the use of prototypes desirable.

We represent each image using a vector of 11 numbers describing general image properties, such as color histograms, as described in [5]. We consider the Euclidean and $L^{0.5}$ distances, and their corresponding prototypes: the Euclidean mean and the $L^{0.5}$-prototype computed according to the result above. Given the first 5 classes, each containing 100 images, we found their corresponding prototypes; we then computed the percentage of images in each class which are closest to their own prototype, using either the Euclidean or the $L^{0.5}$ distance and one of the two prototypes. The results are the following:

prototype:	mean	11 existing features
$L^{0.5}$ distance	68%	80%
Euclidean distance	72%	73%

In the first column, the prototype is computed using the Euclidean mean. In the second column the centroid is computed using the $L^{0.5}$ distance. In each row, a different function is used to compute the distance from each item to the cluster prototype. Best results are indeed obtained with the non-metric $L^{0.5}$ distance and the correct centroid for this particular distance as prototype. While performance in absolute terms depends on how well this data clusters using distances derived from a simple feature vector, relative performance of different methods reveals the advantage of using a prototype computed with a non-metric distance.

minimum of E^p. But E^p is separable in the features, thus the minimum can be computed for each feature separately. Finally, since for $p < 1$ x^p is concave, the minimum for each feature must equal the feature value at some point.

Another important distance function is the generalized Hamming distance: given two vectors of features, their distance is the number of features which are different in the two vectors. This distance was assumed in psychophysical experiments which used artificial objects (Fribbles) to investigate human categorization and object recognition [28]. In agreement with experimental results, the prototype $\bar{q}$ for this distance computed according to the definition above is the vector of "modal" features - the most common feature value computed independently at each feature.

6 Conclusions

We feel that our paper makes two main contributions. First, we reassessed the relevance of existing supervised classification techniques to application-based and human-like classification. We argued that classification systems that can model human performance, or use robust matching methods that are typically used in successful applications, make use of similarity judgments that are non-metric, and in particular, do not obey the triangle inequality. In this case, most existing pattern recognition techniques are not relevant. Exemplar-based methods, however, can be applied naturally when using a wide class of non-metric similarity functions. The key issue, however, in applying exemplar-based methods in such settings is to find methods for choosing good representatives of a class, that accurately characterize it.

We then focused on developing techniques for solving this problem, emphasizing two points: First, we showed that the distance between two images is not a good measure of how well one image can represent another in non-metric spaces. Instead, we suggested considering the correlation between the distances from each image to other previously seen images. Second, we showed that in non-metric spaces, boundary points are less significant for capturing the structure of a class than they are in the Euclidean space. We suggested that atypical points may be more important in describing classes. We demonstrated the importance of these ideas in greatly improving classification results using both synthetic and real images.

Finally, we have suggested ways of applying parametric techniques to supervised learning problems that involve a specific, non-metric distance function. We have shown how to generalize the idea of linear discriminant functions in a way that may be more useful in non-metric spaces. And we have shown a "proof-of-concept" for classification using the centroid of a class as determined

by the specific distance, with L^p distances for $p < 1$.

A Proof: distance predicts redundancy in the Euclidean space

This appendix shows that in the Euclidean space, proximity of points implies that they are more likely to have similar distances to new points.

Let P_1, P_2, and P_3 denote three training points in the Euclidean space. Assume $d(P_1, P_2) < d(P_1, P3)$. Let Q denote a new, test point, and let d_i denote the distance from Q to P_i. If we consider Q to be a random variable, then we wish to know the probability that d_1 is more similar to d_2 than to d_3, that is, that $|d_1 - d_2| < |d_1 - d_3|$. We address this question by first dividing the space into two disjoint subsets, $\mathcal{G}$ and $\mathcal{H}$, where

$$\begin{aligned} Q \in \mathcal{G} &\iff |d_1 - d_2| < |d_1 - d_3| \\ Q \in \mathcal{H} &\iff |d_1 - d_2| > |d_1 - d_3| \end{aligned}$$

We then show that for many reasonable distributions of Q, $Pr(Q \in \mathcal{G}) > \frac{1}{2}$.

Some notations

The following definitions are illustrated in Fig. 9. First, without loss of generality we may assume that P_1, P_2, and P_3 lie in the X-Y plane. Let Q be a point in $\mathcal{R}^n$. Define l_{ij} to be the line connecting P_i and P_j. Define n_{ij} to be the hyperplane normal to l_{ij} and passing through the point on l_{ij} equidistant to P_i and P_j. Note that n_{12}, n_{23}, and n_{13} generally all intersect the X-Y plane at a common point, O. To see this, note that n_{ij} describes the set of points equidistant between P_i and P_j. Also, n_{ij} intersects the X-Y plane in a line (normal to l_{ij}). Therefore, n_{12} and n_{23} generally intersect the X-Y plane in a point[2] that is equidistant between P_1 and P_2, and between P_2 and P_3. This point is also obviously equidistant between P_1 and P_3, and so lies on n_{13}.

Next we define our coordinate system so that O is at the origin. To simplify our exposition, we first consider a compact subset of the Euclidean space, consistent of a huge hypersphere centered at the origin.

[2] In a non-generic case, n_{12} and n_{23} may be parallel when P_1, P_2 and P_3 are collinear. In that case, we may say that they intersect at a point at infinity in the projective plane. This may be handled in a manner similar to the reasoning given below, but we do not explicitly consider it here.

Without loss of generality, we may assume that the distance from O to $P_i, 1 \leq i \leq 3$ is 1, i.e., all P_i lie on the unit circle. We define the X axis to be the line bisecting P_2 and P_3, i.e., the intersection of n_{23} and the X-Y plane. Let the coordinates of P_i be (x_i, y_i), then $x_2 = x_3$, and $y_3 = -y_2$. Again without loss of generality we suppose $x_2 > 0, y_2 < 0$. The fact that P_1 is closer to P_2 than P_3 implies that $y_1 < 0$. Next, consider the triangle formed by P_1, P_2, P_3. Define the three angles of this triangle to be a, b, c, where a is the angle at the vertex P_1, and b and c are angles at the vertices P_2 and P_3 (see Fig. 9).

Each hyperplane n_{ij} divides the space into two regions, one in which points are nearer to P_i than P_j, the other its complement. Together, the 3 hyperplanes divide the space into six regions. The regions are defined as follows:

$$
\begin{aligned}
d_2 < d_1 < d_3 &\equiv Q \in \mathcal{A}_1 \\
d_3 < d_1 < d_2 &\equiv Q \in \mathcal{A}_2 \\
d_1 < d_2 < d_3 &\equiv Q \in \mathcal{B}_1 \\
d_3 < d_2 < d_1 &\equiv Q \in \mathcal{B}_2 \\
d_1 < d_3 < d_2 &\equiv Q \in \mathcal{C}_1 \\
d_2 < d_3 < d_1 &\equiv Q \in \mathcal{C}_2
\end{aligned}
$$

We then note that in the X-Y plane each of these regions is triangle shaped with a vertex at O bounded by two lines meeting at O and by a circular arc at the boundary of the space. For example, $\mathcal{A}_1$ is just the wedge of space defined by the intersection of one half-space bounded by n_{12} and one bounded by n_{13}. Note that the angle of this intersection is a. In general, the angle at O of $\mathcal{A}_1$ and $\mathcal{A}_2$ is a, of $\mathcal{B}_1$ and $\mathcal{B}_2$ is b, and of $\mathcal{C}_1$ and $\mathcal{C}_2$ is c. Therefore the volume of $\mathcal{A}_1$, relative to the volume of the compact space, is $\frac{a}{2\pi}$.

Now, we note that $\mathcal{B}_1, \mathcal{B}_2 \subseteq \mathcal{G}$, and $\mathcal{C}_1, \mathcal{C}_2 \subseteq \mathcal{H}$. This follows directly from their definition. For example, if $Q \in \mathcal{B}_1$, then d_2 is in between d_1 and d_3, and must be closer to d_1 than d_3 is. However, $\mathcal{A}_1$ and $\mathcal{A}_2$ are divided between $\mathcal{G}$ and $\mathcal{H}$, depending on whether d_1, which is in between d_2 and d_3, is closer to one or the other. Furthermore, note that $b > c$. This follows from the law of sines, and the fact that $d(P_1, P_2) < d(P_1, P_3)$. Note also that this implies that $c < \frac{\pi}{2}$, which further implies that the volume of $\mathcal{C}_1 \cup \mathcal{C}_2$ must consume less than half the space.

Derivation: lower bound on the relative volume of $\mathcal{G}$

We now prove that the volume of $\mathcal{G}$ consists of at least half the space. This implies, for example, that if we sample Q from the space with a uniform distribution (or any distribution rotationally symmetric about the origin) than $Pr(Q \in \mathcal{G}) > \frac{1}{2}$.

The following definitions are illustrated in Fig. 9. Suppose for the moment that $x_1 < x_2$, and define $P_{1,min} = (-1, 0)$. Note that $P_{1,min}$ is equidistant to P_2 and P_3. Define $P_{1,max} = (-x2, y2)$. We say that a point is in between two other points on the circle if it lies in the smaller portion of the circle between them. First, we note that if P_1 lies between $P_{1,max}$ and P_2, then the angle b is greater than $\frac{\pi}{2}$. Since the region $\mathcal{B}_1 \cup \mathcal{B}_2 \subseteq \mathcal{G}$ has angle $2b$, $\mathcal{G}$ will occupy more than half the space, which completes the proof for this case. Therefore w.l.o.g. suppose that P_1 lies between $P_{1,min}$ and $P_{1,max}$.

Our proof consists of two steps: first we show that in the limit when $P_1 = P_{1,min}$, the volume of $\mathcal{G}$ is exactly half the space. Then we show that as we move P_1 away from $P_{1,min}$ and towards $P_{1,max}$, the volume of $\mathcal{G}$ can only grow.

case 1: $P_1 = P_{1,min}$

When $P_1 = P_{1,min}$, by symmetry we can see that the space must divide evenly between $\mathcal{H}$ and $\mathcal{G}$. Specifically, suppose $Q = (x, y, z, ...) \in \mathcal{H}$. Define $Q_{opp} = (x, -y, z, ...)$, i.e., Q with the sign reversed on its y coordinate. Then $d(Q, P_1) = d(Q_{opp}, P_1), d(Q, P_2) = d(Q_{opp}, P_3)$ and $d(Q, P_3) = d(Q_{opp}, P_2)$; this implies that $Q_{opp} \in \mathcal{G}$. Therefore, when $P_1 = P_{1,min}$, $\mathcal{G}$ and $\mathcal{H}$ are symmetric about the hyperplane normal to the Y axis - n_{23}, and the relative volume of each is exactly $\frac{1}{2}$.

Define $\mathcal{G}'$ to be the region $\mathcal{G}$ when $P_1 = P_{1,min}$. Also define $\mathcal{A}'_1$ to be the region $\mathcal{A}_1$ when $P_1 = P_{1,min}$, and similarly define $\mathcal{A}'_2, \mathcal{B}'_1, \mathcal{B}'_2, C'_1, C'_2$ and n'_{ij}. Next we will show that $\mathcal{G}' \subseteq \mathcal{G}$ when P_1 has any value between $P_{1,min}$ and $P_{1,max}$.

Some geometrical orderings:

We start by ordering the various hyperplanes according to their line of intersection with the X-Y plane. We specifically order by the angle that this line makes with the X-axis in counterclockwise

direction, giving us the following 2 possible orders (see Fig. 9):

$$\begin{aligned} X \;=\; 0^o \;=\; n'_{23} \;=\; n_{23} \;\leq\; n'_{12} \;\leq\; n_{12} \;\leq\; n'_{13} \;\leq\; n_{13} \;\leq\; 180^o \\ X \;=\; 0^o \;=\; n'_{23} \;=\; n_{23} \;\leq\; n'_{12} \;\leq\; n'_{13} \;\leq\; n_{12} \;\leq\; n_{13} \;\leq\; 180^o \end{aligned} \tag{2}$$

Define a new hyperplane, M, which passes through the midpoint between $P_{1,min}$ and P_1, and is normal to the vector between these points. M divides the space between points closer to P_1 and those closer to $P_{1,min}$ (see Fig. 9). Since M is perpendicular to the vector connecting $P_{1,min}$ and P_1 while n'_{12} is perpendicular to the vector connecting $P_{1,min}$ and P_2, and since n'_{23} is the x-axis, M must lie between n'_{23} and n'_{12} which bound the regions $\mathcal{B}'_1, \mathcal{B}'_2$. From (2) and $n'_{23} \;\leq\; M \;\leq\; n'_{12}$ it follows that M lies completely inside $\mathcal{B}'_1$ and $\mathcal{B}'_2$. It subsequently follows that all points in $\mathcal{A}'_2$, the region bounded by n'_{12} and n'_{13} and lying above n'_{23} (above the X axis), are also above M and thus further from P_1 than they are from $P_{1,min}$.

case 2: $P_1 \in [P_{1,min}, P_{1,max}]$

We now show if $Q \in \mathcal{G}'$ then $Q \in \mathcal{G}$.

First note that (2) implies that $\mathcal{A}'_2 \subseteq \mathcal{A}_2 \cup \mathcal{B}_2$, $\mathcal{B}'_2 \subseteq \mathcal{B}_2$ and $\mathcal{B}'_1 \subseteq \mathcal{B}_1$. Thus if $Q \in \mathcal{B}'_1$ or $Q \in \mathcal{B}'_2$ then $Q \in \mathcal{G}$. If $Q \in \mathcal{A}'_2 \cap \mathcal{G}'$ then Q is in either $\mathcal{A}_2$ or $\mathcal{B}_2$ (and not $\mathcal{C}_1$ or $\mathcal{C}_2$). If $Q \in \mathcal{B}_2$ then $Q \in \mathcal{G}$. Thus we only need to consider the cases where $Q \in \mathcal{A}'_2 \cap \mathcal{A}_2 \cap \mathcal{G}'$ and $Q \in \mathcal{A}'_1 \cap \mathcal{A}_1 \cap \mathcal{G}'$:
$Q \in \mathcal{A}'_2$ implies that:

$$d(Q, P_3) < d(Q, P_{1,min}) < d(Q, P_2)$$

$Q \in \mathcal{G}'$ implies that:

$$d(Q, P_{1,min}) - d(Q, P_3) > d(Q, P_2) - d(Q, P_{1,min})$$

and so

$$2d(Q, P_{1,min}) > d(Q, P_2) + d(Q, P_3)$$

We have shown that $Q \in \mathcal{A}'_2$ also implies:

$$d(Q, P_1) > d(Q, P_{1,min})$$

Thus

$$2d(Q, P_1) > d(Q, P_2) + d(Q, P_3) \implies d(Q, P_1) - d(Q, P_3) > d(Q, P_2) - d(Q, P_1)$$

Since $Q \in \mathcal{A}_2$, it follows that $Q \in \mathcal{G}$. Similar reasoning shows that $Q \in \mathcal{A}_1' \cap \mathcal{G}' \Rightarrow Q \in \mathcal{G}$. In addition, although our reasoning so far has assumed that $x_1 < x_2$, similar arguments show that the same is true for $x_1 > x_2$.

This proves our result, that $\mathcal{G}$ always contains at least half of the space. Thus, if the test points Q are sampled from a uniform distribution on our compact space, then d_2 will be closer to d_1 than d_3 is to d_1 with probability larger than $\frac{1}{2}$, and P_2 will provide a better substitute for P_1 than does P_3. This conclusion remains true if the test points are sampled from a distribution which is spherically symmetric about O on the whole Euclidean space, or any distribution that does not assign to region $\mathcal{H}$ probability larger than its relative volume.

B Detailed description of condensing algorithms

B.1 Boundary detection:

For every class $\mathcal{C}$, compute its representative set of examples $\mathcal{S}$ of size Q_c, using the following algorithm:

1. compute cover

   (0) Initialization:

   - set of representative examples $\mathcal{S}$ is empty
   - set of examples which are incorrectly classified $\mathcal{A} \leftarrow \mathcal{C}$

   (1) For every datapoint c in $\mathcal{C} \setminus \mathcal{S}$, compute N_c - the number of points in $\mathcal{A}$ closer to c than to any point in any other class (i.e., the number of points covered by c).

   (2) Find $\bar{c}$ with the largest value of N_c over all $c \in \mathcal{C} \setminus \mathcal{S}$. Let $T_{\bar{c}}$ denote the group of points in $\mathcal{A}$ closer to $\bar{c}$ than to other classes (i.e., the points covered by $\bar{c}$).

   (3) update:

   - $\mathcal{A} \leftarrow \mathcal{A} \setminus T_{\bar{c}}$
   - $\mathcal{S} \leftarrow \mathcal{S} \cup \{\bar{c}\}$

   $\Longleftarrow$ while size of $\mathcal{S}$ is smaller than Q_c and $\mathcal{A}$ not empty, return to (1)

2. Add boundary points: while size of $\mathcal{S}$ is smaller than Q_c, repeat:

   (1) $\forall c \in \mathcal{C}$, compute

   - D_c - the distance of c to the nearest datapoint in *another* class
   - d_c - the distance of c to the nearest datapoint in the representative set $\mathcal{S}$
   - $\Delta_c = d_c - D_c$

   (2) Find $\bar{c}$ which produces the largest Δ_c over all $c \in \mathcal{C}$.

   (3) update:

   - $\mathcal{S} \leftarrow \mathcal{S} \cup \{\bar{c}\}$

B.2 Atypical selection:

For every class $\mathcal{C}$, compute its representative set of examples $\mathcal{S}$ of size Q_c, using the following algorithm:

1. compute cover, same as described in B.1 for the boundary algorithm.

2. Add atypical points: while size of $\mathcal{S}$ is smaller than Q_c, repeat:

   (1) $\forall c \in \mathcal{C} \setminus \mathcal{S}$, compute

   - d_c - the distance of c to the nearest datapoint in the representative set $\mathcal{S}$

   (2) Find $\bar{c}$ which obtains the largest d_c over all $c \in \mathcal{C} \setminus \mathcal{S}$.

   (3) update:

   - $\mathcal{S} \leftarrow \mathcal{S} \cup \{\bar{c}\}$

B.3 Selection based on correlation:

For every class $\mathcal{C}$, compute its representative set of examples $\mathcal{S}$ of size Q_c, using the following algorithm:

(0) Initialization:

- set of representative examples $\mathcal{S}$ is empty

- set of examples which are incorrectly classified $\mathcal{A} \leftarrow \mathcal{C}$

(1) For every datapoint c in $\mathcal{C} \setminus \mathcal{S}$, compute

- N_c - the number of points in $\mathcal{A}$ closer to c than to any point in any other class
- $Corr_c$ - the maximal correlation of c with points in $\mathcal{S}$
- $V_c = (N_c + 1) \cdot \frac{1 - Corr_c}{2}$

(2) Find $\bar{c}$ which obtains the largest V_c over all $c \in \mathcal{C} \setminus \mathcal{S}$. Let $T_{\bar{c}}$ denote the group of points in $\mathcal{A}$ closer to $\bar{c}$ than to other classes (i.e., the points covered by $\bar{c}$).

(3) update:

- $\mathcal{A} \leftarrow \mathcal{A} \setminus T_{\bar{c}}$
- $\mathcal{S} \leftarrow \mathcal{S} \cup \{\bar{c}\}$

$\Longleftarrow$ while size of $\mathcal{S}$ is smaller than Q_c and $V_c > t$ for some threshold t, return to (1)

C A bounded triangle inequality for non-metric L^p distances

It is known that for $p < 1$, the p-distance defined in (1) does not satisfy the triangle inequality: given 3 points $q_1, q_2, q_3 \in \mathcal{R}^n$, $d(q_1, q_3) > d(q_1, q_2) + d(q_2, q_3)$. More specifically, let $\mathbf{x}$ denote the vector connecting q_1, q_2, and $\mathbf{y}$ denote the vector connecting q_2, q_3; then

$$(\sum_{1}^{n} |x_i + y_i|^p)^{\frac{1}{p}} > (\sum_{1}^{n} |x_i|^p)^{\frac{1}{p}} + (\sum_{1}^{n} |y_i|^p)^{\frac{1}{p}}$$

We will now show that there exists $\kappa > 1$ such that the κ-bounded triangle inequality is satisfied: $d(q_1, q_3) \leq \kappa(d(q_1, q_2) + d(q_2, q_3))$; more specifically,

$$(\sum_{1}^{n} |x_i + y_i|^p)^{\frac{1}{p}} \leq \kappa((\sum_{1}^{n} |x_i|^p)^{\frac{1}{p}} + (\sum_{1}^{n} |y_i|^p)^{\frac{1}{p}}) \tag{3}$$

First, from the concavity of $f(t) = t^p$ for $p < 1$, and for $a, b \geq 0$:

$$\frac{1}{2^{1-p}}(a^p + b^p) \leq (a + b)^p \leq a^p + b^p$$

where the first inequality follows from $\frac{(a^p+b^p)}{2} \leq (\frac{a+b}{2})^p$. Let $a = (\sum_1^n |x_i|^p)^{\frac{1}{p}}$, $b = (\sum_1^n |y_i|^p)^{\frac{1}{p}}$; then

$$\begin{aligned}
& ((\sum_1^n |x_i|^p)^{\frac{1}{p}} + (\sum_1^n |y_i|^p)^{\frac{1}{p}})^p = (a+b)^p \\
\geq \; & \frac{1}{2^{1-p}}(a^p + b^p) = \frac{1}{2^{1-p}}(\sum_1^n |x_i|^p + \sum_1^n |y_i|^p) = \frac{1}{2^{1-p}} \sum_1^n (|x_i|^p + |y_i|^p) \\
\geq \; & \frac{1}{2^{1-p}} \sum_1^n (|x_i| + |y_i|)^p \\
\geq \; & \frac{1}{2^{1-p}} \sum_1^n |x_i + y_i|^p
\end{aligned}$$

Thus (3) holds with $\kappa = 2^{\frac{1-p}{p}}$. Furthermore, this inequality cannot be improved; a worst case, for which equality is obtained, is the case where n is assumed even, $\mathbf{x}$ is the vector whose even components are 1 and the odd ones are 0, and $\mathbf{y}$ is the vector whose odd components are 0 and the even ones are 1.

Note that the smaller p is the larger the bound κ is, and the further from metric the corresponding p-distance is.

Acknowledgments: this work benefited from many inspiring discussions with Ronen Basri. We are grateful to Liz Edlind for Figure 3.

References

[1] Basri, R., L. Costa, D. Geiger, and D. Jacobs, 1998, "Determining the Similarity of Deformable Objects", *Vision Research*, **38**(15/16):2365-2385.

[2] Black, M. and Anandan, P., 1996, "The Robust Estimation of Multiple Motions: Parametric and Piecewise-Smooth Flow Fields," *Computer Vision and Image Understanding* **63**(1):75–104.

[3] Blatt, M., Wiseman, S., and Domany, E., 1996, "Clustering Data through an Analogy to the Potts Model," *Advances in Neural Information Processing Systems, 8*:416–422.

[4] Brand, M., 1996, "A Fast Greedy Pairwise Distance Clustering Algorithm and its Use in Discovering Thematic Structures in Large Data Sets," MIT Media Lab TR #406.

[5] Cox, I., Miller, M., Omohundro, S., and Yianilos, P., 1996, "PicHunter: Bayesian Relevance Feedback for Image Retrieval," *Int. Conf. on Patter Recognition*, C:361–369.

[6] Dasarathy, B., 1994, "Minimal Consistent Set (MCS) Identification for Optimal Nearest Neighbor Decision Systems Design," *IEEE Transactions on Systems, Man and Cybernetics*, **24**(3):511–517.

[7] Donahue, M., Geiger, D., Hummel, R., and Liu, T., 1996, "Sparse Representations for Image Decompositions with Occlusions," *IEEE Conf. on Comp. Vis. and Pat. Rec.*:7–12.

[8] Friedman, J., Bently, J., Finkel, R., 1977, "An Algorithm for Finding Best Matches in Logarithmic Expected Time," *ACM Transactions on Mathematical Software*, **3:3** 209–226.

[9] Y. Gdalyahu and D. Weinshall, "Flexible Syntactic Matching of Curves", *Proc.: Fifth European Conference of Computer Vision*, Freiburg, June 1998.

[10] Gowda, K. and Krishna, G., 1979, "The Condensed Nearest Neighbor Rule Using the Concept of Mutual Nearest Neighbor," *IEEE Trans. on Information Theory*, **25**(4):488–490.

[11] Graepel, T., Herbrich, R., Bollmann-Sdorra, P., and Obermayer, K., (forthcoming), "Classification on Pairwise Proximity Data," NIPS.

[12] Fukunaga, K. and Mantock, J., 1984, "Nonparametric Data Reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**(1):115–118.

[13] Haralick, R. and L. Shapiro, 1993, *Computer and Robot Vision, Vol. 2*, Addison-Wesley Publishing.

[14] Hart, P., 1968, "The Condensed Nearest Neighbor Rule," *IEEE Trans. on Information Theory*, **14**(3):515–516.

[15] Hinton, G., C. Williams, and M. Revow, "Adaptive Elastic Models for Hand-Printed Character Recognition," *NIPS-4*:512–519.

[16] Hofman, T. and J. M. Buhman, 1997, "Pairwise Data Clustering by Deterministic Annealing", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **19**(1):1–14.

[17] Huttenlocher, D., G. Klanderman, and W. Rucklidge, 1993, "Comparing Images Using the Hausdorff Distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15**(9):850-863.

[18] Huttenlocher, D., J. Noh, and W. Rucklidge, 1993, "Tracking Non-Rigid Objects in Complex Scenes," *4th Int. Conf. on Computer Vision*:93–101.

[19] Jain, A. and Zongker, D., "Representation of Handwritten Digits Using Deformable Templates," (Forthcoming).

[20] Meer, P., D. Mintz, D. Kim and A. Rosenfeld, 1991, "Robust Regression Methods for Computer Vision: A Review," *Int. J. of Comp. Vis.* **6**(1):59-70.

[21] Ornstein, L., 1965, "Computer Learning and the Scientific Method: A Proposed Solution to the Information Theoretical Problem of Meaning," *Journal of Mount Sinai Hospital*, **32**(4):437-494.

[22] Poggio, T. and Girosi, F., 1990, "Regularization algorithms for learning thar are equivalent to multilayer networks," *Science*, **247**:978–982.

[23] Royden, H., 1968, *Real Analysis*, MacMillan Publishing Co., New York, NY.

[24] Santini, S. and Jain, R., 1996, "Similarity Queries in Image Databases," *CVPR*:646–651.

[25] Tappert, C., 1982, "Cursive Script Recognition by Elastic Matching," *IBM Journal of Res. Develop.* **26**(6):765–771.

[26] Tsai, W. and S. Yu, 1985, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **7**:453–462.

[27] Tversky, A., 1977, *Psychological Review*, **84(4)**:327–352.

[28] Williams, P., "Prototypes, Exemplars, and Object Recognition", submitted.

[29] Yianilos, P., 1993, "Data Structures and Algorithms for Nearest Neighbor Search in General Metric Spaces", *Proc. of 4th Annual ACM-SIAM Symposium on Discrete Algorithms*:311–321.

[30] Yoshida, K. and H. Sakoe, 1982, "Online Handwritten Character Recognition for a Personal Computer System," *IEEE Transactions on Consumer Electronics* **CE-28**(3):202–209.

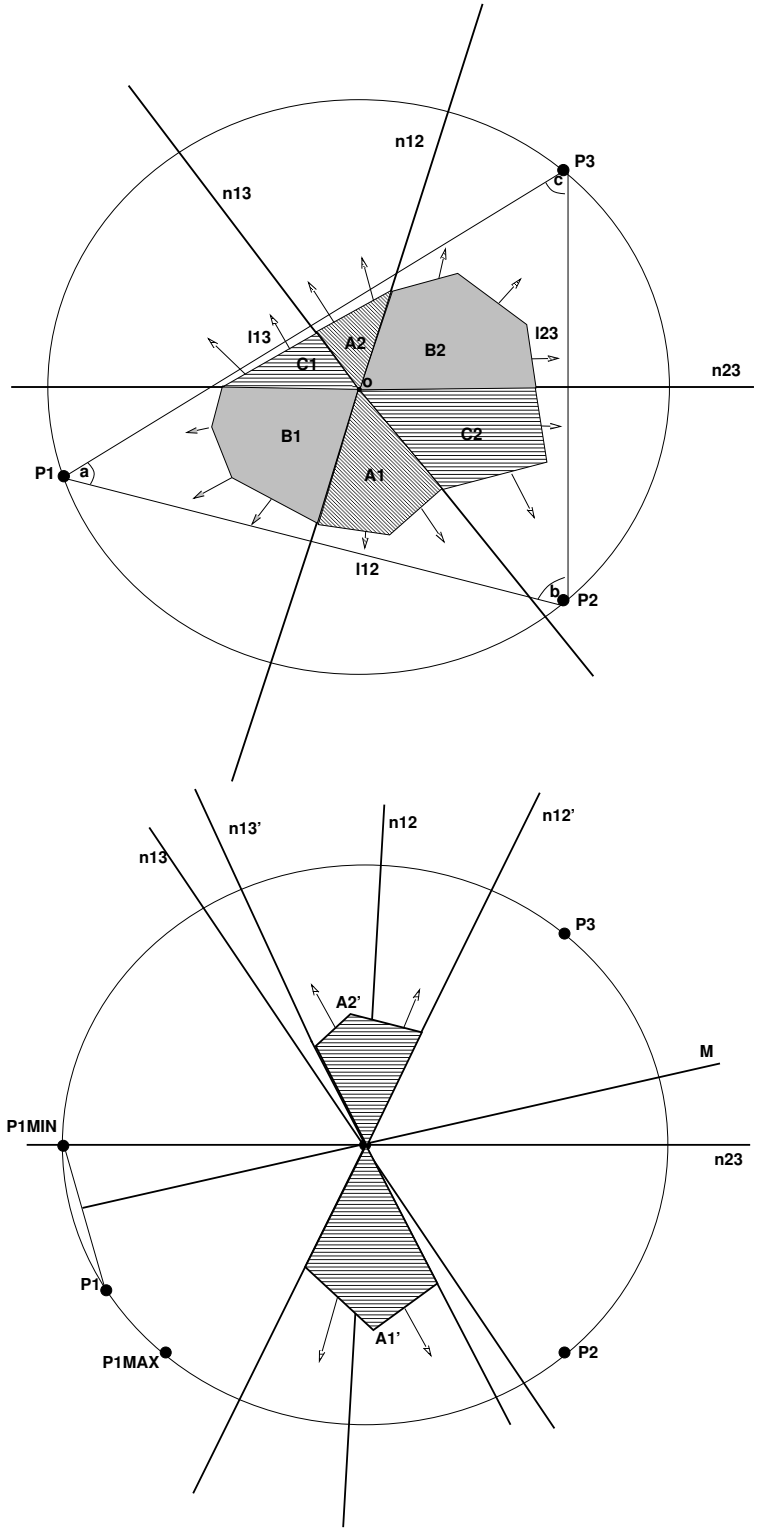


Figure 9: Illustration of definitions given in the text.