Making a low-dimensional representation suitable for diverse tasks

Nathan Intrator
School of Mathematical Sciences
Sackler Faculty of Exact Sciences
Tel Aviv University
Tel Aviv 69978, Israel
nin@math.tau.ac.il

Shimon Edelman
Dept. of Applied Mathematics
and Computer Science
The Weizmann Institute of Science
Rehovot 76100, Israel
edelman@wisdom.weizmann.ac.il

January 2, 1996

Abstract

We introduce a new approach to the training of classifiers for performance on multiple tasks. The proposed hybrid training method leads to improved generalization via a better low-dimensional representation of the problem space. The quality of the representation is assessed by embedding it in a 2D space using multidimensional scaling, allowing a direct visualization of the results. The performance of the approach is demonstrated on a highly nonlinear image classification task.

1 Introduction

The ultimate goal of machine learning is to mimic the human ability to learn a task from a limited set of examples and to generalize learning to new circumstances in a reasonable way. A fundamental question here concerns the type of data to be used in training classifiers for best generalization. Asymptotically, optimal results are obtained when the distribution of the training data is similar to that of the test (generalization) data, and the capacity of the learning machine is adjusted to the amount of the training data. In practice however, very often the amount of data is far smaller than what is assumed for the required generalization task. In such cases, innovative use of training data becomes essential. Methods for data reuse such as cross-validation (Stone, 1974) and bootstrap (Efron and Tibshirani, 1993) can help in obtaining confidence intervals (Baxt and White, 1995) and improved performance (Breiman, 1992; Breiman, 1994; LeBlanc and Tibshirani, 1994; Raviv and Intrator, 1995).

Unlike data, class labels are not often reused (see however, (Grossman and Lapedes, 1993)), in particular, multiple-class labels. Humans make natural and extensive use of the fact that objects may have several class associations (say, at different category levels). In contrast, in machine learning, it is not clear how one should proceed given hierarchical class labels, and whether such information can be used effectively or at all.

We believe that through the use of multiple-class associations learning can be constrained – biased – towards a better solution, and that innovative use of multiple-class labels may be a practical way to introduce prior knowledge into a high-capacity learning machine. We present a method for introducing such prior information during training, while avoiding the need to construct different low-level representations for different tasks defined on the same data. This approach naturally facilitates generalization across tasks, also known as *transfer* of skill — a hallmark of human cognitive prowess (see section 1.1).

Connectionist approaches to the problem of transfer, or cross-task generalization (Pratt, 1993) tend to offer as a solution some kind of weight sharing between networks trained on different tasks. Along these lines, Baxter has proposed recently to use *different* data sets with the same class association for constructing a rich internal representation (Baxter, 1995). His argument is that "a representation that is appropriate for learning a single face should be appropriate for learning all faces." Our approach extends this idea by observing that a representation has to be suitable for many different tasks on the same level of categorization, and for different category levels as well. The problems of *catastrophic interference* and *hypertransfer* (Martin, 1988; Murre, 1995), which are both manifestations of the learning machine's finding a suboptimal representation space, are easily addressed by our algorithm, as discussed in section 4.

It has been observed in the past that training a classifier on multiple tasks (using the same data) may be an efficient way to introduce desirable bias into the solution (Caruana, 1993). Our motivation for multiple-task training is, however, fundamentally different from the subsequent development of that idea by Caruana (1995), who implicitly assumes that the different tasks are on the same level of categorization. In comparison, our approach calls for internal representation to be constructed using a combination of various tasks, including tasks at different levels of categorization. For example, if required to create a classifier to distinguish gender in faces, we would use face identity information (i.e., more specific labels), along with facial expression information (i.e., different labels at the same level of specificity), in addition to the gender labels proper, for constructing the internal representation for the task.[1] Thus, we would use the **same data** with many class associations, to learn a common low-dimensional representation (LDR).

If an LDR is shared between tasks in this manner, a classifier of considerably lower complexity can be constructed for each of these tasks, compared to the usual approach of learning a separate representation for each task. In support of this claim, we show that:

- An LDR that is sufficiently general to support several classification tasks can be acquired in a difficult learning scenario;
- A hybrid approach, according to which a multi-purpose LDR is first found, then used to train a classifier for a specific task, can outperform straightforward training of a specific-task classifier from scratch.

We demonstrate our approach on a highly nonlinear image discrimination task, involving parameterized fractal patterns. This allows us not only to examine the viability of the proposed approach, but also to compare the performance of the implemented algorithm to that of human subjects in a recent series of psychophysical experiments that explored related issues in the representation of complex 3D shapes (Edelman, 1995a; Cutzu and Edelman, 1995).

1.1 Psychological motivation

In psychology, the notion of transfer of learning between tasks encompasses behavioral phenomena ranging from simple (e.g., generalization of conditioned response between familiar and novel stimuli) to extremely complex (e.g., carrying over a solution to a problem in arithmetics to a novel class of problems). In the former example, the degree of generalization between stimuli is governed by their perceptual similarity (Shepard, 1987), while the latter transfer is usually hypothesized to be mediated by more complex cognitive structures or schemata (Reder and Klatzky, 1994). This ubiquity of transfer makes a claim that "all demonstrations of learning and memory involve transfer" (Hintzman, 1994)

[1] A complementary approach here is to learn, instead of a variety of labeling schemes for a given data set, the *transformations* which leave its members invariant (Lando and Edelman, 1995), or the *invariances* of the individual data items (Simard et al., 1992; Thrun and Mitchell, 1995). We do not consider this approach in the present paper.

easily understood. Here, we concentrate on a particular kind of circumstances under which transfer is known to occur, namely, on tasks that involve perceptual classification of complex visual stimuli.[2]

1.1.1 Transfer of perceptual classification

A typical case of transfer in which we are interested is one in which the subject's prior training in the classification of patterns belonging to some well-defined category facilitates his or her learning of the classification of a different set of patterns from the same category, or of patterns from a somewhat different category. A favorite example is the own-race effect in face recognition: people perform much better in various face perception tasks when the faces that serve as stimuli belong to the same race as the subjects (Brigham, 1986). This kind of transfer has been reported recently also with random-dot patterns, generated according to a set of complex statistical criteria (McLaren et al., 1994).

A recent review (Reder and Klatzky, 1994) lists a number of issues regarding transfer on which experimental evidence can be brought to bear. Of particular interest to us are the first two conclusions of that study:

1. "There is broad consensus that transfer is typically very specific to the context in which training has occurred."

2. "As a general principle, having identical elements between the training and performance context facilitates transfer."

Thus, transfer between classification tasks is expected to occur in a situation involving a fixed set of stimuli (patterns) to be classified; this observation influenced our choice of the experimental testbed, described in section 2. Our working hypothesis is that this kind of transfer is supported by learning a low-dimensional representation (LDR) of the set of patterns that is necessarily common to all classification tasks defined on those patterns. As usually in theories of classification, one may ask here whether the postulated LDR encodes the specific patterns in question, or the relevant subspace of the pattern space (cf. Logan, 1988; Maddox and Ashby, 1993). Although transfer is facilitated by having identical elements in the two tasks, as noted in (Reder and Klatzky, 1994), there is evidence that exemplar substitution affects transfer to a much smaller degree than context (rule) substitution (Kramer et al., 1990). Thus, the LDR of the stimulus space should be defined for all patterns in the vicinity of the familiar ones,[3] although its performance may be expected to be better for the familiar patterns proper; cf. the distinction between persistent and ephemeral entities in the representational scheme proposed in (Edelman, 1995b).

1.1.2 Low-dimensional representation as a substrate for transfer in visual perception

The hypothesis, stated in the preceding section, that transfer across classification tasks is supported by learning a common LDR for the set of patterns, has been entertained for decades in the context of perceptual generalization (e.g., in the discrimination of tones or hues, and in the judgment of similarity of 2D outline shapes). In particular, it has been observed that the human visual system performs as if it represents the stimuli in a low-dimensional metric psychological space (see (Shepard, 1987), for a review). Recently Edelman and colleagues have investigated the ability of human subjects to form low-dimensional representations in the context of complex 3D shape classification (Edelman, 1995a; Cutzu

[2] I.e., high-resolution patterns or images of objects, as opposed to uniform-color patches or repeating textures. An example of a nonvisual task of a parallel level of complexity is morphological inflection (Gasser, 1995). We do not consider here transfer of higher cognitive skills such as arithmetics.

[3] That is, patterns close enough to the familiar ones in the underlying high-dimensional space.

and Edelman, 1995). The subjects were confronted with several classes of solid shapes, arranged in a complex pattern in a common underlying high-dimensional parameter space. The experimental task was delayed match to sample, involving images of computer-rendered 3D animal-like objects, jointly parameterized by 70 variables controlling various details of the object shape. In each trial, the subjects (who received no prior training) had to decide whether two images, shown briefly and consecutively, belonged to the same animal shape. In each experiment, the response time and error rate data were combined into a measure of view similarity, and the resulting proximity matrix was submitted to nonmetric multidimensional scaling, or MDS (this technique is discussed below in section 1.3; for an overview, see Shepard, 1980). Examining the configuration of points corresponding to the various views in a 2D MDS solution revealed that (1) different views of the same shape were correctly clustered together, and (2) in each experiment, the relative geometrical arrangement of the view clusters of the different objects reflected the structure of the parameter-space pattern (respectively, a star, a triangle, a square, and a line) that defined the relationships between the stimuli classes.

It should be noted that although the subjects in these experiments were required to discriminate between two objects at a time, the pattern of their performance, as revealed by MDS, indicated an involvement of a common LDR of all the objects. Specifically, the perceptual distances between the objects corresponded closely to the distances in the low-dimensional parameter space used to create the objects, which is a natural low-dimensional representation of the data. This surprising finding prompted us to ask whether machine classification methods can be applied to a similar problem of low-dimensional structure discovery from complicated image data.

1.2 Statistical motivation

The stress on the importance of lowering the dimensionality of the classification problem as a first step towards the development of a versatile representational substrate for classification is a major lesson from psychological studied: Reder and Klatzky (1994) actually state that "transfer is a problem of high dimensionality." Thus, although many current approaches to the training of statistical classifiers lack the provisions for supporting transfer of classification, we claim that this shortcoming can be amended by building on the notion of dimensionality reduction, or, in statistical language, the discovery of structure in the data.

The existence of structure in the data is the key assumption behind statistical data modeling. The underlying structure must be sufficiently simple to permit its recovery from the usually limited amount of data. Because of the scarcity of the data, and the lack of information regarding the details of the model, it is important to use any available knowledge to facilitate the discovery of low-dimensional structure in the data. In particular, the modeling process can be greatly assisted if it is known that the data obey a simple geometrical pattern (e.g., reside in a low-dimensional manifold in the measurement space).

The discovery of structure, sometimes called dimensionality reduction, feature extraction and in some special cases Projection Pursuit, aims to find a sufficiently low-dimensional model of the data, so that robust classification becomes possible given the available (usually small) data set. Thus, it is reasonable to ask whether a given method is capable of finding useful low-dimensional structure, in addition to being able to generalize on a limited amount of data. In other words, a good LDR of the data (if such can be found) is highly likely to capture its true underlying model, which can lead to optimal generalization results (that is, optimal generalization for infinite test data). Thus, as long as the amount of test data is limited, it seems worthwhile to study the quality of the LDR.

The recovery of low-dimensional structure can be performed in an exploratory (unsupervised) manner, or using class labels for data points. Unfortunately, the recovery of LDR by training a classifier with the class labels alone is highly nontrivial, because, in general, the class labels do not

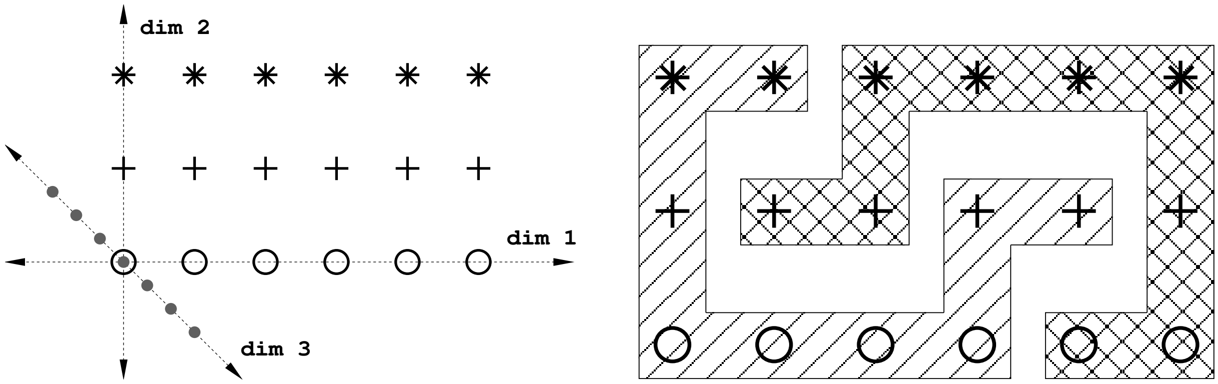


Figure 1: *Left:* The parametric representation from which the high dimensional images were created. Dim 3 is the dimension on which generalization was sought for the simple and the difficult classification tasks. *Right:* The dichotomy classification task, used in testing the LDR (see Figure 3 and section 2).

possess enough structure to direct the classifier to the correct solution. This is another manifestation of the *curse of dimensionality* (Bellman, 1961), which explains why in general there is not enough data to recover the true model (the underlying representation) directly from the classification task. In fact, searching for LDR using a combination of exploratory and class-label approaches together yields in some cases better results (Intrator, 1993). In this paper, we explore the possibility of learning LDR using an excessive number of class labels, thus building more structure into the space of the potential solutions.

1.3 Multidimensional scaling

While finding a good LDR is generally a highly nontrivial problem (Huber, 1985; Intrator and Cooper, 1992), the assessment of the quality of an LDR is a much simpler task. Unfortunately, the LDR often still resides in a space of more than two or three dimensions, thus making direct visual inspection difficult. Multidimensional scaling, applied in the analysis of the human performance in the psychophysical experiments described above, constitutes a useful visualization method in such cases.

MDS has been originally developed in psychometrics, as a method for the recovery of the coordinates of a set of points from measurements of the pairwise distances between those points (Young and Householder, 1938). In a typical application, the experimenter would attempt to characterize a subject's performance by placing a point corresponding to each stimulus perceived by the subject in a coordinate space, derived from subjective similarity ratings of pairs of stimuli. The power of MDS as a tool for the study of internal representations was revealed when Shepard discovered in 1962 that fixing the *relative* distances of a set of points effectively determines their coordinates (Shepard, 1966). This discovery led to the development of the nonmetric MDS algorithm (Kruskal, 1964), which employs gradient descent to seek a monotonic transformation between measured distances and distances computed from the hypothesized point configuration, which would minimize stress (defined as the discrepancy between the ranks of the measured and the computed distances). Recent improvements of the MDS procedure include an implementation using deterministic annealing (Hofmann and Buhmann, 1994), which may prove to be better in avoiding local minima in the search for an optimal configuration. In the present work, we used a modern implementation of nonmetric MDS, available in version 6 of the SAS statistical analysis software (Sas, 1989).

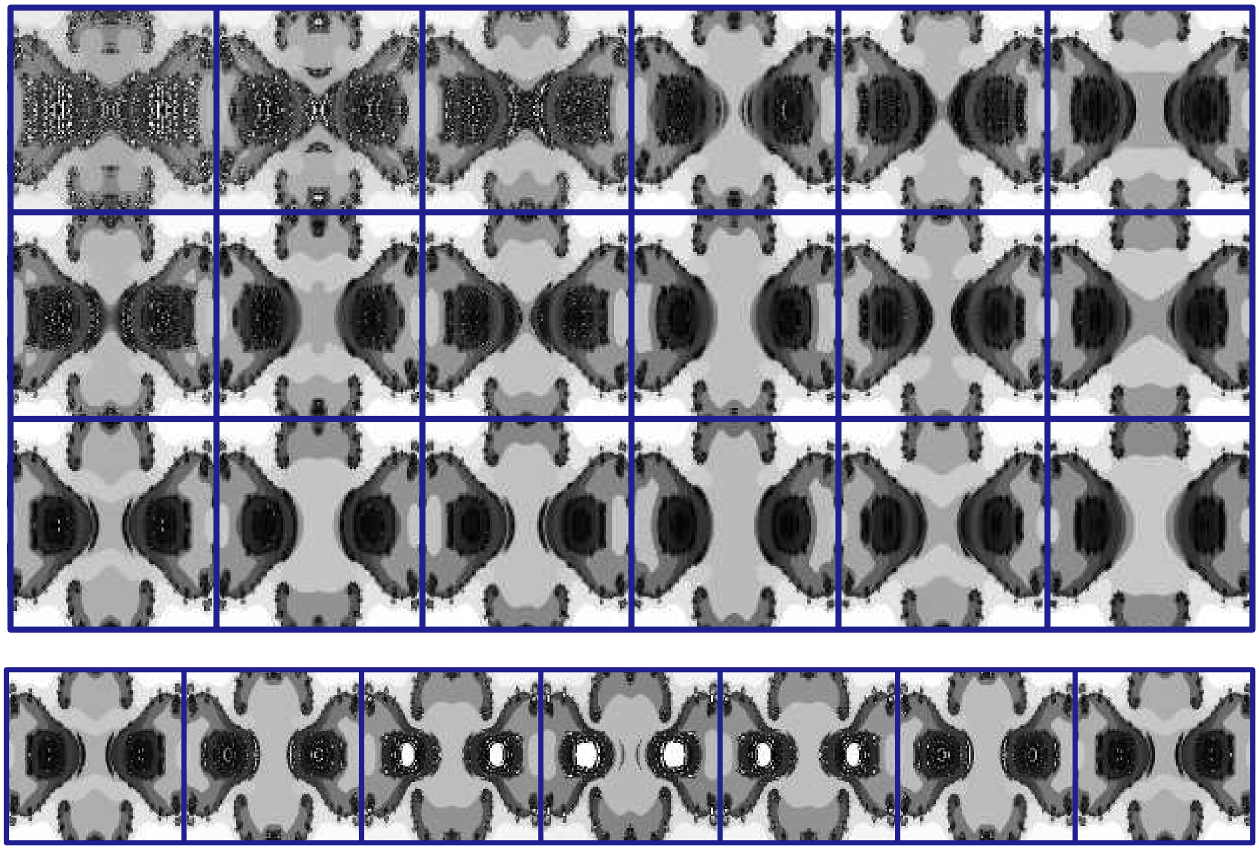


Figure 2: *Top:* the 18 images, each of size 256×256 pixels, corresponding to the 18 points in parameter space defined in Figure 1. The value of the Dim 3 parameter here is held constant. *Bottom:* the 7 images corresponding to the possible values of the Dim 3 parameter; the values of Dim 1 and Dim 2 correspond to the leftmost image in the bottom row.

2 Methodology

Our aim is to test the ability of a neural network to discover simple structure embedded in high-dimensional data, in a situation where the discovery requires a highly nonlinear transformation from the input to the low-dimensional space. We set out to create a difficult LDR discovery task by generating a dataset of fractal images through a nonlinear transformation of a three-dimensional parametric space. The three dimensions of the parametric representation, and the generalization task, are illustrated in Figure 1.

In the experimental design, all the classification tasks were defined in two parametric dimensions; the performance of the classifiers designed to discover and to test the LDR was assessed by computing their generalization ability along the third dimension, orthogonal to the first two. Starting with this parametric representation, we created a high-dimensional collection of image data. Each image (see Figure 2) corresponded to a transformation from the 3-parameter space to a 256×256 pixel space. Before being fed to the classification modules, the images were preprocessed by histogram equalization, then convolved with a bank of 28×28 receptive fields (Matlab^{TM} Image Processing toolbox; Laplacian of Gaussian, kernel size 9, $\sigma = 0.6$), reducing the dimension from 65536 to 784. This preprocessing served as a crude approximation of the transformations that a stimulus undergoes on its way to the primary visual area in the cortex.

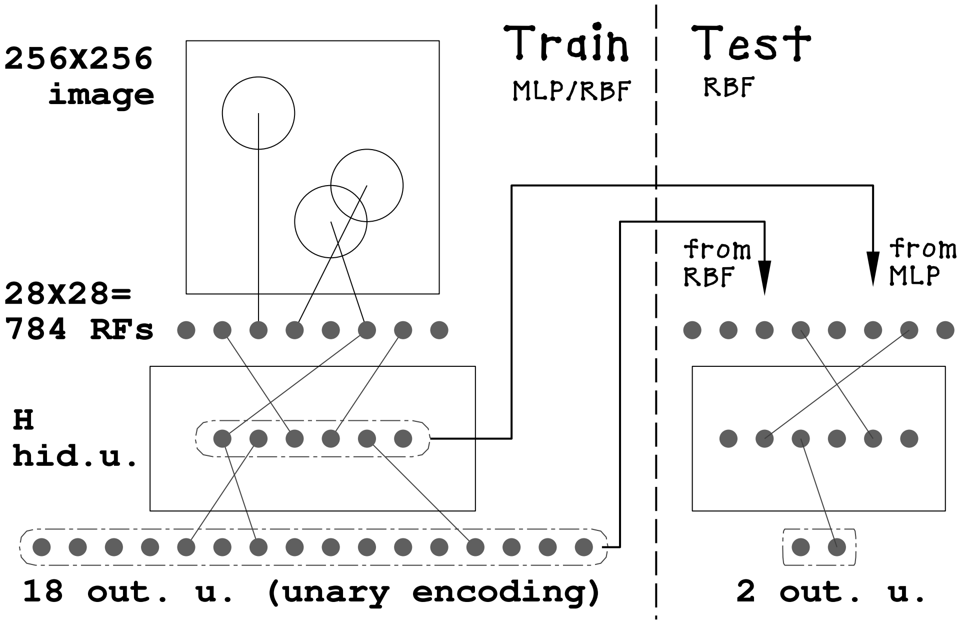


Figure 3: The low-dimensional representation (LDR) extraction scheme (see section 2). The LDR extraction network appears on the left, under the label Train. First, preprocessing is performed on the raw 256×256 fractal pixel images. The preprocessing stage takes the pixel image through 784 receptive fields convolving the inputs with a two dimensional difference of Gaussians filters (center surround cells), to produce a vector of 784 dimensions. Second, a learning module, which can be either multilayer perceptron (MLP), or a radial basis function interpolator (RBF), is trained to produce a unary (1 out of 18) encoding of the input class label. The LDR is extracted from the hidden layer in case of MLP, or from the output layer in case of RBF, and is fed to an RBF network (shown on the right, under Test) for testing. The testing procedure consists of teaching the network a complicated dichotomy on the input space (see Figure 1), and testing its generalization performance. Finally, this performance is compared against the generalization obtained by a classifier trained directly on the dichotomy task using the output of the 784-dimensional RF filters.

2.1 Data generation

The fractal patterns were generated using publicly available software (Xfractint 2.03) (Pickover, 1990, chapter 10), and were imported into Matlab for processing and classification. We chose the quaternion Julia set (entry quatjul in the Xfractint pattern menu), which is parameterized by six variables and is therefore well-suited for generating complicated patterns that depend on up to six parameters. The quatjul iteration formula is

$$\begin{aligned} q(0) &= (xpixel, ypixel, z_j, z_k) \\ q(n+1) &= q(n) * q(n) + c, \end{aligned}$$

where both q and $c = (c_1, c_i, c_j, c_k)$ are quaternions (for further details, see (Pickover, 1990), chapter 10). The three dimensions shown in Figure 1 correspond to the variation of parameters c_1, c_j, and c_k, respectively.

Figure 2 shows a 2D slice through the 3D parametric space; in the making of all these images, the third (generalization) parameter has been kept fixed. As evident from the picture, the mapping from the images back to the parametric representation is far from trivial. Our purpose was to make a neural network classifier learn a mapping from the high dimensional space to various classification tasks, and subsequently to examine the LDR produced by the classifier.

This procedure allowed us to generate data sets of varying degree of difficulty, by controlling the parameters that determine the distances within and between the designated classes of patterns. Three such data sets were produced and used in the experiments we describe below: E (easy), M (moderate), and D (difficult). The M data set is implied in the descriptions of the experimental results, unless otherwise noted.

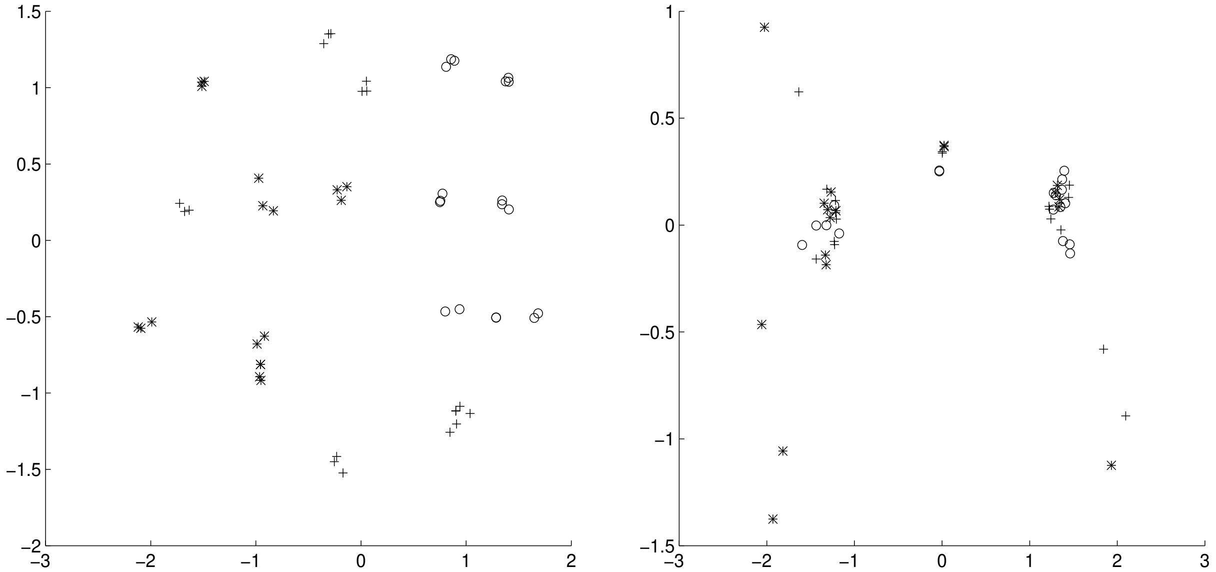


Figure 4: 2D structure derived by MDS from a representation found by an MLP with 5 hidden units, using the M (moderate difficulty) task. *Left:* MLP trained on 18 classes; *Right:* MLP trained on a dichotomy. The dichotomy classification error is 0.056 with both methods.

2.2 Hybrid dimensionality reduction / classification

The two-part hybrid classification method we use is illustrated in Figure 3. The 256×256-pixel images, mapped into a $28 \times 28 = 784$-dimensional RF space, are used to train an LDR-extraction network. This can be a multi-layer perceptron (MLP) or an RBF network; in either case, a 1-out-of-18 unary representation of the class labels (see Figure 1) is enforced at the output. The recovered LDR is used to train an RBF classifier on a two-class problem. The generalization performance of this classifier is compared with that of an identical classifier trained on the raw 784-dimensional RF-space representation of the image set, on the same two-class problem.

3 Results

3.1 A characterization of the resulting LDR

The two-class task is presented in Figure 1 (right). As can be easily seen, this is a nonlinear task even in the low dimensional parametric space, and is thus, highly nontrivial. The results of the non-metric MDS are depicted in Figure 4. While it is easy to believe that structure is not easily apparent in the original 784-dimensional space, it is less trivial to find that MDS is unable to find any structure there

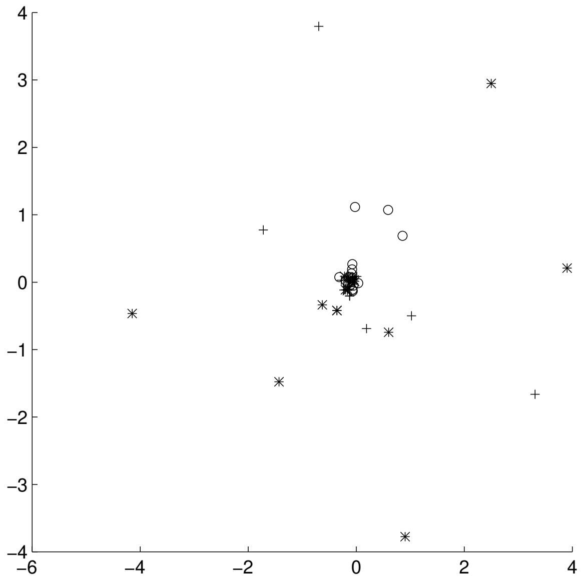


Figure 5: The pattern recovered by MDS from the raw 784-dimensional data. No separation of the 18 classes or the 3 super-classes (each marked by a different symbol) is apparent; most of the points are concentrated in the middle of the plot. Compare with Figure 4.

as well (see Figure 5). This is probably due to bad scaling up of the algorithm with the dimensionality and to the high nonlinearity of the fractal transformation which takes the 3D parametric space to the high dimensional pixel space.

3.2 RBF vs. MLP as an LDR extractor

3.2.1 Advantages in training

The structure derived by MDS from the output units of the RBF network is summarized in Figure 6. These results are not directly comparable to those obtained with the MLP network, because in the present case the useful representation emerged at the output and not at the hidden layer. This is due to the high non-linearity of the task, because of which good performance was only possible with a large number of hidden units. In fact, a basis function has been placed on each of the training patterns, leading to no dimensionality reduction at the hidden layer. We therefore looked at the output as the new LDR; there, the representation was 18-dimensional in the 18-way classification task, and 2-dimensional one in the dichotomy task.

In Figure 6, the top row presents results obtained with a relatively easy data set; the RBF network captures well the structure inherent in the data, which corresponds nicely to the original parameter-space representation of the images. The top right pane shows a two-cluster structure corresponding to the two classes imposed during training; the six points in between the clusters correspond to images which the network found more difficult to classify. The bottom row presents results from the D (difficult) data set. Here, the performance of the hybrid RBF-based LDR extractor was lower. In this case, the performance of the MLP-based LDR extractor deteriorated nearly to chance. The low-dimensional structure of the data is not very prominent here, but still much more structure appears on the left than on the right; correspondingly, the classification performance on the dichotomy problem is slightly better for the hybrid-extracted LDR than on the raw representation. Thus, the RBF network found better structure and obtained better results with the more difficult data we explored, compared

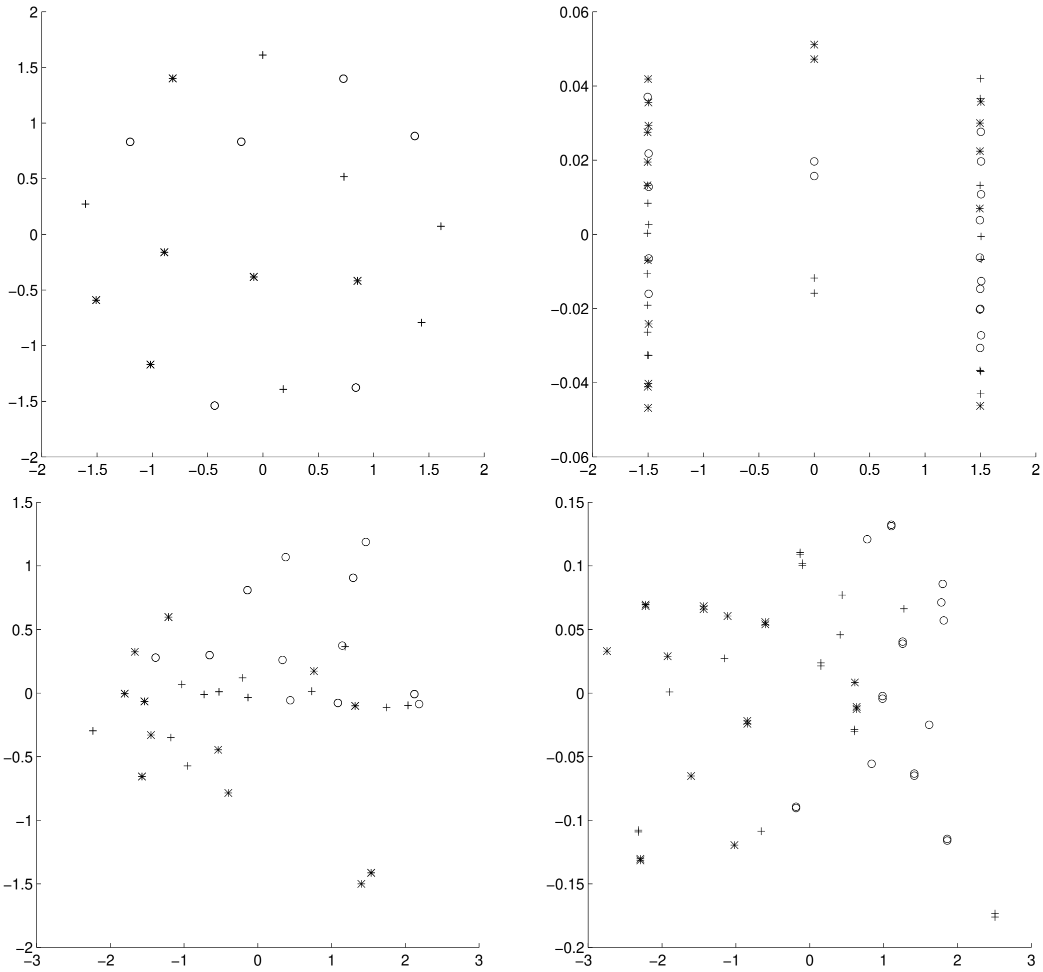


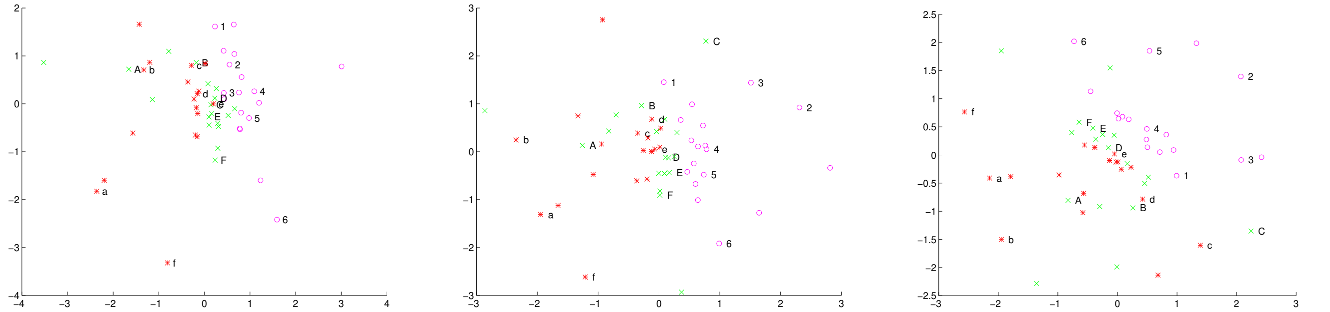
Figure 6: *Top Left:* LDR found by a 72-center RBF (MDS on output units); error is 0.056; *Top Right:* control; LDR found by a 72-center RBF trained on a dichotomy; error is 0.056 (M data set). *Bottom Left:* RBF, on a more difficult data set (D); error is 0.27. *Bottom Right:* control; error is 0.35 (same as on raw data).

to the MLP network.

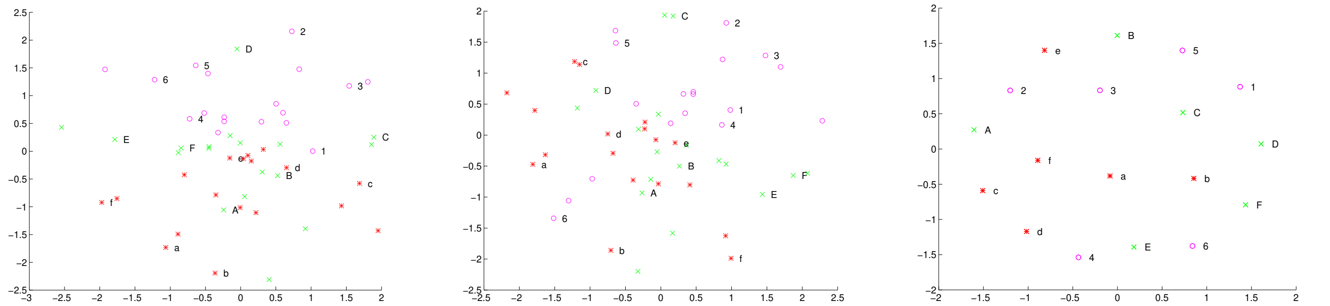
3.2.2 Controlling the quality of the LDR

A natural question that arises in conjunction with our LDR extraction method is how to control the quality of the resulting LDR. The rapid learning of the RBF-based LDR extractor made it possible to explore this issue extensively. We found that one can trade off quality for computational resources, simply by varying the number of RBF centers (Figure 7). RBFs with fewer than six centers could not learn the task at all, while those with 72 centers or more (i.e., those which assigned at least one basis function per training example) produced essentially perfect LDRs. The quality of the LDRs obtained with an intermediate number of centers (as measured by the error rate achieved with the LDR in the dichotomy task) varied, as indicated by the data in Figure 7.

A plot of the error rates (in the 18-way, or training, and the dichotomy, or testing, classification tasks) appears in Figure 8, top pane. Note that the error rates obtained with the different LDRs are correlated with the visually apparent quality of each LDR, whereas the two measures of the



ER_{train}	= 0.64	ER_{test}	= 0.63	ER_{train}	= 0.19	ER_{test}	= 0.24	ER_{train}	= 0.10	ER_{test}	= 0.11
ER_{raw}	= 0.19	ER_{proc}	= 0.17	ER_{raw}	= 0.19	ER_{proc}	= 0.11	ER_{raw}	= 0.19	ER_{proc}	= 0.13
BOF	= 0.15	$DCORR$	= 0.96	BOF	= 0.22	$DCORR$	= 0.91	BOF	= 0.25	$DCORR$	= 0.86



ER_{train}	= 0.04	ER_{test}	= 0.04	ER_{train}	= 0.00	ER_{test}	= 0.00	ER_{train}	= 0.00	ER_{test}	= 0.00
ER_{raw}	= 0.19	ER_{proc}	= 0.09	ER_{raw}	= 0.19	ER_{proc}	= 0.00	ER_{raw}	= 0.19	ER_{proc}	= 0.00
BOF	= 0.28	$DCORR$	= 0.80	BOF	= 0.30	$DCORR$	= 0.73	BOF	= 0.37	$DCORR$	= 0.44

Figure 7: The goodness of the low-dimensional representation extracted by an RBF network depends on its number of hidden units. The six panes in this figure show the configurations derived by MDS from networks having 6, 18, 24, 30, 36, and 72 units (top left to bottom right, respectively). The table below each plot shows the training and testing error rates on the 18-class task (ER_{train} and ER_{test}), the test error rates on the 2-class task using the raw data and the RBF-extracted LDR (ER_{raw} and ER_{proc}), and the badness of fit (stress) and distance correlation values obtained by MDS (BOF and $DCORR$). See also Figure 8.

MDS performance (stress and distance correlation, plotted in the lower pane of that figure) actually deteriorate while the quality of the LDR improves. This observation illustrates the point made by Borg and Lingoes, who caution against using the MDS stress as a sole indicator of the goodness of the configuration (Borg and Lingoes, 1987). When other, independent, measures of the goodness of a configuration are available, they should be consulted to avoid basing the judgment on the location of a (possibly local) minimum in the stress optimization landscape. In the present case, the classification error and the visual quality of the configuration are such independent measures of the goodness of the configuration recovered by MDS from the LDR found by the hybrid classifier.

We have also found that with a multitask MLP classifier we could obtain a richer internal representation by increasing the number of hidden units to a larger extent than what was possible with a regular single-task classifier. The reason for this was overfitting, which affected the generalization performance of a single-task classier much more than that of a multi-task classifier. We have explored this phenomenon by training 2-class MLPs and 18-class MLPs with the same number of hidden units. The 2-MLP classifiers achieved 0 error rate on the training data very easily; their training-set performance, however, was a very poor predictor for the test-set (generalization) performance. For example, for the D data set, a typical value of the correlation between training and testing error rates was near 0 for the 2-MLPs, compared to about 0.3 for the 18-MLPs (six hidden units, 50-epoch training). A natural interpretation of this finding is that the constraints imposed by multitask training bias the internal

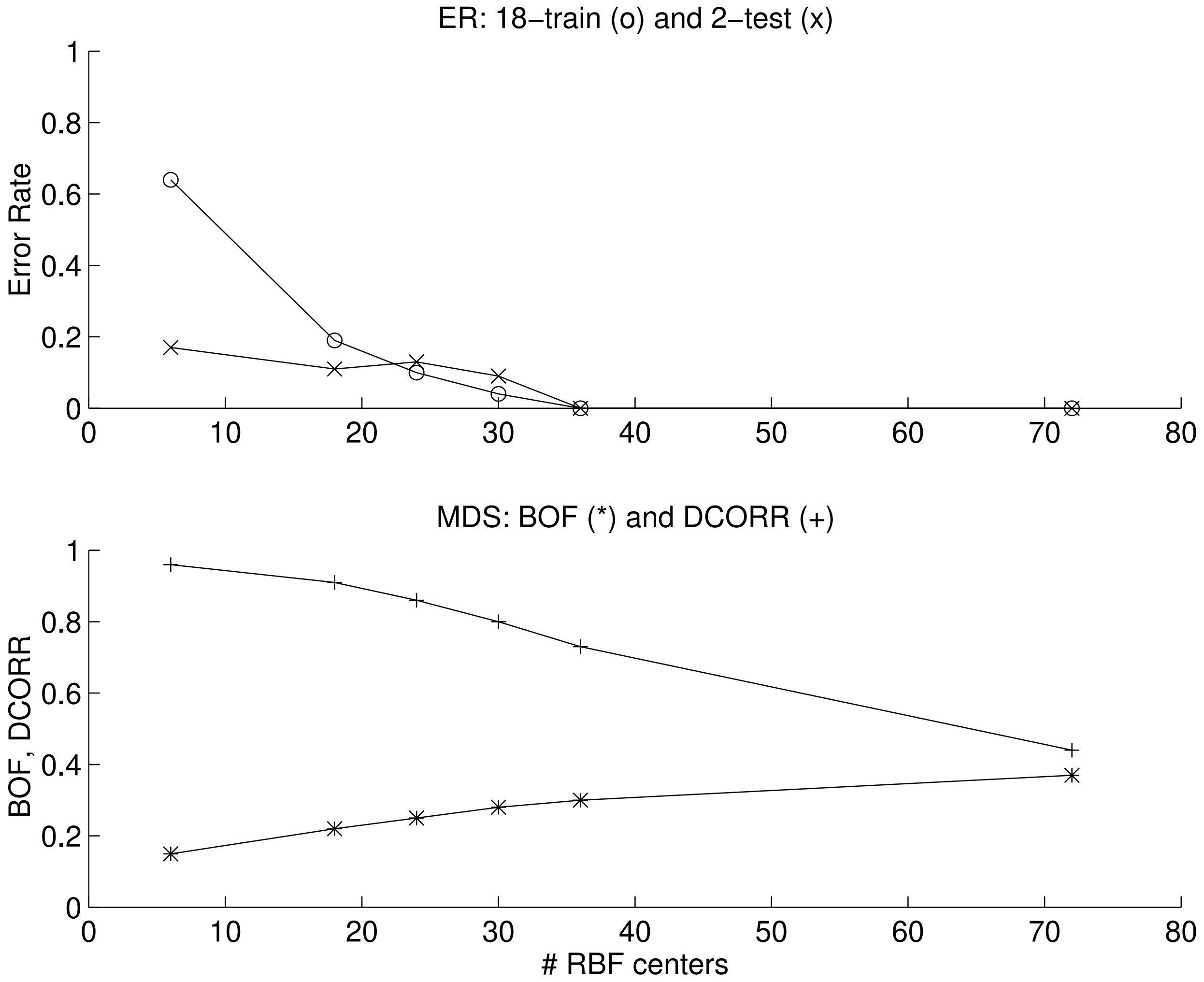


Figure 8: *Top:* error rates in the 18-class and the 2-class tasks, plotted vs. the number of hidden units in the RBF LDR extraction module. *Bottom:* MDS badness of fit (stress) and distance correlation values, plotted vs. the number of hidden units. The data are from the tables that appear in Figure 7.

representation of the MLP to a better solution and thus serve to reduce the variance due to the limit on the capacity of the learning machine.

3.3 MDS as a tool for representation visualization

Nonmetric MDS, which is relatively widely used in exploratory (Shepard, 1980) as well as confirmatory (Edelman, 1995a; Cutzu and Edelman, 1995) data analysis in experimental psychology, has been only rarely applied in the study of representations produced by neural networks. Most studies that did attempt to visualize the representations by embedding them in a metric space used the Sammon mapping, which operates on a principle similar to that of metric MDS (Sammon, 1969). It is important to realize that, as indicated by the plot of Figure 5, direct approaches of this kind, powerful as they may be, are not applicable to the extraction of complicated nonlinear structure directly from high-dimensional data. One may assume that the nonmetric version of MDS would be more successful when applied to the raw data. It turns out, however, that both metric and nonmetric MDS were similar in

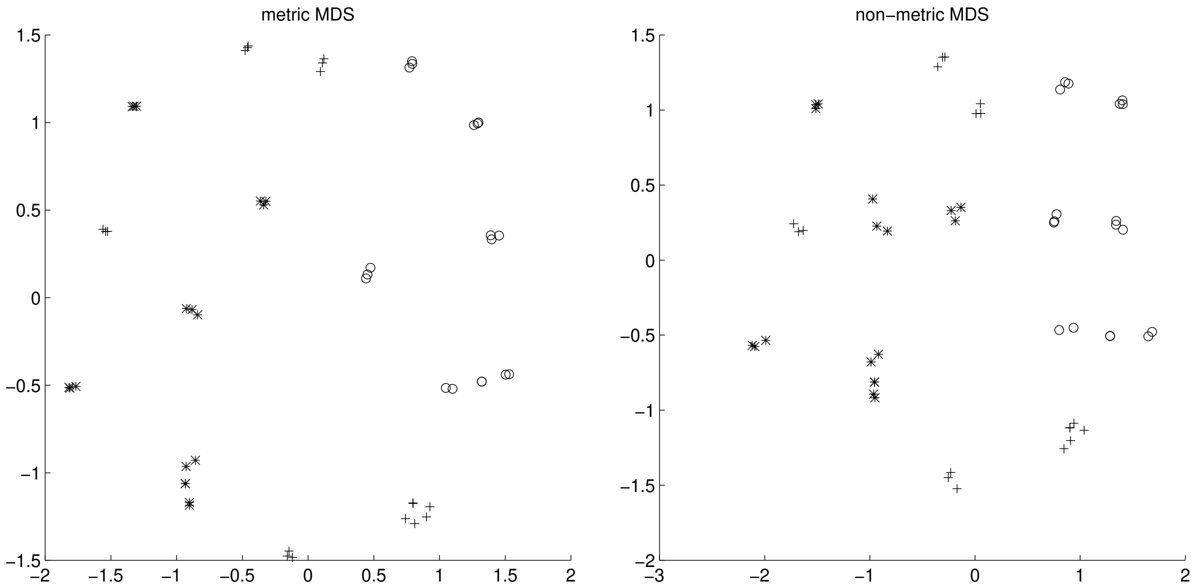


Figure 9: Metric (left) vs. nonmetric (right) MDS, applied to the same distance data. A discussion of the reason behind the similarity of the two configurations may be found in Borg and Lingoes, 1987, ch.1.

their performance in the present data visualization task (see Figure 9). In particular, both versions of MDS were incapable of extracting the structure directly from the raw data.

3.4 Incorporation of prior knowledge via multi-level class labels

Incorporating more prior knowledge into the LDR extractor during training further improved the hidden unit structure and the performance on the two-class (test) problem. To illustrate the ability of our hybrid method to assimilate prior knowledge in a natural manner, we performed two experiments. In the first one, three higher-level class labels were added to the set of 18 labels normally used in the training stage. For each data point, the higher-level label indicated the row to which it belonged (see Figure 1). In the resulting configuration, the 18 clusters were separated, on a coarser level, into three groups, corresponding to the three higher-level class labels (see Figure 10, left). In the second experiment, the LDR extractor was taught three labels corresponding to the rows and six labels corresponding to the columns of the parameter-space configuration. The resulting configuration depended to a significant degree on the relative weights given to the row and column labels. Under nearly equal weights, the points were separated into three clusters by the row label (see Figure 10, middle); when the column weight predominated, the separation was into six clusters (i.e., by column), with some additional structure within each cluster (see Figure 10, right).

We note that a natural extrapolation of this strategy would be to teach the network many possible dichotomies, in the hope that the structure of the underlying LDR can be recovered from the multiple two-way classifications (Price et al., 1995). The advantage of operating at the level of 18 classes (or of three classes, with six subclasses each) is in the much shorter training procedure. On the other hand, training on multiple dichotomies may have the advantage of forcing the LDR extractor to consider multiple, hopefully disjoint, sets of features relevant to the collection of tasks, and not letting it zero in on distinctive features specific for each one-vs-all discrimination. Finding an optimal compromise between these considerations is a subject we leave for future research.

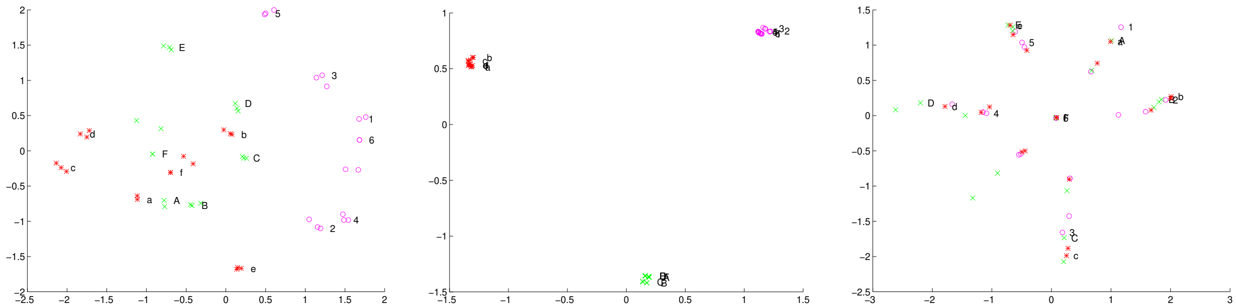


Figure 10: Incorporation of prior knowledge into the LDR extraction process. *Left:* MDS-derived configuration for a 5-HU MLP trained to produce a unary encoding of the 18-way ("identity") label set, to which a coarser set of 3-way class labels, corresponding to the row number in Figure 1, has been appended (also in a unary format). The row variables were given a weight of $w_c = 0.1$ relative to the identity variables. The 2-class error rate on the resulting LDR was 0.056. *Middle:* a 5-HU MLP trained to produce a unary encoding of the 3 row and the 6 column numbers of the stimulus. The 2-class error rate on the resulting LDR was 0.056. the relative weights of the row and the column variables were $w_r = 1.0$, $w_c = 0.75$. *Right:* same 3×6 class structure as before, but the relative weights of the row and the column variables were $w_r = 0.05$, $w_c = 1.0$. The 2-class error rate on the resulting LDR was 0.20. The 2-class error rate on the raw data in all three cases was 0.28.

4 Discussion

4.1 Significance of the present results

The present work touches upon one of the central questions in the study of categorization and learning: can a single internal representation developed by a classifier module as result of training be made useful for multiple tasks? To address this question, we constructed an artificial problem with easily controlled parameters. This allowed us to develop and test a hybrid method for the extraction of low-dimensional representations (LDRs), according to which an LDR is found by training a standard neural network classifier. The versatility of the resulting LDR can then be tested by putting it to use in a substantially different classification scenario, and by comparing the performance of the test-stage classifier on the LDR with that on the raw data.

We note that the assessment of the quality of the LDR is a nontrivial task by itself, even in a situation where the initial data are generated parametrically and are known to reside in a low-dimensional space (as is the case here). Specifically, even when one starts with, say, a two-dimensional parametric representation, it is difficult to expect that, by chance, this representation will be immediately apparent in the pattern of responses of the network, due to the high nonlinearity of the transformation from the underlying parameter space to the high-dimensional stimulus space. Furthermore, if the network is forced to operate in a very low-dimensional space, the chances are poor that it will stumble on the proper parametric representation of the data (typically, each unit in a network has a limited degree of nonlinearity, and it is difficult to expect that highly nonlinear structure hidden in the data can be captured by a few units). Thus, the LDR cannot be too low-dimensional if it is to be learnable, and its very evaluation becomes a nontrivial data exploration task. To overcome this dilemma, we used nonmetric multidimensional scaling (MDS) to embed the (relatively) low-dimensional representation in a two-dimensional space, for visualization and evaluation.

Using the hybrid LDR extraction method in conjunction with the MDS-based approach to visualization, we found that:

- Simultaneous multiple-label classification is an effective way to introduce bias (i.e., prior knowledge) into a flexible classifier, effectively controlling the excess variance that normally plagues flexible classification methods.

- Subordinate-level labels help to construct LDR for a basic-level task. Training a network for 18-way classification resulted in a better LDR for the two-class problem, both in terms of generalization performance, and in terms of similarity between the extracted LDR and the parametric representation built into the data.

  This is surprising, because one expects 18-way classification to be much more difficult than a dichotomy. In statistical terms, the construction of an LDR suitable for the 18-class task requires a more detailed model, involving more parameters, whose estimation is not robust under the usually very limited amount of training data.

- Basic-level information helps to construct LDR for a subordinate-level task. The addition of the basic-class label (row identity in Figure 1) improved the quality of the LDR and its generalization performance.

  This is surprising, because one expects that structure useful for 18-way classification should be sufficient for the simpler task of determining the row to which the input belongs. Yet, we show that the inclusion of this prior knowledge in the form of an additional classification subtask, improves both the LDR similarity to the original parametric space and its generalization performance.

4.2 Relationship to prior work on the transfer of learning

Within the framework that deals with the transfer of learning between tasks, several sets of training data are often used sequentially, to refine the network's performance on a preset task. Because only one of the datasets is used at each training stage, there is a danger that the network will converge to a representation which may be good for a particular data set, but may generalize poorly on new data, or, for that matter, on older data used in previous rounds of training. This leads to the problems of *catastrophic interference* and *hypertransfer* (Martin, 1988; Murre, 1995), which are both manifestations of the learning machine's finding a suboptimal representation space. In other words, the learning machine becomes over-biased to the current training dataset.

Our approach is very different from the sequential-training transfer framework: we use all the available data for training at all times, resulting in less susceptibility to hypertransfer. In fact, our training procedure reduces the possibility of over-bias by imposing multiple classification tasks on the learning machine, effectively forcing a solution that should be equally good for all the tasks. This approach can be especially effective when a low-dimensional representation common to all the tasks at hand is actually known to exist, as in the following example. Consider the problem of learning to navigate a taxicab in downtown New York City. A beginner taxi driver may start by learning a number of point-to-point routes (say, the three routes from the Port Authority Terminal to the Museum of Modern Art, to the Empire State Building, and to Washington Square). The usual problems of transfer of knowledge in sequential learning are expected to arise at this stage. For example, the driver will probably find it quite difficult to navigate to the Museum from Washington Square (rather than from the Port Authority). However, problems of this kind can be avoided if the driver attempts to learn the *map* of Manhattan, that is, the underlying low-dimensional representation of all possible routes between any two points in the region of interest. Thus, our LDR-based approach to the transfer of learning may be called *Manhattan Transfer*.

We point out that, in principle, multiple *conflicting* tasks may actually degrade performance, by causing interference between internal representations that are good for one task and those that are good for another. In our case, this problem is partially alleviated during the second stage of training, when a new classifier is trained on the LDR found by the first-stage multiple-task classifier. Clearly, if the number of hidden units is large enough, an internal representation that causes little interference can be found; this LDR is further transformed in the next stage of training into a representation that is needed for the simpler task at hand. Thus, under conflicting tasks the hybrid training method proposed here should not have a significant advantage over training a single classifier in the traditional manner; a simple remedy in that case is to increase the number of hidden units, so as to achieve a richer internal representation.

4.3 Conclusion

The essence of our approach is that a good LDR can be computed if several tasks are involved in its construction. As an outcome, one obtains a rich internal representation that can be useful for all these tasks. The internal representation can be improved if tasks from different categorical levels are used concurrently. The performance on each of the tasks used in the multi-task training can be improved subsequently, by training a classifier on a particular single task, using the rich LDR as the underlying representation.

Acknowledgments

We thank Dan Roth for comments on a draft of this manuscript, and the participants in the workshop on the transfer of learning, held in Vail in December 1995, for valuable feedback.

References

Baxt, W. G. and White, H. (1995). Bootstrapping confidence intervals for clinical input variable effects in network trained to identify the presence of acute myocardial infraction. *Neural Computation*, 7(3):624–638.

Baxter, J. (1995). Learning internal representations. In *Proc. COLT'95*.

Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.

Borg, I. and Lingoes, J. (1987). *Multidimensional Similarity Structure Analysis*. Springer, Berlin.

Breiman, L. (1992). Stacked regression. Technical Report TR-367, Department of Statistics, University of California, Berkeley.

Breiman, L. (1994). Bagging predictors. Technical Report TR-421, Department of Statistics, University of California, Berkeley.

Brigham, J. C. (1986). The influence of race on face recognition. In Ellis, H. D., Jeeves, M. A., and Newcombe, F., editors, *Aspects of face processing*, pages 170–177. Martinus Nijhoff, Dordrecht.

Caruana, R. (1993). Multitask connectionist learning. In *Proceedings of the 1993 Connectionist Models Summer School*, pages 372–379, San Mateo, CA.

Caruana, R. (1995). Learning many related tasks at the same time with backpropagation. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7, pages 657–664. Morgan Kaufmann, San Mateo, CA.

Cutzu, F. and Edelman, S. (1995). Explorations of shape space. CS-TR 95-01, Weizmann Institute of Science.

Edelman, S. (1995a). Representation of similarity in 3D object discrimination. *Neural Computation*, 7:407–422.

Edelman, S. (1995b). Representation, Similarity, and the Chorus of Prototypes. *Minds and Machines*, 5:45–68.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman and Hall, London.

Gasser, M. (1995). Transfer in a connectionist model of the acquisition of morphology. CogSci TR 147, Indiana University, Bloomington, IN. an expanded version of a paper presented at the Morphology Workshop, Nijmegen, June 13, 1995.

Grossman, T. and Lapedes, A. (1993). Use of bad training data for better prediction. In J. D. Cowan, G. T. and Alspector, J., editors, *Advances in Neural Information Processing Systems*, volume 6, pages 342–350. Morgan Kaufmann.

Hintzman, D. L. (1994). Twenty-five years of learning and memory: was the cognitive revolution a mistake? In Umiltá, C. and Moscovitch, M., editors, *Attention and Performance*, volume XV, chapter 16, pages 360–391. MIT Press.

Hofmann, T. and Buhmann, J. (1994). Multidimensional scaling and data clustering. In J. D. Cowan, G. T. and Alspector, J., editors, *Neural Information Processing Systems*, volume 7, pages 459–466. Morgan Kaufmann.

Huber, P. J. (1985). Projection pursuit (with discussion). *The Annals of Statistics*, 13:435–475.

Intrator, N. (1993). Combining exploratory projection pursuit and projection pursuit regression with application to neural networks. *Neural Computation*, 5(3):443–455.

Intrator, N. and Cooper, L. N. (1992). Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5:3–17.

Kramer, A. F., Strayer, D. L., and Buckley, J. (1990). Development and transfer of automatic processing. *Journal of Experimental Psychology: Human Perception and Performance*, 16:505–522.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Lando, M. and Edelman, S. (1995). Receptive field spaces and class-based generalization from a single view in face recognition. *Network*, 6:551–576.

LeBlanc, M. and Tibshirani, R. (1994). Combining estimates in regression and classification. Preprint.

Logan, G. (1988). Towards an instance theory of automatization. *Psychological Review*, 95:492–527.

Maddox, W. T. and Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics*, 53:49–70.

Martin, G. (1988). The effects of old learning on new in hopfield and backpropagation nets. Technical Report ACA-HI-019, Microelectronics and Computer Technology Corporation (MCC).

McLaren, I. P. L., Leevers, H. J., and Mackintosh, N. J. (1994). Recognition, categorization, and perceptual learning (or, how learning to classify things together helps one to tell them apart). In Umiltá, C. and Moscovitch, M., editors, *Attention and Performance*, volume XV, chapter 35, pages 889–909. MIT Press.

Murre, J. M. J. (1995). Transfer of learning in backpropagation networks and in related neural network models. In Levy, Bairaktaris, Bullinaria, and Cairns, editors, *Connectionist Models of Memory and Language*. UCL Press, London. To appear.

Pickover, C. (1990). *Computers, Pattern, Chaos, and Beauty*. St. Martin's Press.

Pratt, L. Y. (1993). Transferring previously learned back-propagation neural networks to new learning tasks. Technical report ml-tr-37, Rutgers University, CS Dept.

Price, D., Knerr, S., Personnaz, L., and Dreyfus, G. (1995). Pairwise neural network classifiers with probabilistic outputs. In G. Tesauro, D. S. T. and Leen, T. K., editors, *Advances in Neural Information Processing 7*, pages 1109–1116. MIT Press.

Raviv, Y. and Intrator, N. (1995). Bootstrapping with noise: An effective regularization technique. Preprint.

Reder, L. and Klatzky, R. L. (1994). Transfer: training for performance. In Druckman, D. and Bjork, R. A., editors, *Learning, remembering, believing: enhancing human performance*, chapter 3, pages 25–56. National Academy Press, Washington, DC. Also available as TR CMU-CS-94-187; The effect of context on training: is learning situated?

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, 18:401–409.

Sas (1989). *SAS/STAT User's Guide, Version 6*. SAS Institute Inc., Cary, NC.

Shepard, R. N. (1966). Metric structures in ordinal data. *J. Math. Psychology*, 3:287–315.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 210:390–397.

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

Simard, P., Victorri, B., LeCun, Y., and Denker, J. (1992). Tangent prop – a formalism for specifying selected invariances in an adaptive network. In Moody, J., Lippman, R., and Hanson, S. J., editors, *Neural Information Processing Systems*, volume 4, pages 895–903. Morgan Kaufmann, San Mateo, CA.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Royal Statistics Society B*, 36:111–147.

Thrun, S. and Mitchell, T. (1995). Learning one more thing. In Mellish, C., editor, *Proc. 14th IJCAI*, volume 2, pages 1217–1223, San Mateo, CA. Morgan Kaufmann.

Young, G. and Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3:19–22.