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 discriminability 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 \(\mathcal{X} = \{(x_i, y_i)\}_{i=1}^n\), 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 \(\mathcal{X}_T\). 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 \(P(x, y)\), and the optimal classifier \(f\) we seek for should minimize the following expected loss (Vapnik, 1995)

\[
\mathcal{J} = \mathbb{E}_{P(x,y)} L(y, f(x)),
\]

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

\[
\hat{\mathcal{J}} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))
\]

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)

\[
\mathcal{J} \leq \hat{\mathcal{J}} + \varepsilon(n, d_F),
\]

where \(\varepsilon(n, d_F)\) measures the complexity of the hypothesis model. Therefore we can minimize the following structural risk to get an optimal \(f\)

\[
\tilde{\mathcal{J}} = \hat{\mathcal{J}} + \lambda \Omega(f),
\]

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 \(\tilde{\mathcal{J}}\) 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 discriminability 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 discriminability 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

\[
\mathcal{J} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x_i), y_i) + \lambda \Omega(f) + \gamma \mathcal{I}(f) \tag{1}
\]

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

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

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

\[
L_{ij} = \begin{cases} 
\sum_j w_{ij}, & \text{if } i = j \\
-w_{ij}, & \text{otherwise}
\end{cases}
\]

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 \(L\) as

\[
L = \begin{bmatrix}
L_1 & & \\
& L_2 & \\
& & \ddots \\
L_C & & & L_C
\end{bmatrix}
\]

where \(C\) is the number of classes, and \(L_i\) 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 \(f\) as

\[
f = [f_1^T, f_2^T, \cdots, f_C^T]^T
\]

where \(f_1\) is the classification vector on class 1, we can adopt the following term as \(\mathcal{I}(f)\)

\[
\mathcal{I}(f) = \sum_i f_i^T L_i f_i = f^T L f
\]

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 discriminability of \(f\), rather than the smoothness of \(f\) directly when constructing it. Generally, the discriminability 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 Discriminability

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

\[
\mathcal{D}_1(f) = \mathcal{W}(f) - \mathcal{B}(f) \tag{2}
\]

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

\[
\mathcal{W}(f) = \sum_{c=1}^{C} \sum_{x_i \in \pi_c} \left| f_i - \frac{1}{n_c} \sum_{x_j \in \pi_c} f_j \right|^2 \tag{3}
\]

\[
\mathcal{B}(f) = \sum_{c=1}^{C} n_c \left| \frac{1}{n_c} \sum_{x_j \in \pi_c} f_j - \frac{1}{n} \sum_j f_j \right|^2 \tag{4}
\]

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

Define the total scatterness of \(f\) as

\[
\mathcal{T}(f) = \mathcal{W}(f) + \mathcal{B}(f) = \sum_i \left| f_i - \frac{1}{n} \sum_j f_j \right|^2 \tag{5}
\]
\[ W(f) = \sum_{c=1}^{C} \sum_{x_i \in \pi_c} \left\| f_i - \frac{1}{n_c} \sum_{x_j \in \pi_c} f_j \right\|^2 \]
\[ = \sum_{c=1}^{C} \sum_{x_i \in \pi_c} \left( f_i^2 - \frac{2}{n_c} \sum_{x_j \in \pi_c} f_j f_i + \frac{1}{n_c^2} \sum_{x_j \in \pi_c} f_k f_l \right) \]
\[ = \sum_{i j} f_i^2 - \sum_{i j} g_{i j} f_i f_j \]
\[ = \frac{1}{2} \sum_{i j} g_{i j} (f_i - f_j)^2 = \frac{1}{2} f^T L_W f \]

where
\[ g_{i j} = \begin{cases} 1/n_c, & x_i, x_j \in \pi_c \\ 0, & \text{otherwise} \end{cases} \]
and
\[ f = \begin{bmatrix} f_1, f_2, \ldots, f_n \end{bmatrix}^T \in \mathbb{R}^{n \times 1} \]
\[ L_W = \text{diag}(L_{W1}, L_{W2}, \ldots, L_{Wc}) \in \mathbb{R}^{n \times n} \]
\[ L_{Wc} = \begin{bmatrix} 1 & -1/n_c \end{bmatrix}, \quad L_{Wc}(i, i) = -1/n_c. \]

Since
\[ T(f) = f^T \left( I - \frac{1}{n} ee^T \right) f = f^T L_T f, \]

where \( L_T = I - \frac{1}{n} ee^T \) can be viewed as a centralization operator, \( I \) is the \( n \times n \) identity matrix, \( e \in \mathbb{R}^{n \times 1} \) is a column vector with all its elements equal to 1. Then
\[ B(f) = T(f) - W(f) = f^T (L_T - L_W) f = f^T L_B f \]

where \( L_B = L_T - L_W \). Then the discriminability of \( f \) (Eq.2) can be rewritten as
\[ D_1 = f^T (L_W - L_B) f \]

Local Margin Based Discriminability

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 \( x_i \), its \( \xi \) nearest homogeneous neighborhood \( N_i^0 \) is the set of \( \xi \) most similar data which are in the same class with \( x_i \).

Definition 2(Heterogeneous Neighborhoods). For a data point \( x_i \), its \( \zeta \) nearest heterogeneous neighborhood \( N_i^c \) is the set of \( \zeta \) most similar data which are not in the same class with \( x_i \).

Then the neighborhood margin \( \gamma_i \) for \( f \) on \( x_i \) is defined as
\[ \gamma_i = \sum_{k, x_k \in N_i^0} \| f_i - f_k \|^2 - \sum_{j, x_j \in N_i^c} \| f_i - f_j \|^2, \]

where \(|\cdot|\) represents the cardinality of a set. The total neighborhood margin of \( f \) can be defined as
\[ \gamma = \sum_{i} \gamma_i \]
\[ = \sum_{i} \left( \sum_{j, x_j \in N_i^0} \| f_i - f_k \|^2 - \sum_{j, x_j \in N_i^c} \| f_i - f_j \|^2 \right). \]

The discriminability of \( f \) can be described by \( \gamma \). Define two weight matrices
\[ W^0(i, j) = \begin{cases} 1, & x_i \in N_j^0 \text{ or } x_j \in N_i^0 \\ 0, & \text{otherwise} \end{cases} \]
\[ W^c(i, j) = \begin{cases} 1, & x_i \in N_j^c \text{ or } x_j \in N_i^c \\ 0, & \text{otherwise} \end{cases} \]

and their corresponding degree matrices
\[ D^0 = \text{diag}(d_1^0, d_2^0, \ldots, d_n^0) \]
\[ D^c = \text{diag}(d_1^c, d_2^c, \ldots, d_n^c) \]

where \( d_i^0 = \sum_j W^0(i, j) \), \( d_i^c = \sum_j W^c(i, j) \). Then we can define the homogeneous graph Laplacian \( L^0 \) and heterogeneous graph Laplacian \( L^c \) as
\[ L^0 = D^0 - W^0 \]
\[ L^c = D^c - W^c \]

With the above definitions, we can rewrite the discriminability of \( f \) can be defined by
\[ D_2(f) = f^T (L^0 - L^c) f \]

Classification by Discriminative Regularization

The optimal \( f \) can be learned by minimizing
\[ J = \frac{1}{n} \sum_{i} \mathcal{L}(f_i, y_i) + \lambda \Omega(f) + \gamma D_1(f) \]

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

Theorem 1. The minimizer of Eq.(19) admits an expansion
\[ f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) \]

Let the coefficient vector \( \alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n]^T \) and the kernel matrix \( K \in \mathbb{R}^{n \times n} \) with its \((i, j)\)th entry \( K(i, j) = k(x_i, x_j) \), then theorem 1 tells us that \( f = K \alpha \). Therefore we can rewrite Eq.(19) as
\[ J = \frac{1}{n} (K \alpha - y)^T (K \alpha - y) + \lambda \alpha^T K \alpha + \gamma \alpha^T KL_D K \alpha \]

where \( L_D \) is the discriminative graph Laplacian such that \( L_D = L_W - L_B \) for global margin based criterion, and \( L_D = L^0 - L^c \) for local margin based criterion. Since
\[ \frac{\partial J}{\partial \alpha} = 2 \left( \frac{1}{n} K (K \alpha - y) + (\lambda K + \gamma KL_D K) \alpha \right) \]
can solve for the optimal vector $\alpha$. Then using our discriminative regularization framework, we can discover the "optimal" decision boundary to discriminate between the two classes.

**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 $x$ by

$$ f(x) = w^T x + b $$

where $w \in \mathbb{R}^d$ 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

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

Let

$$ G = \begin{bmatrix} x_1^T / \sqrt{n} & 1 / \sqrt{n} \\ x_2^T / \sqrt{n} & 1 / \sqrt{n} \\ \vdots & \vdots \\ x_n^T / \sqrt{n} & 1 / \sqrt{n} \\ \sqrt{\lambda} I_d & 0_d \end{bmatrix}, \quad \hat{y} = \begin{bmatrix} y_1 / \sqrt{n} \\ y_2 / \sqrt{n} \\ \vdots \\ y_n / \sqrt{n} \\ 0_d \end{bmatrix}, \quad \hat{w} = \begin{bmatrix} w \\ b \end{bmatrix} $$

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

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

Then we can rewrite $J'$ as

$$ J' = \|G\hat{w} - \hat{y}\|^2 + \gamma \hat{w}^T \tilde{L}_D \hat{w} $$

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

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

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

$$ (G^T G + \gamma \tilde{L}_D) \hat{w} = G^T \hat{y} $$

Then we can get

$$ \hat{w} = (G^T G + \gamma \tilde{L}_D)^{-1} G^T \hat{y} $$

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 $Ax = b$ by repeatedly multiplying a candidate solution $x^*$ by $A$. 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 $L_D$. This is done by using an intermediate vector $Xw$ and appropriately forming sparse matrix-vector products. Thus, the algorithm can employ very well-developed methods for efficiently obtaining the solution.

Table 1 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 criteria.

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 discriminability 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...
obtained by LDR, since GDR maximizes the discriminability of the classification function based on the global information of the data set, while LDR pays more attention on the local discriminability 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 degress. In our experiments, all the images are resized to 32x32.

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-

---

\(^2\)http://mlearn.ics.uci.edu/MLRepository.html
Table 4: Results on the UCI data sets (%).

<table>
<thead>
<tr>
<th>Data</th>
<th>LSVM</th>
<th>SVM</th>
<th>LlapRLS</th>
<th>KLapRLS</th>
<th>LGDR</th>
<th>LLDRI</th>
<th>GDR</th>
<th>LDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balance</td>
<td>87.76</td>
<td>92.04</td>
<td>88.34</td>
<td>91.63</td>
<td>89.20</td>
<td>88.72</td>
<td>91.45</td>
<td>92.87</td>
</tr>
<tr>
<td>BUPA</td>
<td>66.99</td>
<td>72.43</td>
<td>68.90</td>
<td>73.06</td>
<td>66.94</td>
<td>70.23</td>
<td>74.09</td>
<td>82.13</td>
</tr>
<tr>
<td>Glass</td>
<td>62.57</td>
<td>72.75</td>
<td>61.65</td>
<td>76.24</td>
<td>62.04</td>
<td>63.68</td>
<td>71.35</td>
<td>74.20</td>
</tr>
<tr>
<td>Iris</td>
<td>94.53</td>
<td>92.04</td>
<td>98.27</td>
<td>98.63</td>
<td>89.32</td>
<td>97.64</td>
<td>97.34</td>
<td>98.77</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>87.78</td>
<td>95.14</td>
<td>91.63</td>
<td>98.30</td>
<td>85.20</td>
<td>88.20</td>
<td>95.24</td>
<td>98.75</td>
</tr>
<tr>
<td>Lenses</td>
<td>74.62</td>
<td>79.25</td>
<td>82.65</td>
<td>81.54</td>
<td>82.09</td>
<td>85.64</td>
<td>87.35</td>
<td>87.68</td>
</tr>
<tr>
<td>Pima</td>
<td>77.08</td>
<td>79.23</td>
<td>69.50</td>
<td>77.45</td>
<td>77.24</td>
<td>79.82</td>
<td>79.12</td>
<td>87.70</td>
</tr>
<tr>
<td>Soybean</td>
<td>100.00</td>
<td>97.75</td>
<td>100.00</td>
<td>100.00</td>
<td>99.04</td>
<td>99.68</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Wine</td>
<td>95.57</td>
<td>97.87</td>
<td>97.11</td>
<td>96.54</td>
<td>99.00</td>
<td>99.00</td>
<td>98.35</td>
<td>98.24</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>PIE(20 train)</th>
<th>COIL(2 train)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSVM</td>
<td>36.92</td>
<td>94.30</td>
</tr>
<tr>
<td>K SVM</td>
<td>33.27</td>
<td>97.24</td>
</tr>
<tr>
<td>LLapRLS</td>
<td>30.18</td>
<td>96.78</td>
</tr>
<tr>
<td>KLapRLS</td>
<td>29.49</td>
<td>98.37</td>
</tr>
<tr>
<td>LGDR</td>
<td>30.12</td>
<td>97.12</td>
</tr>
<tr>
<td>LLDR</td>
<td>28.77</td>
<td>99.02</td>
</tr>
<tr>
<td>GDR</td>
<td>27.93</td>
<td>98.93</td>
</tr>
<tr>
<td>LDR</td>
<td>26.49</td>
<td>99.14</td>
</tr>
</tbody>
</table>

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 discriminability of the classification function in our method. Finally the experiments show that our algorithm is more effective than some conventional 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 discriminability of the classification function in our method. Finally the experiments show that our algorithm is more effective than some conventional methods.

References


