Learning Distance Functions using Equivalence Relations

Aharon Bar-Hillel AHARONBH@CS.HUJI.AC.IL
Tomer Hertz TOMBOY@CS.HUJI.AC.IL
Noam Shental FENOAM@CS.HUJI.AC.IL
Daphna Weinshall DAPHNA@CS.HUJI.AC.IL
School of Computer Science and Engineering and the Center for Neural Computation, The Hebrew University of Jerusalem, Jerusalem, Israel 91904

Abstract

We address the problem of learning distance metrics using side-information in the form of groups of "similar" points. We propose to use the RCA algorithm, which is a simple and efficient algorithm for learning a full ranked Mahalanobis metric (Shental et al., 2002). We first show that RCA obtains the solution to an interesting optimization problem, founded on an information theoretic basis. If the Mahalanobis matrix is allowed to be singular, we show that Fisher's linear discriminant followed by RCA is the optimal dimensionality reduction algorithm under the same criterion. We then show how this optimization problem is related to the criterion optimized by another recent algorithm for metric learning (Xing et al., 2002), which uses the same kind of side information. We empirically demonstrate that learning a distance metric using the RCA algorithm significantly improves clustering performance, similarly to the alternative algorithm. Since the RCA algorithm is much more efficient and cost effective than the alternative, as it only uses closed form expressions of the data, it seems like a preferable choice for the learning of full rank Mahalanobis distances.

Keywords: Learning from partial knowledge, semi-supervised learning, feature selection, clustering

1. Introduction

Many learning algorithms use a distance function over the input space as a principal tool, and their performance critically depends on the quality of the metric. Learning a "good" metric from examples may therefore be the key to a successful application of these algorithms. In many cases choosing the right metric may be more important than the specific algorithm which is later used.

Choosing the right metric is especially important in the unsupervised setting of clustering tasks, for such clustering algorithms as K-means and graph based methods. There are also supervised classification techniques which are distance based such as K-Nearest-Neighbors. Kernel machines use inner-product functions which are closely related to the Euclidean distance metric. In this wide variety of algorithms the problem of finding a good metric is equivalent to the problem of finding a good *representation function* $f : X \rightarrow Y$, transferring the data X into representation Y. We will therefore discuss the two problems interchangeably. Our main goal in this paper is to design a simple method for learning a metric, in order to improve the subsequent performance of unsupervised learning techniques. This is accomplished using side-information in the form of equivalence relations. Equivalence relations provide us with small groups of data points that are known to be similar (or dissimilar).

A key observation is that in many unsupervised learning tasks, such groups of similar points may be extracted from the data with minimal effort and possibly automatically, without the need for labels. This occurs when the data originates from a natural sequence that can be modeled as a Markovian process. Consider for example the task of movie segmentation, where the objective is to find all the frames in which the same actor appears. Due to the continuous nature of most movies, faces extracted from successive frames in roughly the same location can be assumed to come from the same person. This is true as long as there is no scene change, which can be automatically and robustly detected (Boreczky & Rowe, 1996). Another analogous example is speaker segmentation and recognition, in which a conversation between several

Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

speakers needs to be segmented and clustered according to the speaker identity. Here, it may be possible to automatically identify small segments of speech which are likely to contain data points from a single *unknown* speaker.

In this paper we discuss the problem of learning linear *representation functions*, or equivalently an optimal Mahalanobis distance between data points, using equivalence relations. Specifically, we focus here on the Relevant Component Analysis (RCA) algorithm, which was first introduced in (Shental et al., 2002); the algorithm is reviewed in Section 2. In Section 3 we present a new analysis, based on a novel information theoretic optimality criterion. RCA is shown to be an optimal learning procedure in this sense. We show that Fisher's linear discriminant function followed by RCA optimizes the same criterion if dimensionality reduction is allowed.

In Section 4 we show that RCA can be presented as an optimal solution to a problem of minimizing inner class distances. Viewed this way, RCA can be directly compared with the approach proposed in (Xing et al., 2002), which is another recent algorithm for metric learning with side information. The comparison shows that the optimality criteria of the two algorithms are similar, but some arbitrary aspects of the criterion presented in (Xing et al., 2002) do not exist in RCA. Our empirical study also shows that the results of the algorithms are comparable: We empirically tested the RCA algorithm on a number of databases from the UCI repository, showing significant improvement in clustering performance which is similar or better than the improvement reported in (Xing et al., 2002). The major difference between the two algorithms is computational: RCA is robust and efficient since it only uses closed-form expressions of the data; the algorithm described in (Xing et al., 2002), on the other hand, uses iterative methods which are sensitive to parameter tuning and which are very demanding computationally.

Related work

There has been much work on learning representations and distance functions in the supervised learning setting, and we can just briefly mention some examples. (Hastie & Tibshirani, 1996) and (Jaakkola & Haussler, 1998) use labeled data to learn good metrics for classification. In (Thrun, 1996) a distance function (or a representation function) is learned for classification using a "leaning-to-learn" paradigm. In this setting several related classification tasks are learned using several labeled data sets, and algorithms are proposed which learn representations and distance functions in a way that allows for the transfer of knowledge between the tasks. In (Tishby et al., 1999) the joint distribution of two random variables X and Y is assumed to be known, and the problem is reduced to the learning of a compact representation of X which bears high relevance to Y. This work, which is further developed in (Chechik & Tishby, 2002), can be viewed as supervised representation learning. Information theoretic criteria for unsupervised learning in neural networks were first suggested by (Linsker, 1989), and has been used since in several tasks in the neural network literature, e.g., (Bell & Sejnowski, 1995).

In recent years some work has been done using equivalence relations as side information. In (Wagstaff et al., 2001) equivalence relations were introduced into the K-means clustering algorithm. Both positive ('a is similar to b') and negative ('a is dissimilar from b') relations were used. The problem of finding a better Mahalanobis metric using equivalence relations was addressed in (Xing et al., 2002), in conjunction with the constrained K-means algorithm. We compare this algorithm to our current work in Section 4, and compare our empirical results with the results of both algorithms in section 6. We have also recently developed a way to introduce both positive and negative equivalence relations into the EM algorithm for the estimation of a mixture of Gaussian models (Hertz et al., 2002; Shental et al., 2003).

2. Relevant Component Analysis

Relevant Component Analysis (RCA) is a method that seeks to identify and down-scale global unwanted variability within the data. The method changes the feature space used for data representation, by a global linear transformation which assigns large weights to "relevant dimensions" and low weights to "irrelevant dimensions" (cf. (Tenenbaum & Freeman, 2000)). These "relevant dimensions" are estimated using chunklets. We define a *chunklet* as a subset of points that are known to belong to the same although *unknown* class; chunklets are obtained from equivalence relations by applying a transitive closure. The RCA transformation is intended to reduce clutter, so that in the new feature space, the inherent structure of the data can be more easily unraveled. The method can be used as a preprocessing step for the unsupervised clustering of the data or nearest neighbor classification.

Specifically, RCA does the following (see illustration in Fig. 1a-f):

1. For each chunklet, subtract the chunklet's mean

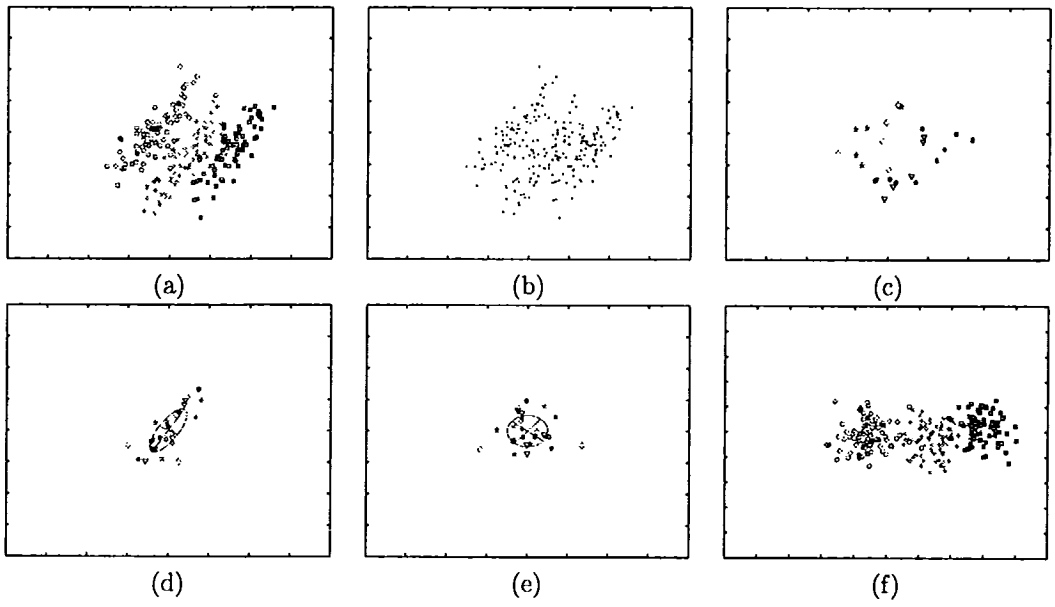

Figure 1. An illustrative example of the RCA algorithm applied to synthetic Gaussian data. (a) The fully labeled data set with 3 classes. (b) Same data unlabeled; clearly the classes' structure is less evident. (c) The set of chunklets that are provided to the RCA algorithm (points that share the same color and marker type form a chunklet). (d) The centered chunklets, and their empirical covariance. (e) The whitening transformation applied to the chunklets. (f) The original data after applying the RCA transformation.

from all of the points it contains (Fig. 1d).

2. Compute the covariance matrix of all the centered data-points in chunklets (Fig. 1d). Assume a total of p points in k chunklets, where chunklet j consists of points $\{x_{ji}\}_{i=1}^{n_j}$ and its mean is $\hat{m}_j$. RCA computes the following matrix:

$$\hat{C} = \frac{1}{p}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji} - \hat{m}_j)(x_{ji} - \hat{m}_j)^t \qquad (1)$$

3. Compute the whitening transformation $W = \hat{C}^{-\frac{1}{2}}$ associated with this covariance matrix (Fig. 1e), and apply it to the original data points: $x_{new} = Wx$ (Fig. 1f). Alternatively, use the inverse of $\hat{C}$ as a Mahalanobis distance.

In effect, the whitening transformation W assigns lower weight to some directions in the original feature space; those are the directions in which the data variability is mainly due to within class variability, and is therefore "irrelevant" for the task of classification.

3. Information maximization under chunklet constraints

In this section we suggest an information theoretic formulation for the problem at hand. The problem is formulated as a constrained search for a good *representation function* . Although it is possible to state the problem for general families of transformations, we treat here only the linear case. In section 3.1 we present and discuss the problem formulation. In 3.2 we show that RCA solves this problem when only linear invertible transformations are considered. In section 3.3 we extend the family of functions considered to include non-invertible linear transformations, which leads to dimensionality reduction. We show that when the data is Gaussian, the solution is given by Fisher's linear discriminant followed by RCA.

3.1. An information theoretic perspective

Following (Linsker, 1989), an information theoretic criterion states that when an input X is transformed into a new representation Y, we should seek to maximize the mutual information $I(X,Y)$ between X and Y under suitable constraints. In the general deterministic case a set $X = \{x_l\}_{l=1}^{n}$ of data points in $\mathcal{R}^N$ is transformed into the set $Y = \{f(x_l)\}_{l=1}^{n}$ of points in $\mathcal{R}^M$. We wish to find a function $f \in F$ that maximizes $I(X,Y)$, where F is the family of allowed transformation functions (the "hypotheses family").

In our case we are also given a set of chunklets of data points from X, $\{x_{ji}\}_{j=1,i=1}^{k,n_j}$, which the repre-

sentation function f is required to keep close to each other. Therefore, we may pose the problem as:

$$\max_{f \in F} I(X,Y) \quad s.t. \quad \frac{1}{p}\sum_{j=1}^{k}\sum_{i=1}^{n_j} ||y_{ji} - m_j^y||^2 \leq K \quad (2)$$

where m_j^y denotes the mean of points in chunklet j after the transformation, P is the total number of points in chunklets, and K is a constant. The mutual information here is the differential mutual information between two continuous variables X and Y, and it depends on their respective densities. One should note that we can only asses these densities using the provided sample of data points.

Since in our case f is deterministic, the maximization of $I(X,Y)$ is achieved by maximizing the entropy $H(Y)$ alone. To see this, recall that

$$I(X,Y) = H(Y) - H(Y|X)$$

Since f is deterministic, there is no uncertainty concerning Y when X is known. Thus $H(Y|X)$ has its lowest possible value at $-\infty$.[1] However, as noted in (Bell & Sejnowski, 1995), $H(Y|X)$ does not depend on f but on the quantization scale. For every finite quantization of the space this term is a constant. Hence maximizing with respect to f can be done by considering only the first term, $H(Y)$.

It should be noted that $H(Y)$ can be increased by simply 'stretching' the data space (e.g. by choosing $f = \lambda x$, where $\lambda > 1$). Therefore, a constraint that keeps certain points close together is required in order to prevent this trivial scaling solution. Also the family F of *representation functions* should be carefully chosen to avoid trivial solutions.

3.2. RCA from an information theoretic perspective

We now look at the problem posed for the family F of invertible linear functions. When f is an invertible function, the connection between the densities of $Y = f(X)$ and X is expressed by $p_y(y) = \frac{p_x(x)}{|J(x)|}$, where $|J(x)|$ is the Jacobian of the transformation. Noting that $p_y(y)dy = p_x(x)dx$, we can relate $H(Y)$ and $H(X)$ as follows:

$$H(Y) = -\int_y p(y)\log p(y)dy =$$

$$-\int_x p(x)\log\frac{p(x)}{|J(x)|}dx = H(X) + \langle \log|J(x)| \rangle_x$$

For a linear function $Y = AX$ the Jacobian is constant and equals $|A|$, and it is the only term in $I(X,Y)$ that depends on the transformation A. Hence problem (2) becomes

$$\max_A |A| \quad s.t. \quad \frac{1}{p}\sum_{j=1}^{k}\sum_{i=1}^{n_j} ||x_{ji} - m_j||^2_{A^tA} \leq K \quad (3)$$

Let $B = A^tA$ denote a Mahalanobis distance matrix, where B is positive definite and $\log|A| = \frac{1}{2}\log|B|$. (3) can now be rewritten as

$$\max_B |B| \quad (4)$$
$$s.t. \quad \frac{1}{p}\sum_{j=1}^{k}\sum_{i=1}^{n_j} ||x_{ji} - m_j||^2_B \leq \mathrm{K}\ ,\ B > 0$$

Writing and solving for the Lagrangian, we get the solution $B = \frac{K}{N}\hat{C}^{-1}$ where $\hat{C}$ is the average chunklet covariance matrix (1) and N is the dimension of the data space. The solution is identical to the Mahalanobis matrix proposed by RCA up to a scale factor.[2] Hence RCA is the solution of (4).

3.3. Dimensionality reduction

In this section we analyze the problem posed in Section 3.1 for the case of general linear transformations, i.e. $Y = AX$ where $A \in \mathcal{M}_{M\times N}$ and $M \leq N$. To simplify the analysis, we assume that X is a multivariate Gaussian. As we saw earlier, maximizing $H(Y)$ is equivalent to maximizing $I(X,Y)$ with respect to f. Since X is assumed to be Gaussian, Y is also Gaussian and its entropy is given by

$$\begin{aligned} H(Y) &= \frac{d}{2}\log 2\pi e + \frac{1}{2}\log|\Sigma_y| \\ &= \frac{d}{2}\log 2\pi e + \frac{1}{2}\log|A\Sigma_x A^t| \end{aligned}$$

so that (2) becomes

$$\max_A \log|A\Sigma_x A^t| \quad (5)$$
$$s.t. \quad \frac{1}{p}\sum_{j=1}^{k}\sum_{i=1}^{n_j} ||x_{ji} - m_j||^2_{A^tA} \leq K$$

For a given target dimension M the solution to the problem is Fisher linear discriminant followed by applying RCA in the reduced dimensional space. A sketch of the proof is given in appendix A.

[1] This non-intuitive divergence is a result of the generalization of information theory to continuous variables; specifically, it is a result of ignoring the discretization constant in the definition of differential entropy.

[2] Such a scale constant is not important in classification tasks, i.e. when using relative distances.

4. RCA also minimizes inner class distances

In order to gain some intuition to the solution provided by the information maximization criterion formalized in Eq. (2), let us look at the optimization problem obtained by reversing the roles of the maximization term and the constraint term:

$$\min_B \frac{1}{p}\sum_{j=1}^{k}\sum_{i=1}^{n_j}||x_{ji}-m_j||_B^2 \quad s.t.\ \ |B|\geq 1 \qquad (6)$$

In (6) a Mahalanobis distance B is sought, which minimizes the sum of all inner chunklet squared distances. Demanding that $|B| \geq 1$ amounts to the demand that minimizing the distances will not be achieved by "shrinking" the entire space. Using Kuhn-Tucker theorem, we can reduce (6) to

$$\min_B \sum_{j=1}^{k}\sum_{i=1}^{n_j}||x_{ji}-m_j||_B^2 - \lambda\log|B| \qquad (7)$$
$$s.t.\ \ \lambda\geq 0,\ \ \lambda\log|B|=0$$

Differentiating the Lagrangian above shows that the minimum is given by $B = |\hat{C}|^{\frac{1}{d}}\hat{C}^{-1}$, where C is the average chunklet covariance matrix. Once again, the solution is identical to the Mahalanobis matrix proposed by RCA up to a scale factor.

It is interesting, in this respect, to compare RCA and the method proposed recently by (Xing et al., 2002). They also consider the problem of learning a Mahalanobis distance using side information in the form of pairwise similarities.[3] They assume knowledge of a set S of pairs of points known to be similar, and a set D of pairs of points known to be dissimilar. Given these sets, they pose the following optimization problem.

$$\min_B \sum_{(x_1,x_2)\in S}||x_1-x_2||_B^2 \qquad (8)$$
$$s.t. \sum_{(x_1,x_2)\in D}||x_1-x_2||_B,\quad B\geq 0$$

This problem is solved using gradient ascent and iterative projection methods.

To allow a clearer comparison of RCA to Eq. (8), we can cast (6) as a minimization of inner chunklet pairwise distances. For each point x_{ji} in chunklet j we have:

$$x_{ji}-m_j = x_{ji}-\frac{1}{n_j}\sum_{k=1}^{n_j}x_{jk} = \frac{1}{n_j}\sum_{\substack{k=1\\k\neq i}}^{n_j}(x_{ji}-x_{jk})$$

[3] Chunklets of size > 2 are not considered.

Problem (6) can now be rewritten as

$$\min_B \sum_{j=1}^{k}\frac{1}{n_j^2}\sum_{i=1}^{n_j}||\sum_{k\neq i}(x_{ji}-x_{jk})||_B^2 \quad s.t.\ \ |B|\geq 1 \qquad (9)$$

When only chunklets of size 2 are given (as in the case studied by Xing et al.), the problem reduces to

$$\min_B \frac{1}{2}\sum_{j=1}^{k}||x_{j1}-x_{j2}||_B^2 \quad s.t.\ \ |B|\geq 1 \qquad (10)$$

Clearly the minimization terms in problems (10) and (8) are identical up to a constant ($\frac{1}{2}$). The difference between the two problems lies in the constraint term they use. The constraint proposed by Xing et al. tries to use information concerning pairs of dissimilar points, whereas the constraint in the RCA formulation can be interpreted as a pure scale constraint, which does not allow the 'volume' of the Mahalanobis neighborhood to shrink.

Although the constraint used by Xing et al. appears to take into consideration further information, closer look shows that it is somewhat arbitrary. The usage of squared distance in the minimization term and the root of square distance for the constraint term is arbitrary and a-symmetric. Most importantly, it should be noted that in most unsupervised applications dissimilar pairs are not explicitly available. In this case (Xing et al., 2002) recommends to take D to be all the pairs of points that are not in S. This is a problematic choice for two reasons: In most practical scenarios pairs of points which are not in S are not necessarily dissimilar. In addition, this definition usually yields a very large set D, which substantially slows the algorithm's running time. In contrast, the RCA distance computation is simple and fast (requiring a single matrix inversion) without any need for an iterative procedure.

In order to further justify the constraint suggested in problem (6), we proceed to suggest a probabilistic interpretation of the RCA algorithm.

5. RCA and Maximum Likelihood

We now analyze the case of data which consists of several normally distributed classes which share the same covariance matrix. Under the assumption that the chunklets are sampled i.i.d and that points within each chunklet are also sampled i.i.d, the likelihood of the chunklets' distribution can be written as:

$$\prod_{j=1}^{k}\prod_{i=1}^{n_j}\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\tfrac{1}{2}(x_{ji}-m_j)^t\Sigma^{-1}(x_{ji}-m_j)\right) \qquad (11)$$

It is easy to see that the RCA Mahalanobis matrix $\hat{C}$ from (1) maximizes (11) over all possible choices of Σ^{-1}, and is therefore the Maximum Likelihood estimator in this setting.

In order to gain further insight into the constraint chosen in (6), we take the log of the likelihood equation (11), drop constant terms and denote $B = \Sigma^{-1}$, to obtain:

$$\hat{C} = \arg\min_{B} \sum_{j=1}^{k} \sum_{i=1}^{n_j} ||x_{ji} - m_j||_B^2 - p \log |B| \quad (12)$$

where p denotes the total number of points in all chunklets. This equation is closely related to the Lagrangian in (7), but here λ (the Lagrange multiplier) is replaced by the constant p. Hence, under Gaussian assumptions, the solution of problem (7) has a probabilistic justification.

The effect of chunklet size

Under Gaussian assumptions, we can define an *unbiased* version of the RCA estimator. Assume for simplicity that there are p constrained data points divided into n chunklets of size k each. The *unbiased* RCA estimator can be written as follows :

$$\hat{C}(n,k) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k-1} \sum_{j=1}^{k} (x_i^j - \hat{m}_i)(x_i^j - \hat{m}_i)^t \quad (13)$$

where x_i^j denotes the data point j in the chunklet i, and $\hat{m}_i$ denotes the empirical mean of chunklet i. $\hat{C}(n,k)$ in (13) is the empirical mean of the covariance estimators produced by each chunklet. It can be shown that the variance of the estimator matrix elements $\hat{C}_{ij}$ is bounded by

$$Var(\hat{C}_{ij}(n,k)) \leq \frac{k}{k-1} Var(\hat{C}_{ij}(1,nk)) \quad (14)$$

where $\hat{C}_{ij}(1,nk)$ is the estimator when all the $p = nk$ points are known to belong to the same class, thus forming the best estimate possible when given p points. For proof see (Hertz et al., 2002). The bound shows that the variance of the RCA estimator using small chunklets rapidly converges to the variance of this best estimator.

6. Experimental Results: Application to clustering

As noted in the introduction, the main goal of our method is to use side information in the form of equivalence relations to improve the performance of unsupervised learning techniques. In order to test our proposed RCA algorithm and to compare it with the work presented by Xing et. al, we used six data sets from the UC Irvine repository which were used in (Xing et al., 2002). As in (Xing et al., 2002) we are given a set S of pairwise similarity constraints (or chunklets of size 2).[4] We used the following clustering algorithms:

1. K-means using the default Euclidean metric (i.e. using no side-information).
2. Constrained K-means: K-means subject to points $(x_i, x_j) \in S$ always being assigned to the same cluster (Wagstaff et al., 2001).
3. Constrained K-means + Metric proposed by (Xing et al., 2002): Constrained K-means using the distance metric proposed in (Xing et al., 2002), which is learned from S.
4. Constrained K-means + RCA: Constrained K-means using the RCA distance metric learned from S.
5. EM: Expectation Maximization of a Gaussian Mixture model (using no side-information).
6. Constrained EM: EM using side-information in the form of equivalence constraints (Hertz et al., 2002; Shental et al., 2003), when using the RCA distance metric as an initial metric.

Following (Xing et al., 2002) we will use a normalized accuracy score to evaluate the partitions obtained by the different clustering algorithms presented above. More formally, in the case of 2-cluster data the accuracy measure used can be written as:

$$\sum_{i>j} \frac{1\{1\{c_i = c_j\} = 1\{\hat{c}_i = \hat{c}_j\}\}}{0.5m(m-1)}$$

where $1\{\cdot\}$ is the indicator function ($1\{True\} = 1), 1\{False\} = 0$), $\{\hat{c}_i\}_{i=1}^m$ is the cluster to which point x_i is assigned by the clustering algorithm, and c_i is the "correct" or desired assignment. The score above is equivalent to computing the probability that the algorithm's assignment $\hat{c}$ of two randomly drawn points x_i and x_j agrees with the "true" assignment c.[5]

[4] To allow for a fair comparison with (Xing et al., 2002), we repeated their exact experimental setup and criteria.

[5] As noted in (Xing et al., 2002), this score needs normalization when the number of clusters is larger than 2. The normalization is achieved by sampling the pairs x_i and x_j from the same cluster (as determined by $\hat{c}$) with probability 0.5 and from different clusters with probability 0.5, so that "matches" and "mismatches" are given the same weight.

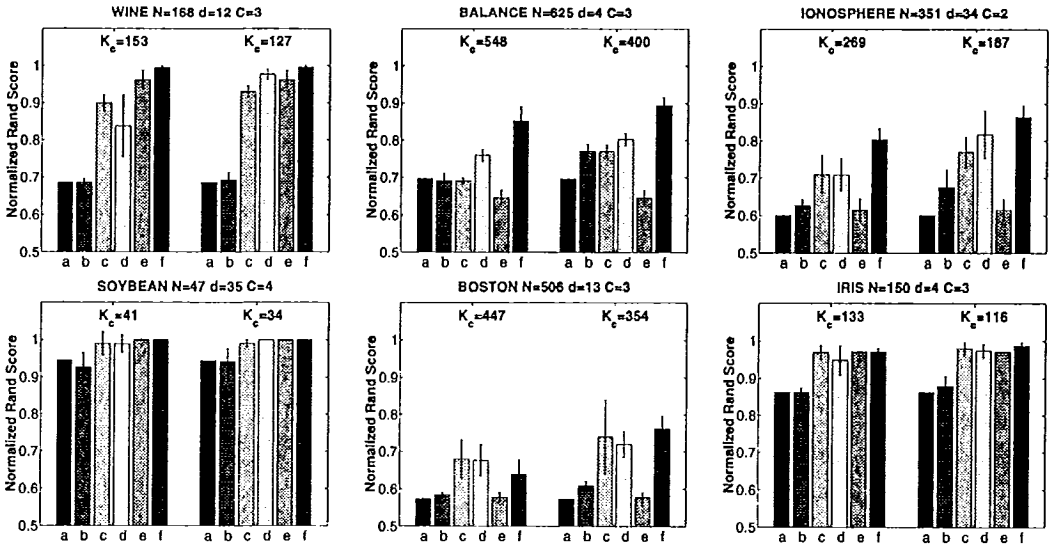

Figure 2. Clustering accuracy on 6 UCI datasets. In each panel, the six bars on the left correspond to an experiment with "little" side-information, and the six bars on the right correspond to "much" side-information. From left to right the six bars are respectively: (a) K-means over the original feature space (without using any side-information). (b) Constrained K-means over the original feature space. (c) Constrained K-means over the feature space suggested by (Xing et al., 2002). (d) Constrained K-means over the feature space created by RCA. (e) EM over the original feature space (without using any side-information). (f) Constrained EM (Shental et al., 2003) over the feature space created by RCA. Also shown are N - the number of points, C - the number of classes, d - the dimension of the feature space, and K_c - the mean number of connected components (see footnote 6). The results were averaged over 20 realizations of side-information.

As in (Xing et al., 2002) we tested our method using two conditions: (1) using "little" side-information S; (2) using "much" side-information.[6] As in (Xing et al., 2002) in all of our experiments we used K-means with multiple restarts.

Fig. 2 shows the results of all algorithms described above when using the two conditions of "little" and "much" side-information.

Clearly using RCA as a distance measure significantly improves the results over the original K-means algorithm. When comparing our results with the results reported in (Xing et al., 2002), we see that RCA achieves similar results. In this respect it should be noted that the RCA metric computation is a single step efficient computation, whereas the method presented in (Xing et al., 2002) requires gradient descent and iterative projections.

[6] S was generated by choosing a random subset of all pairs of points sharing the same class c_i. In the case of little side-information, the size of the subset was chosen so that the resulting number of connected components K_c (using transitive closure over pairs) is roughly 90% of the size of the original dataset. In case of much side information this was changed to 70%.

7. Discussion and Concluding remarks

We have presented an algorithm which makes use of side-information in the form of equivalence relations to learn a Mahalanobis metric. We have shown that our method is optimal under several criteria, and also showed considerable improvement in clustering on several standard datasets.

RCA is one of several techniques which we have developed for using equivalence relations to enhance unsupervised learning. In a related technique, we introduced the constraints into an EM formulation of a Gaussian Mixture Model (Hertz et al., 2002; Shental et al., 2003). This work enhances the power of RCA in two ways: First, it makes it possible to incorporate negative constraints. Second, it allows further improvement of the RCA metric, as may be seen in Fig. 2.

References

Bell, A., & Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*, 1129–1159.

Boreczky, J. S., & Rowe, L. A. (1996). Comparison of video shot boundary detection techniques. *SPIE Storage and Retrieval for Still Images and Video Databases IV, 2664*, 170–179.

Chechik, G., & Tishby, N. (2002). Extracting relevant structures with side information. *NIPS, 15.*

Fukunaga, K. (1990). *Statistical pattern recognition.* San Diego: Academic Press. 2nd edition.

Hastie, T., & Tibshirani, R. (1996). Discriminant adaptive nearest neighbor classification and regression. *Advances in Neural Information Processing Systems* (pp. 409–415). The MIT Press.

Hertz, T., Shental, N., Bar-hillel, A., & Weinshall, D. (2002). Enhancing image and video retrieval: Learning via equivalence constraints. *http://www.cs.huji.ac.il/ daphna/.*

Jaakkola, T., & Haussler, D. (1998). Exploiting generative models in discriminative classifiers.

Linsker, R. (1989). An application of the principle of maximum information preservation to linear systems. *NIPS* (pp. 186–194). Morgan Kaufmann.

Shental, N., Hertz, T., Bar-Hilel, A., & Weinshall, D. (2003). Computing gaussian mixture models with EM using equivalence constraints.

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

Tenenbaum, J., & Freeman, W. (2000). Separating style and content with bilinear models. *Neural Computation, 12*, 1247–1283.

Thrun, S. (1996). Is learning the n-th thing any easier than learning the first? *Advances in Neural Information Processing Systems* (pp. 640–646). The MIT Press.

Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing* (pp. 368–377).

Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained K-means clustering with background knowledge. *Proc. 18th International Conf. on Machine Learning* (pp. 577–584). Morgan Kaufmann, San Francisco, CA.

Xing, E., Ng, A., Jordan, M., & Russell, S. (2002). Distance metric learnign with application to clustering with side-information. *Advances in Neural Information Processing Systems.* The MIT Press.

Appendix A: Information Maximization in the case of non invertible linear transformation

Here we briefly sketch the proof of the claim made in Section 3.3. As before, we denote by C the average covariance matrix of the chunklets. We can rewrite the constrained expression as:

$$\frac{1}{p}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji}-m_j)^t A^t A(x_{ji}-m_j) = tr(A^t AC) = tr(A^t CA)$$

Hence the Lagrangian may be written as:

$$\log|A\Sigma_x A^t| - \lambda(tr(ACA^t) - K)$$

Differentiating the Lagrangian w.r.t A leads to

$$\Sigma_x A(A^t\Sigma_x A)^{-1} = \lambda CA \qquad (15)$$

Multiplying by A^t and rearranging we get: $\frac{I}{\lambda} = A^t CA$. This equation does not give us information concerning the subspace to which the optimal A takes us. However, A whitens the data with respect to the chunklet covariance C in this subspace, similarly to RCA. From $\lambda \neq 0$ it then follows that the inequality constraint is an equality, which can be used to find λ.

$$tr(ACA^t) = tr(\frac{I}{\lambda}) = \frac{M}{\lambda} = K \Longrightarrow \lambda = \frac{M}{K}$$
$$\Longrightarrow ACA^t = \frac{K}{M}I$$

Now, since in our solution space $ACA^t = \frac{K}{M}I$, $\log|ACA^t| = M\log\frac{K}{M}$ holds for all points. Hence we can modify the maximization argument as follows

$$\log|A\Sigma_x A^t| = \log\frac{|A\Sigma_x A^t|}{|ACA^t|} + M\log\frac{K}{M}$$

Now the optimization argument has a familiar form. It is known (Fukunaga, 1990) that maximizing the determinant ratio can be done by projecting the space on the span of the first M eigenvectors of $C^{-1}\Sigma_x$. Denote by B the solution matrix for this unconstrained problem. In order to enforce the constraints we define the matrix $A = \sqrt{\frac{K}{M}}\Lambda_1^{-0.5}B$ and we claim that A is the solution of the constrained problem. Notice that the value of the maximization argument does not change when we switch from A to B since A is a product of B and another full ranked matrix. It can also be shown that A satisfies the constraints and is thus the solution of the problem presented in Eq. (5).