Discriminatively Regularized Least-Squares Classification

Hui Xue[1] Songcan Chen[1*] Qiang Yang[2]

[1] *Department of Computer Science and Engineering, Nanjing University of Aeronautics & Astronautics, 210016, Nanjing, P.R. China*

2 Department of Computer Science and Engineering, Hong Kong University of Science & Technology, Hong Kong

Abstract:

Over the past decades, regularization theory is widely applied in various areas of machine learning to derive a large family of novel algorithms. Traditionally, regularization focuses on smoothing only, and does not fully utilize the underlying *discriminative* knowledge which is vital for classification. In this paper, we propose a novel regularization algorithm in the least-squares sense, called Discriminatively Regularized Least-Squares Classification (DRLSC) method, which is specifically designed for classification. Inspired by several new geometrically motivated methods, DRLSC directly embeds the discriminative information as well as the local geometry of the samples into the regularization term so that it can explore as much underlying knowledge inside the samples as possible and aim to maximize the margins between the samples of different classes in each local area. Furthermore, by embedding equality type constraints in the formulation, the solutions of DRLSC can follow from solving a set of linear equations and the framework naturally contains multi-class problems. Experiments on both toy and real-world problems demonstrate that DRLSC is often superior in classification performance to the classical regularization algorithms, including Regularization Networks, Support Vector Machines and some of the recent studied Manifold Regularization techniques.

Keywords: Classifier design; Discriminative information; Manifold learning; Pattern recognition

1. Introduction

[*] Corresponding author: Tel: +86-25-84896481 Ext. 12106; Fax: +86-25-84498069; E-mail: s.chen@nuaa.edu.cn (S. Chen), xuehui@nuaa.edu.cn (H. Xue) and qyang@cse.ust.hk (Q. Yang)

Regularization methods for machine learning have made great progress recently. Such methods have been extended to several subareas of machine learning, including regression, clustering and classification[1-9].

A related area under extensive development is the manifold learning area, where methods have been developed to take advantage of the locality information while performing dimensionality reduction. In this area, Belkin et al.[5, 10] further introduced the underlying sample distribution information of the data with manifold structures into the traditional regularization, resulting in Manifold Regularization (MR), which aims to retain the manifold structure of the samples in each given class. In the framework of MR, two regularization terms are introduced: one controls the complexity of the classifier, and the other controls the complexity measured by the manifold geometry of the sample distribution[5].

However, when focusing on classification problems, we notice that each of the above methods alone suffers from some deficiencies. First, although the traditional regularization methods have been widely applied to the classifier design, it is essentially derived from multivariate functional fitting or regression problems instead of classification problems[2, 11-13]. It constructs the regularization term by focusing more on the smoothness of the function. However, in classification, similar inputs near the discriminant boundaries are more likely to belong to different classes, implying that just a smoothness constraint may not be sufficient for discrimination among classes. In particular, a classifier may not be always smooth everywhere, especially when we are near the boundaries between classes. Furthermore, the primary goal of classification is to separate the samples of different classes in the output space as far as possible. Hence, the underlying *discriminative* information is crucial for classification. However, since the regularization terms of the traditional regularization methods do not inject more underlying class information in a classifier's design, they may not incorporate all the useful discriminative information for classification.

Second, although MR performs well in semi-supervised learning such as sensor networks[14], for supervised learning, MR suggests constructing a graph or Laplacian matrix for each class, which results in an equal number of the regularization terms that is the same as the number of classes. As a result, dependency on the number of given classes makes MR difficult to scale well. The algorithm may perform badly in cases of small number of classes

(e.g., three or so classes), whereas the computational complexity in the training phase of MR will increase sharply, because making an optimal tuning for the many regularization parameters is impractical.

In this paper, we propose a novel method for classification that, by the well-known "No Free Lunch" Theorem[15], integrates as much underlying knowledge inside the samples as possible, including the discriminative and geometrical information, into a unified regularization framework. We call our method DRLSC, which stands for Discriminatively Regularized Least-Squares Classification. By making the best of the underlying discriminative information rather than only emphasizing the smoothness of the classifier in the traditional regularization methods, DRLSC introduces a new discriminative regularization term in the framework. Furthermore, inspired by the new supervised dimensionality reduction methods, DRLSC also uses two graphs to characterize the intra-class compactness and inter-class separability respectively, and thus can further maximize the margins between the samples of the different classes in each local area. DRLSC integrates the underlying *discriminative* and *geometrical* information into a single regularization term. A major advantage is that it can scale well with the number of the classes. In addition, by introducing the equality constraints in the formulation, the solutions of DRLSC can be found by solving a set of linear equations, which makes the algorithm simpler and more stable. Experiments are conducted to demonstrate the superiority of our DRLSC algorithm compared well with the state-of-the-art regularization methods such as Regularization Networks (RN), Generalized Radial Basis Function Networks (GRBFN), Support Vector Machines (SVM), Least Squares Support Vector Machines (LS-SVM) and Manifold Regularization (MR).

The rest of the paper is organized as follows. Section 2 introduces the related works in regularization. Our contributions are simply described in Section 3. Section 4 presents the proposed Discriminatively Regularized Least-Squares Classification. The analytic solution to DRLSC is derived in Section 5. In Section 6, the experiment analysis is given. Some conclusions are drawn in Section 7.

2. Related Works

Ill-posed problems widely exist in science and engineering regions, which denotes that

given the available input samples, the solution to the problem is nonunique or unstable[2, 16]. Early in the 1960's, Tikhonov had proposed a classical method named *Regularization* to solve these problems[17-19]. By incorporating the right amount of prior information into the formulation, the regularization techniques have been shown to be powerful in making the solution stable[2, 16]. In the past few decades, the regularization theory was introduced to the machine learning community on the premise that the learning can be viewed as a multivariate functional fitting problem[2, 11-13]. Consequently, in the classical Tikhonov regularization, the most common form of prior information involves the assumption that the input-output mapping function, i.e. the solution to the fitting problem, is *smooth*[16]

$$\min_{f \in F} \left\{ \frac{1}{2} \sum_{i=1}^{N} V\left(y_i, f(\mathbf{x}_i)\right) + \frac{\lambda}{2} \|Df\|^2 \right\} \tag{1}$$

where $V(y_i, f(\mathbf{x}_i))$ is the loss function, which indicates the penalty we pay when we see $\mathbf{x}_i$, predict $f(\mathbf{x}_i)$, and the true value is y_i [7]. In the regularization term, D is a linear differential operator that is applied to the function f, in which the prior information about the form of the solution is embedded[16]. D is also referred to as a stabilizer because the smoothness prior involved in it makes the solution stable[2, 16]. Moreover, the regularization parameter λ controls the trade-off between fitting the training samples and the roughness of the solution[2, 7].

Tikhonov[16] presented that when the loss function is designated to be the simple square-loss function

$$V\left(y, f(\mathbf{x})\right) = \left(y - f(\mathbf{x})\right)^2 \tag{2}$$

the solution $f_\lambda(\mathbf{x})$ to the Tikhonov regularization problem can be represented as a linear combination of the Green's function[16]

$$f_\lambda(\mathbf{x}) = \sum_{i=1}^{N} w_i G(\mathbf{x}, \mathbf{x}_i) \tag{3}$$

Poggio and Girosi[11, 12] showed that a regularization algorithm for learning is equivalent to a multilayer neural network with the Green's function as the activation function, resulting in the RN. Haykin[16] indicated that if we select a multivariate Gaussian function as the Green's function, the solution by RN will be an optimal interpolant in the sense that it

minimizes the Tikhonov regularization formula. GRBFN is an approximation of the RN, for its number of the hidden units is typically less than that of the RN's, which is equivalent to the number of the training samples.

In the classical regularization theory, a recent trend in studying the smoothness of the function is to put the function into the Reproducing Kernel Hilbert Space (RKHS)[6, 20], which has been well developed in several areas[2]. In the RKHS, the Tikhonov minimization problem can be rewritten as[21]:

$$\min_{f\in H}\left\{\frac{1}{2}\sum_{i=1}^{N}V\left(y_i, f(\mathbf{x}_i)\right)+\frac{\lambda}{2}\|f\|_K^2\right\} \tag{4}$$

Following the so-call Representer Theorem[20, 22, 23], under very general conditions on the loss function V, the minimizer of (4) will have the form

$$f(\mathbf{x})=\sum_{i=1}^{N}c_i K(\mathbf{x},\mathbf{x}_i) \tag{5}$$

Corresponding to different selections of V, the classical Tikhonov regularization method can be used to derive a large family of the state-of-the-art algorithms in machine learning. When selecting V as the square-loss function, we obtain Regularized Least-Squares Classification (RLSC)[21]. Similarly，we can obtain SVM[1, 24] by choosing V to be the hinge-loss function defined as

$$V(y, f(\mathbf{x}))=\begin{cases}0 & if\ yf(\mathbf{x})\geq 1\\ 1-yf(\mathbf{x}) & otherwise\end{cases} \tag{6}$$

Specifically, if we introduce error terms into the hinge-loss function and consider the equality constraints instead of inequalities in SVM, we obtain the Least Squares Support Vector Machines (LS-SVM) with the formulation in the least-square sense[25]. Though introducing dissimilar loss functions, these regularization algorithms have many inherently similar properties. Evgeniou et al.[26] described a unified framework for RN and SVM. Rifkin[21] indicated that RLSC empirically performs as well as SVM.

Although traditional regularization has been widely applied to the classifier design, it focuses more on the smoothness of the classification function owing to the essential derivation from ill-posed multivariate functional fitting problems as we mentioned above, to enforce the constraint that similar inputs correspond to similar outputs. This constraint is

natural for regression problems. But, it also seems to be too general for classification. Since the regularization terms of the traditional regularization methods do not inject more underlying class information, they may not incorporate all the useful discriminative information for classification.

The well-known "No Free Lunch" Theorem[15] indicates that, there is no pattern classification method that is inherently superior to any other, or even to random guessing. It is the type of problem, underlying information and the amount of training samples that determines the form of classifier to apply. Along this line, Belkin et al.[5, 10] further introduced the underlying sample distribution information of the data with manifold structures into the traditional regularization, resulting in Manifold Regularization (MR)

$$\min_{f\in H}\left\{\frac{1}{N}\sum_{i=1}^{N}V\left(y_i, f(\mathbf{x}_i)\right)+\gamma_A\left\|f\right\|_K^2+\gamma_I\left\|f\right\|_I^2\right\} \tag{7}$$

where the regularization term $\left\|f\right\|_K^2$ controls the complexity of the classifier and the other regularization term $\left\|f\right\|_I^2$ controls the complexity measured by the manifold geometry of the sample distribution[5]. MR naturally contains two extensions of RLSC and SVM, termed as Laplacian RLSC[5] (LapRLSC) and Laplacian SVM (LapSVM)[5] respectively. Though MR performs better in semi-supervised learning such as sensor networks[14], for fully supervised learning, MR suggests constructing a graph for each class, i.e. different $\left\|f\right\|_I^2$ corresponding to different classes, which undoubtedly leads to the appearance of many free regularization parameters in the formulation, especially for the multi-class problems. As a result, the computational complexity in training of MR will increase sharply.

In the previous works of dimensionality reduction, some newly geometrically motivated approaches have been designed to address the issue of utilizing underlying information fully. These approaches aimed to discover the manifold structures inherent in the given data which are embedded in high-dimensional Euclidean space. Isometric Feature Mapping (ISOMAP)[27], Locally Linear Embedding (LLE)[28] and Laplacian Eigenmap[29], are all variations of these nonlinear dimensionality reduction methods. Neighborhood Preserving Embedding (NPE)[30] and Locality Preserving Projection (LPP)[31] are respectively the linear versions for LLE and Laplacian Eigenmap, which seek to the transform matrices

instead of directly computing the embedding in the nonlinear versions, and thus are more easily applied to new samples.

However, these dimensionality reduction algorithms are unsupervised in nature and can not be used to discover the discriminant structure in the data[32]. In contrast, supervised dimensionality reduction can be applied, including Marginal Fisher Analysis (MFA)[33] and Local Discriminant Embedding (LDE)[34] that both exploit supervised class information to construct two graphs which are respectively used to characterize the intra-class compactness and inter-class separability of the samples. Locality Sensitive Discriminant Analysis (LSDA)[32] further adds an adjustable parameter to the former to balance the relative importance of the two graphs. However, these methods have not been applied to regularization techniques to date.

3. Contributions

In this paper, we propose a novel regularization method in the least-squares sense in the context of classification, DRLSC. Different from the recent regularization methods, DRLSC directly introduces *the underlying discriminative information* into the regularization term. On one hand, it minimizes the empirical loss between the desired and actual outputs. On the other hand, it minimizes the intra-class compactness of the outputs while simultaneously maximizing the inter-class separation. We show that this method is more likely suitable for classification as compared to the previous methods in regularization. Furthermore, we also introduce *the underlying geometrical information* into DRLSC to make the full use of available knowledge. Different from MR, we expect to utilize as much discrimination information characterized by the manifold structure of the given data as possible to guide the design of classifiers. To our best knowledge, there is no previously known method with a similar approach on this problem. That is, with the direction of the discriminative information, DRLSC constructs two graphs to characterize the intra-class compactness and inter-class separability respectively whose geometric structures are intended to be also reflected in the output space of classifier to be designed. We show that with this approach, we can further maximize the margins between samples of different classes in each local area in the output space[32].

In summary, the contributions of our approach are as follows:

- Different from traditional regularization, DRLSC directly introduces the discriminative information into the regularization term. DRLSC is currently designed for classification in this paper, but can be extended to other areas such as clustering in the future works. For the primary goal of classification is to separate the samples belonging to the different classes, DRLSC pays more attention to the underlying discriminative information than the general smoothness assumption on the classifier in the previous methods.
- Different from MR, DRLSC further introduces the local manifold structure directly into the new discriminative regularization term instead of several additional ones. It only has one adjustable regularization parameter in the formulation, even for the multi-class problems, which makes it easier to optimize. Thus DRLSC effectively avoids the potential "Curse of Dimensionality" of MR in the optimization. Following the "No Free Lunch" Theorem, DRLSC considers the underlying discriminative information as well as the geometric structure, and thus can be potentially better in classification performance.
- Different from many existing learning machines, such as classical SVM, which decompose the multi-class problems into multiple binary-class classification problems, DRLSC can naturally contain binary-class and multi-class problems in a unified framework in terms of the simple analytic solution, except only a subtle difference in the formulation.
- The framework of DRLSC covers many feasible approaches to further improve regularization. Table 1 shows a common classification for the most popular regularization algorithms from the viewpoint of the loss function, the regularization term and the dependence on the number of classes of the samples, where $R_{disreg}(f,\eta)$ denotes the new discriminative regularization term in our proposed DRLSC. As we can see, DRLSC has the most compact form among all these algorithms. It needs neither the smoothness term $\|f\|_K^2$ nor the manifold term $\|f\|_I^2$. Furthermore, it can solve the multi-class problems in the binary-class complexity[35] and thus is independent of the number of classes. The framework of DRLSC is general. From the table, we can see that with different combinations of the loss function and regularization term, we can obtain

different regularization algorithms. Consequently, once we integrate $R_{disreg}(f,\eta)$ with other loss functions or regularization terms, we can immediately gain a large family of new regularization learning algorithms, which extends our scopes and deserves future study.

Table 1. The common classification from the viewpoint of the loss function, the regularization term and the dependence on the number of classes of the samples for the most popular regularization algorithms

Regularization	Loss Function		Regularization Term			Dependence on the number of classes
	Square-Loss Function	Hinge-Loss Function	$\|f\|_K^2$	$\|f\|_I^2$	$R_{disreg}(f,\eta)$	
RLSC	✓		✓			
LapRLSC	✓		✓	✓		✓
SVM		✓	✓			
LS-SVM	✓	✓	✓			
LapSVM		✓	✓	✓		✓
DRLSC	✓				✓	

4. Discriminatively Regularized Least-Squares Classification (DRLSC)

Suppose that we are given the training samples

$$(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N) \in \mathcal{X} \times \{C_1, \cdots, C_c\} \tag{8}$$

where the domain $\mathcal{X} \in R^n$ is some nonempty set that the pattern $\boldsymbol{x}_i$ are taken from, and the y_is are class labels. The task of classification is, by studying the training samples, to obtain a classifier which can separate samples of different classes in the output space as far as possible. For this general purpose, DRLSC aims to directly embed the underlying discriminative information in the regularization term

$$\min_{f \in F} \left\{ \frac{1}{2} \sum_{i=1}^{N} [y_i - f(\mathbf{x}_i)]^2 + \frac{1}{2} R_{disreg}(f,\eta) \right\} \tag{9}$$

where $R_{disreg}(f,\eta)$ is the new discriminative regularization term in DRLSC.

A general definition for $R_{disreg}(f,\eta)$ can be given by

$$R_{disreg}(f,\eta) = \eta A(f) - (1-\eta) B(f) \tag{10}$$

$A(f)$ and $B(f)$ are the metrics defined in the output space, which measure the intra-class

compactness and inter-class separability of the outputs respectively. η is the parameter that regulates the relative significance of the intra-class compactness versus the inter-class separability, $0 \le \eta \le 1$. It is noteworthy that $R_{disreg}(f,\eta)$ can no longer be guaranteed its nonnegativity, which does not accord with the traditional regularization requirement. However, we still abuse the terminology, i.e. regularization, to name it as the discriminative regularization term.

The common thought of defining *A*(*f*) and *B*(*f*) is the generalized variance in statistics, which is similar to Maximum Margin Criterion (MMC)[36] but defined in the output space. In order to differentiate from DRLSC, we call the corresponding method as DRGV. That is,

$$A(f) = S_w = \sum_{k=1}^{c} \frac{1}{N_k} \sum_{i=1}^{N_k} \left\| f(\mathbf{x}_i^{(k)}) - \frac{1}{N_k} \sum_{j=1}^{N_k} f(\mathbf{x}_j^{(k)}) \right\|^2 \tag{11}$$

where N_k is the number of the samples $\mathbf{x}_i^{(k)}$ belonging to the *k*th class, $k = 1, \cdots, c$, and $\|\cdot\|$ is chosen to be L_2 norm throughout the paper. Similarly,

$$B(f) = S_b = \sum_{k=1}^{c} \sum_{l \ne k} \left\| \frac{1}{N_k} \sum_{i=1}^{N_k} f(\mathbf{x}_i^{(k)}) - \frac{1}{N_l} \sum_{j=1}^{N_l} f(\mathbf{x}_j^{(l)}) \right\|^2 \tag{12}$$

Here, a small S_w implies that every class has a small scatter; meanwhile, a large S_b implies that the class mean vectors scatter in a large space[36]. However, these variances focus on the global class relationship between the samples and thus fail to sufficiently characterize the local manifold structure of the data. Lafon et al.[37] indicated that most of the samples in real-world are highly correlated, at least locally, or equivalently, that the samples distribute in a low intrinsic manifold. Thus, following the "No Free Lunch" Theorem, we present Discriminatively Regularized Least-Squares Classification (DRLSC), which considers the underlying discriminative information as well as the geometric structure of the samples. By further embedding the geometry of the samples into the discriminative regularization term, DRLSC aims to reflect the intrinsic neighbor relations of the samples with the same class labels while separate the nearby samples with the different labels far from each other in the output space. As a result, DRLSC can further maximize the margins between the samples of different classes in each local neighborhood.

For the given training samples, we can build a nearest neighbor graph *G* to model the

intrinsic local geometrical structure. Specifically, for each sample $\boldsymbol{x}_i$, we first seek for its k nearest neighbors $ne(i)=\left\{\mathbf{x}_i^1,\cdots,\mathbf{x}_i^k\right\}$, and then put an edge between $\boldsymbol{x}_i$ and its neighbors. Thus, the weight matrix of G can be defined as follows[32]:

$$W_{ij}=\begin{cases}1 & if\ \mathbf{x}_j\in ne(i)\ or\ \mathbf{x}_i\in ne(j)\\ 0 & otherwise\end{cases} \tag{13}$$

By the well-known spectral graph theory[38], the nearest neighbor graph G with the weight matrix W characterizes the local geometry of the sample manifold. However, only one overall graph can not sufficiently reflect the discriminative structure of the samples. Hence, inspired by the new graph-based supervised dimensionality reduction methods, such as MFA, LDE and LSDA, we also construct two graphs in the input space, i.e. intra-class graph G_w and inter-class graph G_b, whose geometric structures are intended to be also reflected in the output space in which the classifier is designed.

For each sample $\boldsymbol{x}_i$, in terms of LSDA, we first divide the nearest neighborhood $ne(i)$ into two nonoverlapping subsets

$$ne_w(i)=\left\{\mathbf{x}_i^j\,\middle|\ if\ \ \mathbf{x}_i^j\ \ and\ \ \mathbf{x}_i\ \ belong\ \ to\ \ same\ \ class,\ \ 1\le j\le k\right\}$$

$$ne_b(i)=\left\{\mathbf{x}_i^j\,\middle|\ if\ \ \mathbf{x}_i^j\ \ and\ \ \mathbf{x}_i\ \ belong\ \ to\ \ different\ \ classes,\ \ 1\le j\le k\right\}$$

Then we define the two weight matrices of G_w and G_b respectively

$$W_{w,ij}=\begin{cases}1 & if\ \mathbf{x}_j\in ne_w(i)\ or\ \mathbf{x}_i\in ne_w(j)\\ 0 & otherwise\end{cases} \tag{14}$$

$$W_{b,ij}=\begin{cases}1 & if\ \mathbf{x}_j\in ne_b(i)\ or\ \mathbf{x}_i\in ne_b(j)\\ 0 & otherwise\end{cases} \tag{15}$$

Obviously, the nearest neighbor graph G is the combination of the intra-class graph G_w and the inter-class graph G_b.

The goal of DRLSC is to keep the neighboring samples of G_w stay as close as possible while the connected samples of G_b stay as far as possible in the output space. Hence, we redefine the measure to characterize the intra-class compactness from the intra-class graph

$$\tilde{S}_w=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|f(\mathbf{x}_i)-f(\mathbf{x}_j)\right\|^2 W_{w,ij} \tag{16}$$

Likewise, the measure of characterizing the inter-class separability from the inter-class

graph is also redefined as:

$$\tilde{S}_b = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|f(\mathbf{x}_i) - f(\mathbf{x}_j)\right\|^2 W_{b,ij} \tag{17}$$

By embedding $\tilde{S}_w$ and $\tilde{S}_b$ into the discriminative regularization term as the measures $A(f)$ and $B(f)$, we arrive at the following new optimization problem in DRLSC

$$\min_{f\in F}\left\{\frac{1}{2}\sum_{i=1}^{N}\left[y_i - f(\mathbf{x}_i)\right]^2 + \frac{1}{2}\left[\eta\tilde{S}_w - (1-\eta)\tilde{S}_b\right]\right\} \tag{18}$$

Note that the discriminative regularization term in DRLSC is actually a difference of two convex functions. Consequently, the optimization problem (18) is equivalent to

$$\min_{f\in F}\left\{\frac{1}{2}\left\{\sum_{i=1}^{N}\left[y_i - f(\mathbf{x}_i)\right]^2 + \eta\tilde{S}_w\right\} - \frac{1}{2}(1-\eta)\tilde{S}_b\right\}$$

which can be viewed as a special convex difference optimization. Recently, many researchers have devoted to seek for mathematical methods to solve the convex difference optimization problem[39, 40], which generally involve iterative processes and converge to a local minimum. Fortunately, by introducing the equality constraints in the formulation, the optimization problem of DRLSC can be solved analytically. We will discuss the analytical solutions in details in the next section.

To gain more insight into (18), we simply assume that the classifier has a linear form

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{19}$$

By simple algebra formulation, (16) can be reduced to

$$\begin{aligned}\tilde{S}_w &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[f(\mathbf{x}_i) - f(\mathbf{x}_j)\right]^2 W_{w,ij}\\ &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right)^2 W_{w,ij}\\ &= \mathbf{w}^T\mathbf{X}(\mathbf{D}_w - \mathbf{W}_w)\mathbf{X}^T\mathbf{w}\\ &= \mathbf{w}^T\mathbf{X}\mathbf{L}_w\mathbf{X}^T\mathbf{w}\end{aligned} \tag{20}$$

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

Likewise, (17) can be reformulated as:

$$\begin{aligned}\tilde{S}_b &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[f(\mathbf{x}_i)-f(\mathbf{x}_j)\right]^2 W_{b,ij} \\ &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\mathbf{w}^T\mathbf{x}_i-\mathbf{w}^T\mathbf{x}_j\right)^2 W_{b,ij} \\ &= \mathbf{w}^T\mathbf{X}(\mathbf{D}_b-\mathbf{W}_b)\mathbf{X}^T\mathbf{w} \\ &= \mathbf{w}^T\mathbf{X}\mathbf{L}_b\mathbf{X}^T\mathbf{w} \end{aligned} \tag{21}$$

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

Therefore, we can further reformulate the objective function (18) as:

$$\min_{\mathbf{w},b}\left\{\frac{1}{2}\sum_{i=1}^{N}\left[y_i-\left(\mathbf{w}^T\mathbf{x}_i+b\right)\right]^2+\frac{1}{2}\mathbf{w}^T\mathbf{X}\left[\eta\mathbf{L}_w-(1-\eta)\mathbf{L}_b\right]\mathbf{X}^T\mathbf{w}\right\} \tag{22}$$

Different from the traditional regularization methods that we have surveyed in Section 2, DRLSC embeds not only the underlying discriminative information but also the geometrical structure of the samples in the construction of the regularization term. The new discriminative regularization term focuses on a trade-off of the relative significance between the intra-class compactness and the inter-class separability in each local neighborhood. By the "No Free Lunch" Theorem, these should lead to a further improvement in the classification performance of regularization. Furthermore, DRLSC more likely provides us a brand-new viewpoint to combine regularization with supervised dimensionality reduction methods effectively. The general goal of supervised dimensionality reduction methods, such as Maximum Margin Criterion (MMC)[36], Marginal Fisher Analysis (MFA)[33], Local Discriminant Embedding (LDE)[34] and Locality Sensitive Discriminant Analysis (LSDA)[32], is to find an orientation for which the projected samples are well separated[15]. This is much similar to the intuitive motivation in our proposed DRLSC. Hence, in DRLSC, we construct the new discriminative regularization term referring to these new dimensionality reduction methods to some extent. Due to the generality of DRLSC learning framework, any similar supervised dimensionality reduction methods can be also embedded in DRLSC as the regularization term. Through such incorporation with these methods, the designed classifier

can more likely further separate the samples effectively in the output space as shown in our experiments below.

5. Analytic Solutions to DRLSC

Many traditional regularization methods usually solve the optimization problems by using conjugate gradient algorithms. However, these algorithms generally converge slowly and sometimes can not guarantee a convergence to the global optimum. Here, just as LS-SVM[25], we introduce equality constraints for DRLSC with the formulation in least-squares sense. Consequently, the solution can follow directly from solving a set of linear equations instead of the iterative processes in the common convex difference optimizations, which makes DRLSC simpler and more stable. Furthermore, many existing learning machines, such as classical SVM, were developed for binary-class classification and thus generally decompose the multi-class problem into multiple binary-class classification problems, by using the strategies such as one-against-one, one-against-all[41]. However, this is computationally expensive due to the large number of binary classifiers that are to be trained for handling each binary sub-problem[42] and especially in the one-against-all strategy, an imbalance class problem is unavoidably incurred. Different from these methods, DRLSC can simultaneously solve both binary-class and multi-class problems in a unified framework in terms of the simple analytic solution, except only a subtle difference between binary-class and multi-class in the formulation. So, in this section, we will simply discuss the implementation of DRLSC.

Without loss of generalization, we first consider the linear version to DRLSC. For the nonlinear version, we explicitly map the samples into the Empirical Feature Space[43-45]. Xiong et al.[43] indicated that, the comparison between the explicit map and the classical implicit map[1] shows that the former is easier to access and to study the adaptability of a kernel to the input samples than the latter. In fact, Kernel Principal Component Analysis (KPCA)[46] and Kernel Linear Discriminant Analysis (KLDA)[46] both *essentially* first map the samples into a empirical feature space, and then implement PCA and LDA in the feature space respectively. Furthermore, after mapping by the empirical kernel, the classification function will have the same form just as in the linear version. Therefore, here we uniformly

assume that the classifier has the linear form below

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$$

We firstly discuss the binary-class case. Let the class labels $y_i \in \{-1,1\}$, $i = 1, \cdots, N$. We introduce the equality constraints into DRLSC by reformulating the optimization problem (18) as

$$\min_{\mathbf{w},b} \left\{ \frac{1}{2} \sum_{i=1}^{N} e_i^2 + \frac{1}{2} \left[\eta \tilde{S}_w - (1-\eta) \tilde{S}_b \right] \right\}$$

subject to

$$\mathbf{w}^T \mathbf{x}_i + b + e_i = y_i, \quad \forall i = 1, \cdots, N \tag{23}$$

Theorem 1. (Binary-class case) Given the parameter $\eta \in [0,1]$, the minimum norm solution to the problem (23) is characterized by the linear system with the variables $\boldsymbol{\gamma} \in R^N$

$$\begin{bmatrix} 0 & \mathbf{1}_N^T \\ \mathbf{1}_N & \boldsymbol{\Omega}_\eta + \mathbf{I}_N \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{y} \end{bmatrix} \tag{24}$$

Proof: The Lagrangian of this constrained optimization problem becomes

$$\mathcal{L}_\eta(\mathbf{w}, b, e_i; \gamma_i) = \frac{1}{2} \sum_{i=1}^{N} e_i^2 + \frac{1}{2} \left[\eta \tilde{S}_w - (1-\eta) \tilde{S}_b \right] - \sum_{i=1}^{N} \gamma_i \left(\mathbf{w}^T \mathbf{x}_i + b + e_i - y_i \right) \tag{25}$$

where γ_i are Lagrange multipliers.

Following the deduction in the above section, in the linear version

$$\tilde{S}_w = \mathbf{w}^T \mathbf{X} \mathbf{L}_w \mathbf{X}^T \mathbf{w}$$

$$\tilde{S}_b = \mathbf{w}^T \mathbf{X} \mathbf{L}_b \mathbf{X}^T \mathbf{w}$$

Let $\mathbf{S}_\eta = \mathbf{X} \left[\eta \mathbf{L}_w - (1-\eta) \mathbf{L}_b \right] \mathbf{X}^T$, then (25) can be reformulated as

$$\mathcal{L}_\eta(\mathbf{w}, b, e_i; \gamma_i) = \frac{1}{2} \sum_{i=1}^{N} e_i^2 + \frac{1}{2} \mathbf{w}^T \mathbf{S}_\eta \mathbf{w} - \sum_{i=1}^{N} \gamma_i \left(\mathbf{w}^T \mathbf{x}_i + b + e_i - y_i \right) \tag{26}$$

The conditions for optimality *w.r.t.* $\mathbf{w}, b, e_i, \gamma_i$ for the training respectively become

$$\partial \mathcal{L}_\eta / \partial \mathbf{w} = 0 \rightarrow \mathbf{w} = \sum_{i=1}^{N} \gamma_i \left(\mathbf{S}_\eta \right)^+ \mathbf{x}_i$$

$$\partial \mathcal{L}_\eta / \partial b = 0 \to \sum_{i=1}^{N} \gamma_i = 0$$

$$\partial \mathcal{L}_\eta / \partial e_i = 0 \to e_i = \gamma_i$$

$$\partial \mathcal{L}_\eta / \partial \gamma_i = 0 \to \mathbf{w}^T \mathbf{x}_i + b + e_i = y_i \tag{27}$$

where "$(\cdot)^+$"denotes the generalized inverse of a matrix.

Hence, the optimization problem can be written immediately as the linear equations (24) after eliminating the variables $\mathbf{w}$ and e_i, where $\mathbf{\Omega}_\eta \in R^{N\times N}$ with $\mathbf{\Omega}_{\eta,ij} = \mathbf{x}_j^T \left(\mathbf{S}_\eta\right)^+ \mathbf{x}_i$, $\mathbf{1}_N = [1,\cdots,1]^T$, $\boldsymbol{\gamma} = [\gamma_1,\cdots,\gamma_N]^T$, $\mathbf{y} = [y_1,\cdots,y_N]^T$, and $\mathbf{I}_N \in R^{N\times N}$ is an identity matrix. Then we can obtain the minimum norm solution to (23). This proves the theorem. ■

For the multi-class case, we introduce the vector labeled outputs into the solution framework of DRLSC, which can make the computational complexity independent of the number of classes and require no more computation than a single binary classifier[35, 47-49]. Furthermore, Szedmak and Shawe-Taylor[35] presented that this technique does not diminish classification performance but in some cases can improve it, relatively to one-against-one and one-against-all. Therefore, here we code the class labels following the one-of-*c* rule, i.e. if $\mathbf{x}_i$ belongs to the *k*th class, then $\mathbf{y}_i = [0,\cdots,1,\cdots,0]^T \in R^c$, where the *k*th element is 1 and the other elements are 0, $\forall i = 1,\cdots,N$. Then the classifier has the linear form as

$$f(\mathbf{x}) = \mathbf{W}^T \mathbf{x} + \mathbf{b} \tag{28}$$

Note that here $\mathbf{W} \in R^{n\times c}$, $\mathbf{b} \in R^c$.

We introduce the equality constraints into the optimization problem as

$$\min_{\mathbf{W},\mathbf{b}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left\| \mathbf{e}_i \right\|^2 + \frac{1}{2} \left[\eta \tilde{S}_w - (1-\eta) \tilde{S}_b \right] \right\}$$

subject to

$$\mathbf{W}^T \mathbf{x}_i + \mathbf{b} + \mathbf{e}_i = \mathbf{y}_i, \quad \forall i = 1,\cdots,N \tag{29}$$

Similarly to the binary-class case, we have the following theorem in the multi-class case:

Theorem 2. (Multi-class case) Given the parameter $\eta \in [0,1]$, the minimum norm solution to

the problem (29) is characterized by the linear system with the variables $\boldsymbol{\gamma} \in R^{c \times N}$

$$\begin{bmatrix} \mathbf{b} & \boldsymbol{\gamma} \end{bmatrix} \begin{bmatrix} 0 & \mathbf{1}_N^T \\ \mathbf{1}_N & \boldsymbol{\Omega}_\eta + \mathbf{I}_N \end{bmatrix} = \begin{bmatrix} \mathbf{0}_c & \mathbf{Y} \end{bmatrix} \tag{30}$$

where $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_N]$, $\mathbf{0}_c = [0, \cdots, 0]^T$, $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N]$. Note that the other notations, including $\boldsymbol{\Omega}_\eta$, are the same as the ones in Theorem 1.

The proof is similar to Theorem 1. Thus we omit it here.

In summary, the algorithmic procedure of DRLSC algorithm can be formally stated as follows:

Table 2. Algorithm DRLSC

Input: The data $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$; the parameter η.

Output: **w** and b.

1. Construct the intra-class compactness and inter-class separability graphs;
2. Compute the two Laplacian matrices for the two graphs;
3. Analytic solutions: solve *b* and $\boldsymbol{\gamma}$ by Equation (24) or (30), then compute **w** by the corresponding Lagrangian constraints.

6. Experiments

To evaluate the proposed Discriminatively Regularized Least-Squares Classification (DRLSC) algorithm, in this section we systematically compare it with the state-of-the-art regularization algorithms as shown in Table 3, and DRGV discussed at the beginning of Section 4 which defines the Discriminative Regularization term with the Generalized Variances, both on artificial and real-world classification problems. Firstly, we present three synthetic datasets for clearly comparing the classification performances of the traditional regularization algorithm RN with DRGV, DRLSC in terms of the different distribution of samples with different complexities. On the real-world problems, several datasets in the UCI database (the UCI Machine Learning Repository) are used to evaluate the classification accuracies derived from DRLSC in comparison to RN, GRBFN, SVM, LS-SVM, MR and DRGV, both in linear and nonlinear versions. Then, we further apply DRLSC to the image recognition problems.

Table 3. The acronyms, full names and citations for the compared regularization algorithms in the experiments

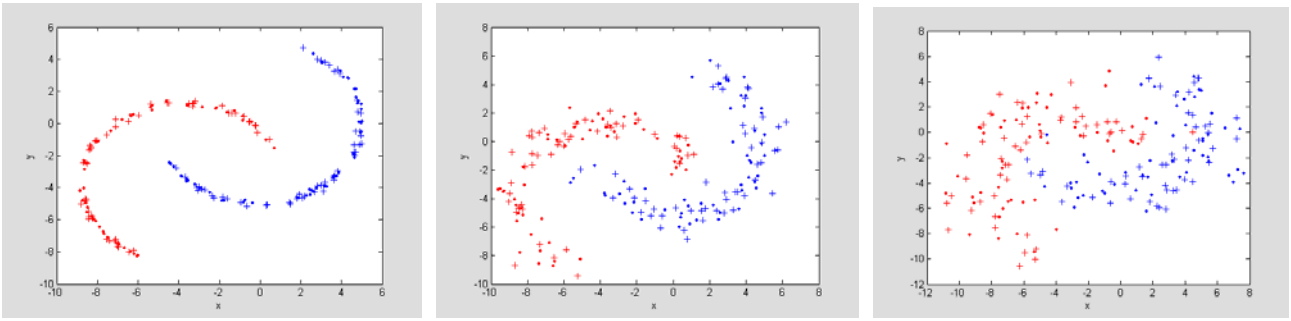
Acronym	Full Name	Citation
RN	Regularization Networks	[16]
GRBFN	Generalized Radial Basis Function Networks	[16]
SVM	Support Vector Machines	[1]
LS-SVM	Least Squares Support Vector Machines	[25]
MR	Manifold Regularization	[5, 10]
DRGV	Discriminative Regularization term defined with the Generalized Variances	Section 4

6.1. Toy Problem

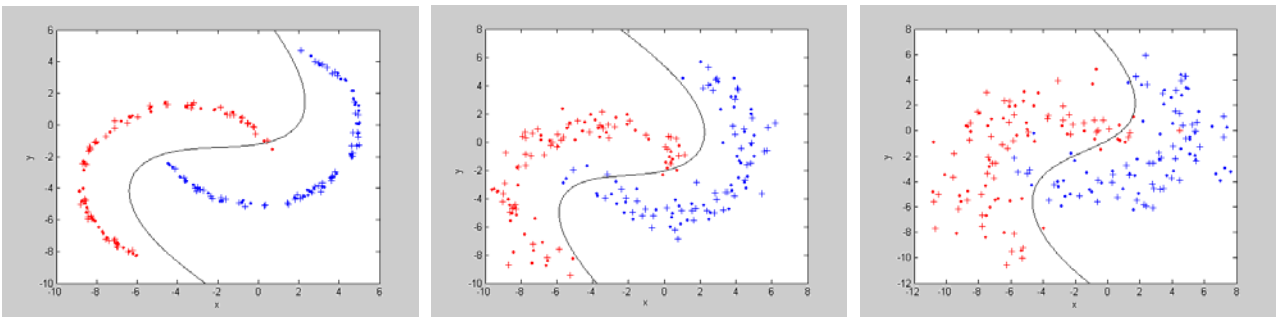
In the toy problems, three two-moon datasets with different complexities are discussed (Figure 1. (A), (B) and (C)). Each dataset contains one hundred samples in each class. As shown in Figure 1, '•' denotes the training samples and '+' denotes the testing samples. We compare RN ((A.1), (B.1), (C.1)) with DRGV ((A.2), (B.2), (C.2)) and DRLSC ((A.3), (B.3), (C.3)). In DRLSC, the number of the k nearest neighbors is fixed to 10. The nine subfigures show the discriminant boundaries of the three methods in each dataset. Furthermore, the respective training and testing accuracies are labeled in Tables 4 and 5 respectively.

From Figure 1 and Tables 4 to 5, it can be seen that: (1) As a traditional regularization method, the discriminant boundaries derived from RN is smooth all along in the three two-moon datasets ((A.1), (B.1), (C.1)). However, with the increase of the complexity in the datasets, the classification accuracies in RN descend relatively sharply, and are always worse than those in DRGV and DRLSC. Especially, RN is more likely (locally) over-smooth in the datasets (B) and (C). (2) When the two classes are far from each other ((A)), the boundaries of DRGV and DRLSC ((A.2), (A.3)) are adequately smooth as well as RN. When the classes get nearer ((B), (C)) and the complexity of classification increases, the discriminant boundaries derived from DRGV and DRLSC become nonsmooth ((B.2), (C.2), (B.3), (C.3)). However, the classification performances of DRGV and DRLSC are both much better than RN, which clearly justifies the intuition that, the smoothness assumption in the traditional

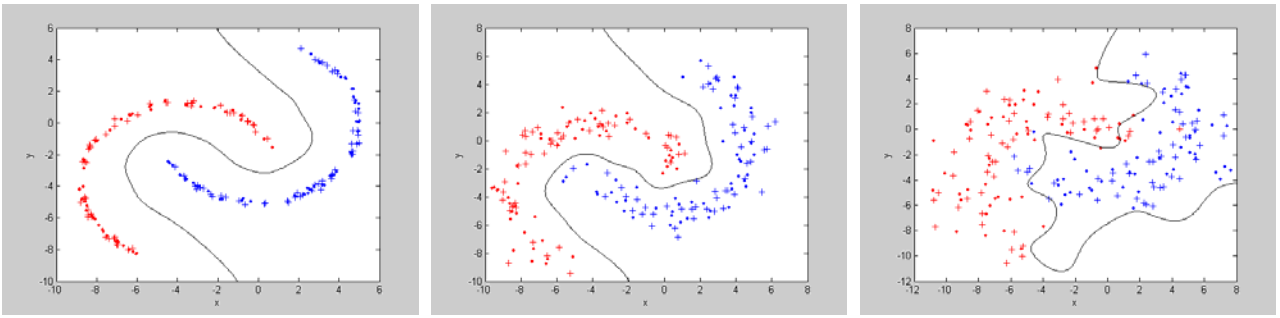
regularization methods is too general for classification. Furthermore, the underlying discriminative information is vital for classification. Consequently, owing to focusing more on the underlying class discriminative information than the smoothness of the classifier, DRGV and DRLSC are naturally superior to RN. (3) For simply using the generalized variance S_w and S_b to characterize the intra-class compactness and inter-class separability from the global viewpoint, the discriminant boundaries of DRGV get nearer the relatively dense distribution regions of the samples belonging to the different classes ((B.2), (C.2)). Thus, when the samples between the classes overlap relatively heavily, just as in (C), DRGV cannot correctly classify the samples near the boundary. On the contrary, thanks to considering the local manifold structure of the samples, the discriminant boundaries of DRLSC accord with more the geometry of the samples, and thus keep high training and testing accuracies even in (C) ((B.3), (C.3)), which validates that DRLSC can outperform DRGV in relatively more complex cases.



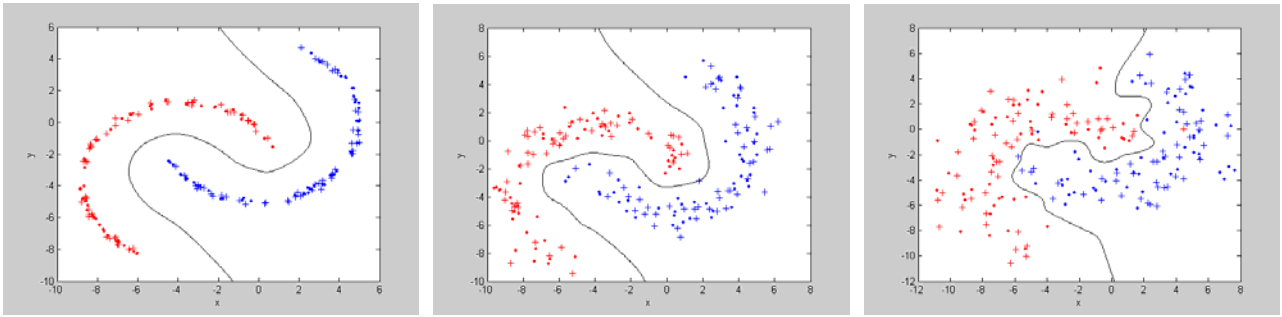
(A) (B) (C)



(A.1) (B.1) (C.1)



(A.2) (B.2) (C.2)



(A.3) (B.3) (C.3)

Figure 1. The discriminant boundaries in the three Two-Moon datasets ((A), (B), (C)): RN ((A.1), (B.1), (C.1)), DRGV ((A.2), (B.2), (C.2)) and DRLSC ((A.3), (B.3), (C.3))

Table 4. Training accuracies (%) compared between RN, DRGV and DRLSC in the three Two-Moon datasets

	(A)	(B)	(C)
RN	99.00	95.00	90.00
DRGV	**100.0**	**100.0**	98.00
DRLSC	**100.0**	**100.0**	**99.00**

Table 5. Testing accuracies (%) compared between RN, DRGV and DRLSC in the three Two-Moon datasets

	(A)	(B)	(C)
RN	**100.0**	98.00	93.00
DRGV	**100.0**	**100.0**	93.00
DRLSC	**100.0**	**100.0**	**98.00**

6.2. UCI Database

To further investigate the classification performance of our DRLSC, we also systematically compare it with many state-of-the-art regularization algorithms and DRGV in

several real-world datasets in the UCI database. These datasets contain seven binary-class datasets and thirteen multi-class datasets. For each dataset, we divide the samples into two non-overlapping training set and testing set, and each set contains almost half of samples in each class respectively. This process is repeated ten times to generate ten independent runs for each dataset, and then the average results are collected and reported.

We report the experimental results both in linear and nonlinear versions, which are listed in Tables 6 and 7. In the linear version, we compare DRLSC to SVM, LS-SVM, MR and DRGV, with the linear kernel as the kernel function in these methods. We apply the one-against-all strategy for SVM in the multi-class cases. In the nonlinear version, we further compare DRLSC with RN, GRBFN, SVM, LS-SVM, MR and DRGV. Multivariate Gaussian function is chosen to be the Green's function, i.e. the activation function of the individual hidden units in the two networks, and also to be the kernel function in the other methods. We apply Fuzzy *c*-means clustering algorithm (FCM) to obtain the center set in GRBFN. The Gaussian parameter σ in the seven methods and the regularization parameters in RN, GRBFN, SVM, LS-SVM, and MR are selected from the set $\left\{2^{-10}, 2^{-9}, \cdots, 2^{9}, 2^{10}\right\}$. Moreover, the parameters η in DRLSC and DRGV are chosen in the interval [0, 1]. And throughout the experiments, we choose the best *k* between two and $\left(\min_{c}\left\{number(N_c)\right\}-1\right)$ in DRLSC and MR. Parameter selection is done by the cross-validation.

From these results, we can make several interesting observations as follows:

- The kernel trick[1] can improve the classification performance for SVM, LS-SVM, MR, DRGV and DRLSC in most datasets. For example, in Iris, the classification accuracies of DRLSC improve more than ten percent in the RBF kernel than in the linear kernel. However, for some datasets, such as Soybean_small for SVM, Water for SVM and LS-SVM, and Wine for SVM, LS-SVM and MR, there is more likely overfitting to some extent. On the contrary, DRLSC still demonstrates high classification accuracy for these datasets, which validates that it is more stable than SVM, LS-SVM and MR in both binary-class and multi-class cases.
- The results show that the traditional regularization methods RN and GRBFN perform relatively poorly almost in all datasets, which clearly justifies that the only emphasis on

the smoothness of the classifier is far from sufficiency for classification. Thus, we should introduce more underlying information, such as discriminative information and local manifold structure, into the regularization framework, to further guide better classification.

- DRLSC outperforms the other six algorithms in most datasets, especially in the RBF kernel. In order to find out whether DRLSC is significantly better than the other methods, we perform the *t*-test on the classification results of the ten runs to calculate the statistical significance of DRLSC. The null hypothesis H_0 demonstrates that there is no significant difference between the mean number of patterns correctly classified by DRLSC 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 Tables 6 and 7 will be denoted "*". Consequently, as shown in Tables 6 and 7, it can be clearly found that DRLSC possesses significantly superior classification performance to the other methods in most datasets. This just accords with our conclusions.
- Specially, MR and DRLSC both introduce the geometry of the samples into the regularization term. The difference is that MR makes a trade-off between the smoothness of the classifier and the maintenance of the local manifold structure, and DRLSC focuses more on utilizing as much the underlying discriminative information as possible to further direct keeping the geometry. The experimental results demonstrate that DRLSC is superior to MR in most cases. This fact further validates that, for classification, emphasizing the class information is more important than doing the smoothness of the classifier.
- Another interesting observation is that, although the general assumption in manifold learning is that the samples should be sufficient, in small datasets such as Lenses and Soybean_small, DRLSC still performs better, which seems to indicate that the assumption in DRLSC can be relaxed. Consequently, DRLSC can have wider application scopes in real-world problems.

Table 6. Classification performance (%) compared between SVM, LS-SVM, MR, DRGV and DRLSC in the 20 UCI datasets with the linear kernel

Dataset	Number of classes	Dimension	Classification accuracy				
			SVM	LS-SVM	MR	DRGV	DRLSC
Ionosphere	2	34	87.78	85.74*	85.00*	85.17*	**88.01**
Sonar	2	60	77.88	77.50	**79.87***	70.77*	76.83
Water	2	38	**98.64***	96.10	97.98*	94.58*	96.27
Wdbc	2	30	94.98*	96.04	94.42*	95.65*	**96.21**
Bupa	2	6	66.99*	66.76*	68.90*	66.94*	**70.23**
Pid	2	8	73.96*	76.88*	69.19*	76.93*	**78.26**
Diabetes	2	8	75.05*	77.29*	69.40*	77.24*	**78.72**
Wine	3	13	95.67*	97.11*	97.11*	**99.00**	**99.00**
Lenses	3	4	74.62*	76.15*	82.85*	82.31*	**84.62**
Tae	3	5	50.39*	51.32*	51.58*	51.97*	**56.58**
New_thyroid	3	5	**95.65***	93.59*	90.00	89.81*	91.11
Iris	3	4	94.53*	93.67*	**96.53***	86.13*	87.07
Cmc	3	9	**55.68***	51.22	52.38	50.93*	51.96
Balance_scale	3	4	87.86*	87.83*	**89.20***	87.83*	88.75
Soybean_small	4	35	**100.0**	**100.0**	**100.0**	99.17	99.58
Vehicle	4	18	**79.81***	79.08*	79.72*	77.48*	78.16
Dermatology	6	33	96.74*	97.34*	98.21	96.63*	**98.26**
Ecoli	6	6	**86.96***	85.77	83.04*	85.48	85.71
Glass	6	9	62.57	59.17*	61.65*	61.93*	**63.76**
Yeast	10	8	52.35*	52.58*	55.54*	**56.66***	56.27

'*' Denotes that the difference between DRLSC and the other four methods is significant at 5% significance level, i.e., t-value >1.7341

Table 7. Classification performance (%) compared between RN, GRBFN, SVM, LS-SVM, MR, DRGV and DRLSC in the 20 UCI datasets with the RBF kernel

Dataset	Classification accuracy						
	RN	GRBFN	SVM	LS-SVM	MR	DRGV	DRLSC
Ionosphere	89.60*	86.88*	95.11*	95.34*	98.30*	94.26*	**99.43**
Sonar	82.88*	77.02*	85.00*	85.10*	92.31*	87.50*	**94.23**
Water	95.59*	91.02*	90.51*	89.83*	98.31*	98.31*	**99.32**
Wdbc	93.12*	94.28*	94.25*	94.74*	95.09*	95.51*	**96.60**
Bupa	72.43*	73.64*	73.06*	73.01*	78.03*	73.29*	**81.73**
Pid	76.25*	77.84	76.56*	76.61*	**80.73***	77.42*	78.36
Diabetes	77.08*	75.16*	77.08*	77.40*	77.37*	78.91*	**79.07**
Wine	73.67*	76.11*	77.78*	77.00*	83.56*	96.45*	**97.56**
Lenses	75.38*	70.00*	79.23*	78.62*	81.54*	86.15	**87.69**
Tae	52.63*	47.37*	54.34*	52.89*	58.29*	56.58*	**61.32**
New_thyroid	93.33*	90.83*	96.02	95.56	**97.31***	94.81*	96.02
Iris	96.80*	96.80*	98.27	97.47*	98.67	96.67*	**98.80**
Cmc	55.33*	56.29	56.41	56.25	56.36	55.43*	**56.82**
Balance_scale	91.28*	91.21*	**92.04**	91.25	91.63	91.25*	91.82

Soybean_small	**100.0**	82.92*	62.50*	97.92*	**100.0**	**100.0**	**100.0**
Vehicle	73.35*	70.66*	74.76*	72.44*	73.70*	81.82	**82.61**
Dermatology	97.37*	96.36*	97.28*	97.77*	98.70	98.53	**98.91**
Ecoli	88.81	89.11	89.17	88.33	**89.70***	88.39	88.63
Glass	70.37	65.69*	72.75	71.10	**76.24***	70.73	71.65
Yeast	60.56	59.22*	60.58*	60.81	**61.57***	60.28*	60.56

'*' Denotes that the difference between DRLSC and the other six methods is significant at 5% significance level, i.e., *t*-value >1.7341

6.3. Image Recognition

Many researchers have suggested that the image data can often be characterized by a low-dimensional intrinsic manifold[27, 31, 50, 51]. That is, the image data often have an underlying invariant and associated transformations, like shifts, rotation, changes of expression, etc, that naturally imply a manifold on which those neighboring points are small transformations of one another[51]. Therefore, in this subsection, we apply our proposed method to image recognition. Three well-known and publicly available databases corresponding to typical image classification problems, i.e. recognition of faces (AR), objects (COIL-20) and handwritten digits (USPS), are used to evaluate DRLSC with RN, GRBFN, SVM, LS-SVM, MR and DRGV.

6.3.1. Dataset Description and Experimental Setting

The AR database[52] contains 100 subjects and each subject has 26 face images taken in two sessions. For each session, there are 13 face images. Here 14 faces of natural expression from the two sessions of each person are chosen for experiments, which are illustrated in Figure 2. The 1400 images are all cropped into the same size of 66×48 pixels.

COIL-20[53] is a database of gray-scale images of 20 objects, as shown in Figure 3[54]. The objects were placed on a motorized turntable against a black background. The turntable was rotated through 360 degrees to vary the object poses with respect to a fixed camera, as shown in Figure 4[54]. Images of the objects were taken at pose intervals of 5 degrees, which corresponds to 72 images per object. For our experiments, we have resized each of the original 1440 images down to 32×32 pixels.

The USPS database[1] consists of grayscale handwritten digit images from 0 to 9, as shown in Figure 5. Each digit contains 1100 images, and the size of each image is 16×16 pixels with 256 gray levels. Similarly to [55], here we select five pairwise digits of varying difficulty for odd vs. even digit classification.

Each image database is partitioned into the different gallery and probe sets where G*m*/P*n* indicates that *m* images per object are orderly in AR and randomly in COIL-20, USPS selected for training and the remaining *n* images are used for testing[33]. Moreover, the parameter selection in the seven algorithms is the same as the subsection 6.2 in the nonlinear version with the RBF kernel.

Figure 2. An illustration of 14 images of one subject from the AR face database

Figure 3. An illustration of 20 subjects in the COIL-20 database

[1] Available at: http://www.cs.toronto.edu/~roweis/data.html

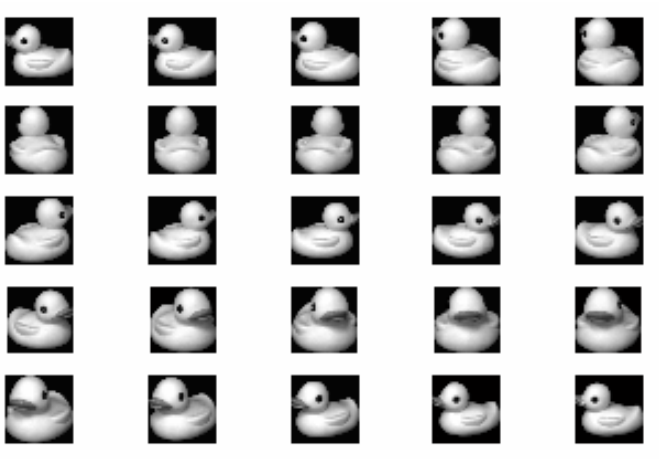


Figure 4. An illustration of partially 25 images of one subject from the COIL-20 database

Figure 5. An illustration of 10 subjects in the USPS database

6.3.2. Evaluation of Classification Performance

Tables 8 to 12 report the experimental results of the seven algorithms in the AR, COIL-20 and USPS databases respectively, in terms of different sampling in the training and testing sets. From these results, we can also obtain several attractive insights as follows:

- It can clearly be seen that, from AR to USPS, the number of the classes decreases meanwhile the number of the samples in the same class increases. The experimental results demonstrate that, when the dataset distribution is complex and the training set is too small to characterize the manifold structure underlying the data well, such as the three cases in the AR database, in which there are 100 classes and the most number of the samples in the same class for training is seven, RN, GRBFN, SVM, LS-SVM, MR and DRGV all appears to be less effective than DRLSC. This fact justifies that, only emphasis on the smoothness of the classifier as RN and GRBFN, the discriminative information of the data as SVM, LS-SVM and DRGV, or the geometrical structure underlying the data as MR, seems far from being sufficient for image classification in these complex cases.

But by the "No Free Lunch" Theorem, DRLSC further considers both the underlying discriminative and geometrical information simultaneously which is both vital for classification, and thus achieves better classification accuracies.

- Due to the decrease of the number of classes and the increase of the samples, in the COIL-20 database, all the algorithms perform much better than in the AR database. The classification accuracies of DRLSC are better than those of the other algorithms in the three cases.
- In the USPS database, when the training sample size is small, such as for the case of G10/P1090, DRLSC also shows better performance relative to the other algorithms. When the training sample size is large enough to represent the data distribution, such as for the cases G100/P1000 and G550/P550, most algorithms can achieve similar classification accuracy. However, in the recognitions of 2 vs. 3 and 3 vs. 8 which are relatively difficult than the other pairwise digits, DRLSC always keeps obvious superiority to the other methods.
- Another observation is that, GRBFN performs poorly in all the cases, especially in the AR and COIL-20 databases. The reason may be that we apply FCM to obtain the center set of the hidden nodes in GRBFN. Fern and Brodley[56, 57] demonstrated that when the data is sparse in the high-dimensional space, it is difficult for any unsupervised clustering algorithm to find any clustering structure in the data. In our experiments, in these two databases, FCM necessarily obtains all the clustering centers in the same point, which leads to the poor performance of GRBFN.

Table 8. Classification performance (%) compared between RN, GRBFN, SVM, LS-SVM, MR, DRGV and DRLSC in the AR face database

	G3/P11	G5/P9	G7/P7
RN	71.73	77.56	92.14
GRBFN	10.73	12.00	24.43
SVM	64.18	71.78	91.43
LS-SVM	64.09	71.89	91.14
MR	68.45	72.22	91.00
DRGV	70.00	78.78	92.00
DRLSC	**74.27**	**80.89**	**93.57**

Table 9. Classification performance (%) compared between RN, GRBFN, SVM, LS-SVM, MR, DRGV and DRLSC in the COIL-20 database

	G9/P63	G18/P54	G36/P36
RN	96.03	97.41	98.19
GRBFN	59.92	66.39	58.47
SVM	96.98	98.98	99.44
LS-SVM	97.06	98.89	99.44
MR	97.38	98.33	98.75
DRGV	98.10	99.07	99.44
DRLSC	**98.33**	**99.17**	**99.72**

Table 10. Classification performance (%) compared between RN, GRBFN, SVM, LS-SVM, MR, DRGV and DRLSC in G10/P1090 in the USPS database

G10/P1090	Classification accuracy						
	RN	GRBFN	SVM	LS-SVM	MR	DRGV	DRLSC
1 vs. 7	94.77	87.39	95.69	95.64	95.73	95.78	**96.97**
2 vs. 3	94.54	94.04	95.69	95.73	95.00	94.45	**96.79**
2 vs. 7	96.61	95.83	96.65	96.74	96.83	96.38	**97.71**
3 vs. 8	92.57	92.43	92.75	92.66	92.98	91.47	**93.58**
4 vs. 7	98.35	94.95	98.62	98.53	98.53	98.39	**99.08**

Table 11. Classification performance (%) compared between RN, GRBFN, SVM, LS-SVM, MR, DRGV and DRLSC in G100/P1000 in the USPS database

G100/P1000	Classification accuracy						
	RN	GRBFN	SVM	LS-SVM	MR	DRGV	DRLSC
1 vs. 7	99.75	96.90	99.85	99.85	**99.95**	99.85	**99.95**
2 vs. 3	98.00	96.45	98.10	98.15	98.35	98.15	**98.40**
2 vs. 7	99.55	98.95	**99.70**	99.60	**99.70**	99.60	**99.70**
3 vs. 8	97.70	95.85	98.40	98.25	97.90	98.20	**98.50**
4 vs. 7	99.30	98.30	**99.70**	99.50	99.50	99.40	**99.70**

Table 12. Classification performance (%) compared between RN, GRBFN, SVM, LS-SVM, MR, DRGV and DRLSC in G550/P550 in the USPS database

G550/P550	Classification accuracy						
	RN	GRBFN	SVM	LS-SVM	MR	DRGV	DRLSC
1 vs. 7	99.82	97.00	**99.91**	**99.91**	**99.91**	**99.91**	**99.91**
2 vs. 3	99.36	97.18	99.55	99.36	99.45	99.55	**99.73**
2 vs. 7	**99.73**	99.55	**99.73**	**99.73**	**99.73**	**99.73**	**99.73**
3 vs. 8	99.09	98.00	99.00	98.64	98.91	99.36	**99.55**

4 vs. 7	**99.91**	99.64	**99.91**	**99.91**	**99.91**	**99.91**	**99.91**

7. Conclusion

In this paper, we propose a novel regularization framework called Discriminatively Regularized Least-Squares Classification (DRLSC). By making the best of the underlying discriminative and geometrical information rather than only emphasizing the smoothness of the classifier in the traditional regularization methods, DRLSC introduces a new discriminative regularization term in the framework. Inspired by the new graph motivated methods, the algorithm relies on two graphs to characterize the intra-class compactness and inter-class separability respectively, and thus can further maximize the margins between the samples of the different classes in each local area. Through introducing equality constraints in the formulation, the solutions of DRLSC can follow from solving a set of linear equations. As a result, the algorithm is simpler and more stable. The experimental results have demonstrated the superiority of our proposed DRLSC compared with the state-of-the-art regularization methods.

There are several directions of future study:

- **Additional generalization:** In this paper, we have introduced the new discriminative regularization term into the regularization framework and incorporated it with the square-loss function. In future, we will systematically compare all possible combinations of the discriminative regularization term with other loss functions or regularization terms mentioned in this paper. This will lead to a large family of new algorithms and we believe that there should be a lot of interesting observations.
- **Theoretical Foundation:** In the signal and image processing region, researchers have proved that using nonsmooth and/or nonconvex regularization terms can frequently yield good estimates[58, 59]. In our DRLSC, the new discriminative regularization term is nonconvex and nonnegative. However, DRLSC has showed superior classification performance to many regularization methods. Therefore, seeking for the theoretical foundation of DRLSC is our main future research direction.
- **Tensorization:** Many researches have showed that representing the objects as tensors of arbitrary order can further improve the performance of algorithms in most cases. We

intend to further investigate this issue in our proposed DRLSC framework in both theory and practice.

- **Preprocessing:** In this paper, we apply DRLSC to the image classification problems. Due to the generality of DRLSC learning framework, it can be further combined with any preprocessing image-based dimensionality reduction methods, and image detailed-preserving regularization approaches[60, 61]. We believe these combinations should lead to better classification performance.
- **Parameter selection:** The selection of the neighbor number k is an open problem in manifold learning. And the difficulty also appears in the selection of the regularization parameters in regularization. More systematic researches are needed.
- **Sparse solutions:** Although the solutions to DRLSC are much simpler than the ones to classical SVM, especially in multi-class problems, the algorithm could be slow when the dataset is large. Tsang and Kwok[62] have presented a sparse solution to MR for large scale problems. Hence how to develop a fast algorithm for the sparse solution to DRLCS is another interesting topic for future study.

Acknowledgment

This work was supported by National Natural Science Foundation of China (60773061).

References

[1] V. Vapnik. Statistical Learning Theory. Wiley, 1998.

[2] Z. Chen and S. Haykin. On different facets of regularization theory. Neural Computation, vol.14(12), 2791-2846, 2002.

[3] D. Schuurmans and F. Southey. Metric-based methods for adaptive model selection and regularization. Machine Learning, vol.48, 51-84, 2002.

[4] O. Bousquet, O. Chapelle, and M. Hein. Measure based regularization. NIPS, 2003.

[5] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from examples. Department of Computer Science,University of Chicago,Tech.Rep, TR-2004-06, 2004.

[6] C.A. Micchelli and M. Pontil. Learning the kernel function via regularization. J. Machine Learning Research, vol.6, 1099-1125, 2005.

[7] R.M. Rifkin and R.A. Lippert. Value regularization and fenchel duality. J. Machine Learning Research, vol.8, 441-479, 2007.

[8] A. Corduneanu and T. Jaakkola. On information regularization. UAI 2003.

[9] H. Xue, S. Chen, and X. Zeng. Classifier learning with a new locality regularization method. Pattern

Recognition, vol.41(5), 1496-1507, 2008.

[10] M. Belkin, P. Niyogi, and V. Sindhwani. On mainfold regularization. In *Proc. Intl. Workshop on Artificial Intelligence and Statistics*, 2005.

[11] T. Poggio and F. Girosi. Networks for approximation and learning. Proc. of the IEEE. vol.78, 1481-1497, 1990a.

[12] T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. Science, vol.247, 978-982, 1990b.

[13] A.R. Barron. Complexity regularization with application to artificial neural networks. In G. Roussas (Ed.), Nonparametric functional estimation and related topics, 561-576, 1991.

[14] J.J. Pan, Q. Yang, H. Chang, and D.-Y. Yeung. A manifold regularization approach to calibration reduction for sensor-network based tracking. Proc. of the Twenty-First National Conference on Artificial Intelligence (AAAI 06), 988-993, 2006.

[15] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. Wiley, 2001.

[16] S. Haykin. Neural Networks: A Comprehensive Foundation. Tsinghua University Press, 2001.

[17] A.N. Tikhonov. On solving incorrectly posed problems and method of regularization. Doklady Akademii Nauk USSR, vol.151, 501-504, 1963.

[18] A.N. Tikhonov and V.Y. Aresnin. Solutions of Ill-posed Problems. Washington, DC:W.H. Winston, 1977.

[19] V.A. Morozov. Methods for Solving Incorrectly Posed Problems. Springer-Verlag, 1984.

[20] G. Wahba. Spline Models for Observational Data. volumn 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial & Applied Mathematics, 1990.

[21] R.M. Rifkin. Everything Old Is New Again: A Fresh Look at Historical Approaches to Machine Learning. PhD thesis, Massachusetts Institute of Technology, 2002.

[22] F. Girosi. An equivalence between sparse approximation and support vector machines. Neural Computation, vol.10(6), 1455-1480, 1998.

[23] B. Scholkopf, R. Herbrich, and A.J. Smola. A generalized representer theorem. Proc. of the 14th Annual Conference on Computational Learning Theory, 416-426, 2001.

[24] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Beijing, Publishing House of Electronics Industry, 2004.

[25] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. Neural Processing Letters, vol.9, 293-300, 1999.

[26] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. Advances in Computational Mathematics, vol.13(1), 1-50, 2000.

[27] J. Tenenbaum, V. Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, vol.290(22), 2319-2323, 2000.

[28] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, vol.290(22), 2323-2326, 2000.

[29] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral technique for embedding and clustering. NIPS, vol.15, Vancouver, British Columbia, Canada, 2001.

[30] X. He, D. Cai, S. Yan, and H. Zhang. Neighborhood preserving embedding. Proc. of the 10th IEEE International Conference on Computer Vision, 2005.

[31] X. He and P. Niyogi. Locality preserving projection. NIPS, 2003.

[32] D. Cai, X. He, K. Zhou, J. Han, and H. Bao. Locality sensitive discriminant analysis. IJCAI, 708-713, 2007.

[33] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: a general

framework for dimensionality reduction. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.29(1), 40-51, 2007.
[34] H. Chen, H. Chang, and T. Liu. Local discriminant embedding and its variants. CVPR, 2005.
[35] S. Szedmak and J. Shawe-Taylor. Muticlass learning at one-class complexity. Technical Report No: 1508, School of Electronics and Computer Science, Southampton, UK, 2005.
[36] H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximun margin criterion. IEEE Trans. on Neural Networks, vol.17(1), 157-165, 2006.
[37] S. Lafon, Y. Keller, and R.R. Coifman. Data fusion and multicue data matching by diffusion maps. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.28(11), 1784-1797, 2006.
[38] F.R.K. Chung. Spectral Graph Theory. Regional Conference Series in Mathematics, number 92, 1997.
[39] A.L. Yuille and A. Rangarajan. The concave-convex procedure. Neural Computation, vol.15, 915-936, 2003.
[40] R. Collobert, F. Sinz, J. Weston, and L. Bottou. Trading convexity for scalability. ICML, 2006.
[41] G. Ou and Y.L. Murphey. Multi-class pattern classification using neural networks. Pattern Recognition, vol.40(1), 4-18, 2007.
[42] S. Asharaf, M.N. Murty, and S.K. Shevade. Multiclass core vector machine. ICML, Corvallis, OR, 2007.
[43] H. Xiong, M.N.S. Swamy, and M.O. Ahmad. Optimizing the kernel in the empirical feature space. IEEE Trans. on Neural Networks, vol.16(2), 460-474, 2005.
[44] O. Mangasarian and E. Wild. Multisurface proximal support vector machine classification via generalized eigenvalues. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.28(1), 69-74, 2006.
[45] Z. Wang, S. Chen, and T. Sun. MultiK-MHKS: a novel multiple kernel learning algorithm. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.30(2), 348-353, 2008.
[46] A. Martinez and A. Kak. PCA versus LDA. IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.23(2), 228-233, 2001.
[47] T. Evgeniou, C.A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. J. Machine Learning Research, vol.6, 615-637, 2005.
[48] C.A. Micchelli and M. Pontil. Kernels for multi-task learning. NIPS, 2004.
[49] C.A. Micchelli and M. Pontil. On learning vector-valued functions. Neural Computation, vol.17, 177-204, 2005.
[50] L.K. Saul and S.T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. J. Machine Learning Research, vol.4, 119-155, 2003.
[51] A. Ghodsi, J. Huang, F. Southey, and D. Schuurmans. Tangent-corrected embedding. CVPR, 2005.
[52] A.M. Martinez and R. Benavente. "The AR Face Database". CVC Technical Report #24, June, 1998.
[53] S.A. Nene, S.K. Nayar, and H. Murase. Columbia Object Image Library (COIL-20). Technical Report CUCS-005-96, February, 1996.
[54] T. Sun and S. Chen. Locality preserving CCA with applications to data visualization and pose estimation. Image and Vision Computing, vol.25(5), 531-543, 2007.
[55] A. Argyriou, M. Herbster, and M. Pontil. Combing graph Laplacians for semi-supervised learning. NIPS, 2005.
[56] X.Z. Fern and C.E. Brodley. Random projection for high dimensional data clustering: a cluster ensemble approach. ICML, 186-193, 2003.
[57] X.Z. Fern and C.E. Brodley. Cluster ensembles for high dimensional clustering : an empirical study. Technical Report, CS06-30-02, Oregon State University, 2006.
[58] S. Durand and M. Nikolova. Stability of minimizers of regularized least squares objective functions I: study

of the local behavior. Technical Report, TSI-ENST, Paris, France, 2001.

[59] S. Durand and M. Nikolova. Stability of minimizers of regularized least squares objective functions II: study of the global behavior. Technical Report, TSI-ENST, Paris, France, 2001.

[60] R.H. Chan, C.-W. Ho, C.-Y. Leung, and M. Nikolova. Minimization of detail-preserving regularization functional by Newton's method with continuation. . ICIP, 125-128, 2005.

[61] R.H. Chan, C.-W. Ho, and M. Nikolova. Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization. IEEE Trans. on Image Processing, vol.14(10), 1479-1485, 2005.

[62] I.W. Tsang and J.T. Kwok. Large-scale sparsified manifold regularization. NIPS, Vancouver, Canada, 2006.