Classification by Discriminative Regularization

Bin Zhang[1*], Fei Wang[2], Ta-Hsin Li[3], Wen jun Yin[1], Jin Dong[1]
[1]IBM China Research Lab, Beijing, China
[2]Department of Automation, Tsinghua University, Beijing, China
[3]IBM Watson Research Center, Yorktown Heights, NY, USA
[*]Corresponding author, email: zbin@cn.ibm.com

Abstract

Classification is one of the most fundamental problems in machine learning, which aims to separate the data from different classes as far away as possible. A common way to get a good classification function is to minimize its empirical prediction loss or structural loss. In this paper, we point out that we can also enhance the discriminality of those classifiers by further incorporating the discriminative information contained in the data set as a prior into the classifier construction process. In such a way, we will show that the constructed classifiers will be more powerful, and this will also be validated by the final empirical study on several benchmark data sets.

Introduction

Classification, or supervised learning with discrete outputs (Duda *et al*., 2000), is one of the vital problems in machine learning and artificial intelligence. Given a set of training data points and their labels $\mathcal{X} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the classification algorithms aim to construct a good classifier $f \in \mathcal{F}$ ($\mathcal{F}$ is the hypothesis space) and use it to predict the labels of the data points in the testing set $\mathcal{X}_\mathcal{T}$. Most of the real world data analysis problems can finally be reformulated as a classification problem, such as text categorization (Sebastiani, 2002), face recognition (Li & Jain, 2005) and genetic sequence identification (Chuzhanova *et al*., 1998).

More concretely, in classification problems, we usually assume that the data points and their labels are sampled from a joint distribution $P(\boldsymbol{x}, y)$, and the optimal classifier f we seek for should minimize the following *expected loss* (Vapnik, 1995)

$$\mathcal{J} = E_{P(\boldsymbol{x},y)}\mathcal{L}(y, f(\boldsymbol{x})),$$

where $\mathcal{L}(\cdot,\cdot)$ is some loss function. However, in practice, $P(\boldsymbol{x}, y)$ is usually unknown. Therefore, we should minimize the following *empirical loss* instead

$$\hat{\mathcal{J}} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, f(\mathbf{x}_i))$$

When deriving a specific classification algorithm by minimizing $\hat{\mathcal{J}}$, we may encounter two common problems:

- Generally achieving a good classifier needs a sufficiently large amount of training data points (such that $\hat{\mathcal{J}}$ may have a better approximation of $\mathcal{J}$), however, in real world applications, it is usually expensive and time consuming to get enough labeled points.
- Even if we have enough training data points, and the constructed classifier may fit very well on the training set, it may not have a good generalization ability on the testing set. This is the problem we called *overfitting*.

Regularization is one of the most important techniques to handle the above two problems. First, for the overfitting problem, Vapnik proposed that $\mathcal{J}$ and $\hat{\mathcal{J}}$ satisfies the following inequality (Vapnik, 1995)

$$\mathcal{J} \leqslant \hat{\mathcal{J}} + \varepsilon(n, d_\mathcal{F}),$$

where $\varepsilon(n, d_\mathcal{F})$ measures the complexity of the hypothesis model. Therefore we can minimize the following *structural risk* to get an optimal f

$$\bar{\mathcal{J}} = \hat{\mathcal{J}} + \lambda\Omega(f),$$

where $\Omega(f)$ is a *structural regularization term* which measures the complexity of f. In such a way, the resulting f could also have a good generalization ability on testing data.

Second, for the first problem, we can apply semi-supervised learning methods, which aims to learn from partially labeled data (Chapelle *et al*., 2006). Belkin *et al*. proposed to add an additional regularization term to $\bar{\mathcal{J}}$ to punish the geometric smoothness of the decision function with respect to the intrinsic data manifold (Belkin *et al*., 2006), and such smoothness is estimated from the whole data set. Belkin *et al*. also pointed out that such a technique can also be incorporated into the procedure of supervised classification, and the smoothness of the classification function can be measured class-wise.

In this paper, we propose a novel regularization approach for classification. Instead of using geometrical smoothness of f as the penalty criterion as in (Belkin *et al*., 2006), we directly use the discriminality of f as the penalty function, since the final goal of classification is just to discriminate the data from different classes. Therefore we call our method *Discriminant Regularization* (*DR*). Specifically, we derive two margin based criterion to measure the discriminality of f, namely the global margin and local margin. We show

that the final solution of the classification function will be achieved by solving a linear equation system, which can be efficiently implemented with many numerical methods, and the experimental results on several benchmark data sets are also presented to show the effectiveness of our method.

One issue should be mentioned here is that independent of this work, (Xue *et al*., 2007) also proposes a similar discriminative regularization technique for classification. However, there is no detailed mathematical derivations in their technical report.

The rest of this paper is organized as follows. In section 2 we will briefly review the basic procedure of manifold regularization, since it is clearly related to this paper. The algorithm details of *DR* will be derived in section 3. In section 4 we sill present the experimental results, followed by the conclusions and discussions in section 5.

A Brief Review of Manifold Regularization

As mentioned in the introduction, *manifold regularization* (Belkin *et al*., 2006) seeks for an optimal classification function by minimizing the following loss

$$\bar{\mathcal{J}} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(f(\mathbf{x}_i), y_i) + \lambda\Omega(f) + \gamma\mathcal{I}(f) \qquad (1)$$

where $\Omega(f)$ measures the complexity of f, and $\mathcal{I}(f)$ penalizes the *geometrical property* of f. Generally, in *semi-supervised learning*, $\mathcal{I}(f)$ can be approximated by

$$\mathcal{I}(f) = \mathbf{f}^T\mathbf{L}\mathbf{f} = \sum_{ij} w_{ij}(f_i - f_j)^2$$

where w_{ij} measures the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, $f_i = f(\mathbf{x}_i)$, $\mathbf{f} = [f_1, f_2, \cdots, f_n]^T$. $\mathbf{L}$ is an $n \times n$ graph Laplacian matrix with its (i, j)-th entry

$$Ł_{ij} = \begin{cases} \sum_j w_{ij}, & if\ i = j \\ -w_{ij}, & otherwise \end{cases}$$

Therefore $\mathcal{I}(f)$ reflects the *variation*, or the *smoothness* over the intrinsic data manifold. For supervised learning, we can define a block-diagonal matrix $\mathbf{L}$ as

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_1 & & & \\ & \mathbf{L}_2 & & \\ & & \ddots & \\ & & & \mathbf{L}_C \end{bmatrix}$$

where C is the number of classes, and $\mathbf{L}_i$ is the *graph Laplacian* constructed on class i, *i.e.*, if we construct a supervised data graph on $\mathcal{X}$, then there only exists connections among the data points in the same class (see Fig.1)). Then by rearranging the elements in $\mathbf{f}$ as

$$\mathbf{f} = [\mathbf{f}_1^T, \mathbf{f}_2^T, \cdots, \mathbf{f}_C^T]^T$$

where $\mathbf{f}_1$ is the classification vector on class 1, we can adopt the following term as $\mathcal{I}(f)$

$$\mathcal{I}(f) = \sum_i \mathbf{f}_i^T\mathbf{L}_i\mathbf{f}_i = \mathbf{f}^T\mathbf{L}\mathbf{f}$$

which is in fact the summation of the smoothness of f within each class.

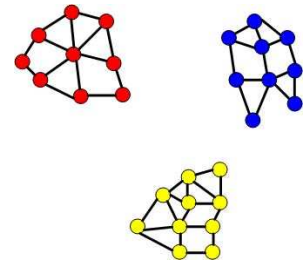


Figure 1: The data graph constructed in supervised manifold regularization. There are only connections among the data points in the same class.

Discriminative Regularization

Although (Belkin *et al*., 2006) proposed an elegant framework for learning by taking the smoothness of f into account, however, the final goal of learning (*e.g.*, classification or clustering) is to discriminate the data from different classes. Therefore, a reasonable classification function should be smooth within each classes, but simultaneously produce great contrast between different classes. Therefore, we propose to penalize the discriminality of f, rather than the smoothness of f directly when constructing it. Generally, the discriminality of f can be measured by the difference (or quotient) of the within-class scatterness and between-class scatterness, which can either be computed globally or locally.

Global Margin Based Discriminality

Similar to the *maximum margin criterion* (Li *et al*., 2006), we measure the discriminality of f by

$$\mathcal{D}_1(f) = \mathcal{W}(f) - \mathcal{B}(f) \qquad (2)$$

where $\mathcal{W}(f)$ and $\mathcal{B}(f)$ measure the *within-class scatterness* and *between-class scatterness* of f which can be computed by

$$\mathcal{W}(f) = \sum_{c=1}^{C}\sum_{\mathbf{x}_i\in\pi_c}\left\| f_i - \frac{1}{n_c}\sum_{\mathbf{x}_j\in\pi_c} f_j \right\|^2 \qquad (3)$$

$$\mathcal{B}(f) = \sum_{c=1}^{C} n_c \left\| \frac{1}{n_c}\sum_{\mathbf{x}_j\in\pi_c} f_j - \frac{1}{n}\sum_j f_j \right\|^2 \qquad (4)$$

where we use π_c to denote the c-th class, and $n_c = |\pi_c|$.

Define the *total scatterness* of f as

$$\mathcal{T}(f) = \mathcal{W}(f) + \mathcal{B}(f) = \sum_i \left\| f_i - \frac{1}{n}\sum_j f_j \right\|^2 \qquad (5)$$

$$
\begin{aligned}
\mathcal{W}(f) &= \sum_{c=1}^{C} \sum_{\mathbf{x}_i \in \pi_c} \left\| f_i - \frac{1}{n_c} \sum_{\mathbf{x}_j \in \pi_c} f_j \right\|^2 \\
&= \sum_{c=1}^{C} \sum_{\mathbf{x}_i \in \pi_c} \left(f_i^2 - \frac{2}{n_c} \sum_{\mathbf{x}_j \in \pi_c} f_j f_i + \frac{1}{n_c^2} \sum_{\mathbf{x}_k, \mathbf{x}_l \in \pi_c} f_k f_l \right) \\
&= \sum_i f_i^2 - \sum_{ij} g_{ij} f_i f_j \\
&= \frac{1}{2} \sum_{ij} g_{ij} (f_i - f_j)^2 = \frac{1}{2} \mathbf{f}^T \mathbf{L}_{\mathcal{W}} \mathbf{f}
\end{aligned}
$$

where

$$
g_{ij} = \begin{cases} 1/n_c, & \mathbf{x}_i, \mathbf{x}_j \in \pi_c \\ 0, & otherwise \end{cases}, \tag{6}
$$

and

$$
\begin{aligned}
\mathbf{f} &= [f_1, f_2, \cdots, f_n]^T \in \mathbb{R}^{n \times 1} && (7) \\
\mathbf{L}_{\mathcal{W}} &= diag(\mathbf{L}_{\mathcal{W}1}, \mathbf{L}_{\mathcal{W}2}, \cdots, \mathbf{L}_{\mathcal{W}C}) \in \mathbb{R}^{n \times n} && (8)
\end{aligned}
$$

and $\mathbf{L}_{\mathcal{W}c} \in \mathbb{R}^{n_c \times n_c}$ is a square matrix with its entries $\mathbf{L}_{\mathcal{W}c}(i,i) = 1 - \frac{1}{n_c}$, $\mathbf{L}_{\mathcal{W}c}(i,j) = -\frac{1}{n_c}$. Since

$$
\mathcal{T}(f) = \mathbf{f}^T \left(\mathbf{I} - \frac{1}{n} \mathbf{e}\mathbf{e}^T \right) \mathbf{f} = \mathbf{f}^T \mathbf{L}_{\mathcal{T}} \mathbf{f}, \tag{9}
$$

where $\mathbf{L}_{\mathcal{T}} = \mathbf{I} - \frac{1}{n}\mathbf{e}\mathbf{e}^T$ can be viewed as a centralization operator, $\mathbf{I}$ is the $n \times n$ identity matrix, $\mathbf{e} \in \mathbb{R}^{n \times 1}$ is a column vector with all its elements equal to 1. Then

$$
\mathcal{B}(f) = \mathcal{T}(f) - \mathcal{W}(f) = \mathbf{f}^T (\mathbf{L}_{\mathcal{T}} - \mathbf{L}_{\mathcal{W}}) \mathbf{f} = \mathbf{f}^T \mathbf{L}_{\mathcal{B}} \mathbf{f} \tag{10}
$$

where $\mathbf{L}_{\mathcal{B}} = \mathbf{L}_{\mathcal{T}} - \mathbf{L}_{\mathcal{W}}$. Then the discriminality of f (Eq.(2) can be rewritten as

$$
\mathcal{D}_1 = \mathbf{f}^T (\mathbf{L}_{\mathcal{W}} - \mathbf{L}_{\mathcal{B}}) \mathbf{f} \tag{11}
$$

Local Margin Based Discriminality

Before we go into the details of our *Local Margin Based Discriminative Regularization*, first let's define two types of neighborhoods(Wang, 2007):

Definition 1(*Homogeneous Neighborhoods*). For a data point $\mathbf{x}_i$, its ξ *nearest homogeneous neighborhood* $\mathcal{N}_i^o$ is the set of ξ most similar[1] data which are in the same class with $\mathbf{x}_i$.

Definition 2(*Heterogeneous Neighborhoods*).For a data point $\mathbf{x}_i$, its ζ *nearest heterogeneous neighborhood* $\mathcal{N}_i^e$ is the set of ζ most similar data which are not in the same class with $\mathbf{x}_i$.

Then the *neighborhood margin* γ_i for f on $\mathbf{x}_i$ is defined as

$$
\gamma_i = \sum_{\substack{k: \mathbf{x}_k \in \mathcal{N}_i^e \\ or \mathbf{x}_i \in \mathcal{N}_k^e}} \|f_i - f_k\|^2 - \sum_{\substack{j: \mathbf{x}_j \in \mathcal{N}_i^o \\ or \mathbf{x}_i \in \mathcal{N}_j^o}} \|f_i - f_j\|^2,
$$

[1]In this paper two data vectors are considered to be similar if the Euclidean distance between them is small, two data tensors are considered to be similar if the Frobenius norm of their difference tensor is small.

where $|\cdot|$ represents the cardinality of a set. The *total neighborhood margin* of f can be defined as

$$
\begin{aligned}
\gamma &= \sum\nolimits_i \gamma_i \\
&= \sum_i \left(\sum_{\substack{k: \mathbf{x}_k \in \mathcal{N}_i^e \\ or \mathbf{x}_i \in \mathcal{N}_k^e}} \|f_i - f_k\|^2 - \sum_{\substack{j: \mathbf{x}_j \in \mathcal{N}_i^o \\ or \mathbf{x}_i \in \mathcal{N}_j^o}} \|f_i - f_j\|^2 \right).
\end{aligned}
$$

The discriminality of f can be described by γ. Define two weight matrices

$$
\mathbf{W}^o(i,j) = \begin{cases} 1, & \mathbf{x}_i \in \mathcal{N}_j^o \ or \ \mathbf{x}_j \in \mathcal{N}_i^o \\ 0, & otherwise \end{cases} \tag{12}
$$

$$
\mathbf{W}^e(i,j) = \begin{cases} 1, & \mathbf{x}_i \in \mathcal{N}_j^e \ or \ \mathbf{x}_j \in \mathcal{N}_i^e \\ 0, & otherwise \end{cases} \tag{13}
$$

and their corresponding degree matrices

$$
\begin{aligned}
\mathbf{D}^o &= diag(d_1^o, d_2^o, \cdots, d_n^o) && (14) \\
\mathbf{D}^e &= diag(d_1^e, d_2^e, \cdots, d_n^e) && (15)
\end{aligned}
$$

where $d_i^o = \sum_j \mathbf{W}^o(i,j)$, $d_i^e = \sum_j \mathbf{W}^e(i,j)$. Then we can define the *homogeneous graph Laplacian* $\mathbf{L}^o$ and *heterogeneous graph Laplacian* $\mathbf{L}^e$ as

$$
\begin{aligned}
\mathbf{L}^o &= \mathbf{D}^o - \mathbf{W}^o && (16) \\
\mathbf{L}^e &= \mathbf{D}^e - \mathbf{W}^e && (17)
\end{aligned}
$$

With the above definitions, we can rewrite the discriminality of f can be defined by

$$
\mathcal{D}_2(f) = \mathbf{f}^T (\mathbf{L}^o - \mathbf{L}^e) \mathbf{f} \tag{18}
$$

Classification by Discriminative Regularization

The optimal f can be learned by minimizing

$$
J = \frac{1}{n} \sum_i^n \mathcal{L}(f_i, y_i) + \lambda \Omega(f) + \gamma \mathcal{D}_i(f) \tag{19}
$$

where $i = 1, 2$. Assume f lies in a *Reproducing Kernel Hilbert Space* (*RKHS*) $\mathcal{H}$, which is associated with a *Mercer kernel* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then the complexity of f can be measured by the induced norm $\|\cdot\|_k$ of f in $\mathcal{H}$. We have the following representor theorem

Theorem 1. *The minimizer of Eq.(19) admits an expansion*

$$
f(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i) \tag{20}
$$

Let the coefficient vector $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_n]^T$ and the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ with its (i,j)-th entry $\mathbf{K}(i,j) = k(\mathbf{x}_i, \mathbf{x}_j)$, then theorem 1 tells us that $\mathbf{f} = \mathbf{K}\boldsymbol{\alpha}$. Therefore we can rewrite Eq.(19) as

$$
J = \frac{1}{n} (\mathbf{K}\boldsymbol{\alpha} - \mathbf{y})^T (\mathbf{K}\boldsymbol{\alpha} - \mathbf{y}) + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} + \gamma \boldsymbol{\alpha}^T \mathbf{K} \mathbf{L}_{\mathcal{D}} \mathbf{K} \boldsymbol{\alpha}
$$

where $\mathbf{L}_{\mathcal{D}}$ is the *discriminative graph Laplacian* such that $\mathbf{L}_{\mathcal{D}} = \mathbf{L}_{\mathcal{W}} - \mathbf{L}_{\mathcal{B}}$ for global margin based criterion, and $\mathbf{L}_{\mathcal{D}} = \mathbf{L}^o - \mathbf{L}^e$ for local margin based criterion. Since

$$
\frac{\partial J}{\partial \boldsymbol{\alpha}} = 2 \left(\frac{1}{n} \mathbf{K}(\mathbf{K}\boldsymbol{\alpha} - \mathbf{y}) + (\lambda \mathbf{K} + \gamma \mathbf{K} \mathbf{L}_{\mathcal{D}} \mathbf{K}) \boldsymbol{\alpha} \right)
$$

Table 1: Discriminative Regularization for Classification

Inputs:
Training data set $\mathcal{X} = \{(\mathbf{x}_i, y_i\}_{i=1}^n\}$;
Testing data set $\mathcal{X}_\mathcal{T} = \{\mathbf{z}_i\}_{i=1}^m$;
Kernel type and parameters;
Regularization parameters λ, γ.

Output:
The predicted labels of the data in $\mathcal{X}_\mathcal{T}$.

Procedure:
1. Construct the kernel matrix $\mathbf{K}$ for the data set $\mathcal{X}$;
2. Construct the discriminative graph Laplacian $\mathbf{L}_\mathcal{D}$;
3. Compute $\boldsymbol{\alpha}$ using Eq.(21);
4. Construct the kernel matrix $\mathbf{K}_\mathcal{T} \in \mathbb{R}^{m\times n}$ between $\mathcal{X}$ and $\mathcal{X}_\mathcal{T}$, with its (i,j)-th entry $K_\mathcal{T}(i,j) = k(\mathbf{x}_i, \mathbf{z}_j)$;
5. Output the data labels by $\mathbf{y}_\mathcal{T} = \mathbf{K}_\mathcal{T}\boldsymbol{\alpha}$.

Let $\frac{\partial L}{\partial \boldsymbol{\alpha}} = 0$, we can get that

$$\boldsymbol{\alpha}^* = (\mathbf{K} + n\lambda\mathbf{I} + n\gamma\mathbf{L}_\mathcal{D}\mathbf{K})^{-1}\mathbf{y} \tag{21}$$

Table 1 summarizes the basic procedure of our *discriminative regularization* based classification method.

The Linear Case: To make our algorithm more efficient, we also derive the linear discriminative regularization based classification method. Mathematically, a linear classifier f predicts the label of $\mathbf{x}$ by

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{22}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector and $b \in \mathbb{R}$ is the offset. Then using our discriminative regularization framework, we can solve for the optimal f by minimizing

$$J' = \frac{1}{n}\sum_i (\mathbf{w}^T\mathbf{x}_i + b - y_i)^2 + \lambda\|\mathbf{w}\|^2 + \gamma\mathbf{w}^T\mathbf{X}^T\mathbf{L}_\mathcal{D}\mathbf{X}\mathbf{w}$$

Let

$$\mathbf{G} = \begin{bmatrix} \mathbf{x}_1^T/\sqrt{n} & 1/\sqrt{n} \\ \mathbf{x}_2^T/\sqrt{n} & 1/\sqrt{n} \\ \vdots & \vdots \\ \mathbf{x}_n^T/\sqrt{n} & 1/\sqrt{n} \\ \sqrt{\lambda}\mathbf{I}_d & \mathbf{0}_d \end{bmatrix}, \ \tilde{\mathbf{y}} = \begin{bmatrix} y_1/\sqrt{n} \\ y_2/\sqrt{n} \\ \vdots \\ y_n/\sqrt{n} \\ \mathbf{0}_d \end{bmatrix}, \ \tilde{\mathbf{w}} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$$

where $\mathbf{I}_d$ is the $d \times d$ *identity matrix*, and $\mathbf{0}_d \in \mathbb{R}^d \times 1$ is a zero column vector. Define the *augmented discriminative graph Laplacian* as

$$\tilde{\mathbf{L}}_\mathcal{D} = \begin{bmatrix} \mathbf{X}^T\mathbf{L}_\mathcal{D}\mathbf{X} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}$$

Then we can rewrite J' as

$$J' = \|\mathbf{G}\tilde{\mathbf{w}} - \tilde{\mathbf{y}}\|^2 + \gamma\tilde{\mathbf{w}}^T\tilde{\mathbf{L}}_\mathcal{D}\tilde{\mathbf{w}}$$

The partial derivative of J' with respect to $\tilde{\mathbf{w}}$ is

$$\frac{\partial J'}{\partial \tilde{\mathbf{w}}} = 2\left(\mathbf{G}^T(\mathbf{G}\tilde{\mathbf{w}} - \tilde{\mathbf{y}}) + \gamma\tilde{\mathbf{L}}_\mathcal{D}\tilde{\mathbf{w}}\right)$$

By setting $\frac{\partial J'}{\partial \tilde{\mathbf{w}}} = 0$, we can get the optimal $\tilde{\mathbf{w}}$ by solving the following linear equation system

$$(\mathbf{G}^T\mathbf{G} + \gamma\tilde{\mathbf{L}}_\mathcal{D})\tilde{\mathbf{w}} = \mathbf{G}^T\tilde{\mathbf{y}} \tag{23}$$

Table 2: Linear Discriminative Regularization

Inputs:
Training data set $\mathcal{X} = \{(\mathbf{x}_i, y_i\}_{i=1}^n\}$;
Testing data set $\mathcal{X}_\mathcal{T} = \{\mathbf{z}_i\}_{i=1}^m$;
Regularization parameters λ, γ.

Output:
The predicted labels of the data in $\mathcal{X}_\mathcal{T}$.

Procedure:
1. Construct the discriminative graph Laplacian $\mathbf{L}_\mathcal{D}$;
2. Compute the augmented weight vector $\tilde{\mathbf{w}}$ using Eq.(24);
3. Output the $\mathbf{z}_i$'s label by $\mathbf{y}_{\mathbf{z}_i} = \mathbf{w}^T\mathbf{z}_i + b$.

Then we can get

$$\tilde{\mathbf{w}} = (\mathbf{G}^T\mathbf{G} + \gamma\tilde{\mathbf{L}}_\mathcal{D})^{-1}\mathbf{G}^T\tilde{\mathbf{y}} \tag{24}$$

It is also worth mentioning that Eq.(23) is a $(d+1)\times(d+1)$ linear equation system which can be easily solved when d is small. When d is large but feature vectors are highly sparse with z number of non-zero entries, we can employ *Conjugate Gradient* (*CG*) methods to solve this system. *CG* techniques are *Krylov* methods that solve a system $\mathbf{A}\mathbf{x} = \mathbf{b}$ by repeatedly multiplying a candidate solution $\mathbf{x}^*$ by $\mathbf{A}$. In the case of *linear Discriminative RLS*, we can construct the matrix-vector product in the LHS of (3) in time $O(n(z+k))$, where k (typically small) is average number of entries per column in $\mathbf{L}_\mathcal{D}$. This is done by using an intermediate vector $\mathbf{X}\mathbf{w}$ and appropriately forming sparse matrix-vector products. Thus, the algorithm can employ very well-developed methods for efficiently obtaining the solution.

Table 2 summarizes the basic procedure of the *linear discriminative regularization* based classification method.

Experiments

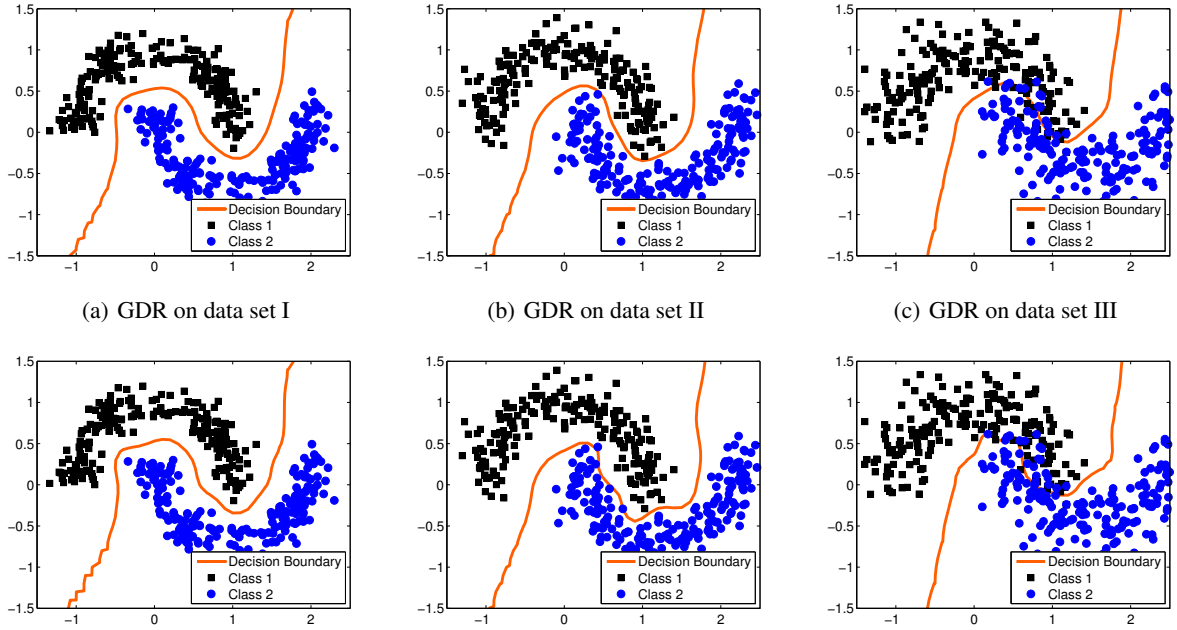
In this section, we will present a set of experimental results to demonstrate the effectiveness of our method. First let's see an illustrative example showing the difference between the global and local margin based criterions.

A Toy Example

In this part we will present a toy example to demonstrate the effectiveness of our discriminative regularization framework as well as the difference between the global margin based discriminative regularization (GDR) and local margin based discriminative regularization (LDR).

In our experiment, we construct three toy data sets based on the two-pattern (Zhou *et al*., 2004), in which the overlapping between the two classes becomes heavier and heavier from data set I to data set III. For both the GDR and LDR methods, we apply the Gaussian kernel with its width being set to 0.1, and the regularization parameters for the complexity and discriminality regularizers are also set to 0.1. The experimental results are shown in Fig.2.

From the figures we can observe that both GDR and LDR can discover the "optimal" decision boundary to discriminate the two classes. As we expected, the decision boundaries achieved by GDR are generally smoother than the ones



(a) GDR on data set I (b) GDR on data set II (c) GDR on data set III

(d) LDR on data set I (e) LDR on data set II (f) LDR on data set III

Figure 2: A toy example showing the difference between global margin based discriminative regularization (GDR) and local margin based discriminative regularization (LDR). In the example we construct three toy data sets using the two-moon pattern, and the overlapping between the two classes becomes heavier from data set I to data set III. From the figures we can observe that the decision boundaries achieved from GDR are generally smoother than the ones achieved from LDR.

obtained by LDR, since GDR maximizes the discriminality of the classification function based on the global information of the data set, while LDR pays more attention on the local discriminality of the decision function.

The Data Sets

We adopt three categories of data sets in our experiments:

- **UCI data**. We perform experiments on 9 UCI data sets[2]. The basic information of those data sets are summarized in table 4.
- **CMU PIE Face Data Set**(Sim *et al.*, 2003). This is a database of 41,368 face images of 68 people, each person under 13 different poses, 43 different illumination conditions, and with 4 different expressions. In our experiments, we only adopt a subset containing five near frontal poses (C05, C07, C09, C27, C29) and all the images under different illuminations and expressions. As a result, there are 170 images for each individual. In our experiments, all the images are resized to 32x32 for computational convenience.
- **COIL-20 Object Image Data Set**(Nene *et al.*, 1996). It is a database of gray-scale images of 20 objects. For each object there are 72 images of size 128x128. The images were taken around the object at the pose interval of 5 degrees. In our experiments, all the images are resized to 32x32.

[2] `http://mlearn.ics.uci.edu/MLRepository.html`

Table 3: Descriptions of the UCI data sets.

Data	Size (n)	Feature (N)	Class
Balance	625	3	2
BUPA	345	6	2
Glass	214	7	9
Iris	150	4	3
Ionosphere	351	34	2
Lenses	24	4	3
Pima	768	8	2
Soybean	307	35	19
Wine	138	13	3

Comparisons and Experimental Settings

In our experiments, we compare the performance of our discriminative regularization algorithm with four other competitive algorithms:

- **Linear Support Vector Machines** (*LSVM*). We use *libSVM* (Fan *et al.*, 2005) to implement the SVM algorithm with a linear kernel, and the cost parameter is searched from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$.
- **Kernel Support Vector Machines** (*KSVM*). We also use *libSVM* (Fan *et al.*, 2005) to im-

Table 4: Results on the UCI data sets (%).

Data	LSVM	SVM	LLapRLS	KLapRLS	LGDR	LLDR	GDR	LDR
Balance	87.76	92.04	88.34	91.63	89.20	88.72	91.45	**92.87**
BUPA	66.99	72.43	68.90	73.06	66.94	70.23	74.09	**82.13**
Glass	62.57	72.75	61.65	76.24	62.04	63.68	71.35	**74.20**
Iris	94.53	92.04	98.27	98.63	89.32	87.86	97.34	**98.77**
Ionosphere	87.78	95.14	91.63	98.30	85.20	88.20	95.24	**98.75**
Lenses	74.62	79.25	82.65	81.54	82.09	85.64	87.35	**87.68**
Pima	77.08	79.23	69.50	77.45	77.24	79.82	79.12	**87.70**
Soybean	**100.00**	97.75	**100.00**	**100.00**	99.04	99.68	**100.00**	**100.00**
Wine	95.57	79.87	97.11	96.54	**99.00**	**99.00**	98.35	98.24

Table 5: Experimental results in image data sets (%).

	PIE(20 train)	COIL(2 train)
LSVM	36.92	94.30
KSVM	33.27	97.24
LLapRLS	30.18	96.78
KLapRLS	29.49	98.37
LGDR	30.12	97.12
LLDR	28.77	99.02
GDR	27.93	98.93
LDR	**26.49**	**99.14**

plement the kernel SVM algorithm with a Gaussian kernel with its width tuned from $\{4^{-4}\sigma_0, 4^{-3}\sigma_0, 4^{-2}\sigma_0, 4^{-1}\sigma_0, \sigma_0, 4^1\sigma_0, 4^2\sigma_0, 4^3\sigma_0, 4^4\sigma_0\}$, where σ_0 is the mean distance between any two examples in the training set. The cost parameter is tuned in the same way as in LSVM.

- **Linear Laplacian Regularized Least Squares** (*LLapRLS*) (Belkin *et al*., 2006). The implementation code is downloaded from `http://manifold.cs.uchicago.edu/manifold_regularization/software.html`, in which the linear kernel is adopted in *RKHS*, and variance of the Gaussian in constructing the graph is set by the same way as in (Zelnik-Manor & Perona, 2005), and the extrinsic and intrinsic regularization parameters are searched from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3, 10^4\}$.
- **Kernel Laplacian Regularized Least Squares** (*KLapRLS*). All the experimental settings are the same as in LLapRLS except that we adopt the Gaussian kernel in RKHS with its width tuned from $\{4^{-4}\sigma_0, 4^{-3}\sigma_0, 4^{-2}\sigma_0, 4^{-1}\sigma_0, \sigma_0, 4^1\sigma_0, 4^2\sigma_0, 4^3\sigma_0, 4^4\sigma_0\}$, where σ_0 is the mean distance between any two examples in the training set.

Experimental Results

The experimental results are shown in table 4 and table 5, where we have highlighted the best performances. For the PIE data set, we randomly labeled 20 images for each person for training, while for the COIL data set, we randomly labeled 2 images per object. The performances reported in table 5 are averaged over 50 independent runs. From these results we can clearly see the effectiveness of our methods.

Conclusions

In this paper we propose a novel regularization approach for supervised classification. Instead of punishing the geometrical smoothness as in (Belkin *et al*., 2006), we directly punish the discriminality of the classification function in our method. Finally the experiments show that our algorithm is more effective than some conventional methods.

References

Belkin, M., Niyogi, P., and Sindhwani, V. Manifold Regularization: a Geometric Framework for Learning from Examples. *Journal of Machine Learning Research*. 2006.

Chapelle O, Schölkopf B, Zien A. Semi-supervised Learning. Cambridge, Massachusetts: The MIT Press, 2006

Chuzhanova, N.A., Jones, A.J., and Margetts, S.. Feature Selection for Genetic Sequence Classification. Bioinformatics. 1998.

Duda, R O, Hart, P E and Stork D G. Pattern Classification (2nd ed.). Wiley, 2000

Fukunaga, F. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 2nd edtion. 1990.

Fan, R. -E., Chen, P. -H., and Lin, C.-J. Working Set Selection Using Second Order Information for Training SVM. *Journal of Machine Learning Research* 6, 1889-1918. 2005.

Li, H., Jiang, T. and Zhang, K. Efficient and robust feature extraction by maximunmargin criterion. *TNN* 17(1), 157-165, 2006.

Nene, S. A., Nayar, S. K. and Murase, H. Columbia Object Image Library (COIL-20). *TR-CUCS-005-96*, February 1996.

Sebastiani, F. Machine Learning in Automated Text Categorization. *ACM Computing Surveys* 34(1): 1-47. 2002.

Sim, T., Baker, S. and Bsat, M. The CMU Pose, Illumination, and Expression Database. *IEEE-TPAMI* 25(12): 1615-1618, 2003.

Li, S. Z. & Jain, A. K. *Handbook of Face Recognition*. 2005.

Vapnik V. The Nature of Statistical Learning Theory. 1995.

Wang, F. and Zhang, C. Feature Extraction by Maximizing the Average Neighborhood Margin CVPR2007.

Xue H, Chen SC, Yang Q: Discriminative regularization: A new classifier learning method. NUAA-TR. 2007.

Zelnik-Manor, L. and Perona, P. Self-Tuning Spectral Clustering. In *NIPS 17*, 1601-1608. 2005.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J. and Schölkopf, B. Learning with Local and Global Consistency. *NIPS 16*. 2004.