Learning Distance Function by Coding Similarity

Aharon Bar Hillel AHARON.BAR-HILLEL@INTEL.COM
Intel research, IDC Matam 10, PO Box 1659 Matam Industrial Park, Haifa, Israel 31015

Daphna Weinshall DAPHNA@CS.HUJI.AC.IL
School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel 91904

Abstract

We consider the problem of learning a similarity function from a set of positive equivalence constraints, i.e. 'similar' point pairs. We define the similarity in information theoretic terms, as the gain in coding length when shifting from independent encoding of the pair to joint encoding. Under simple Gaussian assumptions, this formulation leads to a non-Mahalanobis similarity function which is efficient and simple to learn. This function can be viewed as a likelihood ratio test, and we show that the optimal similarity-preserving projection of the data is a variant of Fisher Linear Discriminant. We also show that under some naturally occurring sampling conditions of equivalence constraints, this function converges to a known Mahalanobis distance (RCA). The suggested similarity function exhibits superior performance over alternative Mahalanobis distances learnt from the same data. Its superiority is demonstrated in the context of image retrieval and graph based clustering, using a large number of data sets.

1. Introduction

Similarity functions play a key role in several learning and information processing tasks. One example is data retrieval, where similarity is used to rank items in the data base according to their similarity to some query item. In unsupervised graph based clustering, items are only known to the algorithm via the similarities between them, and the quality of the similarity function directly determines the quality of the clustering results. Finally, similarity functions are employed in several prominent techniques of supervised learning, from nearest neighbor classification to kernel machines. In the latter the similarity takes the form of a kernel function, and its choice is known to be a major design decision.

Good similarity functions can be designed by hand (Belongie et al., 2002; Zhang et al., 2006) or learnt from data (Shental et al., 2002; Cristianini et al., 2002; Xing et al., 2002; Hertz et al., 2004; Bilenko et al., 2004). As in other contexts, learning can help; so far, the utility of distance function learning has been demonstrated in the context of image retrieval (Hertz et al., 2003; Chang & Yeung, 2005) and clustering (Xing et al., 2002; Hertz et al., 2004; Bar-Hillel et al., 2005). Since a similarity function operates on pairs of points, the natural input to a distance learning algorithm consists of equivalence constraints, which are pairs of points labeled as 'similar' or 'not-similar' (henceforth called positive and negative equivalence constraints respectively). Several scenarios have been discussed in which constraints, which offer a relatively weak form of supervision, are readily available, while labels are much harder to achieve (Hertz et al., 2003). For example, given temporal data such as video or surveillance data, constraints may be automatically obtained based on temporal coherence.

In this paper we derive a similarity measure from general principles, and propose a simple and practical similarity learning algorithm. In general the notion of similarity is somewhat vague, involving possibly conflicting intuitions. One intuition is that similarity should be related to commonalities, i.e., two objects are similar when they share many features. This direction was studied by (Lin, 1998), and is most applicable to items described using discrete features, where the notion of common features is natural. Another intuition, suggested by (Kemp et al., 2005), measures similarity by the plausibility of a common generative process. This

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

notion draws attention to models of the hidden sources and the processes generating the visible items, which has two drawbacks: First, these models and processes are relatively complex and hard to estimate from data. Second, if such models are already known, clustering and classification can be readily done using the models directly, and so similarity judgments are not required.

Our approach focuses on the purposes of similarity judgment, which is usually to decide whether two items belong to the same class. Hence, like (Kemp et al., 2005), we relate similarity to the probability of two items belonging to the same cluster. However, unlike (Kemp et al., 2005), we model the joint distribution of 'pairs from the same cluster' directly, and estimate it from positive equivalence constraints. Given this distribution, the notion of similarity is related to the information one point conveys about the other as measured by coding length, see Section 2.1.

The basic idea is that two objects should be judged more similar the more we can 'compress' one given the information in the other. This idea is strongly related to the notion of "information distance" presented in (Bennett et al., 1998) (see also (Ziv & Merhav, 1993)), and the bottleneck method presented in (Tishby et al., 2000). As shown in Section 2.1, in addition to its coding length origin, our method can also be derived from a statistical inference perspective (as a likelihood ratio test). We regard the coding length interpretation as more intuitive, because it relates similarity to the simpler notion of predicting one point from the other via linear regression.

The main contribution of this paper is the specific application of the abstract similarity notion discussed above to continuous variables. Specifically, in Section 2.2 we develop this notion under Gaussian assumptions, deriving a simple similarity formula which is nevertheless different from a Mahalanobis metric. Intuitively, in the Gaussian setting the similarity between two points x and x' is computed by using x' to predict x via linear regression. The similarity is then related to $log\ p(x|x')$, which encodes the error of the prediction. Now learning the similarity requires only the estimation of two correlation matrices, which can be readily estimated from equivalence constraints.

The suggested similarity is strongly related to Fisher Linear Discriminant (FLD). The matrices employed in its computation are those involved in FLD, i.e., the within-class and between-class scatter matrices (Duda et al., 2001). In Section 3 we show that FLD can be derived from our similarity as the optimal linear projection. Specifically, when coding similarity is regarded as a likelihood ratio test, FLD is the projection maximizing the expected margin of the test. In addition, we explore the connection between coding similarity and the Mahalanobis metric. We show that in a certain large sample limit, coding similarity converges to the Mahalanobis metric estimated by the RCA algorithm (Bar-Hillel et al., 2005).

To evaluate our method, in Section 4 we experimentally explore two tasks: semi-supervised graph based clustering, and retrieval from a facial image database. Graph based clustering is evaluated using data sets from the UCI repository (Blake & Merz, 1998), as well as two harder data sets: the MNist data set of hand-written digits (LeCun et al., 1998), and a data set of animal images (Hertz et al., 2004). We used the YaleB data set of facial images (Georghiades et al., 2000) for face retrieval experiments. In both tasks Gaussian Coding Similarity (GCS) usually outperforms Mahalanobis metrics, learnt by three readily available algorithms (Xing et al., 2002; Bar-Hillel et al., 2005; De-Bie et al., 2003). In terms of computational complexity, the method of (Xing et al., 2002) is relatively demanding, as it is based on iterative nonlinear optimization, while the two other methods offer closed-form solutions based on positive constraints alone. The computational cost of coding similarity is low, similar to the methods of (Bar-Hillel et al., 2005; De-Bie et al., 2003) and much smaller than in the method of (Xing et al., 2002).

2. Similarity based on Coding Length

2.1. General definition

Intuitively, two items are similar if they share common aspects, whereby one can be used to predict some details of the other. Learning similarity is learning what aspects tend to be shared more than others, between points which are equivalent w.r.t a certain goal. Such a similarity notion is naturally related to the joint distribution $p(x, x'|H_1)$, where H_1 is the hypothesis stating that the two points share the same label. We estimate $p(x, x'|H_1)$, and define the similarity $codsim(x, x')$ between two items to be the information one conveys about the other. We measure this information using the coding length $cl(x)$, i.e. the negative logarithm of an event (Cover & Thomas, 1991). The similarity is defined as the gain in coding length obtained by encoding x when x' is known.

$$\begin{aligned} codsim(x, x') &= cl(x) - cl(x|x', H_1) \\ &= \log p(x|x', H_1) - \log p(x) \end{aligned} \tag{1}$$

As stated in (Kemp et al., 2005), this measurement

can be also viewed as a log-likelihood ratio statistic:

$$\begin{aligned}&\log p(x|x', H_1) - \log p(x) \\ &= \log \frac{p(x, x'|H_1)}{p(x)p(x')} = \log \frac{p(x, x'|H_1)}{p(x, x'|\mathrm{H}_0)}\end{aligned} \tag{2}$$

where H_0 denotes the hypothesis stating independence between the points. The coding similarity is therefore the optimal statistic for determining whether two points are drawn from the same class or independently. We can see from this last equation that it is a symmetric function.

The exact nature of the distribution $p(x, x'|H_1)$ may vary depending on the application and the oracle from which the equivalence constraints are obtained. Consider data sampled from several sources in R^d, i.e. $p(x) = \sum_{k=1}^{M} \alpha_k p(x|h_i)$, where $p(x|h_i)$ denotes the distribution of the i-th source. A simple form for $p(x, x'|H_1)$ is obtained when the two points are conditionally independent given the hidden source:

$$p(x, x'|H_1) = \sum_{k=1}^{M} \alpha_k p(x|h_k)p(x'|h_k) \tag{3}$$

This distribution, defined over pairs, corresponds to sampling pairs by first choosing the hidden source, followed by the independent choice of two points from this source. In section 3.1 we show that FLD is the optimal similarity-preserving dimensionality reduction when the equivalence constraints are sampled in this manner. In Section 3.2 we discuss a common case in which the conditional independence between the points is violated.

2.2. Gaussian coding similarity

We now develop the coding similarity notion under some simplifying Gaussian assumptions:

- $p(x, x'|H_1)$ is Gaussian (in R^{2d})
- $p(x) = \int\limits_{x} p(x, x'|H_1) = \int\limits_{x'} p(x, x'|H_1)$

The second assumption is the reasonable (though not always trivially satisfied) demand that $p(x)$ should be the marginal distribution of $p(x, x'|H_1)$ w.r.t both arguments. It is clearly satisfied for distribution (3). It follows from the first assumption that $p(x)$ is also Gaussian (in R^d). The first assumption is clearly a simplification of the true data density in all but the most trivial cases. However, its value lies in its simplicity, which leads to a coding scheme that is efficient and easy to estimate. While clearly inaccurate, we propose here that this model can be very useful.

We assume w.l.o.g that the data's mean is 0 (otherwise, we can subtract it from the data), and so we can parameterize the two distributions using two matrices. Denoting the Gaussian distribution by $G(\cdot|\mu, \Sigma)$ we have

$$\begin{aligned}p(x) &= G(x|0, \Sigma_x) \\ p(x, x'|H) &= G(x, x'|0, \Sigma_{2x}) \\ \Sigma_{2x} &= \begin{pmatrix} \Sigma_x & \Sigma_{xx'} \\ \Sigma_{xx'} & \Sigma_x \end{pmatrix}\end{aligned} \tag{4}$$

where $\Sigma_x = E[xx^t]$, $\Sigma_{xx'} = E[x(x')^t]$. The conditional density $p(x|x', H)$ is also Gaussian $G(x|Mx', \Sigma_{x|x'})$, with $M, \Sigma_{x|x'}$ given by

$$M = \Sigma_{xx'}\Sigma_x^{-1} \qquad \Sigma_{x|x'} = \Sigma_x - \Sigma_{xx'}\Sigma_x^{-1}\Sigma_{xx'} \tag{5}$$

Plugging this into Eq. (1), we get the following expression for Gaussian coding similarity:

$$\begin{aligned}&\log p(x|x', H_1) - \log p(x) = \\ &\log G(x|Mx', \Sigma_{x|x'}) - \log G(x|0, \Sigma_x) =\end{aligned} \tag{6}$$

$$\frac{1}{2}[\log \frac{|\Sigma_x|}{|\Sigma_{x|x'}|} + x^t\Sigma_x^{-1}x - (x - Mx')^t\Sigma_{x|x'}^{-1}(x - Mx')]$$

The Gaussian coding similarity can be easily and almost instantaneously learnt from a set of positive equivalence constraints, as summarized in Algorithm 1. Learning includes the estimation of several statistics, mainly the matrices Σ_x,$\Sigma_{xx'}$, from which the matrices $M, \Sigma_{x|x'}$ are computed. Notice that each constraint is considered twice, once as (x, x') and once as (x', x), to ensure symmetry and to satisfy the marginalization demand. Given those statistics, similarity is computed using Eq. (8), which is based on Eq. (6) but with the multiplicative and additive constants removed.

3. Relation to other methods

In this section we provide some analysis connecting Gaussian coding similarity as defined above to other known learning techniques. In Section 3.1 we discuss the underlying connection between GCS and FLD dimensionality reduction. In Section 3.2 we show that under certain estimation conditions, the dominant term in GCS behaves like a Mahalanobis metric, and specifically that it converges to the RCA metric (Bar-Hillel et al., 2005).

3.1. The optimality of FLD

As defined in Eqs. (5)-(8), the coding similarity depends on two matrices only - the data covariance matrix Σ_x and the covariance between pairs from the

Algorithm 1 Gaussian coding similarity

Learning procedure:
Input: a set of equivalence constraints $\{x_i, x'_i\}_{i=1}^N$, and optionally a dimension parameter k.

1. Compute the mean $Z = \frac{1}{2N}\sum_{i=1}^N [x_i + x'_i]$ and subtract it from the training data
2. Estimate $\Sigma_x, \Sigma_{xx'}$

$$\Sigma_x = \frac{1}{2N}\sum_{i=1}^N [x_i x_i^t + x'_i x_i'^t] \quad (7)$$
$$\Sigma_{xx'} = \frac{1}{2N}\sum_{i=1}^N [x_i x_i'^t + x'_i x_i^t]$$

3. If dimensionality reduction is required, find the k eigenvectors with the highest eigenvalues of $\Sigma_x^{-1}\Sigma_{xx'}$ and put them into $A \in M_{d\times k}$.
   Let $\Sigma_x = A^t\Sigma_x A$, $\Sigma_{xx'} = A^t\Sigma_{xx'}A$, $Z = ZA$
4. Compute M and $\Sigma_{x|x'}$ according to Eq. (5).

Return $Z, M, \Sigma_x^{-1}, \Sigma_{x|x'}^{-1}$ and A (if computed).

Similarity computation for a pair (x, x'):
If A is defined $x = xA - Z$, $x' = x'A - Z$
else $x = x - Z$, $x' = x' - Z$.

Return $codsim(x, x') =$

$$x^t\Sigma_x^{-1}x - (x - Mx')^t\Sigma_{x|x'}^{-1}(x - Mx') \quad (8)$$

same source $\Sigma_{xx'}$. To establish the connection to FLD, let us first consider the expected value of $\Sigma_{xx'}$ under distribution (3):

$$E_{p(x,x'|H_1)}[x(x')^t] = \quad (9)$$
$$\int_x\int_{x'}\sum_{k=1}^M \alpha_k p(x|h_k)p(x'|h_k)x(x')^t =$$
$$\sum_{k=1}^M \alpha_k E_{p(x|h_k)}[x]\cdot E_{p(x'|h_k)}[(x')^t] = \sum_{k=1}^M \alpha_k m_k m_k^t$$

The expected value above, which gives the convergence limit of $\Sigma_{xx'}$ as estimated in Eq. (7), is essentially the between-class scatter matrix S_B used in FLD (Duda et al., 2001). The main difference between the estimation of $\Sigma_{xx'}$ in Eq. (7) and the traditional estimation of S_B is the training data, equivalence constraints vs. labels respectively. Also, while S_B is always of rank $k-1$ and so is its estimator based on labeled data, our estimator from Eq. (7) is usually of full rank.

When the data distributions $p(x, x'|H_1)$ and $p(x)$ lie in high dimensional space, in many cases the projection into a lower dimensional space may increase learning accuracy (by dropping irrelevant dimensions) and computational efficiency. We now characterize the notion of *optimal dimensionality reduction* based on the 'natural margin' of the likelihood ratio test. This test gives the optimal rule (Cover & Thomas, 1991) for deciding between two hypotheses H_0 and H_1, where the data comes from a mixture $p(x) = \alpha p(x|H_0) + (1-\alpha)p(x|H_1)$:

$$decide\ H_1 \iff \log\frac{p(X|H_1)}{p(X|H_0)} > \log\frac{1-\alpha}{\alpha} \quad (10)$$

Hypothesis margin: Let the label of point x be 1 if hypothesis H_1 is true, and -1 if H_0 is true. The natural margin of a point x can be defined as $y_i(\log\frac{p(x_i|h_1)}{p(x_i|h_0)} - \log\frac{1-\alpha}{\alpha})$.

Given this definition, the expected margin of the test is

$$E_x[y(x)(\log\frac{p(x|h_1)}{p(x|h_0)} - \log\frac{1-\alpha}{\alpha})] \quad (11)$$
$$= \alpha\int_x p(x|h_1)[\log\frac{p(x|h_1)}{p(x|h_0)} - \log\frac{1-\alpha}{\alpha}]dx$$
$$-(1-\alpha)\int_x p(x|h_0)[\log\frac{p(x|h_1)}{p(x|h_0)} - \log\frac{1-\alpha}{\alpha}]dx$$
$$= \alpha D_{kl}[p(x|h_1)||p(x|h_0)]$$
$$+ (1-\alpha)D_{kl}[p(x|h_0)||p(x|h_1)]$$
$$+ (1-2\alpha)\log\frac{1-\alpha}{\alpha}$$

Optimal dimensionality reduction: $A \in M_{d\times k}$ is the optimal linear projection from dimension d to k if it maximizes the expected margin defined above.

Theorem 1. *Assume Gaussian distributions $p(x, x'|H_1)$ in R^{2d} and $p(x)$ in R^d, and a linear projection $A \in M_{d\times k}$ where $z = A^t x$. For all $0 \le \alpha \le 1$, the optimal A*

$$A^* = \arg\max_{A\in M_{d\times k}}[\ \alpha D_{kl}[p(z, z'|H_1)||p(z, z'|H_0)] + \quad (12)$$
$$(1-\alpha)D_{kl}[p(z, z'|H_0)||p(z, z'|H_1)]\ ]$$

is the FLD transformation. Thus A is composed of the k eigenvectors of $\Sigma_x^{-1}\Sigma_{xx'}$ with the highest eigenvalues.

The proof of this theorem is relatively complex and we only describe here a very general sketch. Since the distributions involved in Eq. (12) are Gaussian, the

$D_{kl}[\cdot||\cdot]$ can be written in closed-form. We can upper bound these terms by using $A^t\Sigma_{x|x'}A$ to approximate $\Sigma_{z|z'}$. The approximate bound, for fixed α, can be shown to obtain its maximal value at the k eigenvectors of $\Sigma^{-1}_{x|x'}\Sigma_x$ with the highest eigenvalues. These vectors, in turn, are identical to the highest eigenvectors of $\Sigma_x^{-1}\Sigma_{xx'}$. Finally, it is shown that the optimum of the upper bound is also obtained by the original expression, using the same matrix A.

3.2. The Mahalanobis limit

Above we have considered specifically coding similarity with pairs distribution of the form (3). However, in practice the source of equivalence constraints often does not produce an unbiased sample from this distribution. Specifically, equivalence constraints which are obtained automatically, e.g. from a surveillance camera, are often biased and tend to include only very similar (close in the Euclidean sense) points in each pair. This happens since constraints are extracted based on temporal proximity, and hence include highly dependent points. When the points in all pairs are very close to each other, the best regression from one to the other is close to the identity matrix. The following theorem states that under these conditions, coding similarity converges to a Mahalanobis metric.

Theorem 2. *Assume that equivalence constraints are generated by sampling the first point x from $p(x)$ and then x' from a small neighborhood of x. Denote $\Delta = (x-x')/2$. Assume that the covariance matrix $\Sigma_\Delta < \epsilon\Sigma_x$, where $\epsilon > 0$ and $A \leq B$ stands for "$B-A$ is a p.s.d matrix". Then*

$$codsim(x,x') \underset{\varepsilon\to 0}{\to} - (x-x')^t(4\Sigma_\Delta)^{-1}(x-x') \quad (13)$$

where the limit $g(x) \to f(x)$ means $g(x)/f(x) \to 1$.

Proof. We concentrate on approximating the second term in Eq. (8), which involves both x and x'. Denote $\bar{x} = (x+x')/2$, so $x = \bar{x}+\Delta$, $x' = \bar{x}-\Delta$. We get the following estimates for Σ_x , $\Sigma_{xx'}$:

$$\begin{aligned}\Sigma_x &= \frac{1}{2}E(\bar{x}-\Delta)(\bar{x}-\Delta)^t + \frac{1}{2}E(\bar{x}+\Delta)(\bar{x}+\Delta)^t \\ &= \Sigma_{\bar{x}} + \Sigma_\Delta \\ \Sigma_{xx'} &= E(\bar{x}+\Delta)(\bar{x}-\Delta)^t = \Sigma_{\bar{x}} - \Sigma_\Delta\end{aligned}$$

We therefore see that $\Sigma_{xx'} = \Sigma_x - 2\Sigma_\Delta$, and obtain the following approximations for M,$\Sigma_{x|x'}$:

$$\begin{aligned}M &= \Sigma_{xx'}\Sigma_x^{-1} = (\Sigma_x - 2\Sigma_\Delta)\Sigma_x^{-1} \\ &= I - 2\Sigma_\Delta\Sigma_x^{-1} \geq I - 2\varepsilon \\ \Sigma_{x|x'} &= \Sigma_x - (I - 2\Sigma_\Delta\Sigma_x^{-1})(\Sigma_x - 2\Sigma_\Delta) \\ &= 4\Sigma_\Delta - 4\Sigma_\Delta\Sigma_x^{-1}\Sigma_\Delta \geq 4\Sigma_\Delta(I-\varepsilon)\end{aligned}$$

These inequalities lower bound M and $\Sigma_{x|x'}$, and since $\Sigma_\Delta\Sigma_x^{-1}$ is p.s.d it is clear from the equalities above that $M \leq I$, $\Sigma_{x|x'} \leq 4\Sigma_\Delta$. We hence get that $M = I+O(\epsilon)$ and $\Sigma_{x|x'} = 4\Sigma_\Delta(I+O(\epsilon))$.

Returning to Eq. (8), we note that the first term $0 < x^t\Sigma_x^{-1}x < \epsilon x^t\Sigma_\Delta^{-1}x$ is negligible w.r.t the second in the limit of $\epsilon \to 0$. Therefore this term can be rigorously omitted, and we get:

$$\begin{aligned}codsim(x,x') \approx &-(x-[I+O(\varepsilon)]x')^t \cdot \\ &[4\Sigma_\Delta(I+O(\varepsilon))]^{-1} \cdot (x-[I+O(\varepsilon)]x') \\ \underset{\varepsilon\to 0}{\to} &-(x-x')^t(4\Sigma_\Delta)^{-1}(x-x')\end{aligned}$$

Note that $codsim(x,x')$ is negative as appropriate, since it measures similarity rather than distance. □

The Mahalanobis matrix $\Sigma_\Delta = E_{p(x,x'|H_1)}[x-(x+x')/2]$ is actually the inner chunklet covariance matrix, as defined in (Bar-Hillel et al., 2005). It is therefore the RCA transformation, estimated from the population of 'near' point pairs.

4. Experimental validation

We first present several experiments with synthetic data sets in section 4.1, testing the potential value of GCS under controlled conditions. In Section 4.2 we compare several methods in a semi-supervised clustering task, where the data is augmented by equivalence constraints. Finally, in Section 4.3 we test Gaussian coding similarity and other methods in a face retrieval task.

Data sets In Section 4.2 we have experimented with nine data sets from the UCI repository (Blake & Merz, 1998) and added two harder data sets, with 10 classes each: A subset of the MNist digits data set (LeCun et al., 1998), and a data set of animals images (Hertz et al., 2004). For MNist, we randomly chose 50 instances from each digit, and represented the data using 50 PCA dimensions. The animals data set includes 565 images taken from a commercial CD. As in (Hertz et al., 2004) we represent the images using Color Coherence Histograms (CCV) (Pass et al., 1996), containing information about color distribution and color continuity in the image. The vectors were then reduced to 100 PCA dimensions. For the evaluation of retrieval performance in section 4.3 we used the YaleB data set (Georghiades et al., 2000). The variability in this data set is mainly due to change of illumination. Images were aligned using optical flow, and then reduced to 60 PCA dimensions.

Constraint oracle Following (Hertz et al., 2003), we obtained equivalence constraints by simulating a

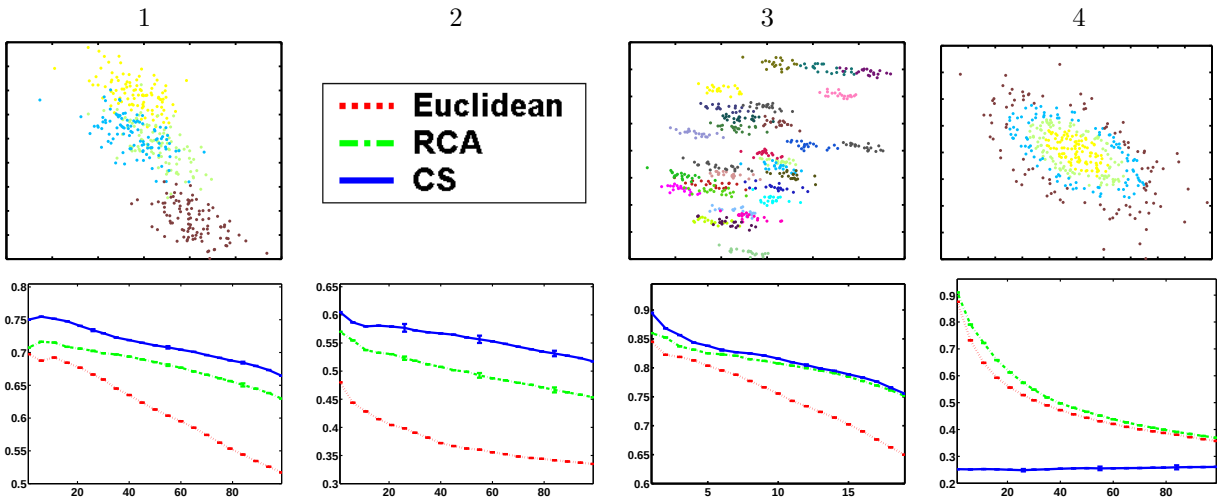

Figure 1. **Top** 3 of the 4 synthetic data sets used (the forth data set (number 2) is the analog of data 1 in R^5). **Bottom** Cumulative neighbor purity curves for the 4 synthetic data sets. The Y axis shows the percentage of correct neighbors vs. the number of neighbors considered (the X axis). In each graph we compare Gaussian coding similarity, RCA and the Euclidean metric. Coding similarity has a considerable edge when the data contains several Gaussian classes. For many small classes (Data 3) GCS approaches RCA, and for the non-convex rings data it fails. Results were averaged over 50 realizations. The figure is best seen in color.

distributed learning scenario, in which small subsets of labels are provided by a number of uncoordinated teachers. Accordingly, we randomly chose small subsets of data points from the dataset, and partitioned each subset into equivalence classes. The constraints obtained from all the subsets are gathered and used for learning. In all the experiments we chose the size of each subset to be $S = 2M$, where M is the number of classes in the data. In each experiment we used N/S subsets, where N is the total number of points in the data. While the number of constraints thus provided is linear in the sample size N, notice that it is a small fraction of all possible pairs of data points, which is $\mathcal{O}(N^2)$. Whenever tested, the method of (Xing et al., 2002) is given both the positive and the negative constraints, while the other tested methods use only the positive constraints.

Compared methods We compare the Gaussian coding similarity to 3 learning techniques which learn a Mahalanobis metric. The method presented in (Xing et al., 2002) learns the metric by non-linear optimization, using iterative projections. The RCA metric, suggested in (Bar-Hillel et al., 2005), is essentially the inverse of the inner class covariance matrix, as estimated from the equivalence constraints. The method of (De-Bie et al., 2003) learns a low-rank Mahalanobis metric based on the FLD dimensionality reduction. The Mahalanobis matrix is A^tA where A is the estimated FLD matrix.

4.1. Synthetic data

We conducted a series of experiments on synthetic data, comparing the performance of GCS to RCA and the Euclidean metric in several interesting conditions. The first three data sets were generated by choosing class centers from a Gaussian distribution $G_1(M|0, \Sigma_1)$, then selecting class points from Gaussians around those centers $G_2(x|M, \Sigma_2)$. When the number of classes is large, $p(x)$ is Gaussian as the convolution of G_1 and G_2, and the GCS assumptions are fully met. The first data set was generated by sampling 400 two-dimensional points in four classes using this protocol. A second data set was produced in a similar manner, but with points in R^5. The third data set includes 600 points in 30 classes with relatively small variance, thus approaching the limit discussed is Theorem 2. Finally we produced a data set of concentric rings, which critically violates class convexity. The data sets are shown in the top row of Figure 1.

The bottom row of Figure 1 shows commulative purity graphs of GCS, RCA and the Euclidean metric on the synthetic data sets. These results give an overview of the strength and weaknesses of GCS. When the data contains several Gaussian data sets (Data 1), GCS has a clear advantage over RCA and the Euclidean metric. This advantage is more pronounced in higher dimension (Data 2). In the case of many small classes with shared covariance (Data 3) the performance of GCS becomes closer to the performance of RCA, as pre-

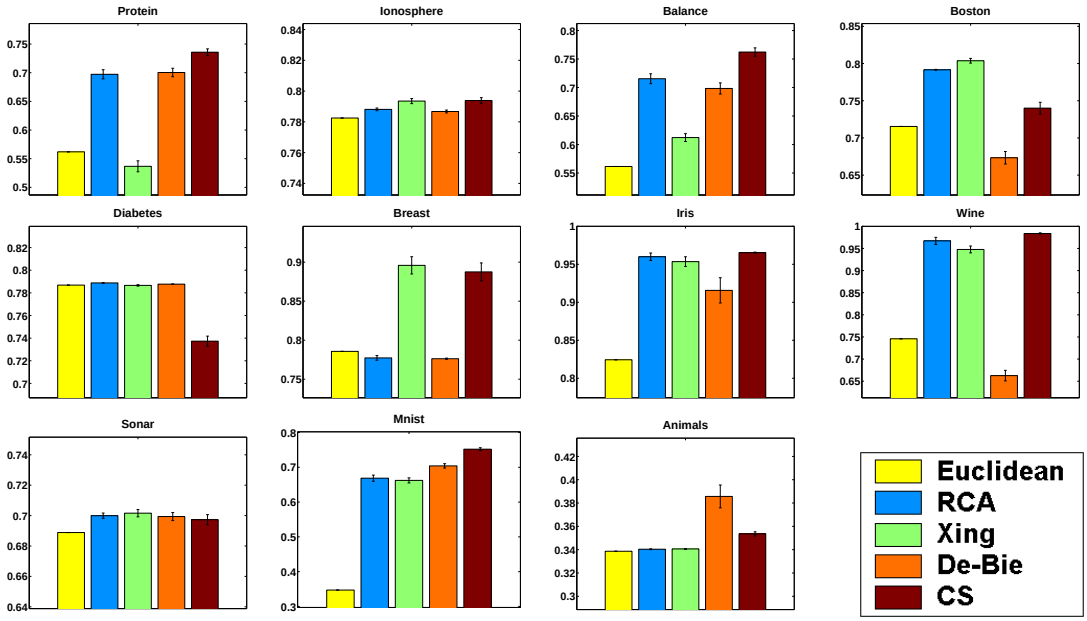

Figure 2. Clustering performance with the average linkage clustering algorithm, using several similarity functions. Performance is measured using the $F_{\frac{1}{2}} = \frac{2PR}{R+P}$ score, where P denotes precision rate and R denotes recall rate. The results are averaged over 50 constraint realizations. The figure is best seen in color.

dicted by Theorem 2. Finally, in data 4 the Gaussian assumptions are severely violated, as class centers are all identical to each other and classes are non convex. In this case, GCS totally fails as expected.

4.2. Clustering with equivalence constraints

Graph based clustering includes a rich family of algorithms, among them are the relatively simple agglomerative linkage algorithms used here (Duda et al., 2001). In graph based clustering, pairwise similarity is the sole source of information regarding the clustered data. Given equivalence constraints, one can adapt the similarity function to the specific problem, and improve clustering results considerably. In our experiments, we have evaluated clustering results obtained using several distance functions: The Euclidean metric, the Mahalanobis metrics learnt using the algorithms mentioned above, and the Gaussian coding similarity. The distance functions were evaluated by applying the agglomerative average linkage algorithm to the similarity graphs produced. Clustering performance was assessed by computing the match between clustering results and the real (known) data labels.

We tested the different similarities in the original space first, and after reducing the data dimension to the number of classes using constrained-based FLD. The results after FLD, which are usually better, are summarized in Figure 2. The ranking of the different algorithms has a large variance, but Coding similarity (rightmost bar, in brown) gives the best average performance, with 5 cases in which it outperforms all the other metrics and 3 cases of being second best. The results in the original space show a similar trend.

4.3. Facial image retrieval

The YaleB data set (Georghiades et al., 2000) contains 64 images per person of 30 people. From each class, we randomly chose 48 images to be part of the 'data base', and used the remaining 16 as queries presented to the data base. We learned the three Mahalanobis distances and coding similarity using constraints obtained from 25 of the 30 classes, and then evaluated retrieval performance for images from both constrained and unconstrained classes. Notice that for unconstrained classes the task is much harder, and any success shows inter-class generalization, since images from these classes were not used during training.

The performance of the four learning methods and the Euclidean metric in the original 60 dimensional space are shown in Figure 3. We can see that coding similarity is clearly superior to other methods in the retrieval of faces from known classes. In contrast to other methods, it operates well even in the original 60 dimensional space. It also has a small advantage in the 'learning-to-learn' scenario, i.e., in the retrieval of faces from unconstrained classes.

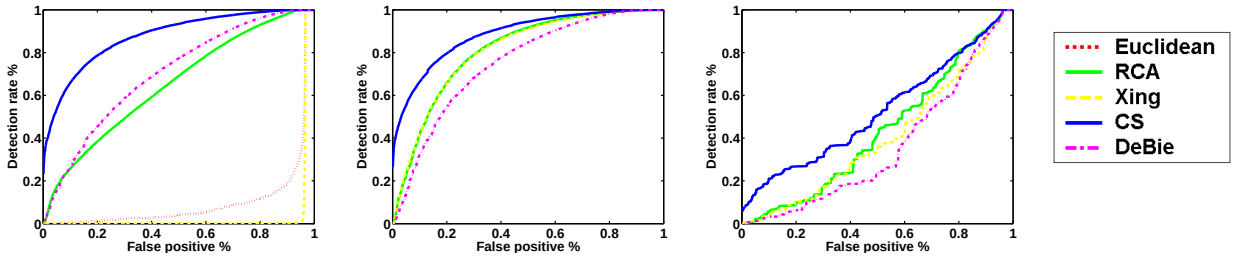

Figure 3. ROC curves for several methods in a face retrieval task. **Left:** Retrieval of test images from constrained classes using 60 PCA dimensions. **Middle:** Retrieval of images from constrained classes using 18 FLD dimensions. **Right:** Retrieval of test queries from unconstrained classes using 18 FLD dimensions. Results were averaged over 20 constraints realizations. The figure is best seen in color.

5. Summary

We described a new measure of similarity between two datapoints, based on the gain in coding length of one point when the other is known. This similarity measure can be efficiently computed from positive equivalence constraints. We showed the relation of this measure to Fisher Linear Discriminant (FLD), and to relevant component analysis (RCA). We demonstrated overall superior performance of the suggested similarity in clustering and retrieval, using a large number of datasets.

Acknowledgements: This research was supported by the EU under the DIRAC integrated project IST-0277.

References

Bar-Hillel, A., Hertz, T., Shental, N., & Weinshall, D. (2005). Learning a mahalanobis metric from equivalence constraints. *JMLR*, *6(Jun)*, 937–965.

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape context. *IEEE PAMI*, *24*, 509–522.

Bennett, C., Gacs, P., Li, M., Vitanyi, P., & Zurek, W. (1998). Information Distance. *IEEE Trans. Information Theory*, *44*, 1407.

Bilenko, M., Basu, S., & Mooney, R. (2004). Integrating constraints and metric learning in semi-supervised clustering. *Proc. ICML* (pp. 81–88).

Blake, C., & Merz, C. (1998). UCI repository of machine learning databases.

Chang, H., & Yeung, D. (2005). Stepwise metric adaptation based on semi-supervised learning for boosting image retrieval performance. *BMVC*.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. John Wiley and Sons Inc.

Cristianini, N., Kandola, J., Elissee, A., & Shawe-Taylor, J. (2002). On kernel target alignment. *Proc. NIPS*.

De-Bie, T., Momma, M., & Cristianini, N. (2003). Efficiently learn the metric with side information. *Lecture Notes in Artificial Intelligence* (pp. 175 – 189).

Duda, R., Hart, P., & Stork, D. (2001). *Pattern Classification*. John Wiley and Sons Inc.

Georghiades, A., Belhumeur, P., & Kriegman, D. (2000). From few to many: Generative models for recognition under variable pose and illumination. *IEEE Int. Conf. on Automatic Face and Gesture Recognition*.

Hertz, T., Bar-Hillel, A., & Weinshall, D. (2004). Boosting margin based distance functions for clustering. *Proc. ICML*.

Hertz, T., Shental, S., Bar-Hillel, A., & Weinshall, D. (2003). Enhancing image and video retrieval: Learning with equivalence constraints. *Proc. CVPR*.

Kemp, C., Bernstein, A., & Tenenbaum, J. (2005). A generative theory of similarity. *The Twenty-Seventh Annual Conference of the Cognitive Science Society*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE*, *86*, 2278–2324.

Lin, D. (1998). An information-theoretic definition of similarity. *Proc. ICML*.

Pass, G., Zabih, R., & Miller, J. (1996). Comparing images using color coherence vectors. *ACM Multimedia*.

Shental, N., Hertz, T., Weinshall, D., & Pavel, M. (2002). Adjustment learning and relevant component analysis. *Proc. ECCV*.

Tishby, N., Pereira, F., & Bialek, W. (2000). The information bottleneck method. *Arxiv preprint physics/0004057*.

Xing, E., Ng, A., Jordan, M., & Russell, S. (2002). Distance metric learnign with application to clustering with side-information. *Proc. NIPS*.

Zhang, H., Berg, A., Maire, M., & Malik, J. (2006). Svm-knn: Discriminative nearest neighbor classification for visual category recognition. *Proc. CVPR*.

Ziv, J., & Merhav, N. (1993). A measure of relative entropy between individual sequences withapplication to universal classification. *IEEE Trans. Information Theory*, *39*, 1270–1279.