Neurocomputing 133 (2014) 209–221

Contents lists available at ScienceDirect

Neurocomputing

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

Discriminality-driven regularization framework for indefinite kernel machine

Hui Xue [a,b,c,*], Songcan Chen [b]

[a] School of Computer Science and Engineering, Southeast University, Nanjing 210096, PR China
[b] School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, PR China
[c] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, PR China

ARTICLE INFO

Article history:
Received 20 January 2013
Received in revised form
8 September 2013
Accepted 19 November 2013
Communicated by J. Kwok
Available online 10 January 2014

Keywords:
Discriminality-driven regularization
Indefinite kernel machine
Semi-supervised learning
Machine learning

ABSTRACT

Indefinite kernel machines have attracted more and more interests in machine learning due to their better empirical classification performance than the common positive definite kernel machines in many applications. A key to implement effective kernel machine is how to use prior knowledge as sufficiently as possible to guide the appropriate construction of the kernels. However, most of existing indefinite kernel machines actually utilize the knowledge involved in data such as discriminative and structural information insufficiently and thus construct the indefinite kernels empirically. Discriminatively regularized least-squares classification (DRLSC) is a recently-proposed supervised classification method which provides a new discriminality-driven regularizer to encourage the discriminality of the classifier rather than the common smoothness. In this paper, we rigorously validate that the discriminative regularizer actually coincides with the definition on the inner product in Reproducing Kernel Kreĭn Space (RKKS) naturally. As a result, we further present a new discriminality-driven regularization framework for indefinite kernel machine based on the discriminative regularizer. According to the framework, we firstly reintroduce the original DRLSC from the viewpoint of the proper indefinite kernelization rather than the empirical kernel mapping. Then a novel semi-supervised algorithm is proposed in terms of different definition on the regularizer. The experiments on both toy and real-world datasets demonstrate the superiority of the two algorithms.

1. Introduction

In the past decades, kernel machines (classifiers) have been widely developed in machine learning with the successful applications of the most famous being support vector machine (SVM) for classification tasks [1]. Different from many traditional methods that require original vectorial representations of the data, these methods embed the data into a high-dimensional (possibly infinite-dimensional) feature space and then choose a problem specific kernel function instead of the inner products on all pairs of the embeddings, called as kernel trick [1,2]. Following the statistical learning theory, the usual kernel functions are required to be (conditionally) positive definite (PD) satisfied the Mercer's conditions, in order to ensure the existence of a Reproducing Kernel Hilbert Space (RKHS) [3] and further lead to convex formulations for the optimization problems in most kernel machines. Consequently, the corresponding global optimal solutions exist [2].

In practice, however, the requirement of a kernel function to be PD turns out to be a very strict assumption [1]. In many situations, standard PD kernels are not applicable [1,4], such as suboptimal optimization procedures for measure derivation [5], partial projections or occlusions [6], and context-dependent alignments or object comparisons [7]. Furthermore, in other situations, standard kernels can be applied, but non-PD kernels, that is, indefinite kernels, frequently emerge when additional problem specific prior knowledge needs to be incorporated in order to improve the performance of the method [1], e.g., when invariance or robustness is incorporated into the construction of the functions [4,8].

In recent years, indefinite kernel machines have attracted more and more research interests due to their empirically better classification performance than the common PD kernel machines [9–14]. However, with loss of the PD-ness of the kernels, the corresponding optimization problems are more likely not convex any longer which leads to difficulty in optimization. Most of the methods that have been developed for solving such problem can fall into one of three basic categories: spectrum transformation, PD kernel proxy, and indefinite kernel extension.

Spectrum transformation methods generate a PD kernel matrix by transforming the spectrum of the indefinite kernel matrix [2]. Pękalska et al. [8] set the negative eigenvalues to be zeros. Gaepel et al. [15] flipped the sign of the negative eigenvalues. Roth et al. [16] shifted the eigenvalues by a positive constant.

[*] Corresponding author at: School of Computer Science and Engineering, Southeast University, Nanjing 210096, PR China. Tel./fax: +86 255 209 0883.
E-mail addresses: hxue@seu.edu.cn (H. Xue), s.chen@nuaa.edu.cn (S. Chen).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.11.016

PD kernel proxy methods consider the indefinite kernel matrix as the noisy observation of some unknown positive semi-definite one [2]. As a result, the optimization of the indefinite kernel machines boils down to the learning of the kernel matrix, which learns the proxy PD kernels to approximate the indefinite kernels [9]. Luss and d′Aspremont [9] proposed a regularized SVM formulation to learn the support vectors as well as the proxy kernel simultaneously. They quadratically smoothed the non-differentiable objective function and then obtained two algorithms including the projected gradient method and the analytic center cutting plan method [2]. Chen and Ye [2] further reformulated the objective function as a semi-infinite quadratically constrained linear program, which can be solved by an iterative algorithm and converge to a global optimum solution. Ying et al. [17] validated that the objective function is continuously differentiable and its gradient is Lipschitz continuous. They developed Nesterov′s smooth optimization approach for indefinite SVM which achieves an optimal convergence rate.

Indefinite kernel extension methods use the indefinite kernel matrix directly and extend the existing PD kernel machines. Lin and Lin [18] proposed an SMO-type method to find stationary points for the non-convex dual formulation of SVM with a non-PD sigmoid kernel. Haasdonk [1] gave a geometric interpretation of indefinite SVM and then executed the corresponding optimization by minimizing the distances between convex hulls in pseudo-Euclidean spaces. Ong et al. [19] extended the common inner product in the RKHS to the Reproducing Kernel Kreĭn Space (RKKS), so as to the product can be negative associated with the indefinite kernel. They also presented a generalized Representer Theorem for constrained stabilization and proved a generalization bound by computing the Rademacher averages of the indefinite kernel class [19].

All of these methods have shown impressive improvements for the learning algorithms of indefinite kernel machines. However, one difficulty with these algorithms is the construction of the kernels when facing usual classification tasks, especially in the indefinite kernel extension methods. Haasdonk [1] has pointed out that the key in the kernel machines is how to introduce some available problem specific prior knowledge into the kernel functions. In the usual classification, there is no prior knowledge about the problem specific similarity measure generally. Consequently, the knowledge involved in the data themselves becomes vital, such as discriminative and structural information. However, most of the existing indefinite kernel machines utilize such information insufficiently and select the indefinite kernels empirically. If the kernels are not appropriate, the performance of the algorithms even becomes poor.

Discriminatively regularized least-squares classification (DRLSC) [20] is a recently-proposed supervised classifier, which presents that relatively speaking, the discriminality of the classifier is more important than the usual smoothness. It defines a novel discriminative regularizer instead of the usual smoothness regularizer, which can be formulated as the difference of the intra-class compactness and inter-class separability metrics. However, the original DRLSC is artificially restricted to the RKHS, which leads such indefinite discriminative regularizer to violating the PD requirement of a regularizer in the RKHS. Furthermore, the solution of DRLSC does not satisfy the Representer Theorem in the RKHS any longer. Consequently, for its nonlinear version, DRLSC has to use the empirical kernel mapping to explicitly map the samples into the empirical feature space [21].

Though DRLSC empirically performs better than many state-of-the-art regularization methods [20], such as regularization networks [22], radial basis function neural network (RBFNN) [22], SVM, least squares support vector machines [23] and manifold regularization [24], its theoretical foundation is still blank. In this paper, we will establish the foundation based on the indefinite kernel theory and further propose a discriminality-driven regularization framework for indefinite kernel machine. The main contributions of this paper include.

- We rigorously validate that the discriminative regularizer can actually be reformulated as an inner product in the RKKS, which testifies that the regularizer is a legal generalized regularizer according to the formulation of the larger RKKS (than RKHS) and thus the solution of DRLSC satisfies the generalized Representer Theorem in the RKKS.
- Based on the regularizer, we propose the discriminality-driven regularization framework for indefinite kernel machine which provides a feasible strategy to embed the prior knowledge into the construction of the indefinite kernel.
- Under the framework, we redesign the original DRLSC to present a new supervised indefinite kernel machine by using the proper indefinite kernelization instead of the non-intrinsic empirical kernel mapping. The algorithm is named as supervised discriminatively regularized least-squares classifier (SupDR), whose performance will be shown much better than DRLSC in the experiment section.
- We further derive a novel semi-supervised indefinite kernel machine by using different definition on the discriminative regularizer according to the semi-supervised learning (SSL) [25,26] scenarios, termed as semi-supervised discriminatively regularized least-squares classifier (SemiDR). To the best of our knowledge, SemiDR is more likely the first attempt to introduce the indefinite kernel theory into the semi-supervised classification, which embeds the local discriminative structure of the labeled data and the global structure of the unlabeled data simultaneously to limit the complexity of our learner. Systematic experiments demonstrate the effectiveness of SemiDR in real-world applications.

The rest of the paper is organized as follows. Section 2 briefly reviews the algorithm of DRLSC. The discriminality-driven regularization framework and the resulting SupDR are derived in Section 3. Section 4 presents the proposed SemiDR. In Section 5, the experimental analysis is given. Some conclusions are drawn in Section 6.

2. Discriminatively regularized least-squares classification (DRLSC)

DRLSC [20] is a unified discriminative regularization framework for supervised learning, which directly concentrates on the discriminative structure of the data and embeds the underlying class information in a discriminative regularizer

$$\min_{\boldsymbol{f}\in K}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i-\boldsymbol{f}(\boldsymbol{x}_i))^2+R_{disreg}(\boldsymbol{f},\eta)\right\} \tag{1}$$

The discriminative regularizer $R_{disreg}(\boldsymbol{f},\eta)$ has a general definition

$$R_{disreg}(\boldsymbol{f},\eta)=\eta A(\boldsymbol{f})-(1-\eta)B(\boldsymbol{f}), \tag{2}$$

where $A(\boldsymbol{f})$ and $B(\boldsymbol{f})$ are the metrics which measure the intra-class compactness and inter-class separability of the outputs respectively. η is the regularization parameter that regulates the relative significance of the intra-class compactness versus the inter-class separability, $0\le\eta\le 1$.

The common definitions on $A(\boldsymbol{f})$ and $B(\boldsymbol{f})$ are the generalized variances in statistics [20]. That is,

$$A(\boldsymbol{f}) = S_w = \sum_{t=1}^{c}\sum_{i=1}^{n_t} \|\boldsymbol{f}(\boldsymbol{x}_i^{(t)}) - \frac{1}{n_t}\sum_{j=1}^{n_t} f(\boldsymbol{x}_j^{(t)})\|^2, \tag{3}$$

where n_t is the number of the samples $\boldsymbol{x}_i^{(t)}$ belonging to class t, $t = 1, \cdots, c$.

$$B(\boldsymbol{f}) = S_b = \sum_{t=1}^{c} n_t \|\frac{1}{n_t}\sum_{i=1}^{n_t} \boldsymbol{f}(\boldsymbol{x}_i^{(t)}) - \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{f}(\boldsymbol{x}_j)\|^2 \tag{4}$$

For the moment, we focus on the linear classifier, i.e., $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$.

Then $A(\boldsymbol{f})$ and $B(\boldsymbol{f})$ can be further formulated as

$$\begin{aligned} A(\boldsymbol{f}) &= \sum_{t=1}^{c}\sum_{i=1}^{n_t} \|\boldsymbol{w}^T\boldsymbol{x}_i^{(t)} - \frac{1}{n_t}\sum_{j=1}^{n_t} \boldsymbol{w}^T\boldsymbol{x}_j^{(t)}\|^2 \\ &= \sum_{t=1}^{c}\sum_{i=1}^{n_t} \boldsymbol{w}^T(\boldsymbol{x}_i^{(t)} - \overline{\boldsymbol{x}}^{(t)})(\boldsymbol{x}_i^{(t)} - \overline{\boldsymbol{x}}^{(t)})^T\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}, \end{aligned} \tag{5}$$

$$\begin{aligned} B(\boldsymbol{f}) &= \sum_{t=1}^{c} n_t \|\frac{1}{n_t}\sum_{i=1}^{n_t} \boldsymbol{w}^T\boldsymbol{x}_i^{(t)} - \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{w}^T\boldsymbol{x}_j\|^2 \\ &= \boldsymbol{w}^T\sum_{t=1}^{c} n_t(\overline{\boldsymbol{x}}^{(t)} - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}^{(t)} - \overline{\boldsymbol{x}})^T\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}, \end{aligned} \tag{6}$$

where $\overline{\boldsymbol{x}}^{(t)}$ denotes the mean of the samples in class t, and $\overline{\boldsymbol{x}}$ is the mean of all samples.

Obviously, $\boldsymbol{S}_w$ and $\boldsymbol{S}_b$ are equivalent to the within-class and between-class scatter matrices in Fisher discriminant analysis (FDA) [27]. Actually, the definitions on $A(\boldsymbol{f})$ and $B(\boldsymbol{f})$ are various. For example, any improvements for FDA can be straightforwardly combined in the regularizer to update $\boldsymbol{S}_w$ and $\boldsymbol{S}_b$.

DRLSC further embeds the local discriminative structure of the data into the regularizer, inspired by some FDA-based supervised manifold dimensionality reduction methods [28]. Concretely, for each sample $\boldsymbol{x}_i$, DRLSC firstly divides the nearest k neighborhood $ne(\boldsymbol{x}_i)$ into two non-overlapping subsets

$$ne_w(\boldsymbol{x}_i) = \left\{ \boldsymbol{x}_i^{(j)} \quad \textit{if} \quad \boldsymbol{x}_i^{(j)} \quad \textit{and} \quad \boldsymbol{x}_i \quad \textit{belong to same class}, \quad 1 \le j \le k \right\},$$

$$ne_b(\boldsymbol{x}_i) = \left\{ \boldsymbol{x}_i^{(j)} \quad \textit{if} \quad \boldsymbol{x}_i^{(j)} \quad \textit{and} \quad \boldsymbol{x}_i \quad \textit{belong to different classes}, \quad 1 \le j \le k \right\}.$$

Then DRLSC defines the intra-class graph G_w and the inter-class graph G_b respectively

$$W_{w,ij} = \begin{cases} 1 & \textit{if } \boldsymbol{x}_j \in ne_w(\boldsymbol{x}_i) \textit{ or } \boldsymbol{x}_i \in ne_w(\boldsymbol{x}_j) \\ 0 & \textit{otherwise} \end{cases},$$

$$W_{b,ij} = \begin{cases} 1 & \textit{if } \boldsymbol{x}_j \in ne_b(\boldsymbol{x}_i) \textit{ or } \boldsymbol{x}_i \in ne_b(\boldsymbol{x}_j) \\ 0 & \textit{otherwise} \end{cases}.$$

Consequently, $A^{(SUP)}(\boldsymbol{f})$ can be redefined to capture the intra-class compactness from the intra-class graph G_w

$$A^{(SUP)}(\boldsymbol{f}) = S_w^{(SUP)} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{w,ij}\|\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{f}(\boldsymbol{x}_j)\|^2. \tag{7}$$

Likewise, $B^{(SUP)}(\boldsymbol{f})$ characterizes the inter-class separability from the inter-class graph G_b

$$B^{(SUP)}(\boldsymbol{f}) = S_b^{(SUP)} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{b,ij}\|\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{f}(\boldsymbol{x}_j)\|^2. \tag{8}$$

In terms of the linear classifier, $A^{(SUP)}(\boldsymbol{f})$ and $B^{(SUP)}(\boldsymbol{f})$ can be further deduced to

$$\begin{aligned} A^{(SUP)}(\boldsymbol{f}) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{w,ij}[\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{f}(\boldsymbol{x}_j)]^2 \\ &= \boldsymbol{w}^T\boldsymbol{X}(\boldsymbol{D}_w - \boldsymbol{W}_w)\boldsymbol{X}^T\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{X}\boldsymbol{L}_w\boldsymbol{X}^T\boldsymbol{w}, \end{aligned} \tag{9}$$

where $\boldsymbol{D}_w$ is a diagonal matrix and its entries $\boldsymbol{D}_{w,ii} = \Sigma_j W_{w,ij}$. $\boldsymbol{L}_w = \boldsymbol{D}_w - \boldsymbol{W}_w$ is the Laplacian matrix of G_w.

$$\begin{aligned} B^{(SUP)}(\boldsymbol{f}) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{W}_{b,ij}[\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{f}(\boldsymbol{x}_j)]^2 \\ &= \boldsymbol{w}^T\boldsymbol{X}(\boldsymbol{D}_b - \boldsymbol{W}_b)\boldsymbol{X}^T\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{X}\boldsymbol{L}_b\boldsymbol{X}^T\boldsymbol{w}, \end{aligned} \tag{10}$$

where $\boldsymbol{D}_b$ is also a diagonal matrix and its entries $\boldsymbol{D}_{b,ii} = \Sigma_j W_{b,ij}$. $\boldsymbol{L}_b = \boldsymbol{D}_b - \boldsymbol{W}_b$ is the Laplacian matrix of G_b.

The final optimization function of DRLSC can be formulated as

$$\min_{\boldsymbol{f} \in \boldsymbol{K}} \left\{ \frac{1}{n}\sum_{i=1}^{n} (y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2 + \boldsymbol{w}^T\boldsymbol{X}[\eta\boldsymbol{L}_w - (1-\eta)\boldsymbol{L}_b]\boldsymbol{X}^T\boldsymbol{w} \right\}. \tag{11}$$

Due to the indefinite property of the discriminative regularizer, the solution of DRLSC does not satisfy the Representer Theorem in the RKHS. As a result, DRLSC has to execute its kernelization by using the explicitly empirical kernel mapping [21], whose dimension has been fixed beforehand equally to the number of the training samples. The kernelization is non-intrinsic and the desired performance of the corresponding empirical kernel DRLSC is difficult to be guaranteed.

3. Discriminality-driven regularization framework for indefinite kernel machine

Indefinite kernel theory is founded on the RKKS which is larger than the traditional RKHS. The key difference between the RKKS and RKHS is that the inner products are indefinite [19]. More formally,

Definition 1. (**Kreĭn space**) [19] An inner product space $(\tilde{\boldsymbol{K}}, \langle\bullet,\bullet\rangle_{\tilde{\boldsymbol{K}}})$ is a Kreĭn space if there exist two Hilbert spaces $\boldsymbol{H}_+$, $\boldsymbol{H}_-$ spanning $\tilde{\boldsymbol{K}}$ such that

(1) All $\boldsymbol{f} \in \tilde{\boldsymbol{K}}$ can be decomposed into $\boldsymbol{f}=\boldsymbol{f}_+ + \boldsymbol{f}_-$, where $\boldsymbol{f}_+ \in \boldsymbol{H}_+$ and $\boldsymbol{f}_- \in \boldsymbol{H}_-$;
(2) $\forall \boldsymbol{f}, \boldsymbol{g} \in \tilde{\boldsymbol{K}}$, $\langle\boldsymbol{f},\boldsymbol{g}\rangle_{\tilde{\boldsymbol{K}}} = \langle\boldsymbol{f}_+,\boldsymbol{g}_+\rangle_{\boldsymbol{H}_+} - \langle\boldsymbol{f}_-,\boldsymbol{g}_-\rangle_{\boldsymbol{H}_-}$.

Furthermore, the solution to the problem of minimizing a regularized risk functional in RKKS still admits a similar representation in terms of an expansion over the training samples to RKHS.

Theorem 1. (***Representer Theorem***) *[19] Let $\tilde{\boldsymbol{K}}$ be an RKKS with kernel $\tilde{K}$. Denote by $V(\boldsymbol{f}, \boldsymbol{X})$ a continuous convex loss functional depending on $\boldsymbol{f} \in \tilde{\boldsymbol{K}}$ only via its evaluations $\boldsymbol{f}(\boldsymbol{x}_i)$ with $\boldsymbol{x}_i \in X$, let $\Omega(\langle \boldsymbol{f}\boldsymbol{f}\rangle)$ be a continuous stabilizer with strictly monotonic $\Omega{:}\boldsymbol{R} \to \boldsymbol{R}$ and let $C\{\boldsymbol{f}, X\}$ be a continuous functional imposing a set of constraints on $\boldsymbol{f}$, that is $C: \tilde{\boldsymbol{K}} \times X \to \boldsymbol{R}$. Then if the optimization problem*

$$\begin{aligned} &\mathop{stabilize}_{\boldsymbol{f} \in \tilde{\boldsymbol{K}}} V(\boldsymbol{f}, X) + \Omega(\langle\boldsymbol{f},\boldsymbol{f}\rangle_{\tilde{\boldsymbol{K}}}) \\ &\text{s.t. } C\{\boldsymbol{f}, X\} \le d, \end{aligned}$$

has a saddle point $\boldsymbol{f}^*$, it admits the expansion

$$\boldsymbol{f}^* = \sum_{i=1}^{n} \alpha_i\tilde{K}(\boldsymbol{x}_i, \bullet),$$

where $\boldsymbol{x}_i \in X$ and $\alpha_i \in \boldsymbol{R}$.

In this section, we will justify that the discriminative regularizer can be formulated as an inner product in the RKKS and the corresponding optimization solution naturally satisfies the generalized Representer Theorem. Consequently, the indefinite kernel theory actually establishes the validity of the regularizer. Furthermore, the regularizer also offers a feasible strategy to fuse more prior knowledge into the indefinite kernel. In terms of the inter-complementarities between the indefinite kernel theory and DRLSC, we further deduce a discriminality-driven regularization framework for indefinite kernel machine.

3.1. Indefinite kernel analysis for discriminality-driven regularization

Firstly, we present the indefinite kernel analysis for the discriminative regularizer, and then give the common discriminality-driven regularization framework which directly embeds classification specific prior knowledge into the construction of the kernel.

Proposition 1. *The discriminative regularizer can be formulated as an inner product in the RKKS, that is,*

$$R_{disreg}(\boldsymbol{f},\eta)=\langle \boldsymbol{f},\boldsymbol{f}\rangle_{\tilde{K}_{disreg}} \tag{12}$$

where $\tilde{\boldsymbol{K}}_{disreg}$ denotes the induced Kreĭn Space by the regularizer.

Proof. Recall that

$$R_{disreg}(\boldsymbol{f},\eta)=\eta A(\boldsymbol{f})-(1-\eta)B(\boldsymbol{f})=\boldsymbol{w}^T[\eta\boldsymbol{S}_w-(1-\eta)\boldsymbol{S}_b]\boldsymbol{w}.$$

Decompose the joint matrix $\eta\boldsymbol{S}_w-(1-\eta)\boldsymbol{S}_b$ into

$$\eta\boldsymbol{S}_w-(1-\eta)\boldsymbol{S}_b=\boldsymbol{XLX}^T \tag{13}$$

Perform the eigenvalue decomposition for $\boldsymbol{L}$

$$\boldsymbol{L}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T.$$

Then we have

$$\eta\boldsymbol{S}_w-(1-\eta)\boldsymbol{S}_b=\boldsymbol{XU}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{X}^T \tag{14}$$

So

$$R_{disreg}(\boldsymbol{f},\eta)=\boldsymbol{w}^T\boldsymbol{XU}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{X}^T\boldsymbol{w}=\boldsymbol{f}^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{f}$$

$$=\boldsymbol{f}^T\left(\sum_{\lambda_i>0}\lambda_i\boldsymbol{u}_i\boldsymbol{u}_i^T+\sum_{\lambda_j<0}\lambda_j\boldsymbol{u}_j\boldsymbol{u}_j^T\right)\boldsymbol{f} \tag{15}$$

Let $\boldsymbol{\Gamma}_+=\boldsymbol{U}_+\boldsymbol{\Lambda}_+\boldsymbol{U}_+^T, \boldsymbol{\Gamma}_-=\boldsymbol{U}_-\boldsymbol{\Lambda}_-\boldsymbol{U}_-^T$, obviously,

$$\boldsymbol{\Gamma}_+^T\boldsymbol{\Gamma}_-=\boldsymbol{U}_+\boldsymbol{\Lambda}_+\boldsymbol{U}_+^T\boldsymbol{U}_-\boldsymbol{\Lambda}_-\boldsymbol{U}_-^T=0,$$

that is, $\boldsymbol{\Gamma}_+$ and $\boldsymbol{\Gamma}_-$ are orthogonal.

Decompose $\boldsymbol{f}=\boldsymbol{f}_++\boldsymbol{f}_-$, where $\boldsymbol{f}_+\in$ *Hilbert*$(\boldsymbol{\Gamma}_+)$, $\boldsymbol{f}_-\in$ *Hilbert* $(-\boldsymbol{\Gamma}_-)$, then

$$\begin{aligned}R_{disreg}(\boldsymbol{f},\eta)&=\boldsymbol{f}^T[\boldsymbol{\Gamma}_+-(-\boldsymbol{\Gamma}_-)]\boldsymbol{f}=(\boldsymbol{f}_++\boldsymbol{f}_-)^T\boldsymbol{\Gamma}_+(\boldsymbol{f}_++\boldsymbol{f}_-)\\&\quad-(\boldsymbol{f}_++\boldsymbol{f}_-)^T(-\boldsymbol{\Gamma}_-)(\boldsymbol{f}_++\boldsymbol{f}_-)\\&=\boldsymbol{f}_+^T\boldsymbol{\Gamma}_+\boldsymbol{f}_+-\boldsymbol{f}_-^T(-\boldsymbol{\Gamma}_-)\boldsymbol{f}_-.\end{aligned} \tag{16}$$

Let

$$\boldsymbol{f}_+^T\boldsymbol{\Gamma}_+\boldsymbol{f}_+=\langle\boldsymbol{f}_+,\boldsymbol{f}_+\rangle_{H_+},\boldsymbol{f}_-^T(-\boldsymbol{\Gamma}_-)\boldsymbol{f}_-=\langle\boldsymbol{f}_-,\boldsymbol{f}_-\rangle_{H_-}. \tag{17}$$

We can formulate the discriminative regularizer as

$$R_{disreg}(\boldsymbol{f},\eta)=\langle\boldsymbol{f}_+,\boldsymbol{f}_+\rangle_{H_+}-\langle\boldsymbol{f}_-,\boldsymbol{f}_-\rangle_{H_-}=\langle\boldsymbol{f},\boldsymbol{f}\rangle_{\tilde{K}_{disreg}}.$$

Consequently, the optimization problem in the discriminality-driven regularization framework can be formulated as

$$\min_{\boldsymbol{f}\in\tilde{K}}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i-\boldsymbol{f}(x_i))^2+\langle\boldsymbol{f},\boldsymbol{f}\rangle_{\tilde{K}_{disreg}}\right\} \tag{18}$$

Proposition 2. *The solution $\boldsymbol{f}^*$ of the discriminality-driven regularization framework admits an expansion*

$$\boldsymbol{f}^*=\sum_{i=1}^{n}\alpha_i^*\tilde{K}_{disreg}(\boldsymbol{x}_i,\bullet) \tag{19}$$

in terms of the indefinite kernel function in the RKKS, where the coefficient

$$\boldsymbol{\alpha}^*=(\hat{\boldsymbol{K}}+n\boldsymbol{I})^+\boldsymbol{Y} \tag{20}$$

$\boldsymbol{\alpha}^*=[\alpha_1^*,\alpha_2^*,\cdots,\alpha_n^*]^T\cdot\hat{\boldsymbol{K}}$ is the $n\times n$ kernel matrix, $\hat{K}_{i,j}=\tilde{K}_{disreg}(\boldsymbol{x}_i,\boldsymbol{x}_j)$. $\boldsymbol{Y}$ is the label vector, $\boldsymbol{Y}=[y_1,y_2,\cdots,y_n]^T$.

Proof. In the optimization problem (18), we have

$$V(\boldsymbol{f},X)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\boldsymbol{f}(\boldsymbol{x}_i))^2,$$

$$\Omega(\langle\boldsymbol{f},\boldsymbol{f}\rangle_{\tilde{K}_{disreg}})=\langle\boldsymbol{f},\boldsymbol{f}\rangle_{\tilde{K}_{disreg}}.$$

So, by the Representer Theorem 1, the solution $\boldsymbol{f}^*$ of (18) has the following form

$$\boldsymbol{f}^*=\sum_{i=1}^{n}\alpha_i\tilde{K}_{disreg}(\boldsymbol{x}_i,\bullet)=\sum_{i=1}^{n}\alpha_i[K_+(\boldsymbol{x}_i,\bullet)-K_-(\boldsymbol{x}_i,\bullet)]. \tag{21}$$

Substituting this form in (18), we arrive at the following differentiable objective function of the coefficient $\boldsymbol{\alpha}$

$$\arg\min 1/n(\boldsymbol{Y}-\hat{\boldsymbol{K}}\boldsymbol{\alpha})^T(\boldsymbol{Y}-\hat{\boldsymbol{K}}\boldsymbol{\alpha})+\boldsymbol{\alpha}^T\hat{\boldsymbol{K}}\boldsymbol{\alpha} \tag{22}$$

The derivative w.r.t to $\boldsymbol{\alpha}$ of the objective function vanishes at the minimizer

$$1/n(-\hat{\boldsymbol{K}})(\boldsymbol{Y}-\hat{\boldsymbol{K}}\boldsymbol{\alpha}^*)+\hat{\boldsymbol{K}}\boldsymbol{\alpha}^*=0, \tag{23}$$

which leads to the following solution

$$\boldsymbol{\alpha}^*=(\hat{\boldsymbol{K}}+n\boldsymbol{I})^+\boldsymbol{Y}.$$

From the two propositions, it can be obviously seen that the discriminality-driven regularization framework naturally lies in the RKKS induced from the discriminative regularizer $R_{disreg}(\boldsymbol{f},\eta)$ itself. Different from many indefinite kernel machines that artificially use the direct indefinite kernels instead of the PD kernels, the discriminality-driven regularization applies the PD kernels K_+ and K_- defined in the two RKHSs $\boldsymbol{H}_+$ and $\boldsymbol{H}_-$ respectively to construct an indefinite kernel following the RKKS formulation. Especially, classification specific prior knowledge involved in the $\boldsymbol{S}_w$ and $\boldsymbol{S}_b$ in $R_{disreg}(\boldsymbol{f},\eta)$ is actually fused into the construction of the kernels by the particular weighted inner products (17) in the $\boldsymbol{H}_+$ and $\boldsymbol{H}_-$. The corresponding kernelization is finally boiled down to the general smoothing problem [19], which is quite simple and mathematically tractable.

3.2. Supervised discriminatively regularized least-squares classifier (SupDR)

Based on the framework, we anew characterize DRLSC by the more proper indefinite kernelization than the original empirical kernel mapping, and term the corresponding algorithm as SupDR. Though its objective function is the same as DRLSC, SupDR actually is a new kernel machine from the viewpoint of the different kernelization.

Following the formulation (11) of DRLSC, SupDR firstly performs the eigenvalue decomposition of $\eta\boldsymbol{L}_w-(1-\eta)\boldsymbol{L}_b$ to form the two RKHSs $\boldsymbol{H}_+$ and $\boldsymbol{H}_-$, and then constructs the PD kernels K_+ and K_- in $\boldsymbol{H}_+$ and $\boldsymbol{H}_-$ respectively to form the indefinite kernel $\tilde{K}_{disreg}$.

Table 1
Pseudo-code for SupDR.

```
Input: The samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$;
    The number $k$ of the nearest neighbors of $\boldsymbol{x}_i$;
    The regularization parameter $\eta$.
Output: The coefficient $\boldsymbol{\alpha}^*$.
for $i = 1, \cdots, n$
$\boldsymbol{x}_i^{(k)} \leftarrow k$th nearest neighbor of $\boldsymbol{x}_i$ among $\{\boldsymbol{x}_j\}_{j=1}^n$;
end
for $i = 1, \cdots, n$
    for $j = 1, \cdots, k$
        if $y_i = y_j$
            $W_{w,ij} \leftarrow 1$;
        else
            $W_{b,ij} \leftarrow 1$;
        end
    end
end
$\boldsymbol{D}_{w,ii} = \sum_j \boldsymbol{W}_{w,ij}$; $\boldsymbol{D}_{b,ii} = \sum_j \boldsymbol{W}_{b,ij}$; $\boldsymbol{L}_w = \boldsymbol{D}_w - \boldsymbol{W}_w$; $\boldsymbol{L}_b = \boldsymbol{D}_b - \boldsymbol{W}_b$
$\boldsymbol{U\Lambda U}^T \leftarrow$ Eigenvalue decomposition of $\eta \boldsymbol{L}_w - (1-\eta)\boldsymbol{L}_b$;
$\boldsymbol{\Gamma}_+ \leftarrow \boldsymbol{U}_+ \boldsymbol{\Lambda}_+ \boldsymbol{U}_+^T$; $\boldsymbol{\Gamma}_- \leftarrow \boldsymbol{U}_- \boldsymbol{\Lambda}_- \boldsymbol{U}_-^T$
Construct the kernel functions $K_+$ and $K_-$ in $\boldsymbol{H}_+$ and $\boldsymbol{H}_-$ respectively;
$\tilde{K}_{disreg} \leftarrow K_+ - K_-$; $\hat{K}_{i,j} \leftarrow \tilde{K}_{disreg}(\boldsymbol{x}_i, \boldsymbol{x}_j)$; $\boldsymbol{Y} \leftarrow [y_1, y_2, \cdots, y_n]^T$;
$\boldsymbol{\alpha}^* = (\hat{\boldsymbol{K}} + n\boldsymbol{I})^+ \boldsymbol{Y}$.
```

Based on the generalized Representer Theorem 1, the solution of SupDR can finally be obtained in terms of (19). The corresponding pseudo-code for SupDR is shown in Table 1.

One issue in the kernelization is the construction of the PD kernels K_+ and K_-. For convenience, we select the kernel functions to be linear kernel or Gaussian kernel in the following experiments. Actually, any Mercer kernels, such as polynomial kernel, exponential kernel, spline kernel, and their generating kernels can all be used as their candidates, which more likely lead to various combinations of the indefinite kernels. It can also further be viewed as a preliminary attempt of the multiple kernel learning [29] without the need of nonnegativity constraint for the combined coefficients, which deserves our future study.

4. Semi-supervised discriminatively regularized least-squares classifier (SemiDR)

In the semi-supervised classification scenarios, the labeled data are generally few but the unlabeled data are abundant [30,31]. Compared to the supervised learning, the structural information about the labeled and unlabeled data is more important for classifier design in SSL. Moreover, the discriminative information involved in the few labeled data is also vital to characterize the data structure appropriately.

Belkin et al. [24] constructed a nearest neighbor graph to characterize the nonlinear structure of data manifold, and then presented the Laplacian regularized least-squares classification (LapRLSC) through combining both the nonnegative Tikhonov and Laplacian regularizers defined on the graph. However, the unsupervised selection of the nearest neighbors in graph construction may lead to the different class samples partitioned into a same neighborhood, which violates the original manifold assumption [32,33] to very great extent and thus would degrade the corresponding classification performance. Wang and Zhang [34] further stressed the discriminative information in the labeled data and defined a nonnegative discriminative kernel regularizer instead of the Laplacian regularizer in LapRLSC, called as semi-supervised discriminative regularization (SSDR). But such discriminative kernel is in fact the inverse of the common Mercer kernel, which is applied some perturbation to avoid singularity. Wu and Schölkopf [35] proposed the local learning regularization (LLReg) with the nonnegative local learning regularizer which makes solution with the property that the label of each sample can be well predicted based on its neighbors and their labels. Wang et al. [36] extended this idea to the construction of the local loss function and presented the local and global regularization (LGReg), which firstly trains a linear classifier in each local neighborhood and then embeds a new nonnegative regularizer into the framework of LapRLSC instead of the Tikhonov regularizer. However, since LLReg and LGReg both must train n classifiers corresponding to the n samples given, their experimental costs are much expensive especially in the large-scale classification. Moreover, Wang et al. [37] proposed the semi-parametric semi-supervised discriminant analysis (SSDA) through designating the loss function as the projection function corresponding to the largest eigenvalue obtained by kernel principal component analysis (KPCA) [38], which can incorporate the geometrical information contained in data into the learning process but still lacks the discriminative power to some extent.

In this section, we try to further derive a novel semi-supervised indefinite kernel machine termed as SemiDR from the proposed framework, which aims to embed the local discriminative structure of the labeled data and the global structure of the unlabeled data simultaneously. However, due to few labeled data in SSL, the original definition on the discriminative regularizer cannot characterize the discriminative structure of the data sufficiently and effectively, at alone precisely, thus leading to the decline of the classifier performance. Therefore, the primary issue in SemiDR is how to redefine the regularizer to adapt the SSL scenarios by making use of not only the labeled data but also the unlabeled data. One possible option is the combination of the Laplacian regularizer in LapRLSC with the discriminative regularizer in SupDR. However, such combination still cannot effectively avoid the insufficient description about the data discriminative structure.

Fortunately, a recently-proposed dimensionality reduction method semi-supervised local FDA (SELF) [39] enlightens us. SELF bridges unsupervised PCA and supervised graph-based local FDA (LFDA) [27], and gives the new definitions on the intra-class and inter-class scatter matrices in SSL. Inspired by SELF, we will redefine the discriminative regularizer.

Given a set of samples

$$X = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_l, x_{l+1}, \cdots, \boldsymbol{x}_n\} \subset \boldsymbol{R}^m,$$

where $X_l = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^l$ are labeled coming from c different classes, and $X_u = \{\boldsymbol{x}_j\}_{j=l+1}^n$ are unlabeled. In terms of PCA, we firstly define the similar total scatter measure which reflects the global structure of the outputs

$$\begin{aligned} S_p &= \sum_{i=1}^n \|\boldsymbol{f}(\boldsymbol{x}_i) - \frac{1}{n}\sum_{j=1}^n \boldsymbol{f}(\boldsymbol{x}_j)\|^2 \\ &= \frac{1}{2}\boldsymbol{w}^T \sum_{i=1}^n \sum_{j=1}^n \Phi_{i,j}^p (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{w} = \boldsymbol{w}^T \boldsymbol{S}_p \boldsymbol{w}, \end{aligned} \tag{24}$$

where $\Phi_{i,j}^p = 1/n$.

According to the pairwise expression of the scatter matrices which is convenient to further integrate with the local manifold geometry [27], we reformulate $A(\boldsymbol{f})$ and $B(\boldsymbol{f})$ of the generalized variances as

$$\begin{aligned} A(\boldsymbol{f}) = S_w &= \sum_{t=1}^c \sum_{i=1}^{n_t} \|\boldsymbol{f}(\boldsymbol{x}_i^{(t)}) - \frac{1}{n_t}\sum_{j=1}^{n_t} \boldsymbol{f}(\boldsymbol{x}_j^{(t)})\|^2 \\ &= \frac{1}{2}\boldsymbol{w}^T \sum_{i=1}^n \sum_{j=1}^n \Phi_{i,j}^w (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{w}, \end{aligned} \tag{25}$$

$$B(\boldsymbol{f}) = S_b = \sum_{t=1}^c n_t \|\frac{1}{n_t}\sum_{i=1}^{n_t} \boldsymbol{f}(\boldsymbol{x}_i^{(t)}) - \frac{1}{n}\sum_{j=1}^n \boldsymbol{f}(\boldsymbol{x}_j)\|^2$$

$$=\frac{1}{2}\boldsymbol{w}^T\sum_{i=1}^{n}\sum_{j=1}^{n}\Phi_{i,j}^{b}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{w}, \tag{26}$$

where

$$\Phi_{i,j}^{w}=\begin{cases}1/n_t & \textit{if} \quad y_i=y_j=t\\ 0 & \textit{if} \quad y_i\neq y_j\end{cases} \tag{27}$$

$$\Phi_{i,j}^{b}=\begin{cases}1/n-1/n_t & \textit{if} \quad y_i=y_j=t\\ 1/n & \textit{if} \quad y_i\neq y_j\end{cases} \tag{28}$$

For more detailed derivations, the interested readers can refer to the literature of LFDA [27].

Similarly to LFDA, we further introduce the weight $\Psi_{i,j}$ of the data graph into the $\Phi_{i,j}^{w}$ and $\Phi_{i,j}^{b}$ as the coefficients, which describes the local discriminative structure of the outputs

$$S_{lw}=\frac{1}{2}\boldsymbol{w}^T\sum_{i=1}^{l}\sum_{j=1}^{l}\Phi_{i,j}^{lw}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{w}=\boldsymbol{w}^T\boldsymbol{S}_{lw}\boldsymbol{w}, \tag{29}$$

$$S_{lb}=\frac{1}{2}\boldsymbol{w}^T\sum_{i=1}^{l}\sum_{j=1}^{l}\Phi_{i,j}^{lb}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T\boldsymbol{w}=\boldsymbol{w}^T\boldsymbol{S}_{lb}\boldsymbol{w}, \tag{30}$$

where

$$\Phi_{i,j}^{lw}=\begin{cases}\Psi_{i,j}/l_t & \textit{if} \quad y_i=y_j=t\\ 0 & \textit{if} \quad y_i\neq y_j\end{cases}, \tag{31}$$

$$\Phi_{i,j}^{lb}=\begin{cases}\Psi_{i,j}(1/l-1/l_t) & \textit{if} \quad y_i=y_j=t\\ 1/l & \textit{if} \quad y_i\neq y_j\end{cases}, \tag{32}$$

$$\Psi_{i,j}=\begin{cases}\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/\sigma^2) & \textit{if } \boldsymbol{x}_j\in ne(\boldsymbol{x}_i) \textit{ or } x_i\in ne(\boldsymbol{x}_j)\\ 0 & \textit{otherwise}\end{cases}.$$

The $\Psi_{i,j}$ weights the value for the output pairs in the same class according to the relative location of the input pairs. While the input pairs in the same class are close, $\Psi_{i,j}$ imposes the corresponding output pairs also being nearby. On the contrary, if the inputs pairs in the same class are far apart, the corresponding value of $\Psi_{i,j}$ would be relatively smaller which tries to keep the pairs away from each other in the output space. Furthermore, for the input pairs in the different classes, there is no weight set since they are expected to separate from each other in the output space irrespective of the affinity in the original input space [27].

Finally, we obtain the new definitions on $A(\boldsymbol{f})$ and $B(\boldsymbol{f})$ in SSL by bridging the measures characterized the global data structure and local discriminative structure. Concretely, on one hand, the total scatter matrix $\boldsymbol{S}_p$ is incorporated with $\boldsymbol{S}_{lb}$ to form the new inter-class scatter matrix $\boldsymbol{S}_{rlb}^{(SSL)}$. Since $\boldsymbol{S}_{lb}$ would be unreliable in SSL in the case of the few labeled data, the global structure is expected to keep simultaneously [39]

$$B^{(SSL)}(\boldsymbol{f})=(1-\gamma)S_{lb}+\gamma S_p=\boldsymbol{w}^T[(1-\gamma)\boldsymbol{S}_{lb}+\gamma\boldsymbol{S}_p]\boldsymbol{w}=\boldsymbol{w}^T\boldsymbol{S}_{rlb}^{(SSL)}\boldsymbol{w}, \tag{33}$$

where γ is the regularization parameter that regulates the relative significance of the two kinds of the structural information, $0\leq\gamma\leq 1$. Here we compute $\boldsymbol{S}_p$ by the both labeled and unlabeled data as in SELF, thus it is more reliable and adding it in $\boldsymbol{S}_{rlb}^{(SSL)}$ would improve the reliability of $B^{(SSL)}(\boldsymbol{f})$.

On the other hand, an identity matrix $\boldsymbol{I}$ is incorporated with $\boldsymbol{S}_{lw}$ as a regularizer to form the new intra-class scatter matrix $\boldsymbol{S}_{rlw}^{(SSL)}$ in order to avoid the ill condition of $\boldsymbol{S}_{lw}$ [39], which can improve the stability of $A^{(SSL)}(\boldsymbol{f})$

$$A^{(SSL)}(\boldsymbol{f})=(1-\gamma)S_{lw}+\gamma\|\boldsymbol{w}\|^2=\boldsymbol{w}^T[(1-\gamma)\boldsymbol{S}_{lw}+\gamma\boldsymbol{I}]\boldsymbol{w}=\boldsymbol{w}^T\boldsymbol{S}_{rlw}^{(SSL)}\boldsymbol{w}, \tag{34}$$

The final optimization objective of SemiDR can be formulated as

$$\min_{\boldsymbol{f}\in\tilde{K}}\left\{\frac{1}{l}\sum_{i=1}^{l}(y_i-\boldsymbol{w}^T\boldsymbol{x}_i)^2+\boldsymbol{w}^T[\eta\boldsymbol{S}_{rlw}^{(SSL)}-(1-\eta)\boldsymbol{S}_{rlb}^{(SSL)}]\boldsymbol{w}\right\}. \tag{35}$$

For the nonlinear version, the procedure of SemiDR is similar to SupDR. Besides the construction of the PD kernels K_+ and K_-, another issue in SemiDR is the matrix decomposition for the joint matrix

$$\eta\boldsymbol{S}_{rlw}^{(SSL)}-(1-\eta)\boldsymbol{S}_{rlb}^{(SSL)},$$

that is, the derivation of (13).

In fact, the pairwise expression of the scatter matrices can be formulated as [27,39,40]

$$\boldsymbol{S}=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}W_{i,j}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T, \tag{36}$$

where $\boldsymbol{W}$ is some $n\times n$ matrix, $\boldsymbol{W}=[W_{i,j}]_{n\times n}$.

Let $\boldsymbol{D}$ be the $n\times n$ diagonal matrix with $D_{i,i}=\sum_{j=1}^{n}W_{i,j}$, and $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$. Then $\boldsymbol{S}$ can be expressed in terms of $\boldsymbol{L}$ as

$$\boldsymbol{S}=\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^T \tag{37}$$

Consequently, $\boldsymbol{S}_{rlw}^{(SSL)}$ and $\boldsymbol{S}_{rlb}^{(SSL)}$ can be reformulated as [39]

$$\boldsymbol{S}_{rlw}^{(SSL)}=\boldsymbol{X}\boldsymbol{L}_{rlw}\boldsymbol{X}^T=\boldsymbol{X}[(1-\gamma)\boldsymbol{L}_{lw}+\gamma(\boldsymbol{X}^T\boldsymbol{X})^{+}]\boldsymbol{X}^T, \tag{38}$$

$$\boldsymbol{S}_{rlb}^{(SSL)}=\boldsymbol{X}\boldsymbol{L}_{rlb}\boldsymbol{X}^T=\boldsymbol{X}[(1-\gamma)\boldsymbol{L}_{lb}+\gamma\boldsymbol{L}_p]\boldsymbol{X}^T, \tag{39}$$

where $\boldsymbol{L}_{lw}=\boldsymbol{D}^{lw}-\boldsymbol{\Phi}^{lw}, \boldsymbol{L}_{lb}=\boldsymbol{D}^{lb}-\boldsymbol{\Phi}^{lb}, \boldsymbol{L}_p=\boldsymbol{D}^p-\boldsymbol{\Phi}^p$.

The corresponding eigenvalue decomposition is

$$\eta[(1-\gamma)\boldsymbol{L}_{lw}+\gamma(\boldsymbol{X}^T\boldsymbol{X})^{+}]-(1-\eta)[(1-\gamma)\boldsymbol{L}_{lb}+\gamma\boldsymbol{L}_p]=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T.$$

where $\boldsymbol{U}$ and $\boldsymbol{\Lambda}$ are defined as before. The pseudo-code for SemiDR is shown in Table 2.

Table 2
Pseudo-code for SemiDR.

```
Input: Labeled samples $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{l}$;
    Unlabeled samples $\{(\boldsymbol{x}_i,y_i)\}_{i=l+1}^{n}$;
    The number $k$ of the nearest neighbors of $\boldsymbol{x}_i$;
The width $\sigma$ in the weights $\Psi_{i,j}$;
    The regularization parameters $\eta$ and $\gamma$
Output: The coefficient $\boldsymbol{\alpha}^*$.
for $i=1,\cdots,l$
    $\boldsymbol{x}_i^{(k)}\leftarrow$ $k$th nearest neighbor of $\boldsymbol{x}_i$ among $\{\boldsymbol{x}_j\}_{j=1}^{n}$;
end
for $i,j=1,\cdots,l$
    if $y_i{=}y_j$
        $\Psi_{i,j}\leftarrow\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/\sigma^2)$; $\Phi_{i,j}^{lw}\leftarrow\Psi_{i,j}/l_{y_i}$; $\Phi_{i,j}^{lb}\leftarrow\Psi_{i,j}(1/l-1/l_{y_i})$;
    else
        $\Phi_{i,j}^{lw}\leftarrow 0$; $\Phi_{i,j}^{lb}\leftarrow 1/l$;
    end
end
for $i,j=1,\cdots,n$
        $\Phi_{i,j}^{p}\leftarrow 1/n$;
end
$\boldsymbol{L}_{lw}\leftarrow\boldsymbol{D}^{lw}-\boldsymbol{\Phi}^{lw}$; $\boldsymbol{L}_{lb}\leftarrow\boldsymbol{D}^{lb}-\boldsymbol{\Phi}^{lb}$; $\boldsymbol{L}_p\leftarrow\boldsymbol{D}^p-\boldsymbol{\Phi}^p$;
$\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\leftarrow$ Eigenvalue decomposition of
    $\eta[(1-\gamma)\boldsymbol{L}_{lw}+\gamma(\boldsymbol{X}^T\boldsymbol{X})^{+}]-(1-\eta)[(1-\gamma)\boldsymbol{L}_{lb}+\gamma\boldsymbol{L}_p]$;
$\boldsymbol{\Gamma}_+\leftarrow\boldsymbol{U}_+\boldsymbol{\Lambda}_+\boldsymbol{U}_+^T$; $\boldsymbol{\Gamma}_-\leftarrow\boldsymbol{U}_-\boldsymbol{\Lambda}_-\boldsymbol{U}_-^T$;
Construct the kernel functions $K_+$ and $K_-$ in $\boldsymbol{H}_+$ and $\boldsymbol{H}_-$ respectively;
$\tilde{K}_{disreg}\leftarrow K_+-K_-$; $\hat{K}_{i,j}\leftarrow\tilde{K}_{disreg}(\boldsymbol{x}_i,\boldsymbol{x}_j)$; $\boldsymbol{Y}\leftarrow[y_1,\cdots,y_l,0,\cdots,0]^T$;
$\boldsymbol{\alpha}^*=(\hat{\boldsymbol{K}}+l\boldsymbol{I})^{+}\boldsymbol{Y}$.
```

5. Experiments

To evaluate the proposed SupDR and SemiDR algorithms, we perform a series of experiments systematically on both toy and real-world classification problems. All the experiments are performed on a server with Xeon(R) X5460 3.16 GHz processor and 32766 MB RAM.

5.1. Toy datasets

Two-moon dataset is a commonly-used toy problem in the comparisons of the classification algorithms. Here we choose the two datasets that contain fifty samples in each class and set the variances of the noise 1.5 and 0.5, corresponding to supervised and semi-supervised classification respectively. As shown in Figs. 1 and 2, '$\cdot$' denotes the training data in the two classes, as well as '$+$' denotes the testing data.

In the supervised dataset, we compare DRLSC, indefinite kernel regularized least-squares classification (IKRLSC) and SupDR. We select the Gaussian kernel as the kernel functions in DRLSC and SupDR, and the indefinite Gaussian combination kernel [19] in IKRLSC. The width parameter σ is selected from the set $\{2^{-10},2^{-9},\ldots,2^{9},2^{10}\}$ as the option of the regularization parameter λ in IKRLSC. The regularization parameters η and γ in DRLSC and SupDR are chosen in the set $\{0,0.01,0.05,0.1,\ldots,0.95,1\}$. All the choices are done by cross-validation [41,42]. Furthermore, in DRLSC and SupDR, the number k of the nearest neighbors is fixed to 7. The three subfigures in Fig. 1 depict the discriminant planes of the three algorithms on the dataset. The corresponding training and testing accuracies are reported below the figures. The results show that IKRLSC and SupDR both perform better than DRLSC on the whole, due to the utilization of the indefinite kernel instead of the unstable empirical kernel mapping. Moreover, the plane obtained by SupDR is more consistent with the structural distribution of the two class data than the plane of IKRLSC and thus gets better classification results, since SupDR further embeds the prior discriminative and structural information into the construction of the indefinite kernel.

In the semi-supervised dataset, we randomly select three samples in each class as the labeled data, illustrated as the filled circle and hollow square symbols in Fig. 2. We compare regularized least-squares classification (RLSC) [43], LapRLSC and SemiDR, where RLSC is acted as a baseline and LapRLSC is the most commonly-used compared algorithm in SSL. The kernel and parameter settings are the same as the supervised case. The classification results are shown in Fig. 2. RLSC only concerns the six labeled data and thus performs poorly. LapRLSC introduces the local structure of the data manifold into RLSC by the Laplacian regularizer, and thus can describe the data distribution to some extent. However, it is relatively sensitive to the local variations of the data due to less emphasis on the global structure. SemiDR gets more reasonable discriminant plane than both the planes of RLSC and LapRLSC, and thus has the best classification accuracies.

5.2. IDA datasets

To further investigate the effectiveness of our SupDR and SemiDR, we also evaluate their performance on the IDA

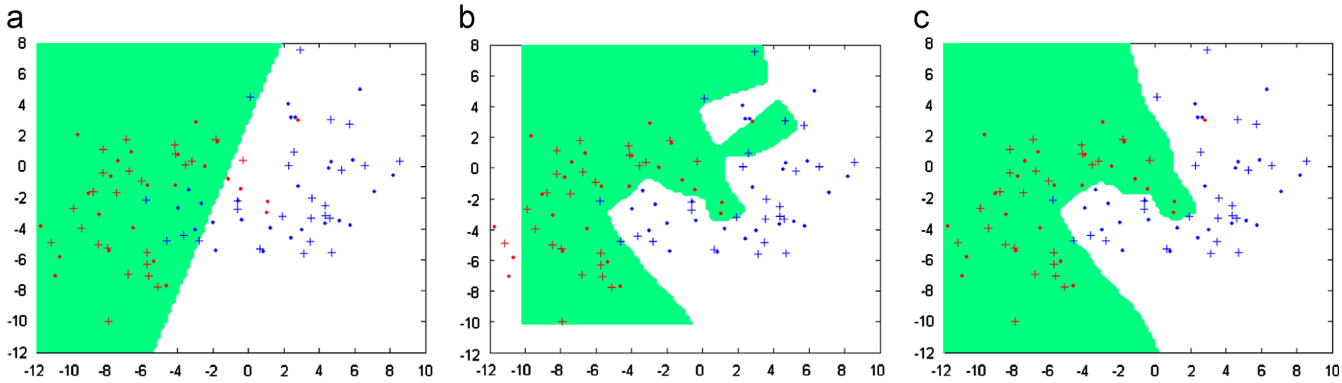

Fig. 1. The illustration of the discriminant planes in supervised classification: DRLSC (a) (Train Accu.=82%, Test Accu.=88%), IKRLSC (b) (Train Accu.=92%, Test Accu.=86%) and SupDR (c) (Train Accu.=98%, Test Accu.=96%).

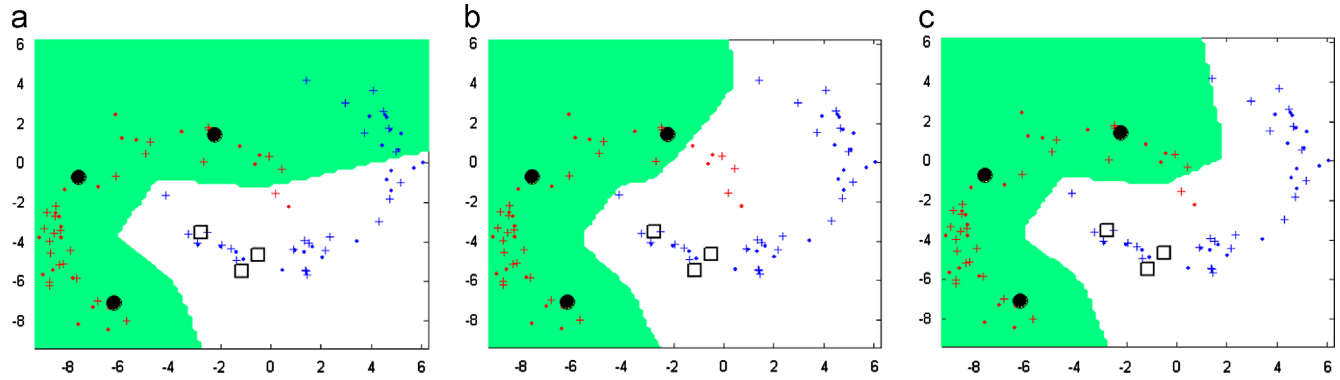

Fig. 2. The illustration of the discriminant planes in semi-supervised classification: RLSC (a) (Train Accu.=100%, Test Accu.=84%), LapRLSC (b) (Train Accu.=100%, Test Accu.=94%) and SemiDR (c) (Train Accu.=100%, Test Accu.=98%).

Table 3
The Attributes of the thirteen datasets in the IDA database.

Dataset	Dimension	Training set size	Testing set size	Run times
Banana	2	400	4900	100
Breast-cancer	9	200	77	100
Diabetis	8	468	300	100
Flare-solar	9	666	400	100
German	20	700	300	100
Heart	13	170	100	100
Ringnorm	20	400	7000	100
Thyroid	5	140	75	100
Titanic	3	150	2051	100
Twonorm	20	400	7000	100
Waveform	21	400	4600	100
Image	18	1300	1010	20
Splice	60	1000	2175	20

Table 4
Acronyms, full names and citations for the algorithms compared in the supervised experiments.

Acronym	Full name	Citation
RBFNN	Radial basis function neural network	[22]
RLSC	Regularized least-squares classification	[43]
SVM	Support vector machine	[46]
IKRLSC	Indefinite kernel regularized least-squares classification	Ours
LapRLSC	Laplacian regularized least-squares classification	[24]
DRLSC	Discriminatively regularized least-squares classification	[20]

database [44].[1] The database consists of thirteen datasets, which all contain two classes. We use the training and testing sets offered by the database. Table 3 presents a brief description of these datasets.

5.2.1. *Supervised classification*

We compare SupDR with RBFNN, RLSC, IKRLSC, LapRLSC and DRLSC, which all develop from the least squares loss function. We also compare it with SVM, which is the state-of-the-art supervised large margin classifier. All the algorithms are listed in Table 4. The kernel and parameter settings are the same as the above toy dataset, except that the number k of the nearest neighbors is selected from {5,10,…,25,30}. The average classification accuracies and standard deviations are reported in Table 5.

RBFNN in nature bases on the cluster assumption of the data [45] which firstly divides the data into several clusters by some clustering strategies. However, cluster assumption seems inapplicable for these datasets since RBFNN performs relatively worse than the other algorithms even RLSC in most datasets. On the contrary, LapRLSC, as a typical paradigm based on the manifold assumption [45], performs much better. However, due to the insufficient utilization of the discriminative information, the accuracies of LapRLSC are also worse than the accuracies of DRLSC and SupDR in several datasets.

IKRLSC and SupDR are both indefinite kernel classifiers, which outperform RLSC and DRLSC respectively. Compared the two algorithms, SupDR is superior to IKRLSC owing to the further introduction of the prior knowledge into the indefinite kernels.

In order to find out whether SupDR is statistically better than the other methods, we perform the t-test on the classification results of the runs to calculate the statistical significance of SupDR. The null hypothesis H_0 demonstrates that there is no significant difference between the mean number of patterns correctly classified by SupDR and the other methods. If the hypothesis H_0 of each dataset is rejected at the 5% significance level, i.e., the t-test value is more than 1.7341, the corresponding results in Table 5 will be denoted "⁎". Consequently, as shown in Table 5, it can be clearly found that SupDR possesses significantly superior classification performance to the other methods in most datasets. This just accords with our conclusions.

5.2.2. *Semi-supervised classification*

We compare SemiDR with RLSC and the PD kernel SSL classifiers LapRLSC, LapSVM, SSDR, SSDA and LGReg, as shown in Table 6. We randomly select 10%, 20%, 30%, 40% and 50% samples in the training set as the labeled data in each dataset, and the remaining samples as the unlabeled data. For LGReg is locally linear [36], here we uniformly adopt the linear kernel. Different from the supervised classification, the number k is selected from {5,10,15,20}. Specially, for the datasets Breast-cancer, Heart, Thyroid and Titanic which have relatively fewer training data, k is selected from {5,10}. This process is repeated ten times to generate ten independent runs for each dataset. Fig. 3 shows the corresponding average classification accuracies and standard deviations of the seven algorithms in the thirteen datasets. In order to avoid the overlapping, each standard deviation is reported by compressing to one-tenth.

On the most datasets, the accuracies of the seven algorithms basically improve step by step with the increase of the labeled data, which validates the well-known "No Free Lunch" Theorem [38]. Comparing the seven algorithms, RLSC performs relatively poorly on the most datasets, which clearly justifies the significance of the unlabeled data in SSL. LapRLSC, LapSVM, SSDR, SSDA and LGReg have comparable classification performance on the datasets, which may imply that these improved algorithms still utilize the latent knowledge in the data insufficiently. Though these algorithms have combined the discriminative information to some extent in various manners, their classification performances are worse than SemiDR′s on many datasets yet.

SemiDR outperforms all the compared algorithms on most datasets, especially in Banana, Breast-cancer, Diabetis, German, Titanic, Waveform and Image. Different from the other algorithms, SemiDR embeds the local and global data structure involved in the labeled and unlabeled data simultaneously and makes use of the discriminative information more sufficiently, which results in its superior performance in the real-world classification tasks.

5.3. *SSL datasets*

We further evaluate SemiDR on the SSL database,[2] which are the benchmark used in the literature [25] and consist of nine semi-supervised learning datasets. The training and testing sets, as well as the labeled and unlabeled data, also have been offered by the database. We select seven datasets, and denote them by SSL*i_j* (i=1,2,…,7), where j is the number of labeled data. In these datasets, SSL6 contains six classes and the other datasets have two classes. Their dimension and total number of data are 241 and 1500 respectively, except 117 and 400 in SSL4.

We compare the algorithms both in the linear and Gaussian kernel versions. In the Gaussian kernel version, we further compare SemiDR under the same optimization objective function with two different kernel tricks, i.e. the indefinite kernel and the empirical kernel in the original DRLSC. We denote the latter as SemiDR_EK. The choices of the parameters are the same as the ones above. The classification results are reported in Tables 7 and 8.

SemiDR excels the other algorithms on almost all the datasets, especially in the ten-labeled data cases. For example, on the

[1] Available from http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm.

[2] Available from http://www.kyb.tuebingen.mpg.de/ssl-book/.

Table 5
Classification accuracies (%) comparisons on the IDA datasets with the *Gaussian* Kernel.

Dataset	Classification				Accuracy		
	RBFNN	RLSC	SVM	IKRLSC	LapRLSC	DRLSC	SupDR
Banana	89.24* ± 1.07	89.25* ± 0.52	89.49* ± 0.49	89.69* ± 0.51	89.31* ± 0.69	89.61* ± 0.62	***91.42*** ± 0.59
Breast-cancer	72.36* ± 4.11	75.62* ± 3.21	75.84* ± 4.02	78.05* ± 3.12	77.31* ± 2.46	77.66* ± 2.27	***80.64*** ± 2.70
Diabetis	75.71* ± 1.54	76.27* ± 1.39	77.33* ± 1.66	77.33* ± 1.43	76.33* ± 1.56	77.97 ± 1.62	***78.73*** ± 1.41
Flare-solar	65.63* ± 1.28	67.83* ± 2.12	68.10* ± 1.50	68.80* ± 1.55	***69.46*** ± 1.07	68.00* ± 1.64	69.24 ± 1.52
German	75.29* ± 1.81	76.57* ± 1.56	78.47* ± 1.37	78.33* ± 2.04	78.83* ± 2.26	78.04* ± 2.15	***79.80*** ± 2.04
Heart	82.45* ± 4.24	83.70* ± 3.89	84.60* ± 3.06	85.00* ± 2.87	84.20* ± 3.88	85.20* ± 3.12	***86.30*** ± 2.21
Ringnorm	98.30 ± 0.34	95.41* ± 0.23	97.55 ± 0.13	96.64* ± 0.07	***98.96**** ± 0.17	97.59 ± 0.14	97.62 ± 0.09
Thyroid	95.48* ± 2.63	94.73* ± 2.56	94.87* ± 1.69	95.60* ± 1.38	95.80* ± 2.96	95.95* ± 1.93	***96.53*** ± 2.28
Titanic	76.74* ± 1.43	77.90* ± 1.79	78.86 ± 1.16	78.47 ± 1.33	78.58 ± 1.57	78.76 ± 1.02	***78.88*** ± 1.15
Twonorm	97.15* ± 0.09	97.67* ± 0.08	97.74* ± 0.12	98.25* ± 0.22	***98.61*** ± 0.26	97.86* ± 0.16	98.39 ± 0.13
Waveform	89.34* ± 0.38	90.31* ± 0.42	90.35* ± 0.36	90.68* ± 0.26	90.48* ± 0.59	90.52* ± 0.32	***91.60*** ± 0.27
Image	96.68 ± 0.87	95.58* ± 0.23	96.60* ± 0.41	96.75 ± 0.38	96.52 ± 0.70	96.96 ± 0.54	***97.36*** ± 0.63
Splice	90.05 ± 0.72	89.11* ± 0.43	89.46* ± 0.68	89.34* ± 0.56	***91.16*** ± 1.22	90.23 ± 0.82	90.64 ± 0.76

Table 6
Acronyms, full names and citations for the algorithms compared in the semi-supervised experiments.

Acronym	Full name	Citation
RLSC	Regularized Least-Squares Classification	[43]
LapRLSC	Laplacian regularized least-squares classification	[24]
LapSVM	Laplacian support vector machine	[24]
SSDR	Semi-supervised discriminative regularization	[34]
SSDA	Semi-parametric semi-supervised discriminant analysis	[37]
LGReg	Local and global regularization	[36]

SSL1_10 and SSL5_10 datasets, the accuracies of SemiDR exceed those of the other algorithms near 10%. Specially, on the multi-class dataset SSL6_10 and SSL6_100, SemiDR in the linear kernel remarkably excels LapRLSC, SSDR, SSDA and LGReg, whose accuracies are even much poorer than those of RLSC. Furthermore, comparing Table 7 with 8, the linear kernel SemiDR sometimes indeed outperforms the Gaussian kernel version in the other algorithms, such as on the SSL1_10, SSL5_10 and SSL7_10 datasets.

SemiDR also exceeds SemiDR_EK on all the datasets in the Gaussian kernel. The superiority of SemiDR further validates the reasonability of the indefinite kernels in practical applications. Though the corresponding kernel matrix is not positive semi-definite any longer as the common Mercer kernel in SemiDR_EK, the utilization of the indefinite kernel can not only provide the theoretical basis for SemiDR, but also actually improve its classification performance experimentally.

5.4. *USPS dataset*

We also compare these algorithms on a relatively large-scale dataset USPS.[3] USPS dataset consists of grayscale handwritten digit images from "0" to "9", and each digit contains 1100 examples. We divide the samples into two non-overlapping training and testing sets, and each set contains half of examples in each digit respectively. Furthermore, we randomly select 5% samples in the training set as the labeled data. The average classification accuracies are shown in Fig. 4.

Due to the similarity of multiple digits, the samples in different classes overlap relatively heavily in the original space. In such circumstance, LapRLSC, SSDR, SSDA and LGReg all fail in the linear kernel case. However, SemiDR performs much better. Its accuracy excels those of LapRLSC, SSDR and SSDA over 40%, and that of LGReg over 30%. Through applying the Gaussian kernelization, the accuracies of all algorithms greatly improve, but SemiDR still performs best.

6. Conclusion

In this paper, we firstly build the theoretical basis for DRLSC by justifying the legitimacy of the discriminative regularizer in the RKKS. Then we further present a new discriminality-driven regularization framework for indefinite kernel machine based on the regularizer, to remedy the insufficient use of classification specific prior knowledge in the construction of the indefinite kernels. Derived from the framework, two indefinite kernel machines SupDR and SemiDR are proposed, where SupDR is the redescription on the original DRLSC by the indefinite kernelization and SemiDR is more likely the first discriminality-driven indefinite kernel method for SSL. The experimental results have demonstrated the superiority of the two methods.

There are several directions for future study:

- ***Theoretical generalization analysis***: Due to the particularity of the indefinite kernel, many state-of-the-art theoretical generalization results for common PD kernel are not applicable. As a result, our framework for indefinite kernel machine mainly focuses on the algorithmic design and solution. The

[3] Available from http://www.cs.nyu.edu/~roweis/data.html.

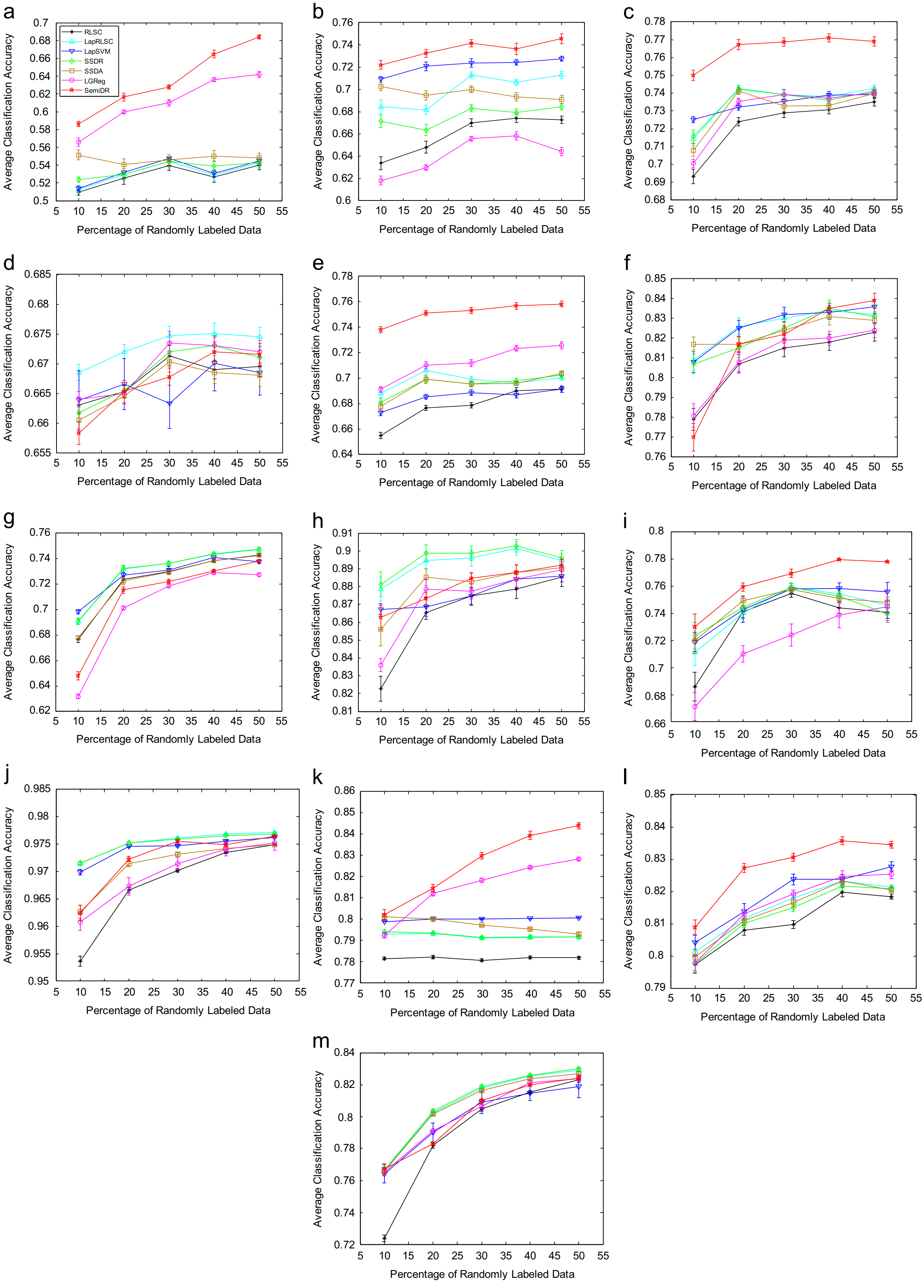

Fig. 3. The classification performance comparisons of RLSC, LapRLSC, LapSVM, SSDR, SSDA, LGReg and SemiDR on the IDA datasets. (a) banana, (b) breast-cancer, (c) diabetis, (d) flare-solar, (e) german, (f) heart, (g) ringnorm, (h) thyroid, (i) titanic, (j) twonorm, (k) waveform, (l) image and (m) splice.

Table 7
Classification accuracies (%) comparisons on the SSL datasets with the *Linear* Kernel.

Dataset	Classification			Accuracy		
	RLSC	LapRLSC	SSDR	SSDA	LGReg	SemiDR
SSL1_10	74.24*	77.78*	77.70*	77.57*	68.53*	***86.91***
	± 6.28	± 6.26	± 6.31	± 6.32	± 6.58	± 3.65
SSL2_10	75.22*	82.48	81.89*	81.12*	80.27*	***82.98***
	± 1.17	± 1.75	± 1.30	± 1.13	± 1.53	± 1.41
SSL3_10	55.13*	***60.79****	60.09	58.90*	56.87*	59.74
	± 4.51	± 3.56	± 3.68	± 3.76	± 4.28	± 3.47
SSL4_10	55.08*	62.88	61.83*	55.33*	62.54*	***63.71***
	± 3.13	± 2.93	± 3.55	± 4.35	± 2.46	± 1.80
SSL5_10	56.52*	61.62*	61.64*	51.43*	55.13*	***72.88***
	± 2.88	± 2.82	± 2.86	± 1.63	± 1.41	± 1.68
SSL6_10	33.00*	19.48*	17.92*	17.90*	17.78*	***33.93***
	± 4.42	± 4.02	± 4.27	± 3.10	± 4.31	± 3.83
SSL7_10	58.79*	60.82*	60.72*	60.97*	62.32*	***68.29***
	± 4.69	± 4.11	± 4.30	± 3.93	± 4.53	± 3.75
SSL1_100	90.43*	93.20	92.80*	92.51*	84.02*	***93.72***
	± 1.42	± 1.46	± 1.40	± 1.46	± 1.56	± 1.32
SSL2_100	86.12*	***88.64***	88.41	88.17	87.40*	88.19
	± 1.28	± 1.30	± 1.25	± 1.28	± 1.34	± 1.21
SSL3_100	69.96*	85.87*	84.21*	83.04*	85.12*	***87.84***
	± 2.24	± 2.54	± 2.38	± 1.69	± 2.32	± 1.53
SSL4_100	73.71*	76.33*	76.58	74.04*	76.57	***77.50***
	± 2.86	± 2.63	± 2.94	± 3.14	± 3.28	± 2.29
SSL5_100	70.83*	78.08*	78.13*	57.09*	77.24*	***79.53***
	± 1.20	± 1.28	± 1.25	± 2.00	± 1.87	± 1.56
SSL6_100	63.96	19.11*	17.51*	19.76*	18.71*	***64.67***
	± 2.91	± 2.36	± 2.72	± 3.95	± 3.27	± 2.93
SSL7_100	70.64*	76.06	76.26	76.89	***76.95***	76.40
	± 2.25	± 2.19	± 2.20	± 1.51	± 2.45	± 1.86

Table 8
Classification accuracies (%) comparisons on the SSL datasets with the *Gaussian* Kernel.

Dataset	Classification			Accuracy		
	RLSC	LapRLSC	SSDR	SSDA	SemiDR_EK	SemiDR
SSL1_10	78.34*	79.28*	80.21*	78.48*	81.53*	***87.91***
	± 5.95	± 7.94	± 6.93	± 6.74	± 4.93	± 3.72
SSL2_10	82.03*	83.18	82.93*	81.73*	80.03*	***83.91***
	± 2.21	± 1.89	± 2.04	± 1.44	± 1.03	± 0.89
SSL3_10	63.83*	67.11*	67.80*	69.84*	63.03*	***72.67***
	± 3.26	± 4.29	± 4.82	± 5.60	± 3.74	± 2.80
SSL4_10	61.42*	57.42*	62.92*	56.46*	56.25*	***72.00***
	± 3.15	± 3.04	± 2.96	± 2.59	± 1.54	± 2.14
SSL5_10	59.39*	58.93*	62.94*	55.40*	59.70*	***73.78***
	± 2.70	± 4.13	± 3.79	± 4.23	± 2.41	± 2.21
SSL6_10	35.71*	41.47	37.42*	35.80*	39.64*	***41.73***
	± 3.85	± 4.16	± 2.59	± 3.65	± 2.63	± 2.86
SSL7_10	61.24*	60.53*	63.58*	57.44*	61.01*	***72.84***
	± 3.14	± 4.25	± 4.13	± 4.46	± 2.69	± 2.37
SSL1_100	94.60*	95.56	95.27	94.98	93.98*	***95.80***
	± 0.42	± 0.77	± 0.93	± 1.36	± 1.52	± 0.65
SSL2_100	90.95*	92.87	92.28*	91.70*	90.18*	***93.22***
	± 1.33	± 1.41	± 1.67	± 1.36	± 1.58	± 1.56
SSL3_100	80.14*	88.23*	87.33*	82.89*	88.47*	***90.67***
	± 1.23	± 1.54	± 1.17	± 0.87	± 1.27	± 0.97
SSL4_100	76.00*	78.10	***79.80****	72.71*	76.42*	78.42
	± 2.20	± 2.92	± 2.89	± 2.35	± 2.64	± 2.18
SSL5_100	75.11*	79.91*	80.12*	77.44*	80.14*	***81.80***
	± 3.04	± 2.45	± 2.03	± 1.83	± 1.14	± 1.36
SSL6_100	78.44	78.53	79.33	78.34	77.35*	***79.40***
	± 2.49	± 2.27	± 2.20	± 2.50	± 1.85	± 1.65
SSL7_100	74.54*	78.09*	79.27*	76.38*	72.50*	***80.73***
	± 2.46	± 2.04	± 2.24	± 2.21	± 1.59	± 1.28

corresponding generalization analysis needs more systematic research.

- ***Optimization on the discriminative regularizer***: The design way of the discriminative regularizer actually provides us a feasible and convenient way to incorporate the classification methods with the dimensionality reduction methods. Take SemiDR as an example. In its implementation in our paper, we just present a simple definition on the discriminative

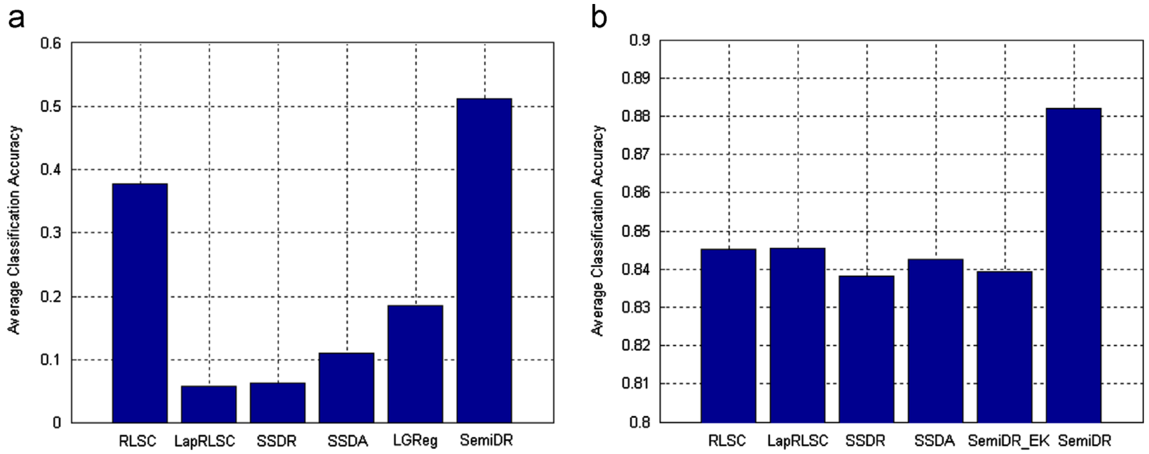

Fig. 4. The classification performance comparisons on the USPS datasets, (a) Linear Kernel and (b) Gaussian Kernel.

regularizer referred to SELF by Sugiyama et al. [39]. Actually, the definition diversity of such regularizer can motivate some further work. For example, we can optimize the regularizer by optimizing the graphs in it to replace the original artificial predefinition in SELF, doing so can desire to mitigate the difficulty in the selection of parameters (k and σ) through cross-validation, where k is the number of the nearest neighbors and σ is the width of the weight. In addition, we can also further introduce the adaptive procedure of the graph construction [47] into the definition of the regularizer to reflect the geometry of the data more faithfully. Certainly, other effective definitions will be one of the directions for our future study.

- ***Optimization on the algorithmic solution***: Although the solving approaches for SupDR and SemiDR are simple, their computational complexity is relatively high especially in large-scale problems. Actually, many algorithmic accelerating techniques, such as Nyström approximation, column sampling and matrix sparsity, can be easily combined with SupDR and SemiDR. Hence how to develop effective and fast solutions for our algorithms is another valuable topic for research.
- ***Indefinite kernel justification framework***: We also provide a feasible theoretical framework of justification for some machine learning algorithms based on the indefinite kernel theory. In fact, the algorithms are not limited to the classification algorithms but can be generalized to dimensionality reduction algorithms. For example, we can also kernelize maximum margin criterion (MMC) [48] in the current new way by the indefinite kernel, where its objective function is the trace difference of the inter-class and intra-class scatter matrices in FDA, and thus the corresponding kernel matrix becomes also indefinite. Furthermore, we can more likely generalize the framework to the algorithms based on the Difference-of-Convex (functions) optimization [49] if the individual convex functions involved can be kernelized.

Acknowledgment

This work was supported in part by National Natural Science Foundations of China (Grant nos. 61375057, 60905002, 61035003, 61170151, 61101202 and 61300165), Natural Science Foundation of Jiangsu Province of China (Grant no. BK20131298), and the Scientific Research Startup Project of New Doctorial Faculties of Southeast University. It was also supported by the Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, and the Fundamental Research Funds for the Central Universities of China (Grant no. NZ2013306).

References

[1] B. Haasdonk, Feature space interpretation of SVMs with indefinite kernels, IEEE Trans. Pattern Anal. Mach. Intell. 27 (4) (2005) 482–492.
[2] J. Chen, J. Ye, Trainging SVM with indefinite kernels. In: Proceedings of the 25th Intl. Conf. on Machine Learning (ICML): 136–143, 2008.
[3] K. Slavakis, P. Bouboulis, S. Theodoridis, Adaptive multiregression in reproducing kernel Hilbert spaces: the multiaccess MIMO channel case, IEEE Trans. Neural Netw. Learn. Syst. 23 (2) (2012) 260–276.
[4] E. Pękalska, B. Haasdonk, Kernel discriminant analysis for positive definite and indefinite kernels, IEEE Trans. Pattern Anal. Mach. Intell. 31 (6) (2009) 1017–1031.
[5] E. Pękalska, A. Harol, R. Duin, B. Spillmann, H. Bunke, Non-Euclidean or non-metric measure can be informative. In: Proceedings of Joint IAPR Workshops Structural and Syntactic Pattern Recognition and Statistical Techniques in Pattern Recognition: 871–880, 2006.
[6] D. Jacobs, D. Weinshall, Y. Gdalyahu, Classification with non-metric distance: image retrieval and class representation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (6) (2000) 583–600.
[7] V. Roth, J. Laub, J. Buhmann, K.-R. Müller, Going metric: denoising pairwise data. In: Proceedings of Advances in Neural Information Processing Systems 16 (NIPS): 841–856, 2003.
[8] E. Pekalska, P. Paclik, R.P. W. Duin, A generalized kernel approach to dissimilarity-based classfication, J. Mach. Learn. Res. 2 (2001) 175–211.
[9] R. Luss, A. d'Aspremont, Support vector machine classification with indefinite kernels, Adv. Neural Inf. Process. Syst. 20 (2007).
[10] Y. Chen, M. R. Gupta, B. Recht, Learning kernels from indefinite similarities. In: Proceedings of the 26th International Conference on Machine Learning (ICML), 2009.
[11] S. Gu, Y. Guo, Learning SVM classifiers with indefinite kernels. In: Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI), 2012.
[12] S. Liwicki, S. Zafeiriou, G. Tzimiropoulos, M. Pantic, Efficient online subspace learning with an indefinite kernel for visual tracking and recognition, IEEE Trans. on Neural Netw. Learn. Syst. 23 (10) (2012) 1624–1636.
[13] Y. Chen, E.K. Garcia, M.R. Gupta, A. Rahimi, L. Cazzanti, Similarity-based classification: concepts and algorithms, J. Mach. Learn. Res. 10 (2009) 747–776.
[14] H. Xue, S. Chen, J. Huang, Discriminative indefinite kernel classifier from pairwise constraints and unlabeled data. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR): 497–500, 2012.
[15] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, K. Obermayer, Classification on pairwise proximity data. In: Proceedings of the Advances in Neural Information Processing Systems 11 (NIPS): 438–444, 1998.
[16] V. Roth, J. Laub, M. Kawanabe, J.M. Buhmann, Optimal cluster preserving embedding of nonmetric proximity data, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1540–1551.
[17] Y. Ying, C. Campbell, M. Girolami, Analysis of SVM with indefinite kernels, Adv. Neural Inf. Process. Syst. 22 (2009).
[18] H.-T. Lin, C.-J. Lin, A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods, National Taiwan University, 2003.
[19] C.S. Ong, X. Mary, S. Canu, A.J. Smola, Learning with non-positive kernels. In: Proceedings of the 21st International Conference on Machine Learning (ICML), 2004.
[20] H. Xue, S. Chen, Q. Yang, Discriminatively regularized least-squares classification, Pattern Recognit. 42 (1) (2009) 93–104.

[21] H. Xiong, M.N.S. Swamy, M.O. Ahmad, Optimizing the kernel in the empirical feature space, IEEE Trans. Neural Netw. 16 (2) (2005) 460–474.
[22] S. Haykin, Neural Networks: A Comprehensive Foundation, Tsinghua University Press, 2001.
[23] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (1999) 293–300.
[24] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: A geometric framework for learning from examples, Department of Computer Science, University of Chicago, Tech. Rep: TR-2004-06, 2004.
[25] O. Chapelle, B. Schölkopf, A. Zien, Semi-Supervised Learning, MIT Press, Cambridge, MA, USA, 2006.
[26] J. Tenenbaum, V. Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (22) (2000) 2319–2323.
[27] M. Sugiyama, Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis, J. Mach. Learn. Res. 8 (2007) 1027–1061.
[28] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[29] Z. Xu, R. Jin, H. Yang, I. King, M.R. Lyu, Simple and efficient multiple kernel learning by group lasso. In: Proceedings of the 27th International Conference on Machine Learning (ICML), Haifa, Israel, 2010.
[30] X. Zhu, Semi-supervised learning literature survey, Department of Computer Sciences, University of Wisconsin, Madison, 2008 (Technical Report, 1530).
[31] Y. Wang, S. Chen, Z.-H. Zhou, New semi-supervised classification method based on modified cluster assumption, IEEE Trans. Neural Netw. Learn. Syst. 23 (5) (2012) 689–702.
[32] Z.-H. Zhou, M. Li, Semi-supervised learning with co-training style algorithm, IEEE Trans. Knowl. Data Eng. 19 (11) (2007) 1479–1493.
[33] L. Chen, I.W. Tsang, D. Xu, Laplacian embedded regression for scalable manifold regularization, IEEE Trans. Neural Netw. Learn. Syst. 23 (6) (2012) 902–915.
[34] F. Wang, C. Zhang, On discriminative semi-supervised classification. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI): 720–725, 2008.
[35] M. Wu, B. Schölkopf, Transductive classification via local learning regularization. In: Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS), 2007.
[36] F. Wang, T. Li, G. Wang, C. Zhang, Semi-supervised classification using local and global regularization. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI): 726–731, 2008.
[37] F. Wang, X. Wang, T. Li, Beyond the graphs: semi-parametric semi-supervised discriminant analysis. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, 2009.
[38] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley, 2001.
[39] M. Sugiyama, T. Idé, S. Nakajima, J. Sese, Semi-supervised local fisher discriminant analysis for dimensionality reduction, Mach. Learn. 78 (1–2) (2010) 35–61.
[40] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (2003) 1373–1396.
[41] M. Kääriäinen, Semi-supervised model seletion based on cross-validation. In: Proceedings of 2006 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada: 1894–1899, 2006.
[42] M. Budka, B. Gabrys, Density-preserving sampling: robust and efficient alternative to cross-validation for error estimation, IEEE Trans. Neural Netw. Learn. Syst. 24 (1) (2013) 22–34.
[43] R.M. Rifkin, Everything Old Is New Again: A Fresh Look at Historical Approaches to Machine Learning. (Ph.D. thesis), Massachusetts Institute of Technology, 2002.
[44] G. Ratsch, T. Onoda, K.-R. Müller, Soft margins for AdaBoost, Mach. Learn. 42 (2001) 287–320.
[45] X. Chen, S. Chen, H. Xue, Large correlation analysis, Appl. Math. Comput. 217 (3) (2011) 9041–9052.
[46] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.
[47] W. Cai, S. Chen, D. Zhang, A multi-objective simultaneous learning framework for clustering and classification, IEEE Trans. Neural Netw 21 (2) (2010) 185–200.
[48] H. Li, T. Jiang, K. Zhang, Efficient and robust feature extraction by maximun margin criterion, IEEE Trans. Neural Netw 17 (1) (2006) 157–165.
[49] F.B. Akoa, Combining DC algorithms (DCAs) and decomposition techniques for the training of nonpositive–semidefinite kernels, IEEE Trans. Neural Netw. Learn. Syst. 19 (11) (2008) 1854–1872.

Hui Xue received the B.Sc. degree in Mathematics from Nanjing Normal University in 2002. In 2005, she received the M.Sc. degree in Mathematics from Nanjing University of Aeronautics & Astronautics (NUAA). And she also received the Ph.D. degree in Computer Application Technology at NUAA in 2008. Since 2009, as an Associate Professor, she has been with the School of Computer Science and Engineering at Southeast University. Her research interests include pattern recognition, machine learning and neural computing.

Songcan Chen received the B.Sc. degree in Mathematics from Hangzhou University (now merged into Zhejiang University) in 1983. In December 1985, he completed the M.Sc. degree in Computer Applications at Shanghai Jiaotong University and then worked at Nanjing University of Aeronautics & Astronautics (NUAA) in January 1986 as an Assistant Lecturer. There he received the Ph.D. degree in Communication and Information Systems in 1997. Since 1998, as a full Professor, he has been with the School of Computer Science and Technology at NUAA. His research interests include pattern recognition, machine learning and neural computing. In these fields, he has authored or coauthored over 130 scientific journal papers.