Information Sciences 282 (2014) 321–331

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Multi-granularity distance metric learning via neighborhood granule margin maximization

Pengfei Zhu [a,b], Qinghua Hu [a,*], Wangmeng Zuo [c], Meng Yang [d]

[a] School of Computer Science and Technology, Tianjin Univerity, Tianjin 300073, China
[b] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[c] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[d] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518000, China

ARTICLE INFO

Article history:
Received 9 July 2013
Received in revised form 6 April 2014
Accepted 8 June 2014
Available online 18 June 2014

Keywords:
Neighborhood granular margin
Metric learning
Neighborhood rough set
Multiple granularity

ABSTRACT

Learning a distance metric from training samples is often a crucial step in machine learning and pattern recognition. Locality, compactness and consistency are considered as the key principles in distance metric learning. However, the existing metric learning methods just consider one or two of them. In this paper, we develop a multi-granularity distance learning technique. First, a new index, neighborhood granule margin, which simultaneously considers locality, compactness and consistency of neighborhood, is introduced to evaluate a distance metric. By maximizing neighborhood granule margin, we formulate the distance metric learning problem as a sample pair classification problem, which can be solved by standard support vector machine solvers. Then a set of distance metrics are learned in different granular spaces. The weights of the granular spaces are learned through optimizing the margin distribution. Finally, the decisions from different granular spaces are combined with weighted voting. Experiments on UCI datasets, gender classification and object categorization tasks show that the proposed method is superior to the state-of-the-art distance metric learning algorithms.

1. Introduction

How to construct or learn a proper distance or similarity measure is a key problem in clustering and classification such as *k*-means, and *k* nearest neighbor searching [10,51,54]. Whereas, the optimal distance metric may be problem-specific and up to the underlying data structure and distributions. To this end, there have been increasing efforts made to learn a distance metric in recent years [6,10,37,39,41,50]. Metric learning methods can be categorized into unsupervised [50], semi-supervised [5] and supervised ones [1,11,25,27,34,53], according to the availability of the labels of training samples. Metric learning has been proved to successfully improve the clustering and recognition performance in information retrieval [29,30], bioinformatics [47] and computer vision tasks [6,10,12,16,39,37].

Generally speaking, metric learning aims to learn an effective distance metric, measured by which the samples from the positive sample pair (i.e., samples with the same class label or similar samples) could be as close as possible, while the samples from the negative sample pair (i.e., samples with the different class labels or dissimilar samples) could be as far as possible. In most cases, a metric learning model has three key components: sample pairs; objective function; regularization. In supervised learning, sample pairs can be generated from k nearest neighbors, e.g., large margin nearest neighbor (LMNN) [6]

* Corresponding author. Tel.: +86 22 27401839.
E-mail address: huqinghua@tju.edu.cn (Q. Hu).

http://dx.doi.org/10.1016/j.ins.2014.06.017
0020-0255/

and neighborhood component analysis (NCA) [15]. In verification tasks, samples pairs could be randomly generated by putting two similarly labeled samples into positive pairs and two differently labeled samples into negative pairs [16,34,37,39,53]. In weakly supervised learning, side information is provided and similar/dissimilar sample pairs are given [50,18]. The objective function is often established by minimizing the distance between two samples in positive pairs and maximizing the distance between two samples in negative pairs [6,16,37]. Besides, Bar et al. proposed to maximize mutual information between the original data and embedded data [3]. To get a stable solution and the expected property for the learned metric, regularizations such as trace of matrix [53], log-determinant regularization [42], sparse regularization [23,32] and nuclear norm [36] are imposed on the learned parameters in different applications [38].

In distance based classification, the performance of local classifiers, e.g., nearest neighbor classifier and neighborhood classifier, is greatly affected by the local distribution of the training samples. Many metric learning methods aim to learn a distance metric to get expected local data structure. The local data structure, i.e., neighborhood, can be evaluated from locality, compactness and consistency. Locality means the neighborhood relationship in the original space, which should be kept in the learning process. Locality preserving is widely applied in dimension reduction [17,40,43], spectral analysis [7,55] and sparse coding [48]. Compactness measures the closeness of samples in the neighborhood, which is the main principle in many clustering algorithms [24]. Consistency is used to measure ratio of the samples that can be recognized with the Bayes rule [9]. A good metric should be capable to preserve locality, lead to compact local data structure and high consistency.

In [21], a neighborhood rough set model is proposed based on neighborhood granulation. The samples in the neighborhood of each sample form a neighborhood granule. Then, a family of neighborhood granules forms an elemental granule system that covers the universe. By computing the consistency of neighborhood granules, the universe is divided into decision positive regions and decision boundary regions. The percentage of samples in the decision positive regions is defined as neighborhood dependency [20,21]. Neighborhood dependency only counts the pure neighborhood granules and does not reflect the real consistency. Then the decision boundary regions are further grouped into recognizable and misclassified subsets based on the class probabilities in the neighborhood. The percentage of misclassified samples is defined as neighborhood decision error rate [19]. Neighborhood decision error reflects the consistency of neighborhood structures. However, it does not consider the locality and compactness. In this work we design a new evaluation index which simultaneously considers the locality, compactness and consistency.

When we learn the distance metric with neighborhood information, a problem appears, i.e., how to set the size of neighborhoods. It is suggested that multi-granularity data analysis may lead to performance improvement. Multiple granularity leads to diverse viewpoints of the world. In different granular spaces, we may view an object differently or get different decisions. This observation has been widely used in feature extraction, feature learning and classifier design. For example, in feature extraction, Gabor feature extracts features in different scales, that is, different down-sampling rate [33]. Additionally, spatial pyramid model in matching uses pooling technique to combine the feature extracted in different patch sizes [52,28]. In feature learning and representation, deep learning, is actually a multi-granularity method. Deep learning learns low-level, middle-level and high-level features, and each level can be interpreted as a granularity [4]. For classifier design, a multi-scale face recognition method is proposed by combining the decision of different scales [57]. In [56], an adaptive neighborhood granularity selection and combination method is proposed to solve the granularity-sensitive problem in neighborhood granular models. Hence, we can learn multiple distance metrics under different granularity and then combine the decisions made from the learned metrics.

In this paper, we propose a multi-granularity neighborhood distance metric learning (MGML) method. Firstly, we propose neighborhood granule margin to evaluate a distance metric. Neighborhood granule margin is defined by maximum log-likelihood of Bayes error. Then we formulate the metric learning problem as a support vector machines (SVM) model, which can be effectively solved by standard SVM solvers. Hence, it is quite efficient and has good scalability. As the optimal neighborhood size may be task-specific, we propose a multi-granularity method to combine the decisions of different granularity. For each neighborhood size, we can learn a distance metric, and then a decision is got. By margin distribution optimization, the granularity weights are learned. Finally, the decisions of different granularity are combined using the learned weights. Experiments on UCI datasets, gender classification, object categorization show that the proposed metric learning method is competent with the state-of-the-art metric learning methods.

The rest of this paper is organized as follows: Section 2 introduces neighborhood granule margin; Section 3 gives the metric learning model by maximizing neighborhood granule margin; Section 4 proposes the multi-granularity distance metric learning method; experimental analysis is described in Section 5, and Conclusions are given in Section 6.

2. Neighborhood granule margin

Given an information system$\langle \boldsymbol{U}, \boldsymbol{A}, \boldsymbol{D} \rangle$, $\boldsymbol{U} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is a non-empty set of objects, $\boldsymbol{A} = \{a_1, \ldots, a_m\}$ is a set of attributes which describe samples, and $\boldsymbol{D}$ is the decision variable.

Definition 1. [21] Given $\boldsymbol{x}_i \in \boldsymbol{U}$, the neighborhood $\delta(\boldsymbol{x}_i)$ of $\boldsymbol{x}_i$ is defined as

$$\delta(\boldsymbol{x}_i) = \{\boldsymbol{x}_j | \boldsymbol{x}_j \in \boldsymbol{U}, \Delta(\boldsymbol{x}_i, \boldsymbol{x}_j) \leqslant \delta\}, \tag{1}$$

where Δ is a distance function defined in feature spaces and δ is the neighborhood size.

Definition 2. Given a metric space $\langle \boldsymbol{U}, \Delta \rangle$, the family of neighborhood granules $\{\delta(\boldsymbol{x}_i)|\boldsymbol{x}_i \in \boldsymbol{U}\}$ forms an elemental granule system that covers the universe. A discriminative neighborhood relation $\boldsymbol{R}$ on the universe can be written as a relation matrix $(r_{ij})_{n\times n}$,

$$r_{ij} = \begin{cases} 1, & \Delta(\boldsymbol{x}_i, \boldsymbol{x}_j) \leqslant \delta,\ y_i = y_j \\ -1, & \Delta(\boldsymbol{x}_i, \boldsymbol{x}_j) \leqslant \delta,\ y_i \neq y_j \\ 0, & \Delta(\boldsymbol{x}_i, \boldsymbol{x}_j) > \delta \end{cases} \tag{2}$$

where y_i and y_j are the labels of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Similar to neighborhood relation [21] and neighborhood graph [40], discriminative neighborhood relation is also a kind of similarity relation, which has the property of reflexivity and symmetry. It takes both the neighborhood relationship and discrimination information into account. Then based on the discriminative neighborhood relation $\boldsymbol{R}$, neighborhood dependency [21] and neighborhood decision error [19] can be reformulated. For neighborhood dependency, we have

$$\gamma = \frac{1}{n}\sum_i \mathrm{f}\left(\sum_{j\in\{j|r_{ij}\neq 0\}} r_{ij}/t_i\right) \tag{3}$$

where $\mathrm{f}(\cdot)$ is an indicator function and t_i is number of samples in the neighborhood of $\boldsymbol{x}_i$. Actually, $\mathrm{f}\left(\sum_{j\in\{j|r_{ij}\neq 0\}} r_{ij}/t_i\right)$ is used to judge whether $\boldsymbol{x}_i$ belongs to the decision positive regions [21]. If $\sum_{j\in\{j|r_{ij}\neq 0\}} r_{ij}/t_i = 1$, then all the samples in the neighborhood of $\boldsymbol{x}_i$ belong to the same class. In this case, $\boldsymbol{x}_i$ belongs to the decision positive regions and $\mathrm{f}\left(\sum_{j\in\{j|r_{ij}\neq 0\}} r_{ij}/t_i\right) = 1$. Otherwise, if $\sum_{j\in\{j|r_{ij}\neq 0\}} r_{ij}/t_i \neq 1$, the samples in the neighborhood of $\boldsymbol{x}_i$ belong to different classes. In this case, $\boldsymbol{x}_i$ does not belong to the decision positive regions and $\mathrm{f}\left(\sum_{j\in\{j|r_{ij}\neq 0\}} r_{ij}/t_i\right) = 0$. As shown in Fig. 1, there are four samples. $\boldsymbol{x}_1$ and $\boldsymbol{x}_4$ belong to the decision positive regions while $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ do not belong to the decision positive regions. Hence, for $i = 1$ and 4, $\mathrm{f}\left(\sum_{j\in\{j|r_{ij}\neq 0\}} r_{ij}/t_i\right) = 1$ while for $i = 2$ and 3, $\mathrm{f}\left(\sum_{j\in\{j|r_{ij}\neq 0\}} r_{ij}/t_i\right) = 0$.

Neighborhood dependency reflects the percentage of pure neighborhood granules while it ignores the fact that the decision boundary samples can be further grouped into recognizable and misclassified subsets based on the probability in the neighborhood [19]. For instance, in Fig. 1, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are both decision boundary samples. Whereas, $\boldsymbol{x}_2$ is recognizable while $\boldsymbol{x}_3$ is misclassified. Then neighborhood decision error is defined to measure the percentage of misclassified samples [19]:

$$NDER = \frac{1}{n}\sum_i \mathrm{g}\left(\sum_j r_{ij}\right) \tag{4}$$

where $\mathrm{g}(\cdot)$ is an indicator function. Here, if $\sum_j r_{ij} \leqslant 0$, then $\mathrm{g}\left(\sum_j r_{ij}\right) = 1$, which means $\boldsymbol{x}_i$ would be misclassified according to the probability in the neighborhood [19]. If $\sum_j r_{ij} > 0,\ \mathrm{g}\left(\sum_j r_{ij}\right) = 0$.

Based on neighborhood dependency and neighborhood decision error, feature selection algorithms have been proposed to find a feature subset that keeps the neighborhood dependency or minimizes the neighborhood decision error [19,21]. Different from feature selection, we can learn the a desired metric Δ to get an expected neighborhood for each sample. Measured by a given distance measure Δ, the differently labeled samples in the neighborhood of $\boldsymbol{x}_i$ should be far away from $\boldsymbol{x}_i$ while the similarly labeled samples should be close to $\boldsymbol{x}_i$.

In Fig. 1, there are four neighborhood granules. According to the local probability in the neighborhood, $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{x}_4$ are correctly classified while $\boldsymbol{x}_3$ is misclassified. Here the probability of $\boldsymbol{x}_i$ being correctly classified is defined as:

$$p(y_i|\boldsymbol{x}_i) = \frac{\prod_{r_{ij}=1}\exp(-d_{ij})}{\prod_{r_{ij}=1}\exp(-d_{ij}) + \prod_{r_{ij}=-1}\exp(-d_{ij})} \tag{5}$$

where d_{ij} is the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and y_i is the label of $\boldsymbol{x}_i$. Then the probability of $\boldsymbol{x}_i$ being misclassified is

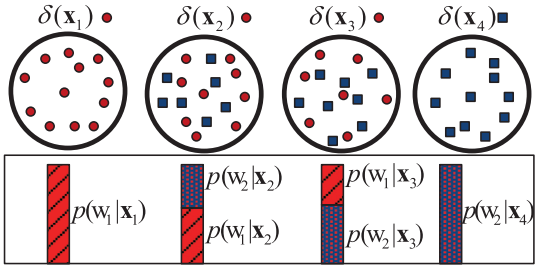

Fig. 1. Different neighborhood granules.

$$1 - p(y_i|\mathbf{x}_i) = \frac{\prod_{r_{ij}=-1} \exp(-d_{ij})}{\prod_{r_{ij}=1} \exp(-d_{ij}) + \prod_{r_{ij}=-1} \exp(-d_{ij})}. \tag{6}$$

A good distance metric Δ should get low Bayes error rate. Here, maximum log-likelihood is explored to optimize the parameter, i.e., Δ. The log-likelihood Ł can be written as:

$$\begin{aligned} Ł &= \sum_{i=1}^{n} \log(p(y_i|\mathbf{x}_i)) - log(1 - p(y_i|\boldsymbol{x}_i)) = \sum_{i=1}^{n} \log \frac{p(y_i|\mathbf{x}_i)}{1 - p(y_i|\boldsymbol{x}_i)} = \sum_{i=1}^{n} \log \frac{d_{r_{ij}=1} \exp(-d_{ij})}{\prod_{r_{ij}=-1} \exp(-d_{ij})} = \sum_{i=1}^{n} \left(\sum_{r_{ij}=-1} d_{ij} - \sum_{r_{ij}=1} d_{ij} \right) \\ &= \sum_{i=1}^{n} \left(-\sum_{j=1}^{n} d_{ij} r_{ij} \right) \end{aligned} \tag{7}$$

In Eq. (7), for sample $\boldsymbol{x}_i$, there are two parts: $dr_1 = \sum_{r_{ij}=-1} d_{ij}$ and $dr_2 = \sum_{r_{ij}=1} d_{ij}$ The first part is the sum of the distance between $\boldsymbol{x}_i$ and differently labeled samples in its neighborhood and the second part is the sum of the distance between $\boldsymbol{x}_i$ and samples with the same class label in the neighborhood. By maximum log-likelihood, in the neighborhood, differently labeled samples are farther while similarly labeled samples are closer.

Definition 3. Given $\boldsymbol{U} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, for sample $\boldsymbol{x}_i$, neighborhood granule margin is defined as:

$$v_{x_i} = -\sum_{j=1}^{n} d_{ij} r_{ij} = dr_1 - dr_2 \tag{8}$$

Larger $v_{\boldsymbol{x}_i}$ can lead to more discrimination ability. Besides, the neighborhood discriminative relationship $\boldsymbol{R}$ considers locality of neighborhood. dr_2 considers compactness of neighborhood. By maximizing neighborhood granule margin, a more compact data representation can be obtained. Finally, the consistency has been hidden in Eq. (7).

A desired distance metric Δ can be learned by maximizing neighborhood granule margin. Fig. 2 shows neighborhood granules measured by the desired distance metric. For $\delta(\boldsymbol{x}_1)$ and $\delta(\boldsymbol{x}_4)$, compared to Fig. 1, we get more compact data representation. For $\delta(\boldsymbol{x}_2)$, although $\boldsymbol{x}_2$ is correctly classified, we get a more consistent neighborhood and the locality is preserved. For $\delta(\boldsymbol{x}_3)$, measured by the learned metric, discriminative neighborhood relationship is now consistent with the spatial distance. Hence, by maximizing neighborhood granule margin, the locality is kept, the compactness is strengthened and the consistency is improved.

3. Neighborhood distance metric learning

In this section, we introduce the distance metric learning method by maximizing neighborhood granule margin.

Compared to other distance metrics, Mahalanobis distance is independent of data distribution and has been widely used in different recognition tasks. Hence, we choose Mahalanobis distance as the distance metric Δ. The Mahalanobis distance between $\boldsymbol{x}_i \in \mathbb{R}^d$ and $\boldsymbol{x}_j \in \mathbb{R}^d$ is defined as:

$$d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)\boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \tag{9}$$

where the matrix $\boldsymbol{M} \succcurlyeq 0$ is required to be positive semidefinite. Sometimes, when there are no constraints imposed on $\boldsymbol{M}$, Eq. (9) becomes a discriminative function in terms of $\boldsymbol{M}$ [31,16]. The matrix $\boldsymbol{M}$ is often estimated from the data's inverse covariance matrix and plays an important role in multivariate statistics. If $\boldsymbol{M}$ is a diagonal matrix, the diagonal element can be used as feature weights to evaluate the feature importance. Then features can be ranked according to the feature weights and a discriminative feature subset can be selected.

We want to learn a $\boldsymbol{M}$ by maximizing neighborhood granule margin, which can be converted to a loss minimization problem. Then we get the following optimization objective:

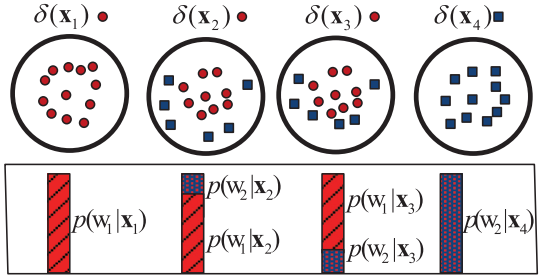

Fig. 2. Expected neighborhood granules.

$$\begin{aligned} \min_{\boldsymbol{M}} \quad & \Re(\boldsymbol{M}) + \lambda_1 \sum_{i=1}^{n} L(v_{\boldsymbol{x}_i}) \\ \text{s.t.} \quad & \boldsymbol{M} \succcurlyeq 0 \end{aligned} \tag{10}$$

where $\Re(\boldsymbol{M})$ is the regularization item imposed on $\boldsymbol{M}$, $L(v_{\boldsymbol{x}_i})$ is the loss function, and λ_1 is a constant that balances the loss and regularization. For the regularization item $\Re(\boldsymbol{M})$, different regularization can be chosen to get expected property, e.g., sparse or low rank regularization. As the work in[6], we choose $\|\boldsymbol{M}\|_F^2$, that is, the Frobenius norm of $\boldsymbol{M}$. For the loss function $L(v_{\boldsymbol{x}_i})$, different loss functions can be chosen, e.g., hinge loss in SVM [46], square loss in SRC [49], or logistic loss in logistic regression [13]. Here linear loss is adopted and the problem in Eq. (10) becomes:

$$\begin{aligned} \min_{\boldsymbol{M}} \quad & \frac{1}{2}\|\boldsymbol{M}\|_F^2 + \lambda_1 \sum_{i=1}^{n} 1 - v_{\boldsymbol{x}_i} \\ \text{s.t.} \quad & \boldsymbol{M} \succcurlyeq 0 \end{aligned} \tag{11}$$

Similar to SVM, by introducing slack variables, the optimization problem in Eq. (11) then becomes:

$$\begin{aligned} \min_{\boldsymbol{M},\xi_{ik},\xi_{ij},b} \quad & \frac{1}{2}\|\boldsymbol{M}\|_F^2 + \lambda_1 \left(\sum_{ij} \xi_{ij} + \sum_{ik} \xi_{ik} \right) \\ \text{s.t.} \quad & d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) + b \leqslant -1 + \xi_{ij}, j \in \{j | r_{ij} \neq 0 \text{ and } y_i = y_j\}; \\ & d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_k) + b \geqslant 1 - \xi_{ik}, k \in \{k | r_{ik} \neq 0 \text{ and } y_i \neq y_k\}; \\ & \boldsymbol{M} \succcurlyeq 0, \forall i, j, k, \xi_{ij} \geqslant 0, \xi_{ik} \geqslant 0. \end{aligned} \tag{12}$$

where ξ_{ij} and ξ_{ik} are slack variables. $j \in \{j | r_{ij} \neq 0 \text{ and } y_i = y_j\}$ means that $\boldsymbol{x}_j$ is in the neighborhood of $\boldsymbol{x}_i$ with the same label. $d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j) + b \leqslant -1 + \xi_{ij}$ represents that the distance between $\boldsymbol{x}_i$ and similarly labeled sample $\boldsymbol{x}_j$ should be decreased. Correspondingly, $k \in \{k | r_{ik} \neq 0 \text{ and } y_i \neq y_k\}$ means $\boldsymbol{x}_k$ is in the neighborhood of $\boldsymbol{x}_i$ with a different label. $d_{\boldsymbol{M}}(\boldsymbol{x}_i, \boldsymbol{x}_k) + b \geqslant 1 - \xi_{ik}$ represents that the distance between $\boldsymbol{x}_i$ and differently labeled sample $\boldsymbol{x}_k$ should be enlarged.

Let us denote by $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \boldsymbol{z}_{i2})$ a generated sample pair. If $\boldsymbol{z}_{i1}$ and $\boldsymbol{z}_{i2}$ have the same label, then we call $\boldsymbol{z}_i$ a positive sample pair and label it as "−1"; otherwise, $\boldsymbol{z}_i$ is a negative sample pair and labelled as "+1"; The covariance matrix of the two samples in $\boldsymbol{z}_i$ is $\boldsymbol{C}_i = (\boldsymbol{z}_{i1} - \boldsymbol{z}_{i2})^T (\boldsymbol{z}_{i1} - \boldsymbol{z}_{i2})$. Suppose that we generated *ns* training sample pairs, and thus we have *ns* covariance matrices $\boldsymbol{C}_i, i = 1, 2, \ldots, ns$. We label $\boldsymbol{C}_i$ as "+1" or "−1" based on the label of $\boldsymbol{z}_i$, and define the following kernel function to measure the similarity between $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$:

$$k(\boldsymbol{C}_i, \boldsymbol{C}_j) = tr(\boldsymbol{C}_i \boldsymbol{C}_j) = \langle \boldsymbol{C}_i, \boldsymbol{C}_j \rangle \tag{13}$$

where $tr(\cdot)$ is the trace operator of a matrix and $\langle \cdot, \cdot \rangle$ means the inner product of matrices.

Suppose that we have a query sample pair, denoted by $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2)$. The covariance matrix of $\boldsymbol{z}$ is denoted by $\boldsymbol{C}$. We introduce the following discriminative function to judge whether $\boldsymbol{z}$ is positive or negative:

$$f(\boldsymbol{C}) = \sum_i \alpha_i l_i k(\boldsymbol{C}_i, \boldsymbol{C}) + b = \sum_i \alpha_i l_i < \boldsymbol{C}_i, \boldsymbol{C} > + b = < \sum_i \alpha_i l_i \boldsymbol{C}_i, \boldsymbol{C} > + b \tag{14}$$

where l_i is the label of pair $\boldsymbol{z}_i$, and α_i is a weight. Let

$$\boldsymbol{M} = \sum_i \alpha_i l_i \boldsymbol{C}_i. \tag{15}$$

Then we have $f(\boldsymbol{C}) = \langle \boldsymbol{M}, \boldsymbol{C} \rangle + b$.

The metric learning problem in Eq. (12) can then be converted into the following problem:

$$\begin{aligned} \min_{\boldsymbol{M},b,\xi} \quad & \frac{1}{2}\|\boldsymbol{M}\|_F^2 + \lambda_1 \sum_i \xi_i \\ \text{s.t.} \quad & l_i(\langle \boldsymbol{M}, \boldsymbol{C}_i \rangle + b) \geqslant 1 - \xi_i, \xi_i \geqslant 0 \end{aligned} \tag{16}$$

The Lagrangian of Eq. (16) can be defined as follows:

$$Lr(\boldsymbol{M}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{M}\|_F^2 + \lambda_1 \sum_i \xi_i - \sum_i \alpha_i [l_i(\langle \boldsymbol{M}, \boldsymbol{C}_i \rangle + b)] - 1 + \xi_i - \sum_i \beta_i \xi_i \tag{17}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the Lagrange multipliers which satisfy $\alpha_i \geqslant 0$ and $\beta_i \geqslant 0, \forall i$. To convert the original problem to its dual, we let the derivative of the Lagrangian with respect to $\boldsymbol{M}, b$ and ξ to be 0:

$$\frac{\partial Lr(\boldsymbol{M}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{M}} = 0 \Rightarrow \boldsymbol{M} - \sum_i \alpha_i l_i \boldsymbol{C}_i = 0 \tag{18}$$

$$\frac{\partial Lr(\boldsymbol{M}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial b} = 0 \Rightarrow \sum_i \alpha_i l_i = 0 \tag{19}$$

$$\frac{\partial Lr(\boldsymbol{M}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \xi_i} = 0 \Rightarrow \lambda_1 - \alpha_i - \beta_i = 0 \Rightarrow 0 \leqslant \alpha_i \leqslant \lambda_1, \forall i \tag{20}$$

Then we substitute Eqs. (18)–(20) back into the Lagrangian, and we get the Lagrange dual problem of the metric learning problem in Eq. (16):

$$\begin{aligned} \max_{\alpha} \quad & -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j l_i l_j k(\boldsymbol{C}_i, \boldsymbol{C}_j) + \sum_i \alpha_i \\ \text{s.t.} \quad & 0 \leqslant \alpha_i \leqslant \lambda_1, \sum_i \alpha_i l_i = 0 \end{aligned} \tag{21}$$

Obviously, the problem in Eq. (21) can be easily solved by the support vector machine (SVM) solvers such as LIBSVM [8]. As $d_{\boldsymbol{M}}(\boldsymbol{z}_1, \boldsymbol{z}_2)$ is required to be a Mahalanobis distance metric, $\boldsymbol{M}$ should be semi-positive definite. Similar to MMC [50] and MCML [14], we can compute the singular value decomposition (SVD) of $\boldsymbol{M} = \boldsymbol{U}\Lambda\boldsymbol{V}$, where Λ is the diagonal matrix of eigenvalues, and then set the negative eigenvalues in Λ to 0, resulting in a new diagonal matrix Λ_+. Finally, we let $\boldsymbol{M}_+ = \boldsymbol{U}\Lambda_+\boldsymbol{V}$ be the learned matrix. The algorithm of neighborhood distance metric learning (NDML) method is given in Table 1. In LIBSVM, sequential minimal optimization (SMO) is used. The time complexity of SMO is $O(n^2 d)$, where n and d are the number of samples and features, respectively. Hence, the time complexity of NDML is $O(ns^2 d)$, where ns and d are the number of sample pairs and features, respectively. As there are quite a lot of SVM solvers for large scale problems, NDML has good property in scalability, especially when the number of samples is very large. Besides, when the feature dimension is quite high, PCA can be adopted to reduce the feature dimension as a preprocessing step for metric learning.

4. Multi-granularity distance metric learning

The learned $\boldsymbol{M}$ is affected by the granularity, i.e., neighborhood size. The neighborhood size decides the number of sample pairs for metric learning. When we assign a very small value to δ, the number of selected sample pairs would be quite small; when the neighborhood size is very large, there will be at most $n(n-1)/2$ sample pairs. As the learned $\boldsymbol{M}$ may be sensitive to δ (refer to Section 5.1), in this section, we propose a multi-granularity distance metric learning method. As shown in Fig. 3, for training samples, given different granularity we can learn different distance metrics $\boldsymbol{M}$. Then inspired by the work in [56], we can learn the weights of different granularity. Given a test sample, we get a decision vector using the learned distance metrics, and then combine the outputs to get the final decision.

Given $\boldsymbol{S} = \{(\boldsymbol{x}_i, y_i)\}, i = 1, 2, \ldots n$, distance metrics $\boldsymbol{M}_1, \boldsymbol{M}_2, \ldots, \boldsymbol{M}_t$ are learned under t granularity. The classification results in t different granularity spaces are $\boldsymbol{H} \in \mathbb{R}^{n\times t}$, where $\boldsymbol{w} = \langle w_1, w_2, \ldots, w_t\rangle$ is the weight vector of different granularity.

Definition 4. [56] For multi-class classification, the classification outputs in t different granular spaces are $\{h_{ij}\}, j = 1, 2, \ldots, t$. The matrix $\boldsymbol{Q} = \left\{q_{ij}\right\}_{n\times t}$ is defined as:

$$q_{ij} = \varsigma(y_i, h_{ij}) = \begin{cases} +1, & \text{if } y_i = h_{ij}, \\ -1, & \text{if } y_i \neq h_{ij}. \end{cases} \tag{22}$$

Definition 5. [56] For $\boldsymbol{x}_i$, the classification outputs in t different granular spaces are $\{h_{ij}\}, j = 1, 2, \ldots, t$. The ensemble margin of $\boldsymbol{x}_i$ is defined as

$$\rho(\boldsymbol{x}_i) = \sum_{j=1}^{t} w_j q_{ij}. \tag{23}$$

Ensemble margin should be enlarged by weight learning and margin maximization is usually transformed to a loss minimization problem [22,44,45].

Definition 6. [56] For each sample $\boldsymbol{x}_i \in \boldsymbol{S}$, ensemble margin of $\boldsymbol{x}_i$ is $\rho(\boldsymbol{x}_i)$. Then the ensemble loss of $\boldsymbol{x}_i$ is

Table 1
The algorithm of neighborhood distance metric learning.

Input: A set of samples $\boldsymbol{S} = \{(\boldsymbol{x}_i, y_i)\}$, i = 1, 2, …n
Output: Distance metric $\boldsymbol{M}$
1: compute neighborhood discriminative relationship $\boldsymbol{R}$;
2: get positive and negative constraints from $\boldsymbol{R}$;
3: get $\boldsymbol{M}$ by solving Eq. (21);
4: project $\boldsymbol{M}$ to PSD cone, $\boldsymbol{M}_+ = \boldsymbol{U}\Lambda_+\boldsymbol{V}$.

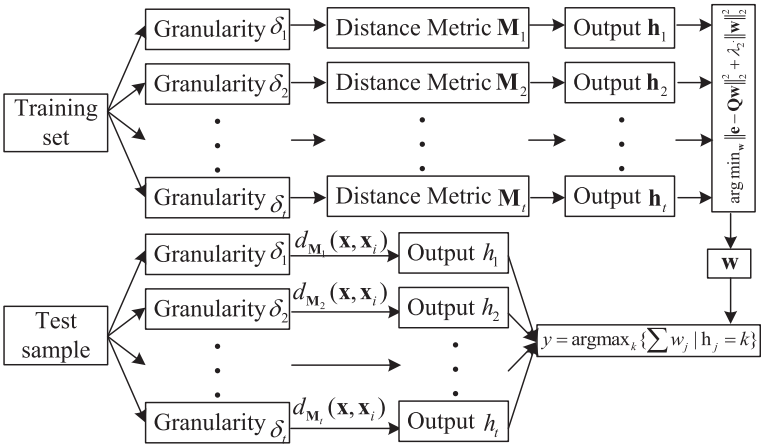

Fig. 3. Framework of multi-granularity metric learning.

$$\varepsilon_{\boldsymbol{x}_i} = \varepsilon(\rho(\boldsymbol{x}_i)) = \varepsilon\left(\sum_{j=1}^{m} w_j q_{ij}\right) \tag{24}$$

where ε is a loss function. For a sample set $\boldsymbol{S}$, the ensemble square loss is

$$\varepsilon(\boldsymbol{S}) = \sum_{i=1}^{n} \varepsilon_{\boldsymbol{x}_i} = \sum_{i=1}^{n} (1 - \rho(\boldsymbol{x}_i))^2 = \|\boldsymbol{e} - \boldsymbol{Q}\boldsymbol{w}\|_2^2 \tag{25}$$

where $\boldsymbol{e} = [1; 1; \ldots; 1], \boldsymbol{e} \in \mathbb{R}^n$.

To get better margin distribution, we should minimize the ensemble square loss with l_p-norm regularization imposed on the weight vector to get a stable solution. Hence, we can construct the optimization objective as follows.

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \|\boldsymbol{e} - \boldsymbol{Q}\boldsymbol{w}\|_2^2 + \lambda_2 \|\boldsymbol{w}\|_{l_p} \text{ s.t. } \sum_{j=1}^{m} w_j = 1, \tag{26}$$

where λ_2 is a constant.

The problem in Eq. (26) can be solved as follows:

$$\hat{\boldsymbol{w}} = \arg\ \min_{\boldsymbol{w}} \{\|\bar{\boldsymbol{e}} - \bar{\boldsymbol{Q}}\boldsymbol{w}\|_2^2 + \lambda_2 \|\boldsymbol{w}\|_{l_p}\}. \tag{27}$$

where $\bar{\boldsymbol{e}} = [\boldsymbol{e}; 1], \bar{\boldsymbol{Q}} = [\boldsymbol{Q}; \boldsymbol{e}]$.

For multi-granularity metric learning problem, we need to seek a sparse solution (as shown in Section 5.1). Hence, l_1 norm is imposed on $\boldsymbol{w}$ and l_1_ls is used to solve this problem [26].

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \{\|\bar{\boldsymbol{e}} - \bar{\boldsymbol{Q}}\boldsymbol{w}\|_2^2 + \lambda_2 \|\boldsymbol{w}\|_1\}. \tag{28}$$

After the weight vector $\boldsymbol{w}$ is learned by solving Eq. (28), given a test sample $\boldsymbol{x}$, using the learned distance metrics $\boldsymbol{M}_1, \boldsymbol{M}_2, \ldots, \boldsymbol{M}_t$, we can get a decision vector $\boldsymbol{h} = [h_1, h_2, \ldots, h_t]$. Then the prediction label of $\boldsymbol{x}$ is:

$$y = argmin_k \left\{\sum w_j | h_j = k\right\} \tag{29}$$

The algorithm of multi-granularity distance metric learning (MGML) method is given in Table 2.

Table 2
The algorithm of multi-granularity distance metric learning.

Input: A set of samples $\boldsymbol{S} = \{(\boldsymbol{x}_i, y_i)\}$, $i = 1, 2, \ldots n$
Output: Distance metrics $\boldsymbol{M}_1, \boldsymbol{M}_2, \ldots, \boldsymbol{M}_t$ and granularity weights $\boldsymbol{w}$
1: Choose t granularity $\delta = \{\delta_1, \delta_2, \ldots, \delta_t\}$;
2: Learn t distance metrics $\boldsymbol{M}_1, \boldsymbol{M}_2, \ldots, \boldsymbol{M}_t$ of t granularity;
3: Get classification outputs $\boldsymbol{H}$ for t granularity;
4: Get the decision matrix $\boldsymbol{Q}$;
5: Learn granularity weight vector $\boldsymbol{w}$ by solving Eq. (28).

Table 3
Data description of UCI datasets.

Data	Sample	Feature	Class
breast	106	9	6
glass	214	9	7
heart	270	13	2
horse	368	22	2
iono	351	33	2
sonar	208	60	2
wine	178	13	3
wpbc	198	33	2

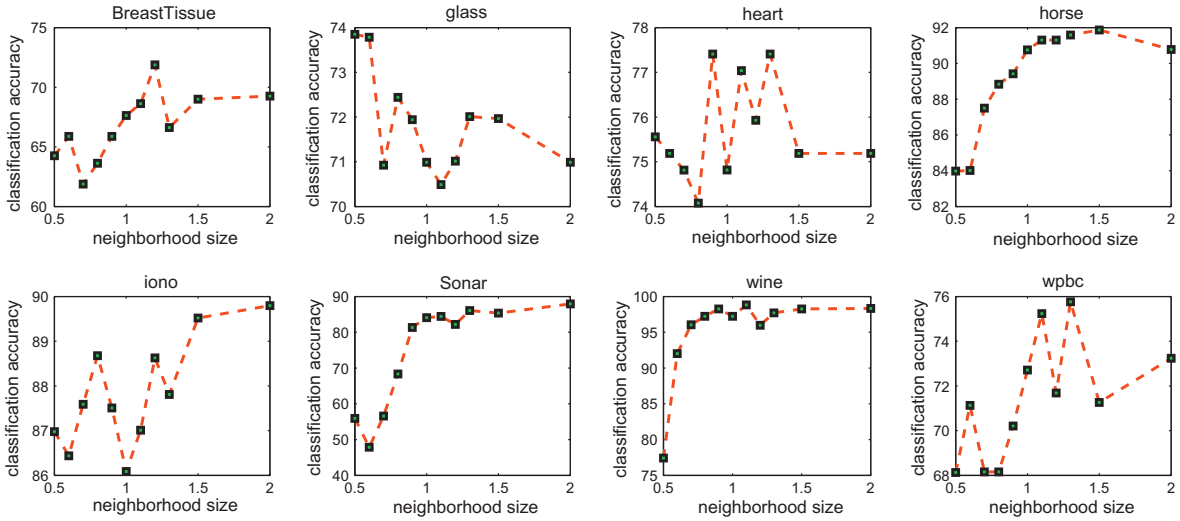

Fig. 4. Metric learning performance with different neighborhood size.

Table 4
Accuracy of different distance metric learning methods.

Data	NN	LMNN [6]	NCA [15]	ITML [10]	NDML	MGML
breast	69.3 ± 15.8	65.6 ± 14.3	69.3 ± 11.1	63.2 ± 14.1	**71.9 ± 12.3**	**70.5 ± 15.0**
glass	66.3 ± 7.9	62.8 ± 16.3	60.6 ± 12.2	65.5 ± 11.2	**73.9 ± 11.6**	**71.9 ± 10.5**
heart	75.6 ± 10.0	77.8 ± 6.1	77.4 ± 5.9	78.9 ± 9.6	**77.4 ± 11.9**	**79.6 ± 6.8**
horse	89.2 ± 3.9	90.5 ± 4.6	90.3 ± 7.4	91.1 ± 4.7	**91.9 ± 3.3**	**92.1 ± 3.9**
iono	86.4 ± 4.9	88.1 ± 5.1	87.5 ± 6.6	89.0 ± 6.9	**89.8 ± 4.7**	**89.2 ± 5.0**
sonar	85.5 ± 9.2	88.1 ± 10.1	85.6 ± 5.5	80.9 ± 7.3	**89.0 ± 5.5**	**88.4 ± 8.8**
wine	94.9 ± 5.1	97.7 ± 3.0	96.7 ± 3.9	97.7 ± 3.0	**98.8 ± 2.5**	**98.9 ± 2.3**
wpbc	70.7 ± 6.7	**78.8 ± 5.6**	72.7 ± 9.2	69.2 ± 10.4	75.3 ± 9.3	75.3 ± 9.8
Average	79.7	81.2	80.0	79.4	**83.5**	**83.2**

Bold means the highest accuracy in all the comparison methods.

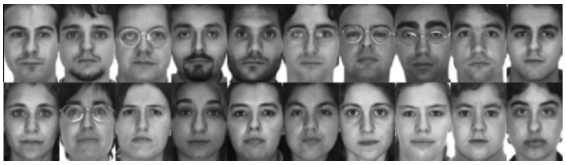

Fig. 5. Face images of gender classification.

5. Experiment analysis

In this section, the performance of the proposed distance metric learning method is evaluated on UCI datasets, gender classification and object categorization tasks. The code of MGML can be downloaded from http://www4.comp.pol-

Table 5
Gender classification accuracy of different methods.

Method	NN	LMNN [6]	NCA [15]	ITML [10]	MGML
Accu.	90.3	91.0	91.4	90.7	**92.1**

Bold means the highest accuracy in all the comparison methods.

Fig. 6. COIL database.

Table 6
Recognition rate on COIL20 database.

Method	NN	LMNN [6]	NCA [15]	ITML [10]	MGML
Accu.	91.1 ± 4.2	91.7 ± 4.0	95.0 ± 3.8	94.5 ± 5.9	**95.4 ± 4.3**

Bold means the highest accuracy in all the comparison methods.

yu.edu.hk/cspzhu/. For parameter setting, there are two parameters in MGML, i.e., λ_1 in Eq. (21) and λ_2 in Eq. (28). In all the experiments, λ_1 is set as 1 and λ_2 is set as 0.1.

5.1. UCI datasets

We collected eight datasets from UCI Machine Learning Repository [2]. Table 3 lists the detailed information of the datasets, including numbers of samples, features and classes. Firstly, we shows the classification accuracy of NDML with granularity. As shown in Fig. 4, the accuracy varies greatly under different granularity. For different datasets the optimal granularity is different. This validates that multi-granularity learning is necessary.

Then we compare the classification accuracy of different metric learning methods, as illustrated in Table 4. We use nearest neighbor classifier to report the performance. NN means that the Euclidean distance is used. The highest accuracy of NDML using different granularity is also reported. From the comparison we see that NDML and MGML outperform other methods on seven datasets except wpbc. Compared with MGML, NDML is superior on some datasets. Whereas, NDML has to suffer from granularity sensitivity. Compared to Euclidean distance, MGML improves the average classification accuracy by 3.5%.

5.2. Gender classification

A non-occluded subset (14 images per subject) of the AR dataset [35] is used, which consists of 50 male and 50 female subjects. We use the images from the first 25 males and 25 females for training, and the remaining images for testing. The images were cropped to 60×43. A subset of faces of 10 men and 10 women are shown in Fig. 5. PCA was used to reduce the dimension of each image to 50. The classification accuracy of different methods is shown in Table 5. Similar to the performance on UCI datasets, MGML also achieves the highest accuracy.

5.3. Object categorization

We choose COIL20 database to validate the performance of the proposed method. COIL20 database contains 20 objects. The images of each objects were taken 5° apart as the object is rotated on a turntable and each object has 72 images. The size of each image is 32×32 pixels, with 256 gray levels per pixel. The sample images of 20 objects are shown in Fig. 6. PCA was used to reduce the dimension of each image from 1024 to 100. The performance of different methods is illustrated in Table 6. Compared to NN, NCA, ITML and MGML improve the accuracy greatly while for LMNN the improvement is unconspicuous. Compared to NCA and ITML, the performance of MGML is better.

6. Conclusions

Learning a desired distance metric from given training samples is quite important in machine learning. In this paper, we proposed a multi-granularity distance metric learning method via maximizing neighborhood granule margin. Firstly, we propose neighborhood granule margin that considers neighborhood locality, compactness and consistency to evaluate the

distance metric. Then we learn a desired metric by maximizing neighborhood granule margin and formulate metric learning as a sample pair classification task, which can be effectively solved by standard SVM solvers. Considering the sensitivity of metric learning to granularity, we propose a multi-granularity metric learning method by margin distribution optimization. Experiment analysis on different classification tasks shows that the proposed method outperforms the state-of-the-art metric learning methods.

Acknowledgments

This work is partly supported by National Program on Key Basic Research Project under Grant 2013CB329304, National Natural Science Foundation of China – China under Grant 61222210 and New Century Excellent Talents in University under Grant NCET-12-0399.

References

[1] V. Ablavsky, S. Sclaroff, Learning parameterized histogram kernels on the simplex manifold for image and action classification, in: International Conference on Computer Vision (ICCV 2011), IEEE, 2011, pp. 1473–1480.
[2] K. Bache, M. Lichman, UCI Machine Learning Repository, 2013.
[3] A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall, Learning distance functions using equivalence relations, in: International Conference on Machine Learning (ICML 2003), ACM, 2003, pp. 11–18.
[4] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (2009) 1–127.
[5] M. Bilenko, S. Basu, R.J. Mooney, Integrating constraints and metric learning in semi-supervised clustering, in: International Conference on Machine Learning (ICML 2004), ACM, 2004, pp. 11–18.
[6] J. Blitzer, K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, in: Advances in Neural Information Processing Systems (NIPS 2005), MIT, 2005, pp. 1473–1480.
[7] D. Cai, X. He, J. Han, Spectral regression for efficient regularized subspace learning, in: International Conference on Computer Vision (ICCV 2007), IEEE, 2007, pp. 1–8.
[8] C.C. Chang, C.J. Lin, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (2011) 27.
[9] M. Dash, H. Liu, Consistency-based search in feature selection, Artif. Intell. 151 (2003) 155–176.
[10] J.V. Davis, B. Kulis, P. Jain, S. Sra, I.S. Dhillon, Information-theoretic metric learning, in: International Conference on Machine Learning (ICML 2007), ACM, 2007, pp. 209–216.
[11] N. Fan, Learning nonlinear distance functions using neural network for regression with application to robust human age estimation, in: International Conference on Computer Vision (ICCV 2011), IEEE, 2011, pp. 249–254.
[12] S.L. France, J. Douglas Carroll, H. Xiong, Distance metrics for high dimensional nearest neighborhood recovery: compression and normalization, Inf. Sci. 184 (2012) 92–110.
[13] J. Friedman, T. Hastie, R. Tibshirani, Special invited paper. additive logistic regression: a statistical view of boosting, Ann. Stat. 28 (2000) 337–374.
[14] A. Globerson, S.T. Roweis, Metric learning by collapsing classes, in: Advances in Neural Information Processing Systems (NIPS 2005), MIT, 2005, pp. 451–458.
[15] J. Goldberger, G.E. Hinton, S.T. Roweis, R. Salakhutdinov, Neighbourhood components analysis, in: Advances in Neural Information Processing Systems (NIPS 2004), MIT, 2004, pp. 513–520.
[16] M. Guillaumin, J. Verbeek, C. Schmid, Is that you? Metric learning approaches for face identification, in: International Conference on Computer Vision (ICCV 2009), IEEE, 2009, pp. 498–505.
[17] X. He, S. Yan, Y. Hu, P. Niyogi, H.J. Zhang, Face recognition using laplacianfaces, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 328–340.
[18] S.C. Hoi, W. Liu, M.R. Lyu, W.Y. Ma, Learning distance metrics with contextual constraints for image retrieval, in: Computer Vision and Pattern Recognition (CVPR 2006), IEEE, 2006, pp. 2072–2078.
[19] Q. Hu, W. Pedrycz, D. Yu, J. Lang, Selecting discrete and continuous features based on neighborhood decision error minimization, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 40 (2010) 137–150.
[20] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inf. Sci. 178 (2008) 3577–3594.
[21] Q. Hu, D. Yu, Z. Xie, Neighborhood classifiers, Expert Syst. Appl. 34 (2008) 866–876.
[22] Q. Hu, P. Zhu, Y. Yang, D. Yu, Large-margin nearest neighbor classifiers via sample weight learning, Neurocomputing 74 (2011) 656–660.
[23] K. Huang, Y. Ying, C. Campbell, GSML: a unified framework for sparse metric learning, in: International Conference on Data Mining (ICDM 2009), IEEE, 2009, pp. 189–198.
[24] A.K. Jain, Data clustering: 50 years beyond *k*-means, Pattern Recogn. Lett. 31 (2010) 651–666.
[25] N. Jiang, W. Liu, Y. Wu, Order determination and sparsity-regularized metric learning adaptive visual tracking, in: Computer Vision and Pattern Recognition (CVPR 2012), IEEE, 2012, pp. 1956–1963.
[26] S. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, An interior-point method for large-scale l1-regularized least squares, IEEE J. Sel. Top. Signal Process. 1 (2007) 606–617.
[27] T. Kozakaya, S. Ito, S. Kubota, Random ensemble metrics for object recognition, in: International Conference on Computer Vision (ICCV 2011), IEEE, 2011, pp. 1959–1966.
[28] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Computer Vision and Pattern Recognition (CVPR 2006), IEEE, 2006, pp. 2169–2178.
[29] G. Lebanon, Metric learning for text documents, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 497–508.
[30] J.E. Lee, R. Jin, A.K. Jain, Rank-based distance metric learning: an application to image retrieval, in: Computer Vision and Pattern Recognition (CVPR 2008), IEEE, 2008, pp. 1–8.
[31] Z. Li, L. Cao, S. Chang, J.R. Smith, T.S. Huang, Beyond mahalanobis distance: Learning second-order discriminant function for people verification, in: Computer Vision and Pattern Recognition Workshops (CVPRW 2012), IEEE, 2012, pp. 45–50.
[32] D. Lim, B. McFee, G.R. Lanckriet, Robust structural metric learning, in: International Conference on Machine Learning (ICML 2013), ACM, 2013, pp. 615–623.
[33] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, IEEE Trans. Image process. 11 (2002) 467–476.
[34] J. Lu, J. Hu, X. Zhou, Y. Shang, Y.P. Tan, G. Wang, Neighborhood repulsed metric learning for kinship verification, in: Computer Vision and Pattern Recognition (CVPR 2012), IEEE, 2012, pp. 2594–2601.
[35] A. Martinez, The AR Face Database, CVC Technical Report 24 (1998).
[36] B. McFee, G.R. Lanckriet, Metric learning to rank, in: International Conference on Machine Learning (ICML 2010), ACM, 2010, pp. 775–782.
[37] A. Mignon, F. Jurie, Pcca: A new approach for distance learning from sparse pairwise constraints, in: Computer Vision and Pattern Recognition (CVPR 2012), IEEE, 2012, pp. 2666–2672.

[38] A. Neumaier, Solving ill-conditioned and singular linear systems: a tutorial on regularization, Siam Rev. 40 (1998) 636–666.
[39] H.V. Nguyen, L. Bai, Cosine similarity metric learning for face verification, in: Asian Conference of Computer Vision (ACCV 2010), Springer, 2010, pp. 709–720.
[40] X. Niyogi, Locality preserving projections, in: Advances in Neural Information Processing Systems (NIPS 2004), MIT, 2004, pp. 153–160.
[41] X. Peng, D. Xu, Twin Mahalanobis distance-based support vector machines for pattern recognition, Inf. Sci. 200 (2012) 22–37.
[42] G.J. Qi, J. Tang, Z.J. Zha, T.S. Chua, H.J. Zhang, An efficient sparse metric learning in high-dimensional space via l1-penalized log-determinant regularization, in: International Conference on Machine Learning (ICML 2009), ACM, 2009, pp. 841–848.
[43] L. Qiao, S. Chen, X. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognition 43 (2010) 331–341.
[44] C. Shen, H. Li, Boosting through optimization of margin distributions, IEEE Trans. Neural Networks 21 (2010) 659–666.
[45] C. Shen, H. Li, On the dual formulation of boosting algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 2216–2231.
[46] D. Tao, X. Tang, X. Li, X. Wu, Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1088–1099.
[47] J. Wang, X. Gao, Q. Wang, Y. Li, ProDis-contSHC: learning protein dissimilarity measures and hierarchical context coherently for protein–protein comparison in protein database retrieval, BMC Bioinf. 13 (2012) S2.
[48] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Computer Vision and Pattern Recognition (CVPR 2010), IEEE, 2010, pp. 3360–3367.
[49] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 210–227.
[50] E.P. Xing, M.I. Jordan, S. Russell, A. Ng, Distance metric learning with application to clustering with side-information, in: Advances in Neural Information Processing Systems (NIPS 2002), MIT, 2002, pp. 505–512.
[51] J. Xu, J. Yang, Z. Lai, *K*-Local hyperplane distance nearest neighbor classifier oriented local discriminant analysis, Inf. Sci. 232 (2013) 11–26.
[52] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Computer Vision and Pattern Recognition (CVPR 2009), IEEE, 2009, pp. 1794–1801.
[53] Y. Ying, P. Li, Distance metric learning with eigenvalue optimization, J. Mach. Learn. Res. 13 (2012) 1–26.
[54] D. Yu, X. Yu, Q. Hu, J. Liu, A. Wu, Dynamic time warping constraint learning for large margin nearest neighbor classification, Inf. Sci. 181 (2011) 2787–2796.
[55] Z. Zhao, H. Liu, Spectral feature selection for supervised and unsupervised learning, in: International Conference on Machine Learning (ICML 2007), ACM, 2007, pp. 1151–1157.
[56] P. Zhu, Q. Hu, Adaptive neighborhood granularity selection and combination based on margin distribution optimization, Inf. Sci. 249 (2013) 1–12.
[57] P. Zhu, L. Zhang, Q. Hu, S. Shiu, Multi-scale patch based collaborative representation for face recognition with margin distribution optimization, in: European Conference on Computer Vision (ECCV 2012), Springer, 2012, pp. 822–835.