Parametric Distance Metric Learning with Label Information*

Zhihua Zhang, James T. Kwok and Dit-Yan Yeung
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
{zhzhang, jamesk, dyyeung}@cs.ust.hk

Abstract

Distance-based methods in pattern recognition and machine learning have to rely on a similarity or dissimilarity measure between patterns in the input space. For many applications, Euclidean distance in the input space is not a good choice and hence more complicated distance metrics have to be used. In this paper, we propose a parametric method for metric learning based on class label information. We first define a dissimilarity measure that can be proved to be metric. It has the favorable property that between-class dissimilarity is always larger than within-class dissimilarity. We then perform parametric learning to find a regression mapping from the input space to a feature space, such that the dissimilarity between patterns in the input space is approximated by the Euclidean distance between points in the feature space. Parametric learning is performed using the iterative majorization algorithm. Experimental results on real-world benchmark data sets show that this approach is promising.

1 Introduction

The notion of similarity or dissimilarity plays a fundamental role in pattern recognition and machine learning. A promising direction to pursue is to learn good (dis)similarity measures from data. Recently, learning distance metrics from data has aroused a great deal of interest from machine learning researchers. One typically wants to embed patterns in a (possibly non-metric) input space into a feature space, in which the Euclidean distance between points accurately reflects the dissimilarity between the corresponding patterns. Therefore the (linear or nonlinear) mapping from the input space to the feature space corresponds to feature extraction. Alternatively, the feature space may be a low-dimensional space for data visualization.

In this paper, we propose a parametric distance metric learning method in the supervised setting. The main ideas of our method are summarized as follows. Using class label information, we define a similarity measure (and hence also the corresponding dissimilarity measure) between patterns in the input space. The dissimilarity measure implicitly induces a metric space for embedding the original patterns. To explicitly represent the mapping from the input space to the feature space, we then approximate the mapping by a regression model to embed the original patterns in an Euclidean space. The regression parameters are estimated from data with the objective that the dissimilarity between patterns in the input space is approximated by the Euclidean distance between points in the feature space. Once the regression model has been found, any new pattern can be mapped to its corresponding location in the feature space. Distance-based methods, such as k-means clustering, nearest neighbor classifiers and support vector machines, can then be applied in the feature space for clustering or classification applications.

The rest of this paper is organized as follows. A modified metric incorporating class label information is proposed in Section 2. Section 3 outlines our regression model for metric learning and the corresponding optimization method. Experimental results are presented in Section 4, and the last section gives some concluding remarks.

2 Modified Metric with Label Information

Denote the input space by $\mathbb{R}^q$ and the set of all C possible class (target) labels by $\mathcal{T}$. A training set $\mathcal{D} \subseteq \mathbb{R}^q \times \mathcal{T}$ has n patterns $\{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_n, t_n)\}$, where $t_i = r$ if pattern i belongs to class r. Here, each pattern is assumed to belong to only one class. In general, a number of similarity measures can be defined on these patterns [Gower and Legendre, 1986]. In this paper, we utilize also the label information in defining the similarity s_{ij} between patterns $\mathbf{x}_i$ and $\mathbf{x}_j$:

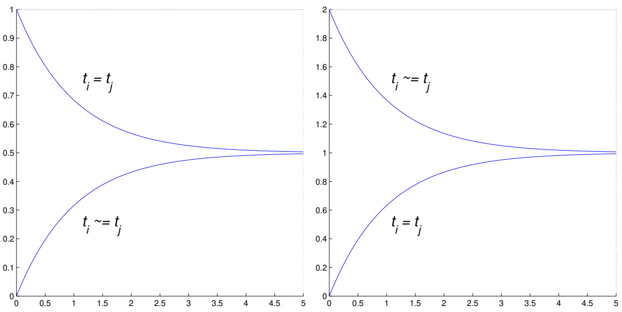
$$s_{ij} = \begin{cases} \frac{1}{2} + \frac{1}{2}\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\beta}\right) & t_i = t_j \\ \frac{1}{2} - \frac{1}{2}\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\beta}\right) & t_i \neq t_j \end{cases}, \quad (1)$$

where $\|\cdot\|$ denotes the Euclidean norm and $\beta > 0$ is a width parameter. The corresponding dissimilarity δ_{ij} is then:

$$\begin{aligned} \delta_{ij} &= s_{ii} + s_{jj} - 2s_{ij} \\ &= \begin{cases} 1 - \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\beta}\right) & t_i = t_j \\ 1 + \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\beta}\right) & t_i \neq t_j \end{cases}. \end{aligned} \quad (2)$$

*This research has been partially supported by the Research Grants Council of the Hong Kong Special Administrative Region under grants DAG01/02.EG28, HKUST2033/00E and HKUST6195/02E.

As illustrated in Figure 1, this (dis)similarity measure enjoys some nice properties for pattern discrimination. For example, the (dis)similarity between any two patterns in the same class is always larger (smaller) than that between any two patterns belonging to different classes. Moreover, the larger the Euclidean distance between the patterns is, the smaller is the within-class similarity while the larger is the between-class similarity.



(a) s_{ij} versus $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ (b) δ_{ij} versus $\|\mathbf{x}_i - \mathbf{x}_j\|^2$

Figure 1: Similarity and dissimilarity in (1) and (2).

In recent years, finite metric spaces and their embeddings have received much attention [Indyk, 2001; Linial *et al.*, 1995]. Among embedding into normed spaces, embedding into an Euclidean space is the most popular.

Given the dissimilarity matrix $\mathbf{\Delta} = [\delta_{ij}]_{n\times n}$, we are interested in the question of whether and how the dissimilarity matrix $\mathbf{\Delta}$ can be embedded. In other words, for the original points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, we attempt to find a configuration of points $\{\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n\}$ in some Euclidean space $\mathbb{R}^l$ such that the squared distances between these points will be equal to the so-defined dissimilarities δ_{ij}.

The following theorem confirms that $\mathbf{\Delta}$ can be embedded.

Theorem 1 *Define $d_{ij} = \sqrt{\delta_{ij}}$, with δ_{ij} as in (2). The matrix $\mathbf{D} = [d_{ij}]$ is metric. In other words, d_{ij} satisfies the following properties:*

1. $d_{ij} \geq 0, \quad \forall i, j,$
2. $d_{ii} = 0, \quad \forall i,$
3. $d_{ij} = d_{ji}, \quad \forall i, j,$
4. $d_{ik} + d_{jk} \geq d_{ij}, \quad \forall i, j, k.$

The proof of this theorem can be found in [Zhang *et al.*, 2003]. The subsequent task is then to find the embedding, i.e., points $\{\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n\} \in \mathbb{R}^l$ such that the inter-point distance is equal to d_{ij}. In general, obtaining an exact solution for these $\widehat{\mathbf{x}}_i$'s is difficult. Nevertheless, because $\mathbf{D}$ is metric, an approximate solution can be easily obtained by using principal coordinate analysis or other multidimensional scaling (MDS) methods [Cox and Cox, 2000]. We will return to this problem in Section 3.

Notice that the resultant Euclidean embedding will still incorporate information from both the input space representation ($\mathbf{x}_i$) of the patterns and their corresponding class labels (t_i). Moreover, the distance metric d_{ij}, like the associated δ_{ij}, enjoys those nice properties useful for pattern discrimination.

3 Metric Learning with Regression Model

As mentioned in Section 2, an approximate solution for $\{\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n\}$ can be obtained by using MDS. However, this may be intractable for large data sets. Moreover, for new patterns with unknown labels, the problem then is on how to determine s_{ij} in the first place. Following [Koontz and Fukunaga, 1972; Cox and Ferry, 1993; Webb, 1995], we attempt to find a mapping from $\mathbf{x}_i$ in the original input space $\mathbb{R}^q$ to $\widehat{\mathbf{x}}_i$ in the embedded Euclidean space $\mathbb{R}^l$. One possibility is to first obtain a MDS configuration, and then construct a regression model from $\mathbf{x}_i$ to $\widehat{\mathbf{x}}_i$ [Cox and Ferry, 1993]. However, this mapping is not determined as part of the MDS procedure [Webb, 1995]. In the following, we will follow the approach of [Webb, 1995].

Denote the mapping from the original input space $\mathbb{R}^q$ to the embedded Euclidean space $\mathbb{R}^l$ by $\mathbf{f} = (f_1, \ldots, f_l)'$. Assume that each f_i is a linear combination of p basis functions:

$$f_i(\mathbf{x}; \mathbf{W}) = \sum_{j=1}^{p} w_{ji}\phi_j(\mathbf{x}), \tag{3}$$

where $\mathbf{W} = [w_{ji}]_{p\times l}$ contains the free parameters, and the $\phi_j(\mathbf{x})$'s are basis functions that can be linear or nonlinear. The regression mapping (3) can be written in matrix form as

$$\mathbf{y}(\mathbf{W}) = \mathbf{f}(\mathbf{x}; \mathbf{W}) = \mathbf{W}'\boldsymbol{\phi}(\mathbf{x}),$$

where $\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_p(\mathbf{x})]'$. Let $\mathbf{X}$ be the target configuration, with $d_{ij}(\mathbf{X}) = \sqrt{\delta_{ij}}$ where δ_{ij} is defined in (2). Using the iterative majorization algorithm, we then minimize the squared error

$$e^2(\mathbf{W}) = \sum_{i=1}^{n}\sum_{j=1}^{n}(d_{ij}(\mathbf{X}) - q_{ij}(\mathbf{W}))^2 \tag{4}$$

w.r.t. $\mathbf{W}$, where $q_{ij}(\mathbf{W}) = \|\mathbf{W}'(\boldsymbol{\phi}(\mathbf{x}_i) - \boldsymbol{\phi}(\mathbf{x}_j))\|$. More details can be found in [Zhang *et al.*, 2003].

4 Experiments

In this Section, we perform experiments on six benchmark data sets (Table 1) from the UCI repository [Murphy and Aha, 1994]. The distance metric is learned using a small subset of the labeled patterns, with $l = p = q$, $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$ and the width β in (1) set to the average distance of the labeled patterns to the class means. The remaining patterns are then used for testing.

Table 2 shows the classification results by the nearest mean and nearest neighbor classifiers, with both the Euclidean and learned metrics. As can be seen, the learned metric almost always outperforms the original metric.

Next, we perform clustering experiments using the k-means clustering algorithm, with the value of k set to the true number of clusters in each data set. The clustered patterns are assigned labels and the clustering accuracy is measured by comparing these labels with the true labels (as in classification problems). As these cluster labels can be permuted without changing the clustering solution, results reported here are based on the labeling with the highest clustering accuracy. As can be seen from Table 3, the learned metric outperforms that with the original metric on all data sets.

Table 1: The six UCI data sets used in the experiments.

data set	total # of patterns		# patterns for metric learning
	class	size	
Pima Indians diabetes (diabetes)	1	500	80
	2	268	50
soybean	1	10	2
	2	10	2
	3	10	2
	4	17	4
wine	1	59	20
	2	71	20
	3	48	20
Wisconsin breast cancer (WBC)	1	212	50
	2	357	50
ionosphere	1	126	50
	2	225	50
iris	1	50	10
	2	50	10
	3	50	10

Table 2: Classification accuracies on the UCI data sets (Numbers in bold indicate the better results).

data set	nearest mean		nearest neighbor	
	Euclidean metric	learned metric	Euclidean metric	learned metric
diabetes	463/638	**475**/638	**432**/638	425/638
soybean	36/37	**37**/37	35/37	**37**/37
wine	86/118	**115**/118	77/118	**117**/118
WBC	430/469	**451**/469	420/469	**453**/469
ionosphere	159/251	**201**/251	212/251	**225**/251
iris	108/120	**110**/120	**114**/120	**114**/120

Table 3: Clustering accuracies on the UCI data sets (Numbers in bold indicate the better results).

data set	Euclidean metric	learned metric
diabetes	459/638	**480**/638
soybean	**37**/37	**37**/37
wine	85/118	**117**/118
WBC	412/469	**446**/469
ionosphere	168/251	**221**/251
iris	107/120	**110**/120

5 Concluding Remarks

In this paper, we proposed a new parametric method for distance metric learning based on class label information. Experiments on UCI data sets show promising results.

The current work can be extended in several directions. First, nonlinear basis functions can be used to improve the approximation power of the regression mapping. Second, although Theorem 1 states that the dissimilarity measure induces a metric, it is not clear whether the matrix is also Euclidean. If this is the case, a new kernel can then be defined on the joint space of the input space and class label space [Schölkopf, 2002]. Third, in addition to using label information, we will also incorporate manifold structure between neighboring patterns into our metric learning process.

References

[Cox and Cox, 2000] T.F. Cox and M.A.A. Cox. *Multidimensional Scaling*. Chapman & Hall/CRC, second edition, 2000.

[Cox and Ferry, 1993] T.F. Cox and G. Ferry. Discriminant analysis using non-metric multidimensional scaling. *Pattern Recognition*, 26(1):145–153, 1993.

[Gower and Legendre, 1986] J.C. Gower and P. Legendre. Metric and Euclidean properties of dissimilarities coefficients. *Journal of Classification*, 3:5–48, 1986.

[Indyk, 2001] P. Indyk. Algorithmic applications of low-distortion geometric embeddings. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, pages 10–33, 2001.

[Koontz and Fukunaga, 1972] W.L.G. Koontz and K. Fukunaga. A nonlinear feature extraction algorithm using distance information. *IEEE Transactions on Computers*, 21(1):56–63, 1972.

[Linial *et al.*, 1995] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

[Murphy and Aha, 1994] P.M. Murphy and D.W. Aha. UCI repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1994.

[Schölkopf, 2002] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.

[Webb, 1995] A.R. Webb. Multidimensional scaling by iterative majorization using radial basis functions. *Pattern Recognition*, 28(5):753–759, 1995.

[Zhang *et al.*, 2003] Z. Zhang, J.T. Kwok and D.Y. Yeung. Parametric distance metric with label information. ftp://ftp.cs.ust.hk/pub/techreport/03/tr03-02.ps.gz, 2003.