Discriminability-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

A R T I C L E   I N F O

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:
Discriminability-driven regularization
Indefinite kernel machine
Semi-supervised learning
Machine learning

A B S T R A C T

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 discriminability-driven regularizer to encourage the discriminability 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 Krein Space (RKKS) naturally. As a result, we further present a new discriminability-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.

© 2014 Elsevier B.V. All rights reserved.

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]. Pekalska et al. [8] set the negative eigenvalues to be zeros. Gaap et al. [15] flipped the sign of the negative eigenvalues. Roth et al. [16] shifted the eigenvalues by a positive constant.
**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 Kreın Space (RKKS), so as to the problem 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 discriminability 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 discriminativity-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 discriminativity-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 discriminativity-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

$$
\min_{f \in \mathcal{K}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + R_{\text{dereg}}(f, \eta) \right\}
$$

(1)

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

$$
R_{\text{dereg}}(f, \eta) = \eta A(f) - (1 - \eta) B(f),
$$

(2)

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

---

2. **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 Kreın Space (RKKS), so as to the problem 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 discriminability 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 discriminativeness-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 discriminativeness-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 discriminativeness-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

$$
\min_{f \in \mathcal{K}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + R_{\text{dereg}}(f, \eta) \right\}
$$

(1)

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

$$
R_{\text{dereg}}(f, \eta) = \eta A(f) - (1 - \eta) B(f),
$$

(2)

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

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

$$A(f) = S_w = \sum_{t=1}^{c} \sum_{i=1}^{n_t} \frac{1}{n_t} \sum_{j=1}^{n_t} \| f(x_i^{(t)}) - \frac{1}{n_t} \sum_{j=1}^{n_t} f(x_j^{(t)}) \|^2,$$

where $n_t$ is the number of the samples $x_i^{(t)}$ belonging to class $t$, $t = 1, \ldots, c$.

$$B(f) = S_b = \sum_{t=1}^{c} n_t \frac{1}{n_t} \sum_{j=1}^{n_t} f(x_i^{(t)}) - \frac{1}{n_t} \sum_{j=1}^{n_t} f(x_j^{(t)}) \|^2$$

(4)

For the moment, we focus on the linear classifier, i.e.,

$$f(x) = w^T x.$$

Then $A(f)$ and $B(f)$ can be further formulated as

$$A(f) = \sum_{t=1}^{c} \sum_{i=1}^{n_t} \| w^T x_i^{(t)} - \frac{1}{n_t} \sum_{j=1}^{n_t} w^T x_j^{(t)} \|^2$$

$$= \sum_{t=1}^{c} n_t \frac{1}{n_t} \sum_{j=1}^{n_t} w^T (x_i^{(t)} - x_j^{(t)}) (x_i^{(t)} - x_j^{(t)})^T w = w^T S_w w.$$  

(5)

$$B(f) = \sum_{t=1}^{c} n_t \frac{1}{n_t} \sum_{j=1}^{n_t} w^T x_i^{(t)} - \frac{1}{n_t} \sum_{j=1}^{n_t} w^T x_j^{(t)} \|^2$$

$$= w^T \sum_{t=1}^{c} n_t (x_i^{(t)} - x_j^{(t)}) (x_i^{(t)} - x_j^{(t)})^T w = w^T S_b w.$$  

(6)

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

Obviously, $S_w$ and $S_b$ are equivalent to the within-class and between-class scatter matrices in Fisher discriminant analysis (FDA) [27]. Actually, the definitions on $A(f)$ and $B(f)$ are various. For example, any improvements for FDA can be straightforwardly combined in the regularizer to update $S_w$ and $S_b$. 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 $x_i$, DRLSC firstly divides the nearest $k$ neighborhood $n(x_i)$ into two non-overlapping subsets

$$n_e(x_i) = \{ x_j \text{ if } x_i \text{ and } x_j \text{ belong to same class}, \ 1 \leq j \leq k \},$$

$$n_n(x_i) = \{ x_j \text{ if } x_i \text{ and } x_j \text{ belong to different classes}, \ 1 \leq j \leq k \}.$$

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

$$W_{w,ij} = \begin{cases} 1 & \text{if } x_j \in n_e(x_i) \text{ or } x_i \in n_e(x_j) \ , \\ 0 & \text{otherwise} \end{cases},$$

$$W_{b,ij} = \begin{cases} 1 & \text{if } x_j \in n_n(x_i) \text{ or } x_i \in n_n(x_j) \ , \\ 0 & \text{otherwise} \end{cases}.$$

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

$$A^{\text{SUP}}(f) = S_w^{\text{SUP}} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{w,ij} \| f(x_i) - f(x_j) \|^2.$$  

(7)

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

$$B^{\text{SUP}}(f) = S_b^{\text{SUP}} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{b,ij} \| f(x_i) - f(x_j) \|^2.$$  

(8)

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

$$A^{\text{SUP}}(f) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{w,ij} \| f(x_i) - f(x_j) \|^2$$

$$= w^T X D_w - W_w^T X w = w^T X L_w X w.$$  

(9)

where $D_w$ is a diagonal matrix and its entries $D_{w,ii} = \sum_{j=1}^{n} W_{w,ij}$.

$$B^{\text{SUP}}(f) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{b,ij} \| f(x_i) - f(x_j) \|^2$$

$$= w^T X D_b - W_b^T X w = w^T X L_b X w.$$  

(10)

where $D_b$ is also a diagonal matrix and its entries $D_{b,ii} = \sum_{j=1}^{n} W_{b,ij}$.

$L_b = D_b - W_b$ is the Laplacian matrix of $G_b$.

The final optimization function of DRLSC can be formulated as

$$\min_{f} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + w^T X [\eta L_w - (1 - \eta) L_b] X w \right\}.$$  

(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. Discriminatory-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. (Krein space) [19]** An inner product space $(\mathcal{K}, \langle \cdot, \cdot \rangle_{\mathcal{K}})$ is a Krein space if there exist two Hilbert spaces $H_-, H_+$, $H_+$ spanning $\mathcal{K}$ such that

1. All $f \in \mathcal{K}$ can be decomposed into $f = f_+ + f_-$, where $f_+ \in H_+$ and $f_- \in H_-$.
2. $\forall f, g \in \mathcal{K}$, $\langle f, g \rangle_{\mathcal{K}} = \langle f_+, g_+ \rangle_{H_+} - \langle f_-, g_- \rangle_{H_-}$.

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 $\mathcal{K}$ be an RKKS with kernel $K$. Denote by $\Omega(f, X)$ a continuous loss function depending on $f \in \mathcal{K}$ only via its evaluations $f(x_i)$ with $x_i \in X$, let $\Omega(f, f)$ be a continuous stabilizer with strictly monotonic $\Omega: \mathcal{K} \rightarrow \mathcal{R}$ and let $C(f, X)$ be a continuous functional imposing a set of constraints on $f$, that is $C : \mathcal{K} \times X \rightarrow \mathcal{R}$. Then if the optimization problem

$$\min_{f} \left\{ \langle f, X \rangle + \Omega(f, f) \right\}$$

s.t. $C(f, X) \leq d$.

has a saddle point $f^*$, it admits the expansion

$$f^* = \sum_{i=1}^{n} \alpha_i \hat{K}(x_i, \cdot),$$

where $x_i \in X$ and $\alpha_i \in \mathcal{R}$.

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 information into the indefinite kernel. In terms of the
inter-complementarities between the indefinite kernel theory and DRLSC, we further deduce a discriminability-driven regularization framework for indefinite kernel machine.

3.1. Indefinite kernel analysis for discriminability-driven regularization

Firstly, we present the indefinite kernel analysis for the discriminative regularizer, and then give the common discriminability-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,

$$R_{\text{discreg}}(f, \eta) = \langle f, J_f \rangle_{K_{\text{discreg}}}$$

(12)

where $K_{\text{discreg}}$ denotes the induced KreIn Space by the regularizer.

**Proof.** Recall that

$$R_{\text{discreg}}(f, \eta) = \eta A(f) - (1-\eta) B(f) = w^T \eta(S_w - (1-\eta)S_b) w.$$  

Decompose the joint matrix $\eta S_w - (1-\eta)S_b$ into

$$\eta S_w - (1-\eta)S_b = XLX^T$$

(13)

Perform the eigenvalue decomposition for $L = U\Lambda U^T$.

Then we have

$$R_{\text{discreg}}(f, \eta) = w^T U\Lambda U^T X^T w = f^T U\Lambda U^T f$$

(15)

$$= f^T \left( \sum_{\lambda > 0} \lambda u_i u_i^T + \sum_{\lambda < 0} \lambda u_i u_i^T \right) f$$

Let $\Gamma_+ = U_+ \Lambda_+ U_+^T$, $\Gamma_- = U_- \Lambda_- U_-^T$, obviously,

$$\Gamma_+^T \Gamma_- = U_+ \Lambda_+ U_+^T U_- \Lambda_- U_-^T = 0,$$

that is, $\Gamma_+$ and $\Gamma_-$ are orthogonal.

Decompose $f = f_+ + f_-$, where $f_+ \in \text{Hilbert}(\Gamma_+)$, $f_- \in \text{Hilbert}(-\Gamma_-)$, then

$$R_{\text{discreg}}(f, \eta) = f_+^T (\Gamma_+ - (-\Gamma_-)) f_+ + f_-^T (\Gamma_- + (-\Gamma_-)) f_-$$

(16)

$$= f_+^T \Gamma_+ f_+ - f_-^T \Gamma_- f_-$$

$$= f_+^T \Gamma_+ f_+ - f_-^T \Gamma_- f_-$$

Let

$$\Gamma_+^T \Gamma_+ f_+ = \Gamma_-^T \Gamma_- f_- = \mathbf{0}$$

(17)

We can formulate the discriminative regularizer as

$$R_{\text{discreg}}(f, \eta) = \langle f, f_+ \rangle_{\mathbf{H}_+} - \langle f, f_- \rangle_{\mathbf{H}_-} = \langle f, f \rangle_{K_{\text{discreg}}}.$$  

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

$$\min_{f \in \mathcal{K}} \left( \frac{1}{n} \sum_{i=1}^n \langle y_i, f(x_i) \rangle^2 + \langle f, J_f \rangle_{K_{\text{discreg}}} \right)$$

(18)

**Proposition 2.** The solution $f^*$ of the discriminability-driven regularization framework admits an expansion

$$f^* = \sum_{i=1}^n a_i^* \hat{K}_{\text{discreg}}(x_i, \cdot)$$

(19)

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

$$a^* = (\hat{K} + nI)^+ Y$$

(20)

$$a^* = [a_1^*, a_2^*, \ldots, a_n^*]^T \hat{K}$$

is the $n \times n$ kernel matrix, $\hat{K}_{ij} = \hat{K}_{\text{discreg}}(x_i, x_j)$. $Y$ is the label vector, $Y = [y_1, y_2, \ldots, y_n]^T$.

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

$$V(f, X) = \frac{1}{n} \sum_{i=1}^n \langle y_i, f(x_i) \rangle^2.$$

$$\Omega(f, J_f) = \langle f, J_f \rangle_{K_{\text{discreg}}}$$

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

$$f^* = \sum_{i=1}^n a_i^* \hat{K}_{\text{discreg}}(x_i, \cdot) = \sum_{i=1}^n a_i [K_+(x_i, \cdot) - K_-(x_i, \cdot)].$$

(21)

Substituting this form in (18), we arrive at the following differentiable objective function of the coefficient $a$

$$\text{arg min } 1/n(Y - Ka)^T(Y - Ka) + a^T Ka$$

(22)

The derivative w.r.t to $a$ of the objective function vanishes at the minimizer

$$1/n(-\hat{K})(Y - Ka) + \hat{K} a = 0,$$

which leads to the following solution

$$a^* = (\hat{K} + nI)^+ Y.$$

From the two propositions, it can be obviously seen that the discriminability-driven regularization framework naturally lies in the RKKS induced from the discriminative regularizer $R_{\text{discreg}}(f)$. Different from many indefinite kernel machines that artificially use the direct indefinite kernels instead of the PD kernels, the discriminability-driven regularization applies the PD kernels $K_+$ and $K_-$ defined in the two RKHSs $\mathbf{H}_+$ and $\mathbf{H}_-$ respectively to construct an indefinite kernel following the RKKS formulation. Especially, classification specific prior knowledge involved in the $S_w$ and $S_b$ in $R_{\text{discreg}}(f)$ is actually fused into the construction of the kernels by the particular weighted inner products (17) in the $\mathbf{H}_+$ and $\mathbf{H}_-$. 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 $\eta L_0 - (1-\eta)L_0$ to form the two RKHSs $\mathbf{H}_+$ and $\mathbf{H}_-$, and then constructs the PD kernels $K_+$ and $K_-$ in $\mathbf{H}_+$ and $\mathbf{H}_-$. respectively to form the indefinite kernel $\hat{K}_{\text{discreg}}$. 


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 (LaplSC) 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

$$X = \{x_1, \cdots, x_i, x_{i+1}, \cdots, x_n\} \subset \mathbb{R}^m,$$

where $x_i = \{(x_i, y_i)\}_i^c$ are labeled coming from c different classes, and $x_{i+n} = \{x_i\}_i^{n-1}$ are unlabeled. In terms of PCA, we firstly define the similar total scatter measure which reflects the global structure of the outputs

$$S_p = \sum_{i=1}^n \|f(x_i) - \frac{1}{n} \sum_{j=1}^n f(x_i)\|_2^2$$

$$= \frac{1}{2} w^T \sum_{i=1}^n \sum_{j=1}^n \phi_i(x_i - x_j)(x_i - x_j)^T w = w^T S_p w,$$

(24)

where $\phi_i = 1/n$.

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

$$A(f) = S_p = \sum_{i=1}^n \sum_{j=1}^n \|f(x_i) - \frac{1}{n} \sum_{j=1}^n f(x_j)\|_2^2$$

$$= \frac{1}{2} w^T \sum_{i=1}^n \sum_{j=1}^n \phi_i(x_i - x_j)(x_i - x_j)^T w,$$

(25)

$$B(f) = S_p = \sum_{i=1}^n \sum_{j=1}^n \|f(x_i) - \frac{1}{n} \sum_{j=1}^n f(x_j)\|_2^2$$

$$= \frac{1}{2} w^T \sum_{i=1}^n \sum_{j=1}^n \phi_i(x_i - x_j)(x_i - x_j)^T w.$$

(25)

Table 1
Pseudo-code for SupDR.

<table>
<thead>
<tr>
<th>Input:</th>
<th>The samples $(x_i, y_i)_{i=1}^n$;</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number $k$ of the nearest neighbors of $x_i$;</td>
<td></td>
</tr>
<tr>
<td>The regularization parameter $q_i$.</td>
<td></td>
</tr>
<tr>
<td>Output:</td>
<td>The coefficient $a_i$.</td>
</tr>
<tr>
<td>for $i = 1, \ldots, n$</td>
<td></td>
</tr>
<tr>
<td>$x_i^{l_k} \leftarrow$ k-th nearest neighbor of $x_i$ among $</td>
<td>x_i</td>
</tr>
<tr>
<td>for $i = 1, \ldots, n$</td>
<td></td>
</tr>
<tr>
<td>for $j = 1, \ldots, k$</td>
<td></td>
</tr>
<tr>
<td>if $y_i = y_j$</td>
<td></td>
</tr>
<tr>
<td>$W_{ii} \leftarrow 1$;</td>
<td></td>
</tr>
<tr>
<td>else</td>
<td></td>
</tr>
<tr>
<td>$W_{ij} \leftarrow 1$;</td>
<td></td>
</tr>
<tr>
<td>end</td>
<td></td>
</tr>
<tr>
<td>end</td>
<td></td>
</tr>
<tr>
<td>$D_{kk} = \sum_{j=1}^n W_{jj}$;</td>
<td></td>
</tr>
<tr>
<td>$D_{k} = \sum_{j=1}^n W_{kj} - D_{kk}$;</td>
<td></td>
</tr>
<tr>
<td>$L_k = D_{kk} - W_{kk}$;</td>
<td></td>
</tr>
<tr>
<td>$L_k \leftarrow D_k - W_k$;</td>
<td></td>
</tr>
<tr>
<td>$\Gamma_{kk} \leftarrow \text{Eigenvalue decomposition of } D_{kk}$;</td>
<td></td>
</tr>
<tr>
<td>$\Gamma_{k} \leftarrow \text{Eigenvalue decomposition of } D_{kk}$;</td>
<td></td>
</tr>
<tr>
<td>Construct the kernel functions $K_+$ and $K_-$ in $H_+$ and $H_-$ respectively;</td>
<td></td>
</tr>
<tr>
<td>$K_{\text{disc}} \leftarrow K_+ - K_- ; K_{ii} \leftarrow K_{\text{disc}}(x_i, x_j)$;</td>
<td></td>
</tr>
<tr>
<td>$Y = [y_1, y_2, \ldots, y_n]^T$;</td>
<td></td>
</tr>
<tr>
<td>$\alpha^* = (k + n)^{-1} Y$.</td>
<td></td>
</tr>
</tbody>
</table>

| $\Lambda_{kk} \leftarrow \sum_{j=1}^n W_{jj}$; |
| $\Lambda_{k} \leftarrow \sum_{j=1}^n W_{kj} - \Lambda_{kk}$; |
| $\Lambda_k \leftarrow D_{kk} - \Lambda_{kk}$; |
| $\Lambda_k \leftarrow D_k - \Lambda_k$; |
| $\Gamma_{kk} \leftarrow \text{Eigenvalue decomposition of } \Lambda_{kk}$; |
| $\Gamma_{k} \leftarrow \text{Eigenvalue decomposition of } \Lambda_{kk}$; |
| Construct the kernel functions $K_+$ and $K_-$ in $H_+$ and $H_-$ respectively; |
| $K_{\text{disc}} \leftarrow K_+ - K_- ; K_{ii} \leftarrow K_{\text{disc}}(x_i, x_j)$; |
| $Y = [y_1, y_2, \ldots, y_n]^T$; |
| $\alpha^* = (k + n)^{-1} Y$. |
\[\frac{1}{2}W^T \sum_{i=1}^{n} \sum_{j=1}^{n} \Phi^T_{ij}(x_i - x_j)(x_i - x_j)^T w.\]  

where

\[\Phi^T_{ij} = \begin{cases} 
1/n_i & \text{if } y_i = y_j = t \\
0 & \text{if } y_i \neq y_j
\end{cases}\\
\]  

\[\Phi^T_{ij} = \begin{cases} 
1/n_i - 1/n_i & \text{if } y_i = y_j = t \\
1/n_i & \text{if } y_i \neq y_j
\end{cases}\\
\]  

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

Similarly to LFDA, we further introduce the weight \(\Psi_{ij}\) of the data graph into the \(\Phi^T_{ij}\) and \(\Phi^T_{ij}\) as the coefficients, which describes the local discriminative structure of the outputs

\[S_{lw} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Psi_{ij}(x_i - x_j)(x_i - x_j)^T w = w^T S_{lw} w.\]  

\[S_{lb} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} x_i(x_i - x_j)(x_i - x_j)^T w = w^T S_{lb} w.\]  

where

\[\Psi_{ij} = \begin{cases} 
\Psi_{ij}/t & \text{if } y_i = y_j = t \\
0 & \text{if } y_i \neq y_j
\end{cases}\\
\]  

\[\Psi_{ij} = \begin{cases} 
\Psi_{ij}(1/t - 1/t) & \text{if } y_i = y_j = t \\
1/t & \text{if } y_i \neq y_j
\end{cases}\\
\]  

\[\Psi_{ij} = \begin{cases} 
\exp(-||x_i - x_j||^2/\alpha^2) & \text{if } x_i \in \text{ne}(x_j) \text{ or } x_j \in \text{ne}(x_i) \\
0 & \text{otherwise}
\end{cases}\\
\]  

The \(\Psi_{ij}\) 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, \(\Psi_{ij}\) 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 \(\Psi_{ij}\) 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 get the new definitions on \(A(f)\) and \(B(f)\) in SSL by bridging the measures characterized the global data structure and local discriminative structure. Concretely, on one hand, the total scatter matrix \(S_p\) is incorporated with \(S_b\) to form the new inter-class scatter matrix \(S_{lw} + S_{lb}\). Since \(S_{lb}\) would be unreliable in SSL in the case of the few labeled data, the global structure is expected to keep simultaneously [39]

\[B^{SSL}(f) = (1 - \gamma)S_{lb} + \gamma S_p = w^T [(1 - \gamma)S_{lb} + \gamma S_p] w = w^T S_{lw}^P w.\]  

where \(\gamma\) 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 \(S_p\) by the both labeled and unlabeled data as in SELF, thus it is more reliable and adding it in \(S_{lw}^P\) would improve the reliability of \(B^{SSL}(f)\).

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

\[A^{SSL}(f) = (1 - \gamma)S_{lw} + \gamma I w^T = w^T [(1 - \gamma)S_{lw} + \gamma I] w = w^T S_{lw}^I w.\]  

The final optimization objective of SemiDR can be formulated as

\[\min_{f, \phi} \left\{ \frac{1}{2} \sum_{i=1}^{l} (y_i - w^T x_i)^2 + w^T [\eta S_{lw}^I - (1 - \eta)S_{lw}] w \right\}.\]  

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

\[\eta S_{lw}^I - (1 - \eta)S_{lw},\]  

that is, the derivation of (13).

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

\[S = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(x_i - x_j)(x_i - x_j)^T,\]  

where \(W\) is some \(n \times n\) matrix, \(W = [W_{ij}]_{n \times n}\).

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

\[S = XLX^T\]  

Consequently, \(S_{lw}^I\) and \(S_{lw}^I\) can be reformulated as [39]

\[S_{lw}^I = XL_{lw}X^T = X[(1 - \gamma)L_{lw} + \gamma(X^T X)^+]X^T,\]  

\[S_{lb}^I = XL_{lb}X^T = X[(1 - \gamma)L_{lb} + \gamma L_{lb}]X^T,\]  

where \(L_{lw} = D - \Phi^T_{lw}, L_{lb} = D - \Phi^T_{lb}, L_{pb} = D - \Phi^T_{pb}\).

The corresponding eigenvalue decomposition is

\[\eta[(1 - \gamma)L_{lw} + \gamma(X^T X)^+] - (1 - \eta)(1 - \gamma)L_{lb} + \gamma L_{lb}] = U \Lambda U^T.\]  

where \(U\) and \(\Lambda\) are defined as before. The pseudo-code for SemiDR is shown in Table 2.

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pseudo-code for SemiDR.</strong></td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td>Input: Labeled samples (</td>
</tr>
<tr>
<td>Unlabeled samples (</td>
</tr>
<tr>
<td>The number (k) of the nearest neighbors of (x_i;)</td>
</tr>
<tr>
<td>The width (\sigma) in the weights (\Psi_{ij};)</td>
</tr>
<tr>
<td>The regularization parameters (\eta) and (\gamma;)</td>
</tr>
<tr>
<td>Output: The coefficient (w,)</td>
</tr>
<tr>
<td>for (i = 1, \ldots, l)</td>
</tr>
<tr>
<td>(x_i^{(k)} = \text{th nearest neighbor of } x_i \text{ among }</td>
</tr>
<tr>
<td>end</td>
</tr>
<tr>
<td>for (i, j = 1, \ldots, n)</td>
</tr>
</tbody>
</table>
| \(\Phi^T_{ij} = \begin{cases} 
\exp(-||x_i - x_j||^2/\alpha^2) & \text{if } x_i \in \text{ne}(x_j) \text{ or } x_j \in \text{ne}(x_i); \\
\Phi^T_{ij}(1/t - 1/t) & \text{if } y_i = y_j = t; \\
\exp(-||x_i - x_j||^2/\alpha^2) & \text{if } y_i \neq y_j; \\
0/t - 1/t & \text{if } y_i = y_j = t; \\
0 & \text{otherwise}
\end{cases}\\
\]  

end |
| end |
| \(\Phi^T_{ij} = \begin{cases} 
\exp(-||x_i - x_j||^2/\alpha^2) & \text{if } x_i \in \text{ne}(x_j) \text{ or } x_j \in \text{ne}(x_i); \\
\Phi^T_{ij}(1/t - 1/t) & \text{if } y_i = y_j = t; \\
\exp(-||x_i - x_j||^2/\alpha^2) & \text{if } y_i \neq y_j; \\
0/t - 1/t & \text{if } y_i = y_j = t; \\
0 & \text{otherwise}
\end{cases}\\
\]  

end |
| \(L_{lw} = D - \Phi^T_{lw}, L_{lb} = D - \Phi^T_{lb}, L_{pb} = D - \Phi^T_{pb};\) |
| \(U \Lambda U^T = \) Eigenvalue decomposition of |
| \(\eta[(1 - \gamma)L_{lw} + \gamma(X^T X)^+] - (1 - \eta)(1 - \gamma)L_{lb} + \gamma L_{lb}] = U \Lambda U^T;\) |
| \(\gamma;\) |
| \(\eta;\) |
| \(U \Lambda U^T;\) |
| Construct the kernel functions \(K_+\) and \(K_-\) in \(H_+\) and \(H_-\) respectively, |
| \(K_{\text{diag}} = K_+ - K_-;\) |
| \(K_{\text{diag}}(x, x);\) |
| \(\gamma = [\gamma_1, \ldots, \gamma_l, 0, \ldots, 0];\) |
| \(\gamma = (K + I)^T Y;\) |
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 ‘cdot’ 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 $\sigma$ is selected from the set {$2^{-10}, 2^{-9}, \ldots, 2^9, 2^{10}$} as the option of the regularization parameter $\lambda$ in IKRLSC. The regularization parameters $\eta$ and $\gamma$ in DRLSC and SupDR are chosen in the set {0,0.01,0.05,0.1,0.15,...,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 datasets.

![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%).](image1)

![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%).](image2)
The Attributes of the thirteen datasets in the IDA database.

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

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

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

database [44]. 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 SemiDR 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 \( 5,10,15,20 \). 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 [43], 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 SemiDR 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, 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 SSLj \( (i=1,2,...,7) \), 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
comparing Table 7 with 8, the linear kernel SemiDR sometimes remarkably excels LapRLSC, SSDR, SSDA and LGReg, whose accuracies of all algorithms greatly improve, but SemiDR still 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 exceed that of SemiDR_EK on all the datasets in the circumstance, LapRLSC, SSDR, SSDA and LGReg all fail in the linear kernel case. However, SemiDR performs much better. Its accuracy exceeds 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.

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifiers</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RBNN</td>
<td>RLSC</td>
</tr>
<tr>
<td>Banana</td>
<td>±1.07</td>
<td>±0.52</td>
</tr>
<tr>
<td>Breast-cancer</td>
<td>±4.31</td>
<td>±3.21</td>
</tr>
<tr>
<td>Diabetis</td>
<td>±1.54</td>
<td>±1.39</td>
</tr>
<tr>
<td>Flare-solar</td>
<td>±1.28</td>
<td>±2.12</td>
</tr>
<tr>
<td>German</td>
<td>±1.81</td>
<td>±1.56</td>
</tr>
<tr>
<td>Heart</td>
<td>±4.24</td>
<td>±1.89</td>
</tr>
<tr>
<td>Ringsnorm</td>
<td>±0.34</td>
<td>±0.23</td>
</tr>
<tr>
<td>Thyroid</td>
<td>±2.63</td>
<td>±2.56</td>
</tr>
<tr>
<td>Titanic</td>
<td>±1.43</td>
<td>±1.79</td>
</tr>
<tr>
<td>Twormnorm</td>
<td>±0.09</td>
<td>±0.08</td>
</tr>
<tr>
<td>Waveform</td>
<td>±0.38</td>
<td>±0.42</td>
</tr>
<tr>
<td>Image</td>
<td>±0.87</td>
<td>±0.23</td>
</tr>
<tr>
<td>Splice</td>
<td>±0.72</td>
<td>±0.43</td>
</tr>
</tbody>
</table>

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 redescriptions 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
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) diabetes, (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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classification</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSL1_10</td>
<td>RLSC 74.24*</td>
<td>SSDA 77.57*</td>
</tr>
<tr>
<td></td>
<td>± 6.28</td>
<td>± 6.32</td>
</tr>
<tr>
<td>SSL2_10</td>
<td>75.22*</td>
<td>81.12*</td>
</tr>
<tr>
<td></td>
<td>± 1.37</td>
<td>± 1.53</td>
</tr>
<tr>
<td>SSL3_10</td>
<td>55.13*</td>
<td>58.90*</td>
</tr>
<tr>
<td></td>
<td>± 4.51</td>
<td>± 3.76</td>
</tr>
<tr>
<td>SSL4_10</td>
<td>55.08*</td>
<td>53.33*</td>
</tr>
<tr>
<td></td>
<td>± 3.13</td>
<td>± 4.35</td>
</tr>
<tr>
<td>SSL5_10</td>
<td>56.52*</td>
<td>51.43*</td>
</tr>
<tr>
<td></td>
<td>± 2.88</td>
<td>± 1.63</td>
</tr>
<tr>
<td>SSL6_10</td>
<td>33.00*</td>
<td>17.90*</td>
</tr>
<tr>
<td></td>
<td>± 4.42</td>
<td>± 3.10</td>
</tr>
<tr>
<td>SSL7_10</td>
<td>58.79*</td>
<td>60.59*</td>
</tr>
<tr>
<td></td>
<td>± 4.69</td>
<td>± 3.93</td>
</tr>
<tr>
<td>SSL1_100</td>
<td>90.43*</td>
<td>92.51*</td>
</tr>
<tr>
<td></td>
<td>± 1.42</td>
<td>± 1.46</td>
</tr>
<tr>
<td>SSL2_100</td>
<td>86.12*</td>
<td>88.17</td>
</tr>
<tr>
<td></td>
<td>± 1.28</td>
<td>± 1.28</td>
</tr>
<tr>
<td>SSL3_100</td>
<td>69.96*</td>
<td>83.04*</td>
</tr>
<tr>
<td></td>
<td>± 2.24</td>
<td>± 1.69</td>
</tr>
<tr>
<td>SSL4_100</td>
<td>73.71*</td>
<td>74.04*</td>
</tr>
<tr>
<td></td>
<td>± 2.86</td>
<td>± 3.14</td>
</tr>
<tr>
<td>SSL5_100</td>
<td>70.83*</td>
<td>57.09*</td>
</tr>
<tr>
<td></td>
<td>± 1.20</td>
<td>± 2.00</td>
</tr>
<tr>
<td>SSL6_100</td>
<td>63.96*</td>
<td>19.76*</td>
</tr>
<tr>
<td></td>
<td>± 2.91</td>
<td>± 3.95</td>
</tr>
<tr>
<td>SSL7_100</td>
<td>70.64*</td>
<td>76.89</td>
</tr>
<tr>
<td></td>
<td>± 2.25</td>
<td>± 1.51</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classification</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSL1_10</td>
<td>RLSC 78.34*</td>
<td>SSDA 78.48*</td>
</tr>
<tr>
<td></td>
<td>± 5.95</td>
<td>± 6.74</td>
</tr>
<tr>
<td>SSL2_10</td>
<td>82.03*</td>
<td>81.73*</td>
</tr>
<tr>
<td></td>
<td>± 2.21</td>
<td>± 1.44</td>
</tr>
<tr>
<td>SSL3_10</td>
<td>63.87*</td>
<td>69.84*</td>
</tr>
<tr>
<td></td>
<td>± 3.26</td>
<td>± 5.60</td>
</tr>
<tr>
<td>SSL4_10</td>
<td>61.42*</td>
<td>56.46*</td>
</tr>
<tr>
<td></td>
<td>± 3.15</td>
<td>± 2.59</td>
</tr>
<tr>
<td>SSL5_10</td>
<td>59.39*</td>
<td>55.40*</td>
</tr>
<tr>
<td></td>
<td>± 2.70</td>
<td>± 4.23</td>
</tr>
<tr>
<td>SSL6_10</td>
<td>35.71*</td>
<td>35.80*</td>
</tr>
<tr>
<td></td>
<td>± 3.85</td>
<td>± 3.65</td>
</tr>
<tr>
<td>SSL7_10</td>
<td>61.24*</td>
<td>57.44*</td>
</tr>
<tr>
<td></td>
<td>± 3.14</td>
<td>± 4.46</td>
</tr>
<tr>
<td>SSL1_100</td>
<td>94.60*</td>
<td>94.98</td>
</tr>
<tr>
<td></td>
<td>± 4.02</td>
<td>± 1.36</td>
</tr>
<tr>
<td>SSL2_100</td>
<td>90.95*</td>
<td>91.70*</td>
</tr>
<tr>
<td></td>
<td>± 1.33</td>
<td>± 1.36</td>
</tr>
<tr>
<td>SSL3_100</td>
<td>80.14*</td>
<td>82.89*</td>
</tr>
<tr>
<td></td>
<td>± 1.23</td>
<td>± 0.87</td>
</tr>
<tr>
<td>SSL4_100</td>
<td>76.00*</td>
<td>72.71*</td>
</tr>
<tr>
<td></td>
<td>± 2.20</td>
<td>± 2.35</td>
</tr>
<tr>
<td>SSL5_100</td>
<td>75.11*</td>
<td>77.44*</td>
</tr>
<tr>
<td></td>
<td>± 3.04</td>
<td>± 1.83</td>
</tr>
<tr>
<td>SSL6_100</td>
<td>78.44*</td>
<td>78.34</td>
</tr>
<tr>
<td></td>
<td>± 2.49</td>
<td>± 2.50</td>
</tr>
<tr>
<td>SSL7_100</td>
<td>74.54*</td>
<td>76.38*</td>
</tr>
<tr>
<td></td>
<td>± 2.46</td>
<td>± 2.21</td>
</tr>
</tbody>
</table>

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
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 pre-definition 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 Doctoral 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


