Pattern Recognition 43 (2010) 2971–2981

Contents lists available at ScienceDirect

Pattern Recognition

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

Generalized iterative RELIEF for supervised distance metric learning ☆

Chin-Chun Chang

Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung 202, Taiwan

ARTICLE INFO

Article history:
Received 14 July 2009
Received in revised form
24 February 2010
Accepted 28 February 2010

Keywords:
Distance metric learning
Iterative RELIEF
Feature weighting

ABSTRACT

The RELIEF algorithm is a popular approach for feature weighting. Many extensions of the RELIEF algorithm are developed, and I-RELIEF is one of the famous extensions. In this paper, I-RELIEF is generalized for supervised distance metric learning to yield a Mahananobis distance function. The proposed approach is justified by showing that the objective function of the generalized I-RELIEF is closely related to the expected leave-one-out nearest-neighbor classification rate. In addition, the relationships among the generalized I-RELIEF, the neighbourhood components analysis, and graph embedding are also pointed out. Experimental results on various data sets all demonstrate the superiority of the proposed approach.

1. Introduction

The RELIEF algorithm is an effective, simple, and widely used approach to feature weighting [1–4]. The weight for a feature of a measurement vector is defined in terms of feature relevance. In [5], a probabilistic interpretation of RELIEF is made, which states that the learned weight for a feature is propositional to the difference between two conditional probabilities. These two probabilities are of the value of a feature being different conditioned on the given *nearest miss* and *nearest hit*, respectively. It has been pointed out that RELIEF is indeed an online algorithm for maximizing a margin-based objective function defined by the nearest neighbors of samples [6]. Thus, RELIEF usually performs better than the filter approach [3] due to the feedback of the nearest-neighbor classifier; in addition, RELIEF is often more efficient than the wrapper approach [3] because RELIEF determines the feature weights through solving a convex optimization problem. Despite the success of RELIEF, some improvements on RELIEF have been come out.

The RELIEF-F algorithm [5] extends RELIEF to deal with multi-classes problems with incomplete and noisy data. The RRELIEF-F [7] algorithm is aimed for determining proper weights for the regression problem. A unified view for a family of the RELIEF algorithms is provided in [8]. As indicated by Sun [6], one of the weakness of RELIEF is that the nearest neighbor of a sample is defined in the original measurement space, which may not be the one in the weighted space. The RELIEF-F algorithm uses an average of the k nearest neighbors of a sample instead of the nearest neighbor of the sample to compute the sample margins. The I-RELIEF algorithm [6] regards the indication of the nearest neighbor of a sample as a latent variable and follows the framework of the Expectation-Maximization (EM) algorithm [9] to calculate the feature weights. In addition, RELIEF assigns high weights to all relevant features regardless of the correlation among them. Thus, in the presence of highly correlated features, RELIEF may misleads the selection of feature subsets [3,10]. In [11], RELIEF is extended to induce a full metric matrix. In [12], the correlated relevant features are removed by the k-means algorithm. In [13], RELIEF with the consideration of the correlation among features is utilized to find an effective feature subset. The O-RELIEF algorithm [10] applies an orthogonal transform to remove the correlations between features, and then weights the uncorrelated features by means of RELIEF.

The RELIEF algorithm is actually a kind of supervised distance metric learning because the learned distance function is a special case of the Mahananobis distance function. In [14], a comprehensive survey of distance metric learning is provided. Generally, the learned metric can be a global metric or a local metric. The local metric approach is aimed for a distance metric proper within a neighborhood of a query point. On the other hand, the global metric approach, such as RELIEF, induces a distance metric appropriate to all samples.

The global distance metric is often determined in a transformed space. The approach of Fukunaga and Flick [15] is based on the aspect of minimizing the difference from the finite sample risk to the asymptotic risk defined by the k-nearest neighbors. The approach of Hastie and Tibshirani [16], referred as to DANN, is based on a different weighting scheme, and refines the distance metric by a few iterations. The neighbourhood components

☆This work was supported financially by National Science Council under the Grant NSC 94-2213-E-019-011 and in part by NTOU-RD981-05-02-04-01.

E-mail address: cvml@mail.ntou.edu.tw

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.02.024

Table 1
A comparison of some variants of the RELIEF algorithm, where the symbol ○ for the column "Neighborhood" represents that the method is based on neighborhood relationships defined in the original measurement space.

Method	Output	Neighborhood	Description
RELIEF-F [5]	Feature weights	○	Use an average of the k nearest neighbors of a sample for the nearest hit/miss
I-RELIEF [6]	Feature weights	×	Use a stochastic-nearest-neighbor model
O-RELIEF [10]	Full metric matrix	○	Remove correlations among features by an orthogonal transformation, and then calculate feature weights by RELIEF
LFE [11]	Full metric matrix	○	Find the metric matrix such that the difference between the total squared distance from samples to their respective nearest misses and the total squared distance from samples to their respective nearest hits is maximized

analysis [17], referred as to NCA, learns distance metrics based on stochastic nearest neighbors. The convergence of DANN and NCA, however, are not ensured. The approach of Yang et al. [18], referred as to LDM, learns distance metrics by optimizing the compactness of a class and the separability among classes in a local sense. Because LDM indeed reweights the components of the principal component analysis (PCA) of the samples, the transformation matrix induced by LDM is not general.

Some global distance metrics are induced based on predefined equivalence and inequivalence constraints on training samples. The equivalence constraint forces the pair of semantically similar samples to be close together, whereas the inequivalence constraint states that the pair of dissimilar samples should not be near. With some kinds of predefined equivalence constraint and inequivalence constraint, the convex programming [19], the semidefinite programming [20,21], the convolutional network [22], and the LogDet divergence [23] have been used to induce distance metrics. However, when the underlying sample distribution of a class is multimodal, determining such constraints on samples for training may be difficult.

As mentioned previously, the RELIEF algorithm may be disadvantageous when the nearest neighbors of samples defined in the original measurement space are inappropriate or when some features have high correlations. Table 1 compares some variants of the RELIEF algorithm discussed previously, in which only I-RELIEF uses updated neighborhood information for every iteration. In addition, I-RELIEF has a well-understood convergence property. Hence, it seems promising to generalize the I-RELIEF algorithm to yield an effective Mahananobis distance function to alleviate the weakness of the RELIEF-like algorithm.

In this paper, through modifying the objective function of I-RELIEF, the batch version of the I-RELIEF algorithm is extended to yield a Mahananobis distance function. This distance function is suitable for the nearest-neighbor classifier. This is because the objective function of the generalized I-RELIEF is based on the expected leave-one-out nearest-neighbor classification rate. In addition, the generalized I-RELIEF is shown to be related to NCA. Besides, it reveals that the generalized I-RELIEF is under the framework of graph embedding [24] and thus also suitable for supervised dimensionality reduction.

This paper is organized as follows. Section 2 presents a probability model for the leave-one-out nearest-neighbor classification rate, and shows the relationships among this model and the objective functions of NCA and I-RELIEF. Section 3 presents the generalized I-RELIEF algorithm. Section 4 shows experimental results. The concluding remarks are in the last section.

2. The distance function for the nearest-neighbor classification

Let $\mathcal{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$ with $\mathbf{x}_i\in\mathbb{R}^n$ be a set of N sample vectors, and $\mathbf{L}$ be an $n\times d$ transformation matrix, where $d\le n$. Without lost of generality, the sample vectors are assumed to have a zero mean. Suppose that these sample vectors could be mapped through $\mathbf{L}^T$ onto a d-dimensional space $\mathcal{F}$ such that in $\mathcal{F}$, almost every sample and the nearest neighbor of the sample both belong to the same class. Thus, the metric space $\mathcal{F}$ is ideal for the nearest-neighbor classification. Since the nearest neighbor is often defined in terms of Euclidean distance, determining an effective Euclidean distance between the projections of two sample vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{F}$, which can be calculated by $\mathtt{dist}_{\mathcal{F}}(\mathbf{x},\mathbf{y})=\|\mathbf{L}^T(\mathbf{x}-\mathbf{y})\|_2$, is the main goal of this study.

In the following, first, a probability model for the leave-one-out nearest-neighbor classification rate, which is based on the model of the stochastic nearest neighbor of NCA, is introduced. Next, a link between this model and the objective function of I-RELIEF is established. This link reveals that although defined in terms of feature relevance, the objective function of I-RELIEF is also closely related to the expected leave-one-out nearest-neighbor classification rate. Thus, it is justified that the distance function induced through a generalization of the objective function of I-RELIEF could be an effective distance function.

2.1. The distance function induced by maximizing the leave-one-out nearest-neighbor classification rate

Define $\mathbf{z}_i=(z_{i1},z_{i2},\ldots,z_{iN})^T$ as a binary N-dimensional latent variable for the i th sample. The latent variable $\mathbf{z}_i$ has a single component equal to one and all others to zero, in which z_{ii} is always zero, and if $z_{ij}=1$, in $\mathcal{F}$, the jth sample is the nearest neighbor of the ith sample. According to Goldberger et al. [17], the conditional probability that the jth sample is the nearest neighbor of the i th sample in $\mathcal{F}$, given $z_{ij}=1$, can be defined as

$$p(\mathcal{X}|z_{ij}=1,\mathbf{L})\propto\exp(-\|\mathbf{L}^T(\mathbf{x}_j-\mathbf{x}_i)\|_2^2/\sigma),$$

where σ is the kernel width. In addition, the prior probabilities of the components of the latent variable $\mathbf{z}_i$ is assumed identical, so that $p(z_{ij}=1|\mathbf{L})=1/(N-1)$.[1] Thus, we have

$$p(\mathcal{X},z_{ij}=1|\mathbf{L})=p(\mathcal{X}|z_{ij}=1,\mathbf{L})p(z_{ij}=1|\mathbf{L})\propto\frac{1}{N-1}\exp(-\|\mathbf{L}^T(\mathbf{x}_j-\mathbf{x}_i)\|_2^2/\sigma),$$

and obtain

$$p(z_{ij}=1|\mathcal{X},\mathbf{L})=\frac{\exp(-\|\mathbf{L}^T(\mathbf{x}_j-\mathbf{x}_i)\|_2^2/\sigma)}{\sum_{k=1,k\neq i}^{N}\exp(-\|\mathbf{L}^T(\mathbf{x}_k-\mathbf{x}_i)\|_2^2/\sigma)}$$

based on the fact that $p(z_{ij}=1|\mathcal{X},\mathbf{L})=p(\mathcal{X},z_{ij}=1|\mathbf{L})/p(\mathcal{X}|\mathbf{L})$. Let $C(\mathbf{x}_i)$ denote the set of the samples belonging to the class of $\mathbf{x}_i$. Thus, the probability $p_h(i|\mathbf{L})$ that the nearest-neighbor classification of the i th sample is correct with given $\mathbf{L}$ can be calculated by

$$p_h(i|\mathbf{L})=\sum_{j\in C(\mathbf{x}_i)-\{i\}}p(z_{ij}=1|\mathcal{X},\mathbf{L}).$$

[1] This is because the ith component of $\mathbf{z}_i$ is always zero.

In addition, the expected leave-one-out classification rate of the nearest-neighbor classifier conditioned on $\mathbf{L}$ is proportional to the sum of $p_h(i|\mathbf{L})$ for all samples:

$$f(\mathbf{L}) = \sum_{i=1}^{N} p_h(i|\mathbf{L}). \tag{1}$$

As described in Goldberger et al. [17], $\mathbf{L}$ can be obtained by directly maximizing the function $f(\mathbf{L})$ or a log likelihood function of $p_h(i|\mathbf{L})$, $i=1,\ldots,N$ defined as follows:

$$g(\mathbf{L}) = \sum_{i=1}^{N} \log(p_h(i|\mathbf{L})). \tag{2}$$

The functions f and g are indeed the objective functions of NCA.

Similarly, the probability $p_m(i|\mathbf{L})$ that the ith sample is miss-classified with given $\mathbf{L}$ can be obtained by

$$p_m(i|\mathbf{L}) = \sum_{j \notin C(\mathbf{x}_i)} p(z_{ij}=1|\mathcal{X},\mathbf{L}).$$

Thus, another objective function, which is also in terms of the miss-classification rate, may be defined by the sum of the log-ratio of $p_h(i|\mathbf{L})$ to $p_m(i|\mathbf{L})$ as follows:

$$\operatorname*{argmax}_{\mathbf{L}} \sum_{i=1}^{N} \log\left(\frac{p_h(i|\mathbf{L})}{p_m(i|\mathbf{L})}\right) = \operatorname*{argmax}_{\mathbf{L}} (g(\mathbf{L}) - \overline{g}(\mathbf{L})), \tag{3}$$

where $\overline{g}(\mathbf{L})$ is a log likelihood function of $p_m(i|\mathbf{L})$, $i=1,\ldots,N$ defined as

$$\overline{g}(\mathbf{L}) = \sum_{i=1}^{N} \log(p_m(i|\mathbf{L})). \tag{4}$$

Thus, the objective function (3) is for seeking $\mathbf{L}$ maximizing the gap between g and $\overline{g}$. In the sequent section, the objective function of I-RELIEF will be shown closely related to the function (3).

2.2. Relations between I-RELIEF and the leave-one-out nearest-neighbor classification rate

The function $g(\mathbf{L})$ can be expressed as

$$g(\mathbf{L}) = \sum_{i=1}^{N} \log\left(\sum_{j \in C(\mathbf{x}_i)-\{i\}} \frac{p(z_{ij}=1|\mathcal{X},\mathbf{L}) \dfrac{q(z_{ij})}{\sum_{j \in C(\mathbf{x}_i)-\{i\}} q(z_{ij})}}{\dfrac{q(z_{ij})}{\sum_{j \in C(\mathbf{x}_i)-\{i\}} q(z_{ij})}} \right),$$

and a lower bound for $g(\mathbf{L})$ can be obtained by Jensen's inequality [25] as follows:

$$\begin{aligned} g(\mathbf{L}) &\geq \sum_{i=1}^{N} \sum_{j \in C(\mathbf{x}_i)-\{i\}} \beta_{ij} \{\log(p(z_{ij}=1|\mathcal{X},\mathbf{L})) - \log(\beta_{ij})\}, \\ &= \sum_{i=1}^{N} \sum_{j \in C(\mathbf{x}_i)-\{i\}} \beta_{ij} \{\log(p(\mathcal{X},z_{ij}=1|\mathbf{L})) - \log(\beta_{ij})\} - N \log(p(\mathcal{X}|\mathbf{L})) \triangleq \mathcal{H}(\mathbf{L}), \end{aligned}$$

where $\beta_{ij} = q(z_{ij}) / \sum_{j \in C(\mathbf{x}_i)-\{i\}} q(z_{ij})$ with $j \in C(\mathbf{x}_i)-\{i\}$ and $\mathcal{H}(\mathbf{L})$ represents this lower bound. By using the Lagrange multiplier [26], the β_{ij} maximizing this lower bound with $\mathbf{L}$ fixed can be determined by

$$\beta_{ij} = \frac{p(z_{ij}=1|\mathcal{X},\mathbf{L})}{\sum_{k \in C(\mathbf{x}_i)-\{i\}} p(z_{ik}=1|\mathcal{X},\mathbf{L})} = \frac{\exp(-\|\mathbf{L}^T(\mathbf{x}_j-\mathbf{x}_i)\|_2^2/\sigma)}{\sum_{k \in C(\mathbf{x}_i)-\{i\}} \exp(-\|\mathbf{L}^T(\mathbf{x}_k-\mathbf{x}_i)\|_2^2/\sigma)}. \tag{5}$$

Similarly, a lower bound for $\overline{g}(\mathbf{L})$, denoted by $\mathcal{M}(\mathbf{L})$, can be obtained as follows:

$$\overline{g}(\mathbf{L}) \geq \sum_{i=1}^{N} \sum_{j \notin C(\mathbf{x}_i)} \alpha_{ij} \{\log(p(\mathcal{X},z_{ij}=1|\mathbf{L})) - \log(\alpha_{ij})\} - N \log(p(\mathcal{X}|\mathbf{L})) \triangleq \mathcal{M}(\mathbf{L})$$

where α_{ij} with $j \notin C(\mathbf{x}_i)$ analogy to β_{ij} for $\mathcal{H}(\mathbf{L})$ is as follows:

$$\alpha_{ij} = \frac{p(z_{ij}=1|\mathcal{X},\mathbf{L})}{\sum_{k \notin C(\mathbf{x}_i)} p(z_{ik}=1|\mathcal{X},\mathbf{L})} = \frac{\exp(-\|\mathbf{L}^T(\mathbf{x}_j-\mathbf{x}_i)\|_2^2/\sigma)}{\sum_{k \notin C(\mathbf{x}_i)} \exp(-\|\mathbf{L}^T(\mathbf{x}_k-\mathbf{x}_i)\|_2^2/\sigma)}. \tag{6}$$

With fixed α's and β's, maximization of the difference between $\mathcal{H}(\mathbf{L})$ and $\mathcal{M}(\mathbf{L})$ with respect to $\mathbf{L}$:

$$\begin{aligned} \max_{\mathbf{L}} \mathcal{H}(\mathbf{L}) - \mathcal{M}(\mathbf{L}) = \max_{\mathbf{L}} \sum_{i=1}^{N} \Bigg\{ & \sum_{j \in C(\mathbf{x}_i)-\{i\}} \beta_{ij} \log(p(\mathcal{X},z_{ij}=1|\mathbf{L})) \\ & - \sum_{j \notin C(\mathbf{x}_i)} \alpha_{ij} \log(p(\mathcal{X},z_{ij}=1|\mathbf{L})) \Bigg\} \end{aligned}$$

is equivalent to maximization of the function $Q(\mathbf{L})$ defined as follows:

$$Q(\mathbf{L}) = \mathtt{tr}\left(\mathbf{L}^T \sum_{i=1}^{N} \left\{ \sum_{j \notin C(\mathbf{x}_i)} \alpha_{ij} \mathbf{x}_{ji} \mathbf{x}_{ji}^T - \sum_{j \in C(\mathbf{x}_i)-\{i\}} \beta_{ij} \mathbf{x}_{ji} \mathbf{x}_{ji}^T \right\} \mathbf{L} \right),$$

where $\mathbf{x}_{ji} = \mathbf{x}_j - \mathbf{x}_i$ and $\mathtt{tr}(.)$ is the trace of a matrix. This is because $\sum_{i=1}^{N} \sum_{j \notin C(\mathbf{x}_i)} \alpha_{ij} \log(\alpha_{ij})$ and $\sum_{i=1}^{N} \sum_{j \in C(\mathbf{x}_i)-\{i\}} \beta_{ij} \log(\beta_{ij})$ are constant, and the term $N \log(p(\mathcal{X}|\mathbf{L}))$ is canceled.

Let $\mathbf{\Lambda}$ be a diagonal matrix with nonnegative elements. Now, it can be seen that $Q(\mathbf{\Lambda}^{1/2})$ is equal to a version of the objective function of I-RELIEF, which is defined in terms of the L_2-norm and has no outlier terms.

The I-RELIEF algorithm finds a proper $\mathbf{\Lambda}$ through maximizing the gap between the lower bounds of g and $\overline{g}$ based on the framework of the EM algorithm as follows.

- *The E-step*: Calculate α's and β's by Eqs. (6) and (5), respectively.
- *The M-step*: Maximize $Q(\mathbf{\Lambda}^{\frac{1}{2}})$, with respect to $\mathbf{\Lambda}$, subject to that the diagonal elements of $\mathbf{\Lambda}$ are nonnegative and $\mathtt{tr}(\mathbf{\Lambda}^2) = 1$.

The E-Step rises up the lower bounds $\mathcal{H}$ and $\mathcal{M}$ of g and $\overline{g}$, and the M-step maximizes the gap between $\mathcal{M}$ and $\mathcal{H}$. The function $Q(\mathbf{L})$ with respect to a general matrix $\mathbf{L}$ is indeed a generalization of the objective function of I-RELIEF. When $\mathcal{H}$ and $\mathcal{M}$ are tight lower bounds, maximizing the gap between $\mathcal{H}$ and $\mathcal{M}$ is approximately maximizing the function (3). As a result, a link between the generalized I-RELIEF and the leave-one-out nearest-neighbor classification rate has been established.

3. The generalized I-RELIEF

3.1. The formulation of the generalized I-RELIEF

The objective function $Q(\mathbf{L})$ can also be expressed in terms of the sample matrix $\mathbf{X}=[\mathbf{x}_1 \ \ldots \mathbf{x}_N]$ as follows:

$$Q(\mathbf{L}) = \mathtt{tr}\left(\mathbf{L}^T \mathbf{X} \sum_{i=1}^{N} \left\{ \sum_{j \notin C(\mathbf{x}_i)} \alpha_{ij} \mathbf{e}_{ji} \mathbf{e}_{ji}^T - \sum_{j \in C(\mathbf{x}_i)-\{i\}} \beta_{ij} \mathbf{e}_{ji} \mathbf{e}_{ji}^T \right\} \mathbf{X}^T \mathbf{L} \right),$$

where $\mathbf{e}_{ji}$ is an N-dimensional vector. The components of $\mathbf{e}_{ji}$ are all zero except for the i th and j th components which are -1 and 1, respectively. Furthermore, by defining two proximity matrices $\mathbf{A}=[a_{ij}]$ and $\mathbf{B}=[b_{ij}]$ as

$$a_{ij} = \begin{cases} 0 & \text{if } i=j; \\ \alpha_{ij}+\alpha_{ji} & \text{if } j \notin C(\mathbf{x}_i); \\ 0 & \text{elsewhere,} \end{cases}$$

and

$$b_{ij}=\begin{cases}0 & \text{if } i=j;\\ \beta_{ij}+\beta_{ji} & \text{if } j\in C(\mathbf{x}_i)-\{i\};\\ 0 & \text{elsewhere,}\end{cases}$$

a compact form of $Q(\mathbf{L})$ can be obtained as follows:

$$Q(\mathbf{L})=\mathtt{tr}(\mathbf{L}^T\mathbf{X}(\mathbf{L}_A-\mathbf{L}_B)\mathbf{X}^T\mathbf{L}), \tag{7}$$

where $\mathbf{L}_A$ and $\mathbf{L}_B$ are the Laplacian matrices of matrices $\mathbf{A}$ and $\mathbf{B}$, respectively. The Laplacian matrices $\mathbf{L}_A$ and $\mathbf{L}_B$ are defined as

$$\mathbf{L}_A=\mathbf{D}_A-\mathbf{A}, \tag{8}$$

$$\mathbf{L}_B=\mathbf{D}_B-\mathbf{B}, \tag{9}$$

where $\mathbf{D}_A=[d_{A;ij}]$ and $\mathbf{D}_B=[d_{B;ij}]$ are diagonal matrices with $d_{A;ii}=\sum_{j=1}^{N}a_{ij}$ and $d_{B;ii}=\sum_{j=1}^{N}b_{ij}$.

In this study, the transformation matrix $\mathbf{L}$ for the generalized I-RELIEF is obtained by solving the constrained optimization problem:

$$\mathrm{P}:\ \arg\max_{\mathbf{L}} Q(\mathbf{L}),\ \text{subject to } \mathbf{L}^T\mathbf{X}\mathbf{L}_B\mathbf{X}^T\mathbf{L}=\mathbf{I}_0,$$

where $\mathbf{I}_0$ denotes a diagonal matrix with elements either 0 or 1. The targets of the constraint on $\mathbf{L}$ are threefold: (1) to put bounds to the maximum value of $Q(\mathbf{L})$; (2) to establish a proper metric system in $\mathcal{F}$; and (3) to make the resultant distance function independent to the direction, following which the expected distance from a sample to the nearest hit is longer than that to the nearest miss. Since the Laplacian matrix $\mathbf{L}_B$ is positive semi-definite [27], $\mathbf{X}\mathbf{L}_B\mathbf{X}^T$ is positive semi-definite as well. For simplicity, $\mathbf{X}\mathbf{L}_B\mathbf{X}^T$ is assumed positive definite, which can be achieved through some regularization techniques if it is necessary. Thus, the matrix $\mathbf{L}$ optimizing the problem P can be obtained by the following theorem.

Theorem 1. *The solution of the constrained optimization problem P can be obtained by*

$$\mathbf{L}=\mathbf{\Pi}\mathbf{\Lambda}, \tag{10}$$

where $\mathbf{\Pi}$ *is an eigenvector matrix of the matrix pair* $(\mathbf{X}\mathbf{L}_A\mathbf{X}^T,\mathbf{X}\mathbf{L}_B\mathbf{X}^T)$ *with* $\mathbf{\Pi}^T\mathbf{X}\mathbf{L}_B\mathbf{X}^T\mathbf{\Pi}=\mathbf{I}$, *and* $\mathbf{\Lambda}=[\lambda_{ij}]$ *is a diagonal matrix with*

$$\lambda_{ii}=\begin{cases}1 & \text{if the ith generalized eigenvalue is larger than 1;}\\ 0 & \text{elsewhere.}\end{cases}$$

In the following, a proof is presented by using the majorization relationship between the main diagonal elements and the eigenvalues of a Hermitian matrix. First, the majorization relationship is stated as follows. For detailed proofs, the reader may refer to [28].

Definition 1. The n-dimensional vector $\mathbf{a}=[a_i]$ is said to majorize the n-dimensional vector $\mathbf{b}=[b_i]$ if $\sum_{j=1}^{k}a_{p_j}\geq\sum_{j=1}^{k}b_{q_j}$ for all $k=1,\ldots,n$, with equality for $k=n$, where $a_{p_1}\leq\cdots\leq a_{p_n}$ and $b_{q_1}\leq\cdots\leq b_{q_n}$ are the elements of $\mathbf{a}$ and $\mathbf{b}$ arranged in nondecreasing order.

Theorem 2. *Let* $\mathbf{A}$ *be a Hermitian matrix. The vector of the main diagonal elements of* $\mathbf{A}$ *majorizes the vector of the eigenvalues of* $\mathbf{A}$.

Next, since $\mathbf{X}\mathbf{L}_B\mathbf{X}^T$ is positive definite and $\mathbf{L}^T\mathbf{X}\mathbf{L}_B\mathbf{X}^T\mathbf{L}$ should be equal to $\mathbf{I}_0$, the columns of $\mathbf{L}$ should be either zero or orthonormal to each other with respect to $\mathbf{X}\mathbf{L}_B\mathbf{X}^T$. Thus, $\mathbf{L}$ can be decomposed as $\mathbf{L}=\mathbf{\Pi}\mathbf{\Gamma}\tilde{\mathbf{\Lambda}}$, where $\mathbf{\Gamma}$ is an orthogonal matrix, and $\tilde{\mathbf{\Lambda}}=[\tilde{\lambda}_{ij}]$ is a diagonal matrix with diagonal elements either 0 or 1. By substituting $\mathbf{\Pi}\mathbf{\Gamma}\tilde{\mathbf{\Lambda}}$ for $\mathbf{L}$, the objective function $Q(\mathbf{L})$ becomes

$$Q(\mathbf{\Pi}\mathbf{\Gamma}\tilde{\mathbf{\Lambda}})=\mathtt{tr}(\tilde{\mathbf{\Lambda}}^T\mathbf{\Gamma}^T(\mathbf{\Theta}-\mathbf{I})\mathbf{\Gamma}\tilde{\mathbf{\Lambda}}),$$

where $\mathbf{\Theta}=[\theta_{ij}]$ denotes $\mathbf{\Pi}^T\mathbf{X}\mathbf{L}_A\mathbf{X}^T\mathbf{\Pi}$. The matrix $\mathbf{\Theta}$ is the eigenvalue matrix, a diagonal matrix, of the matrix pair $(\mathbf{X}\mathbf{L}_A\mathbf{X}^T,\mathbf{X}\mathbf{L}_B\mathbf{X}^T)$ because $\mathbf{\Pi}$ is an eigenvector matrix of this matrix pair with $\mathbf{\Pi}^T\mathbf{X}\mathbf{L}_B\mathbf{X}^T\mathbf{\Pi}=\mathbf{I}$.

Let $\mathbf{H}=[h_{ij}]$ denote $\mathbf{\Gamma}^T(\mathbf{\Theta}-\mathbf{I})\mathbf{\Gamma}$. The function $Q(\mathbf{\Pi}\mathbf{\Gamma}\tilde{\mathbf{\Lambda}})$ becomes

$$Q(\mathbf{\Pi}\mathbf{\Gamma}\tilde{\mathbf{\Lambda}})=\sum_{i=1}^{n}\tilde{\lambda}_{ii}h_{ii}.$$

For given $\mathbf{\Gamma}$, it can be seen that by determining $\tilde{\mathbf{\Lambda}}$ by

$$\tilde{\lambda}_{ii}=\begin{cases}1 & \text{if } h_{ii}>0;\\ 0 & \text{elsewhere,}\end{cases} \tag{11}$$

$Q(\mathbf{\Pi}\mathbf{\Gamma}\tilde{\mathbf{\Lambda}})$ has the maximum value $\sum_{i\in\{k|h_{kk}>0,k=1,\ldots,n\}}h_{ii}$. The maximum value of $Q(\mathbf{\Pi}\mathbf{\Gamma}\tilde{\mathbf{\Lambda}})$ can be shown not greater than $Q(\mathbf{\Pi}\mathbf{\Lambda})$ by using Theorem 2 as follows.

First, the main diagonal elements of $\mathbf{H}$ and $\mathbf{\Theta}-\mathbf{I}$ are arranged in nondecreasing order as $h_{p_1p_1}\leq\cdots\leq h_{p_lp_l}<0\leq\cdots\leq h_{p_np_n}$ and $\theta_{q_1q_1}-1\leq\cdots\leq\theta_{q_{l'}q_{l'}}-1<0\leq\cdots\leq\theta_{r_nr_n}-1$, respectively. Moreover, since $\mathbf{\Gamma}$ is an orthogonal matrix, we have that the main diagonal elements of $\mathbf{\Theta}-\mathbf{I}$ are the eigenvalues of $\mathbf{H}$, and $\mathtt{tr}(\mathbf{H})$ is equal to $\mathtt{tr}(\mathbf{\Theta}-\mathbf{I})$. Thus, we have

$$Q(\mathbf{\Pi}\mathbf{\Gamma}\tilde{\mathbf{\Lambda}})=\sum_{j=l+1}^{n}h_{p_jp_j}=\mathtt{tr}(\mathbf{H})-\sum_{j=1}^{l}h_{p_jp_j}, \tag{12}$$

and

$$Q(\mathbf{\Pi}\mathbf{\Lambda})=\sum_{j=l'+1}^{n}(\theta_{q_jq_j}-1)=\mathtt{tr}(\mathbf{\Theta}-\mathbf{I})-\sum_{j=1}^{l'}(\theta_{q_jq_j}-1). \tag{13}$$

In addition, we have

$$\sum_{j=1}^{l}h_{p_jp_j}\geq\sum_{j=1}^{l}(\theta_{q_jq_j}-1)\geq\sum_{j=1}^{l'}(\theta_{q_jq_j}-1) \tag{14}$$

based on two facts. The first inequality is established because the main diagonal elements of $\mathbf{H}$ majorizes those of $\mathbf{\Theta}-\mathbf{I}$ according to Theorem 2. The second inequality is held because among $\theta_{q_jq_j}-1$, $j=1,\ldots,n$, only the first l' terms are negative. Thus, combining relationships (12)–(14), and $\mathtt{tr}(\mathbf{H})=\mathtt{tr}(\mathbf{\Theta}-\mathbf{I})$ all together, we can obtain $Q(\mathbf{\Pi}\mathbf{\Gamma}\tilde{\mathbf{\Lambda}})\leq Q(\mathbf{\Pi}\mathbf{\Lambda})$, and have proven Theorem 1.

Now, the steps of the generalized I-RELIEF algorithm are as follows.

Step 1: Assign a proper initial value to $\mathbf{L}$. For example,
- (a) if the components of the sample vector are not badly scaled, the initial value of $\mathbf{L}$ may be $\mathbf{I}$;
- (b) otherwise, the initial value of $\mathbf{L}$ may be $\mathbf{W}^{-1/2}$ where $\mathbf{W}$ is the intra-class scatter matrix.

Step 2: Repeat the following steps until the estimate of $\mathbf{L}$ has no significant changes or the number of iterations exceeds a pre-specified number.
- *The E-step*: Form matrices $\mathbf{L}_A$ and $\mathbf{L}_B$ by Eqs. (8) and (9) with respect to the current estimate of $\mathbf{L}$.
- *The M-step*: Solve a generalized eigenvalue problem of the matrix pair $(\mathbf{X}\mathbf{L}_A\mathbf{X}^T,\mathbf{X}\mathbf{L}_B\mathbf{X}^T)$. Determine $\mathbf{L}$ by Theorem 1.

As shown in [6], I-RELIEF can be regarded as a contraction operator. By the Banach fixed point theorem, it is also shown that I-RELIEF would converge toward a fix point if a proper kernel width σ is selected. The same argument can also be applied to the generalized I-RELIEF algorithm. The reader may refer to [6] for detailed proofs. However, this convergence property does not ensure that the gap between $\overline{g}$ and g is always enlarged for every iteration. Despite this negative fact, the generalized I-RELIEF can

produce an effective distance function due to its relationship to graph embedding [24], which will be illustrated in the following.

3.2. Discussions

3.2.1. Determination of kernel widths

In [29], a variety of approaches to determine the kernel width are introduced. It shows that determining the kernel width via the k-nearest neighbors is a simple and effective way. Here, the kernel width σ is determined by the average of the squared Euclidean distances from the samples to their Kth nearest neighbors: $\sigma = (1/N)\sum_{i=1}^{N} \|\mathbf{L}^T(\mathbf{x}_{K;i} - \mathbf{x}_i)\|_2^2$, where $\mathbf{L}^T\mathbf{x}_{K;i}$ denotes the K-th nearest neighbor of $\mathbf{L}^T\mathbf{x}_i$ in $\mathcal{F}$. At the beginning of every iteration of Step 2, the kernel width is re-estimated according to the current estimate of $\mathbf{L}$. If no prior knowledge about K is available, seven can be selected for K according to the suggestion of Zelnik-Manor and Perona [30].

3.2.2. Dimensionality reduction

For regularizing and speeding up the subsequent machine learning algorithm, the dimensionality of the metric space $\mathcal{F}$ is often desired to be less than n. In this study, two approaches to dimensionality reduction are proposed for the generalized I-RELIEF as follows.

(1) *The direct approach with an acceleration option*: An additional step can be added as Step 3 of the generalized I-RELIEF algorithm for dimensionality reduction. This step is to abandon the components of the mapped sample corresponding to small eigenvalues of the matrix pair $(\mathbf{X}\mathbf{L}_A\mathbf{X}^T, \mathbf{X}\mathbf{L}_B\mathbf{X}^T)$. Besides, the E-step for calculating α's and β's may also be based on reduced mapped samples. Although just an approximation, the computation of matrices $\mathbf{L}_A$ and $\mathbf{L}_B$ is accelerated. It should be noticed that the M-step should be still based on the original sample matrix $\mathbf{X}$.

(2) *The linearization approach*: The direct approach may not be applied on higher-dimensional data (i.e. $n \gg N$) because two $n \times n$ matrices should be formed explicitly. Suppose that the columns of the transformation matrix are linear combinations of the sample vectors; that is, $\mathbf{L} = \mathbf{X}\tilde{\mathbf{L}}$. By substituting $\mathbf{X}\tilde{\mathbf{L}}$ for $\mathbf{L}$, the compact form (7) of $Q(\mathbf{L})$ becomes

$$Q(\tilde{\mathbf{L}}) = \mathrm{tr}(\tilde{\mathbf{L}}^T\mathbf{G}(\mathbf{L}_A - \mathbf{L}_B)\mathbf{G}^T\tilde{\mathbf{L}}), \tag{15}$$

where $\mathbf{G} = \mathbf{X}^T\mathbf{X}$ is the Gram matrix of the sample vectors. Thus, $\tilde{\mathbf{L}}$ is the eigenvector matrix of the matrix pair $(\mathbf{G}\mathbf{L}_A\mathbf{G}^T, \mathbf{G}\mathbf{L}_B\mathbf{G}^T)$ such that $\tilde{\mathbf{L}}^T\mathbf{G}\mathbf{L}_B\mathbf{G}^T\tilde{\mathbf{L}} = \mathbf{I}_0$. Then, the policy of the direct approach can be applied.

3.2.3. The relationship between the generalized I-RELIEF and graph embedding

Graph embedding [24] is a general framework for the algorithm of dimensionality reduction. This framework formulates a dimensionality reduction problem as finding a transformation matrix to maximize the trace-ratio or trace-difference between the constraint matrix of a penalty graph and the Laplacian matrix of an intrinsic graph. The objective function of the generalized I-RELIEF is also under this framework. This fact can be observed from Eq. (7), the compact form of $Q(\mathbf{L})$, which is a trace-difference between two Laplacian matrices for the interclass and the intraclass similarity. In fact, the marginal Fisher discriminant analysis (MFDA) [24,31] also has a similar mathematical formulation. However, there exist some difference between the generalized I-RELIEF and MFDA because they are aimed for different problems.

- First, the proximity matrices for the interclass and intraclass similarity are defined in different manners although both are based on the k-nearest neighbors.
- Second, MFDA relies on the relationship defined by the k-nearest neighbors in the original measurement space, whereas the generalized I-RELIEF updates the nearest neighbors for every iteration.
- Third, they have different constraints on the transformation matrix.

Despite the difference between the generalized I-RELIEF and MFDA, the relationship between them provides an aspect for graph embedding from the expected leave-one-out nearest-neighbor classification rate, and a justification of the generalized I-RELIEF for supervised dimensionality reduction.

Table 2
Descriptions of the eleven data sets, where n, c, and N denote the dimensionality of the data, the number of classes,and the number of samples, respectively.

Data set	n	c	N
Iris	4	3	150
Wine	13	3	178
Balance	4	3	625
Ionosphere	34	2	351
Crings	3	5	500
USPS	256	10	9298
Spambase	57	2	4601
Prostate	12 000	2	102
GCM	16 063	14	190
YALE	5120	38	2414
AR	5120	126	1638

4. Experimental results

In the following experiments, the proposed algorithm, referred as to GI-RELIEF, was compared with a variety of approaches of distance metric learning, such as PCA, RCA [32], DANN [16], Fisher's discriminant analysis (FDA) [33], LDM [18],[2] local Fisher's discriminant analysis (LFDA) [29][3], LMNN [20][4], the batch version of I-RELIEF [6], MFDA [24,31], and NCA [17][5]. Here, we mainly focused on the supervised approach and implemented them in the MATLAB programming language. The experiments were conducted on a computer which has two Intel® Xeron 2.0 GHz CPUs and two gigabytes of RAM, and runs the Windows Server 2003 operating system.

Table 2 lists eleven data sets for performance evaluation. These eleven data sets include

- one synthesized data set: Crings;
- five data sets from UCI machine learning repository [34]: Iris, Wine, Balance, Ionosphere, and Spambase;
- one data set of handwritten digit images: USPS[6];
- two gene expression data sets: Prostate [35] and GCM [36]; and
- two face data sets: YALE [37], and AR [38].

[2] Software available at http://www.cs.cmu.edu/~liuy/ldm_scripts_2.zip
[3] http://sugiyama-www.cs.titech.ac.jp/~sugi/software/LFDA/index.html
[4] http://www.weinbergerweb.net/Downloads/LMNN.html
[5] http://www.cs.berkeley.edu/~fowlkes/software/nca/
[6] http://www-stat-class.stanford.edu/~tibs/ElemStatLearn/data.html

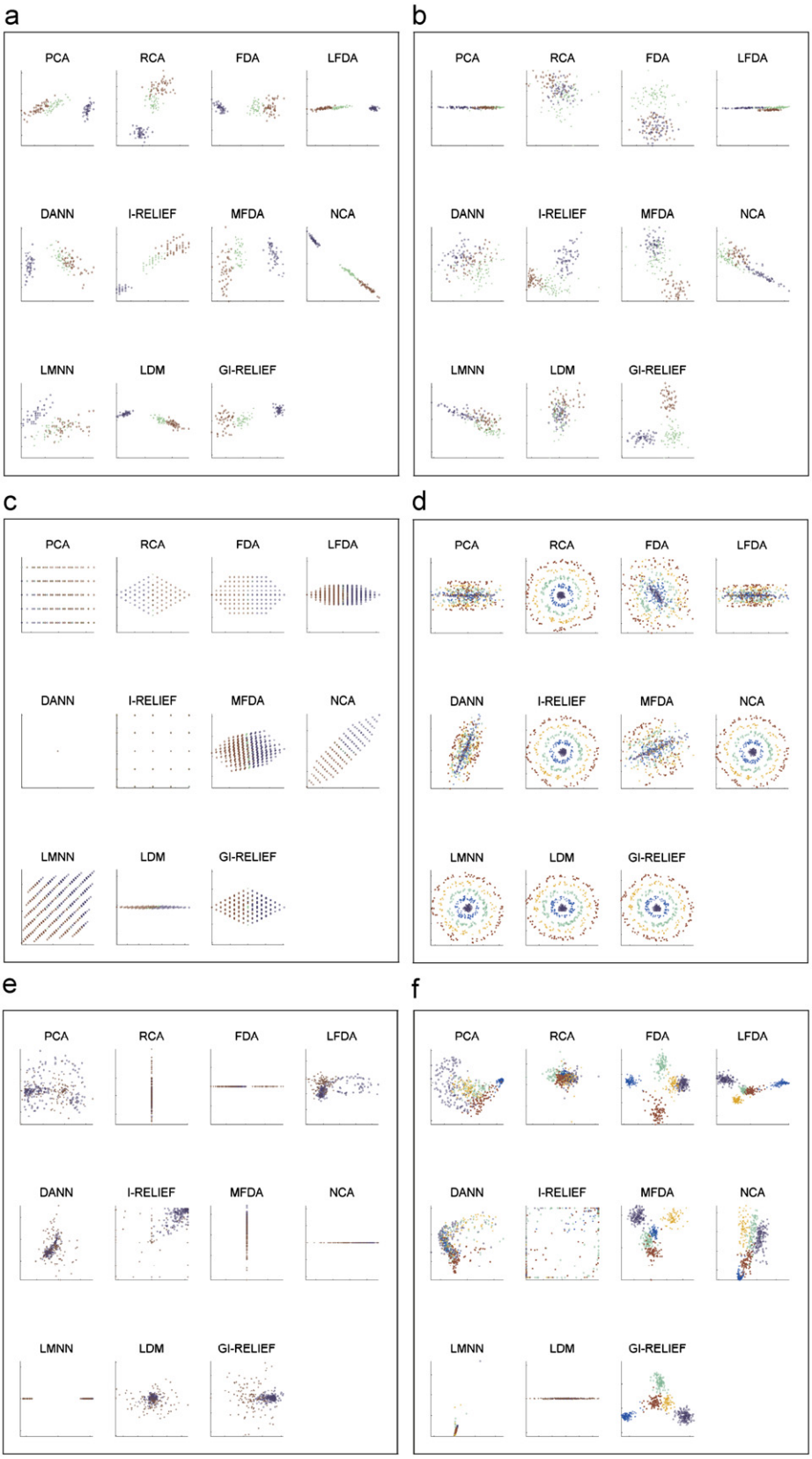

Fig. 1. 2-D visualization of data sets: (a) Iris, (b) Wine, (c) Balance, (d) Crings, (e) Ionosphere and (f) a subset of USPS.

These data sets include five lower-dimensional and small-scale data sets, namely, Iris, Wine, Balance, Ionosphere, and Crings; two large-scale data sets, namely, USPS and Spambase; and four higher-dimensional data sets, namely, Prostate, GCM, YALE, and AR. The synthesized data set Crings consists of three-dimensional sample vectors of five classes. These classes have equal prior probabilities.

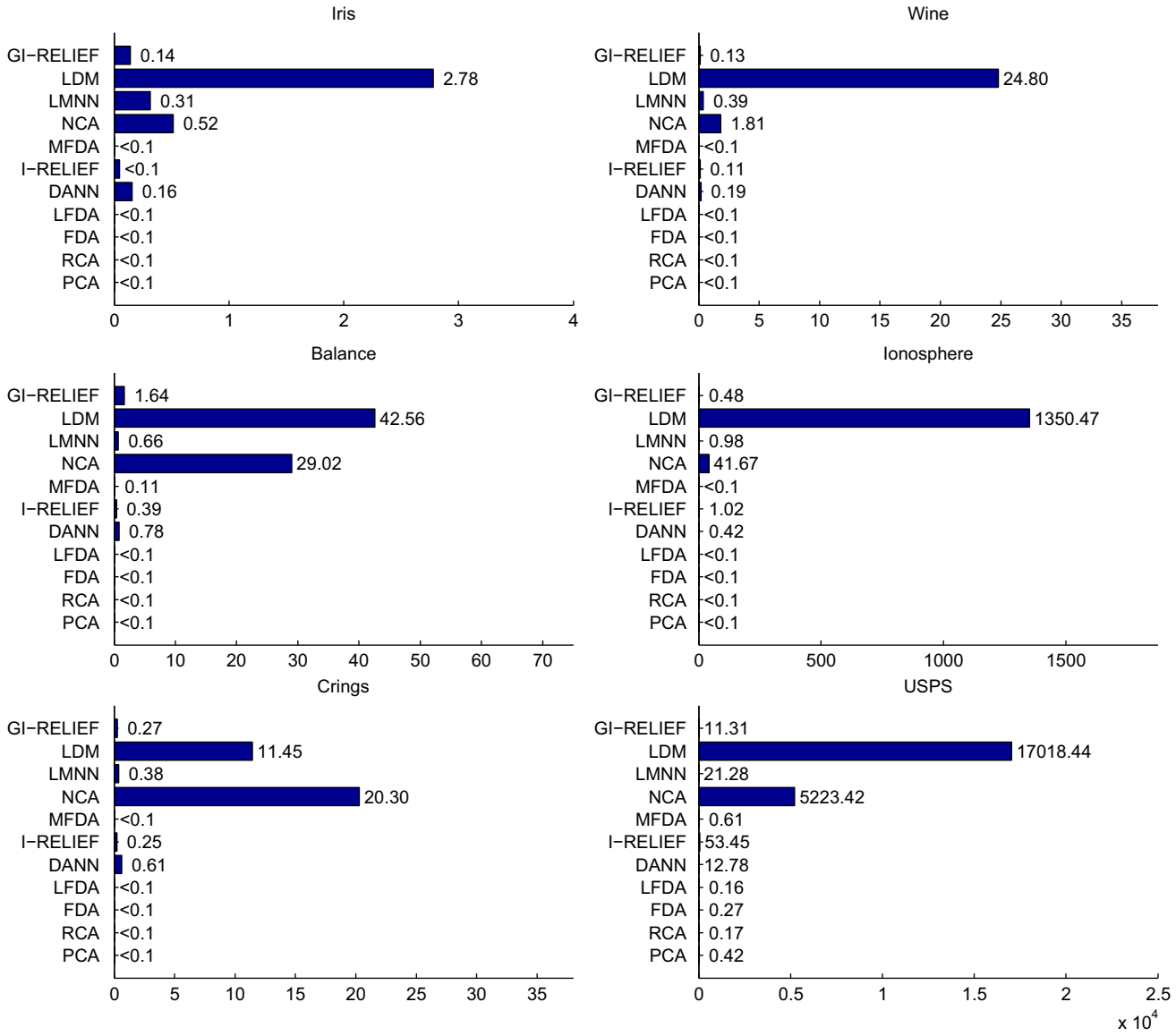

Fig. 2. Computing time for 2-D visualization of data sets Iris, Wine, Balance, Crings, Ionosphere, and a subset of USPS, where the *x*-axis is the computing time in seconds.

In the first two dimensions, these classes are distributed over five co-centric ring-shaped regions such that each region contains a single class. The third dimensions of the samples are Gaussian random noises. The magnitude of the Gaussian random noise is not small so that the main principal components of the samples cover the third dimension. The face data sets contain grayscale face images, which are manually aligned with respect to the two eyes, and resized to 64×80 for the experiments.

4.1. Data visualization

Fig. 1 shows the 2-D visualization of six data sets, namely, Iris, Wine, Balance, Crings, Ionosphere and a subset of the first five classes of the data set USPS. Here, the direct approach without the acceleration option was used for the generalized I-RELIEF for dimensionality reduction. Due to the out of memory error, LDM was only tested against a subset of the first two classes of the USPS data set. It can be seen that these classes are separated much better by the generalized I-RELIEF than the other methods. Fig. 2 reveals that the computing time of the generalized I-RELIEF is also satisfactory.

Some experiments on the convergence of the generalized I-RELIEF were conducted. Fig. 3(a) shows that the generalized I-RELIEF with the proposed kernel width can converge toward a fix point for three data sets but oscillates for the other three. Fig. 3(b) shows that with a larger kernel width, the generalized I-RELIEF can converge for all six data sets. This result justifies that the generalized I-RELIEF inherits the convergence property of I-RELIEF. Although the generalized I-RELIEF with the proposed kernel width may not converge for some data sets, we found that the induced distance functions are still effective as shown in the following experiment.

4.2. Nearest-neighbor classification

The linearization approach for the generalized I-RELIEF was used for the higher-dimensional data sets, while the direct approach without the acceleration option was applied to the others. Since Fig. 2 reveals that LDM and NCA have high computational cost, they were only tested against the five small-scale data sets. Besides, for the higher-dimensional data sets, PCA was applied prior to the six methods: RCA, DANN, MFDA, LFDA, DANN, and LMNN for reducing

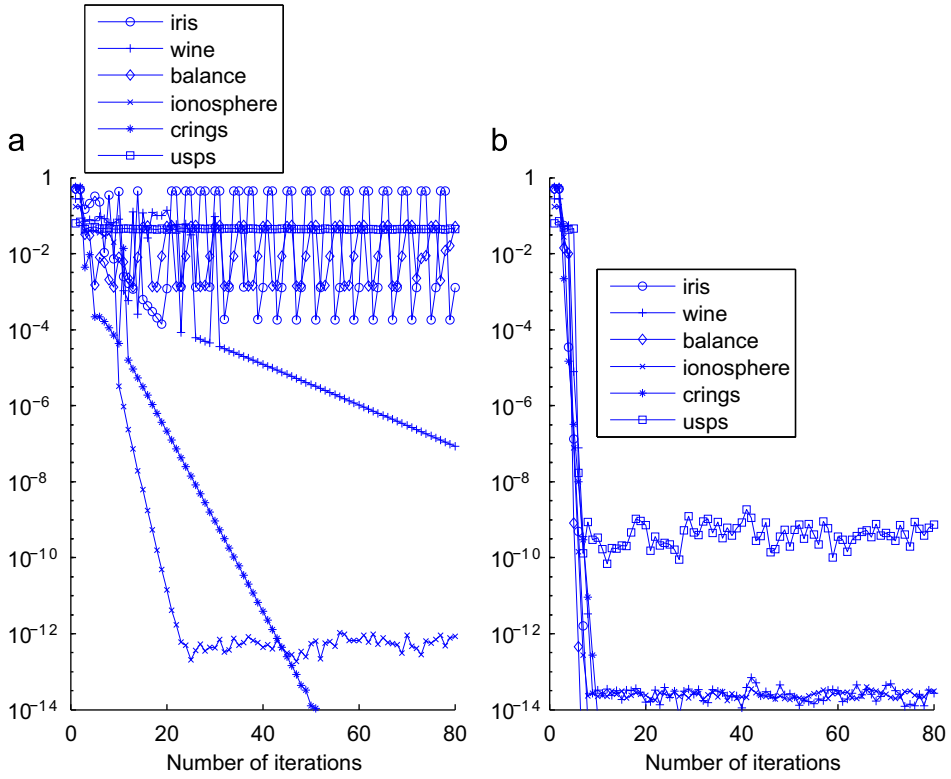

Fig. 3. The convergence analysis of the generalized I-RELIEF, where the y-axis is $\|\mathbf{L}_{\text{previous}}-\mathbf{L}_{\text{current}}\|_F/n$, (a) is for $\sigma=\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{L}^T(\mathbf{x}_{K;i}-\mathbf{x}_i)\|_2^2$, and (b) is for $\sigma=\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{L}^T(\mathbf{x}_{K;i}-\mathbf{x}_i)\|_2^2+1000$.

Table 3
The parameters for estimating the classification rate.

Data set	Training set size	Test set size	Number of runs	Estimation method
Iris	114	36	50	Holdout
Wine	135	43	50	Holdout
Balance	469	156	50	Holdout
Ionosphere	264	87	50	Holdout
Crings	500	165	50	Holdout
USPS	2328	6970	1	Holdout
Spambase	1151	3450	1	Holdout
Prostate	101	1	102	Leave-one-out
GCM	189	1	190	Leave-one-out
YALE	605	1809	10	Holdout
AR	1386	252	10	Holdout

the dimensionality of the sample vector. It should be noticed that no principal components were abandoned here. Then, these six methods worked on the reduced sample as usual. Finally, the transformation matrices of the six methods were formed by multiplying the transformation matrix of PCA with those induced by the six methods for the reduced sample. The nearest neighbor classifier was used to evaluate the accuracy of the induced distance functions. The holdout method or the leave-one-out method [39] was used to estimate the classification rate. The choice between these two methods is dependent on the scale of the data set. Table 3 shows the parameters for this experiment. For each method to be tested, the components of the mapped vector were sorted by the feature importance in nonincreasing order. The dimensionality of the mapped vector was augmented from one to the theoretical limit of the method with increments of one, and the associated classification rate was recorded and averaged over all runs of this experiment.

Fig. 4 plots the classification rates, where the numbers enclosed in the parentheses are the best average classification rate and the associated dimensionality of the mapped vector, respectively. The first number should be large, while the second should be small because it is related to the effectiveness of the induced transformation matrix. It can be seen that the generalized I-RELIEF is the best or the second best in terms of the classification rate except for the data sets USPS and YALE. In addition, the transformation matrix induced by the generalized I-RELIEF is also effective because the mapped vector of the generalized I-RELIEF often has lower dimensionality than that of the method with a comparable classification rate. Besides, except for the data set USPS, the generalized I-RELIEF is better than I-RELIEF in terms of either the classification rate or the effectiveness of the induced transformation matrix. Fig. 5 shows that the generalized I-RELIEF also has feasible computing time for the large-scale data sets, and the higher-dimensional data sets.

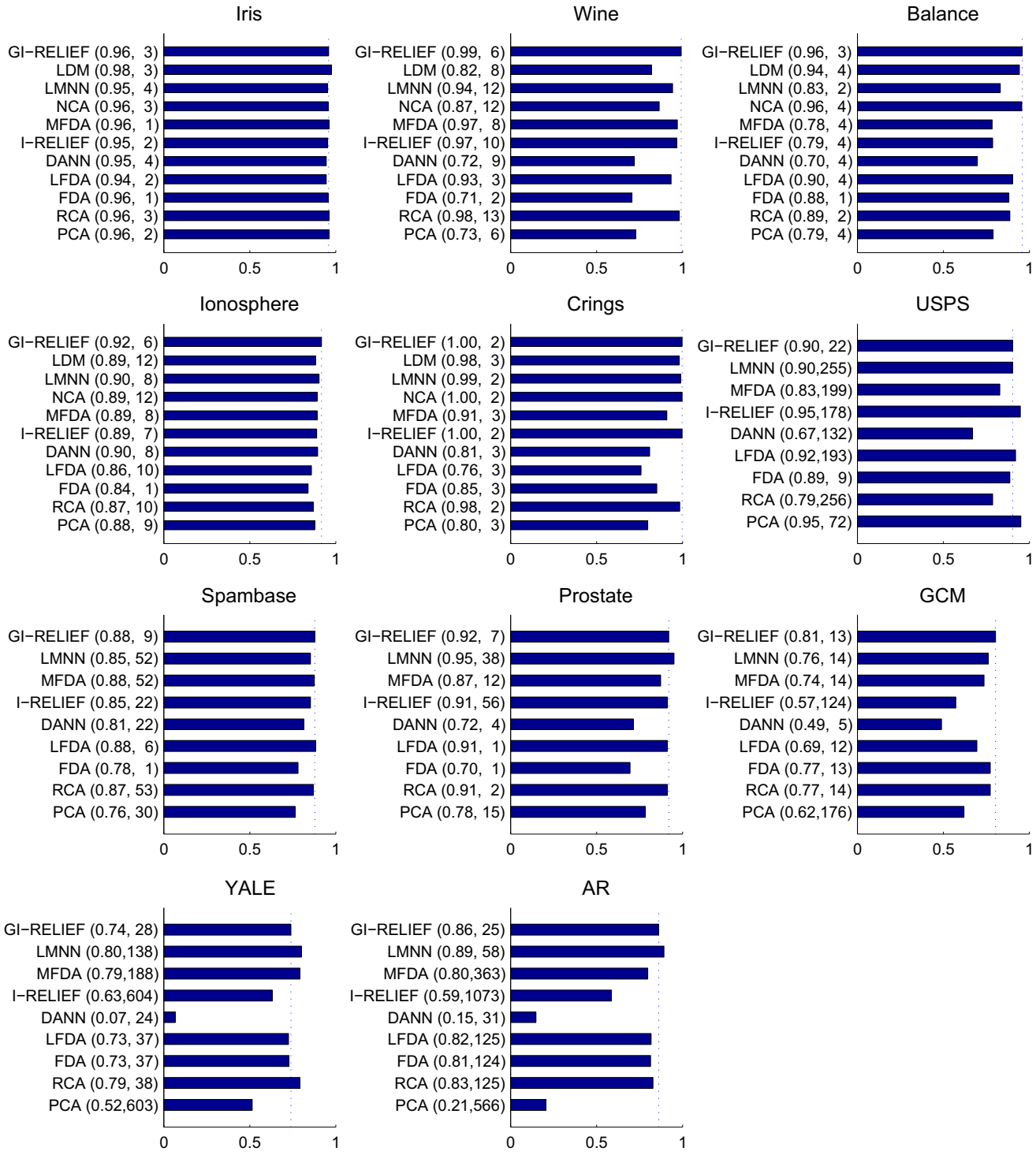

Fig. 4. The classification rates of the methods to be compared with respect to the eleven data sets, where the *x*-axis is the classification rate, the first number in the parentheses is the best average classification rate, and the second is the dimensionality of the mapped vector associated with the best average classification rate.

Overall, among the eleven methods to be compared, the generalized I-RELIEF is the most consistently effective to the eleven test data sets. In addition, these experimental results also confirm that with the close relationships to the nearest-neighbor classification rate and graph embedding, the generalized I-RELIEF is capable of yielding an effective Mahananobis distance function for nearest-neighbor classification.

5. Conclusion

Although designed to weight each feature according to feature relevance, the I-RELIEF algorithm has been shown to have a connection to the expected leave-one-out nearest-neighbor classification rate. This result justifies that the Mahananobis distance function induced by the generalized I-RELIEF is suitable for the nearest-neighbor classification. Since taking the correlations among

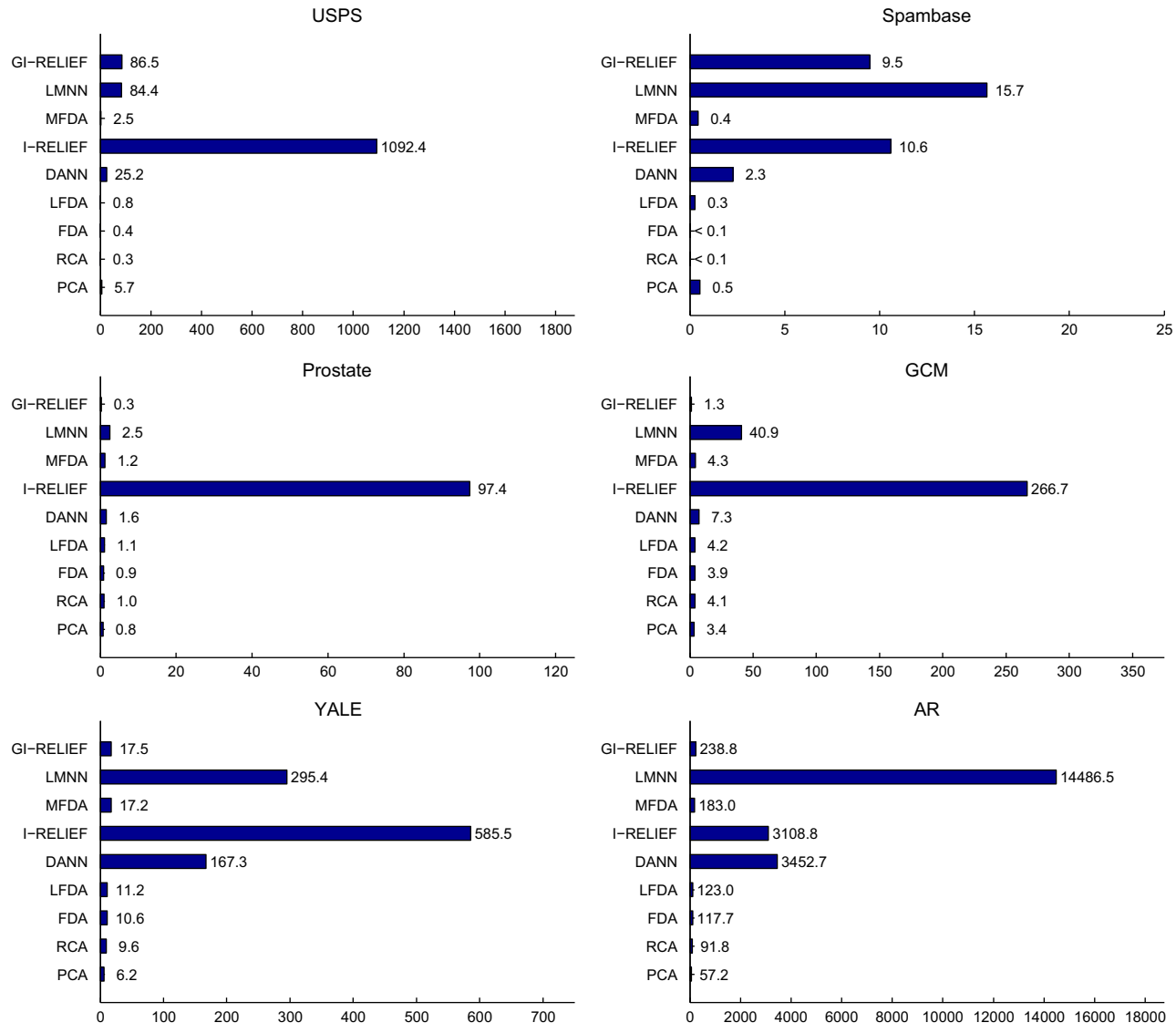

Fig. 5. Average computing time of PCA, RCA, FDA, LFDA, DANN, LMNN, I-RELIEF, MFDA, and GI-RELIEF with respect to data sets USPS, Spambase, Prostate, GCM, YALE, and AR, where the *x*-axis is the computing time in seconds.

features into account, the generalized I-RELIEF is superior to I-RELIEF in the presence of highly correlated features. In addition, it has been pointed out that the objective functions of I-RELIEF and the generalized I-RELIEF are also under the framework of graph embedding. Thus, NCA, I-RELIEF, the generalized I-RELIEF, and graph embedding can be linked together as follows. The distance function induced by NCA is through directly optimizing the expected leave-one-out nearest-neighbor classification rate, whereas those induced by I-RELIEF and the generalized I-RELIEF are by means of the framework of the EM algorithm to optimize a function of the lower bounds of the classification and the miss-classification rate. The objective functions of I-RELIEF and the generalized I-RELIEF are also under the framework of graph embedding. Thus, the generalized I-RELIEF is an algorithm of "repetitive" graph embedding for inducing distance functions, and can do linear dimensionality reduction as well. Since not assuming a parametric model for the underlying sample distribution, the generalized I-RELIEF may be applied when the sample of a class exhibits a multimodal distribution.

Although the proposed kernel width is suitable for a variety of data sets, needing a proper kernel width is one of the possible weakness of the generalized I-RELIEF. This weakness is worth improving. In addition, further investigations on consistently enlarging the gap between functions g and $\overline{g}$ for every iteration are also needed.

References

[1] K. Kira, L.A. Rendell, A practical approach to feature selection, in: Proceedings of the Ninth International Conference on Machine Learning, 1992, pp. 249–256.
[2] T.G. Dietterich, Machine-learning research—four current directions, AI Magazine 18 (1997) 97–136.
[3] R. Kohavi, G. John, Wrappers for feature selection, Artificial Intelligence 97 (1997) 273–324.
[4] D. Wettschereck, D.W. Aha, T. Mohri, A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms, Artificial Intelligence Review 11 (1997) 273–314.
[5] I. Kononenko, Estimating attributes: analysis and extensions of RELIEF, in: European Conference on Machine Learning, 1994, pp. 171–182.
[6] Y. Sun, Iterative RELIEF for feature weighting: algorithms, theories, and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (6) (2007) 1035–1051.
[7] I. Kononenko, E. Simec, M. Robnik-Sikonja, Overcoming the myopia of inductive learning algorithms with RELIEFF, Applied Intelligence 7 (1997) 39–55.
[8] M. Robnik-Šikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Machine Learning 53 (2003) 23–69.

[9] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society Series B 39 (1) (1977) 1–38.
[10] J. Yang, Y.-P. Li, Orthogonal RELIEF algorithm for feature selection, in: Proceedings of 2006 International Conference on Intelligent Computing, Springer, Berlin/Heidelberg, 2006.
[11] Y. Sun, D. Wu, A RELIEF based feature extraction algorithm, in: Proceedings of the SIAM International Conference on Data Mining, 2008, pp. 188–195.
[12] J. Bins, B. Draper, Feature selection from huge feature sets, in: Proceedings of the Eighth IEEE International Conference on Computer Vision, vol. 2, 2001, pp. 159–165.
[13] R. Flórez-López, Reviewing RELIEF and its extensions: a new approach for estimating attributes considering high-correlated features, in: Proceedings of IEEE International Conference on Data Mining, 2002, pp. 605–608.
[14] L. Yang, R. Jin, Distance metric learning: a comprehensive survey, Technical Report 24, Department of Computer Science and Engineering, Michigan State University, 2006.
[15] K. Fukunaga, T.E. Flick, An optimal global nearest neighbor metric, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI 6 (3) (1984) 314–318.
[16] T. Hastie, R. Tibshirani, Discriminant adaptive nearest neighbor classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (6) (1996) 607–616.
[17] J. Goldberger, S. Roweis, G. Hinton, R. Salakhutdinov, Neighbourhood components analysis, in: L.K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems, vol. 17, MIT Press, Cambridge, MA, 2005, pp. 513–520.
[18] L. Yang, R. Jin, R. Sukthankar, Y. Liu, An efficient algorithm for local distance metric learning, in: Proceedings of the 21st National Conference on Artificial Intelligence, 2006.
[19] E.P. Xing, A.Y. Ng, M.I. Jordan, S. Russell, Distance metric learning with application to clustering with side-information, in: S.T.S. Becker, K. Obermayer (Eds.), Advances in Neural Information Processing Systems, vol. 15, MIT Press, Cambridge, MA, 2003, pp. 505–512.
[20] K. Weinberger, J. Blitzer, L. Saul, Distance metric learning for large margin nearest neighbor classification, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Advances in Neural Information Processing Systems, vol. 18, MIT Press, Cambridge, MA, 2006, pp. 1473–1480.
[21] S. Yan, J. Liu, X. Tang, T.S. Huang, A parameter-free framework for general supervised subspace learning, IEEE Transactions on Information Forensics and Security 2 (1) (2007) 69–76.
[22] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 539–546.
[23] J.V. Davis, B. Kulis, P. Jain, I.S. Dhillon, Information-theoretic metric learning, in: Proceedings of the 24th international conference on Machine learning, 2007, pp. 209–216.
[24] S. Yan, D. Xu, B. Zhang, H.J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1) (2007) 40–51.
[25] T.M. Cover, J.A. Thomas, Elements of Information Theory, John & Wiley Sons, Inc, New York, 1991.
[26] M.S. Bazaraa, H.D. Sherali, C.M. Shetty, Nonlinear Programming Theory and Algorithms, second ed., John & Wiley sons, Inc, New York, 1993.
[27] F.R.K. Chung, Spectral Graph Theory, American Mathematical Society, New York, 1997.
[28] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
[29] M. Sugiyama, Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis, Journal of Machine Learning Research 8 (2007) 1027–1061.
[30] L. Zelnik-Manor, P. Perona, Self-tuning spectral clustering, in: L.K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems, vol. 17, MIT Press, Cambridge, MA, 2005, pp. 1601–1608.
[31] D. Xu, S. Yan, D. Tao, S. Lin, H.J. Zhang, Marginal Fisher analysis and its variants for human gait recognition and content-based image retrieval, IEEE Transactions on Image Processing 16 (11) (2007) 2811–2821.
[32] A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall, Learning distance functions using equivalence relations, in: Proceedings of the 20th International Conference on Machine Learning, 2003, pp. 11–18.
[33] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., John Wiley & Sons, Inc, 2001.
[34] A. Asuncion, D. Newman, UCI machine learning repository, 2007. URL 〈http://www.ics.uci.edu/~mlearn/MLRepository.html〉.
[35] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2002) 203–209.
[36] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander, R. Golub, Multiclass cancer diagnosis using tumor gene expression signatures, Proceedings of the National Academy of Science 98 (26) (2001) 15149–15154.
[37] K. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Transactions Pattern Analysis and Machine Intelligence 27 (5) (2005) 684–698.
[38] A.M. Martinez, R. Benavente, The AR face database, Technical Report 24, CVC, June 1998.
[39] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, Boston, 1990.

About the Author: CHIN-CHUN CHANG received the B.S. degree and the M.S. degree in computer science in 1989 and 1991, respectively, and the Ph.D. degree in computer science in 2000, all from National Chiao Tung University, Hsinchu, Taiwan.

From 2001 to 2002, he was a faculty of the Department of Computer Science and Engineering, Tatung University, Taipei, Taiwan. In 2002, he joined the Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, Taiwan, where he is currently an Assistant Professor. His research interests include computer vision, machine learning, and pattern recognition.