Pattern Recognition 44 (2011) 1777–1784

Contents lists available at ScienceDirect

Pattern Recognition

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

Sparse regularization for semi-supervised classification

Mingyu Fan [a], Nannan Gu [b], Hong Qiao [b], Bo Zhang [a,*]

[a] LSEC and Institute of Applied Mathematics, AMSS, Chinese Academy of Sciences, Beijing 100190, China
[b] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 21 June 2010
Received in revised form
29 September 2010
Accepted 15 February 2011
Available online 22 February 2011

Keywords:
Regularization theory
Semi-supervised learning
Regularized least square classification
Dimensionality reduction

ABSTRACT

Manifold regularization (MR) is a promising regularization framework for semi-supervised learning, which introduces an additional penalty term to regularize the smoothness of functions on data manifolds and has been shown very effective in exploiting the underlying geometric structure of data for classification. It has been shown that the performance of the MR algorithms depends highly on the design of the additional penalty term on manifolds. In this paper, we propose a new approach to define the penalty term on manifolds by the sparse representations instead of the adjacency graphs of data. The process to build this novel penalty term has two steps. First, the best sparse linear reconstruction coefficients for each data point are computed by the l^1-norm minimization. Secondly, the learner is subject to a cost function which aims to preserve the sparse coefficients. The cost function is utilized as the new penalty term for regularization algorithms. Compared with previous semi-supervised learning algorithms, the new penalty term needs less input parameters and has strong discriminative power for classification. The least square classifier using our novel penalty term is proposed in this paper, which is called the Sparse Regularized Least Square Classification (S-RLSC) algorithm. Experiments on real-world data sets show that our algorithm is very effective.

1. Introduction

Regularization theory was originally introduced to solve ill-posed inverse problems [18]. In the past decades, regularization has shown great power and been applied in many areas of machine learning, such as regression, clustering, classification and model selection [10]. Many state-of-art machine learning algorithms, including Support Vector Machines (SVMs) [19], Regularized Neural Networks (RNNs) [14] and Regularized Least Square Classifier (RLSC) [17], can be derived from the regularization framework.

Recently, in [4], Belkin et al. proposed a general Manifold Regularization (MR) framework for a full range of learning problems from unsupervised, semi-supervised to supervised. The framework was developed in the setting of Reproducing Kernel Hilbert Spaces (RKHS), and a new Representor theorem was obtained in this setting for the regularization framework. In contrast to the traditional regularization theory, which concentrates on the complexity of functions in the functional space, the MR framework supplements an additional penalty term to the traditional regularization based on the assumption that data lie on an intrinsic low-dimensional manifold. The additional penalty term is used to measure the smoothness of functions on data manifolds, which will be referred to as the (penalty) term on manifolds for short. Such a term can improve the performance of the obtained learner by exploiting the intrinsic structure of data. The MR algorithms, including the Laplacian Regularized Least Square Classification (LapRLSC) and the Laplacian SVM (LapSVM) methods [4], have been shown especially useful and efficient in semi-supervised learning problems when both labeled examples and unlabeled examples are available for learning.

Many semi-supervised learning methods can be unified in the MR framework. The Discriminatively Regularized Least Square Classification (DRLSC) method builds the penalty term on manifolds by integrating both discriminative and geometrical information in each local region [23]. Although the method is proposed as a supervised learning method, it can be applied to semi-supervised classification problems. The MR framework can also unify many of the graph-based semi-supervised learning algorithms by ignoring the complexity of functions, which only have the penalty term on manifolds in the framework. Zhu et al. proposed a semi-supervised learning method called the Gaussian fields and harmonic functions (GFHF) method, based on a Gaussian random field model [24]. Wang and Zhang proposed a semi-supervised learning algorithm by using the local linear reconstruction coefficients, which is similar to the GFHF method [20].

Despite the success of these semi-supervised classification methods, there are still some issues that have not yet been properly addressed. In particular,

(1) *Neighbors selection.* Many graph-based methods, including the MR framework, define the adjacency graphs by using a fixed neighborhood size for all the data points. However, a fixed

* Corresponding author. Tel.: +86 10 6265 1358.
E-mail addresses: fanminyu@amss.ac.cn (M. Fan), gunannan@gmail.com (N. Gu), hong.qiao@ia.ac.cn (H. Qiao), b.zhang@amt.ac.cn (B. Zhang).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.02.013

neighborhood size causes the difficulty of parameter selection and cannot be adaptive to uneven data.

(2) *Manifold assumption.* Many graph-based methods, including the MR framework, assume that high-dimensional data distribute on a low-dimensional manifold. However, for many types of data, we lack convincing evidence for the manifold structure.

(3) *Explicit classifier for new points.* Some graph-based methods do not have an explicit multi-class classifier for novel examples, which limits their application in on-line decision making tasks.

To address the above issues, we propose the sparse regularization (SR) approach for semi-supervised learning. A novel penalty term is defined using the sparse representation [22] of the data. With the novel penalty term, the approach can derive classifiers in the MR framework. Therefore, the proposed SR approach not only inherits the advantages of fewer parameters and highly discriminative ability from the sparse representation, but also has a natural out-of-sample extension for novel examples, which is inherited from the MR framework. Experiments on real-world data sets demonstrate the effectiveness and highly discriminative ability of our approach.

The rest of this paper is organized as follows. Some previous works are introduced in Section 2. The proposed SR approach and the derived Sparse Regularized Least Square Classification (S-RLSC) algorithm are presented in Section 3. Then in Section 4, experiments on benchmark real-world data sets are reported. Finally, we conclude this paper in Section 5.

2. Previous works

In a general semi-supervised classification problem, the training data set is represented as $\{(x_i,z_i),x_{l+j},i=1,\ldots,l,j=1,\ldots,u\}$, where l is the number of labeled data points, u is the number of unlabeled data points, $x_i\in\mathbb{R}^N$ is a data point and $z_i\in\{-1,1\}$ is the class label of x_i.

2.1. Regularization on explicit functions

Belkin et al. [4] proposed the MR framework based on the theory of RKHS. Assuming that f is a real-valued function in the RKHS $\mathcal{H}_K$, the MR framework can be expressed in the form

$$f^*=\underset{f\in\mathcal{H}_K}{\operatorname{argmin}}\left\{\frac{1}{l}\sum_{i=1}^{l}V(x_i,z_i,f)+\gamma_A\|f\|_K^2+\gamma_I\|f\|_I^2\right\}, \tag{1}$$

where V is some loss function, $\|f\|_K^2$ is the norm of the function in $\mathcal{H}_K$ which controls the complexity of the classifier and $\|f\|_I^2$ is the penalty term to regularize the smoothness of the function on manifolds. If

$$\|f\|_I^2=\frac{1}{(l+u)^2}\sum_{i,j=1}^{l+u}(f(x_i)-f(x_j))^2w_{ij}, \tag{2}$$

where w_{ij} are edge weights in the data adjacency graph, then it follows by the Representer Theorem [4, Theorem 2] that the solution of the optimization problem (1) admits the representation

$$f^*(x)=\sum_{i=1}^{u+l}\alpha_i k(x_i,x) \tag{3}$$

in terms of the labeled and unlabeled samples, where $k(\cdot,\cdot)$ is some Mercer kernel function associated with the RKHS $\mathcal{H}_K$. For different choices of loss function V and $\|f\|_I^2$, different MR algorithms can be derived from the MR framework (1). For example, if the loss function V is defined to the square loss function

$$V(x,z,f)=(z-f(x))^2$$

and $\|f\|_I^2$ is defined by (2), then the Laplacian Regularized Least Square Classifier (LapRLSC) can be obtained; if V is chosen as the hinge loss function

$$V(x_i,z_i,f)=\begin{cases}1-z_if(x_i) & \text{if } z_if(x_i)>0\\ 0 & \text{otherwise}\end{cases}$$

and $\|f\|_I^2$ is again given as in (2), then the Laplacian Regularized Support Vector Machines (LapSVMs) can be obtained (see [4]). By using the square loss function as the loss function V and by making the best use of the underlying discriminative and geometrical information of the data manifold to define the penalty term $\|f\|_I^2$, a new MR algorithm called the DRLSC algorithm was obtained in [23] from the MR framework (1). Although the DRLSC algorithm was proposed as a supervised learning algorithm, it is similar with the LapRLSC algorithm and can be used as a semi-supervised learning algorithm.

2.2. Regularization on implicit functions

If the parameter γ_A is set to be zero, then the second term of the MR framework (1) that controls the complexity of the classifier vanishes. As a result, the feasible function f in (1) is not restricted to being in the RKHS $\mathcal{H}_K$. In fact, the feasible function f can be any function; in particular, it can be required to be an unknown or implicit function satisfying that $z_i=f(x_i)$ for $i=1,\ldots,l$, so the error part $(1/l)\sum_{i=1}^{l}V(x_i,z_i,f)$ vanishes. Thus, the MR framework has only the penalty term $\|f\|_I^2$ on manifolds, where f is an unknown or implicit function satisfying that $z_i=f(x_i)$ for $i=1,\ldots,l$. For unlabeled data points x_{l+i} $(i=1,\ldots,u)$ define implicitly $z_{l+i}=f(x_{l+i})$ for $i=1,\ldots,u$, which are unknown and regarded as the labels of the unlabeled data points x_{l+i} (or values of the function f at x_{l+i}) $(i=1,\ldots,u)$. If we further define $\|f\|_I^2=(1/2)\sum_{i,j}^{l+u}w_{ij}(z_i-z_j)^2$, where w_{ij} are edge weights in the data adjacency graph as defined in Subsection 2.1, then the labels z_{l+i} of unlabeled data points can be computed by minimizing $\|f\|_I^2$. Therefore, the minimization of the penalty term on manifolds can also give new semi-supervised learning algorithms.

Belkin and Niyogi proposed a manifold learning based classifier [3], which is built by the eigenvectors of the Laplacian matrix. Zhu et al. introduced the GFHF method based on a random field model [24]. The GFHF method is defined on a weighted graph superimposed on the whole data set, which comprises both labeled and unlabeled data points. The pairwise similarities between the data points are defined as

$$w_{ij}=\exp\left(-\sum_{k=1}^{N}\frac{(x_{ik}-x_{jk})^2}{\sigma_k^2}\right),$$

where x_{ik} is the k-th component of the data point x_i and σ_k is the length-scale hyper-parameter for the k-th component.

Let $W=(w_{ij})$ be the $(l+u)\times(l+u)$ similarity matrix, let D be the diagonal matrix of order $l+u$ with $D_{ii}=\sum_{j=1}^{l+u}w_{ij}$ and let $L=D-W$. Then the GFHF method minimizes the quadratic function

$$E(Z)=\frac{1}{2}\sum_{i,j}w_{ij}(z_i-z_j)^2=Z^TLZ, \tag{4}$$

where $Z=(z_1,\ldots,z_l,z_{l+1},\ldots,z_{l+u})^T$ with $z_{l+i}=f(x_{l+i})$, $i=1,\ldots,u$. The similarity matrix W (and also the diagonal matrix D) can be split into four blocks:

$$W=\begin{pmatrix}W_{ll} & W_{lu}\\ W_{ul} & W_{uu}.\end{pmatrix}$$

Assume that $Z=(Z_l^TZ_u^T)^T$, where $Z_l=(z_1,\ldots,z_l)^T$ and $Z_u=(z_{l+1},\ldots,z_{l+u})^T$. Suppose Z minimizes the function in (4). Then

Z_u can be computed as follows:

$$Z_u = (D_{uu} - W_{uu})^{-1} W_{ul} Z_l.$$

Similarly, Wang and Zhang [20] proposed to minimize the function in (4) with different pairwise similarities w_{ij}. They first compute the locally linear reconstruction coefficients by

$$M_i = \operatorname{argmin} \left\| x_i - \sum_{j \in N_K^i} M_{ij} x_j \right\|^2$$
$$\text{s.t.} \quad \sum_j M_{ij} = 1, \quad M_{ij} \geq 0,$$
$$M_{ij} = 0 \quad \text{if } j \notin N_K^i,$$

which can be easily solved by a quadratic optimization problem. Then the pairwise similarities w_{ij} are given as $w_{ij} = M_{ij} + M_{ji}$. The subsequent classification procedure of the method is the same with the GFHF method.

3. Sparse regularization

Instead of minimizing the smoothness of functions on data manifolds, we propose to keep the strong discriminative ability of sparse representation for data by minimizing a cost function. In this section, we first present the sparse representation method used in this paper and then give the Sparse Regularization (SR) method and the Sparse Regularized Least Square Classification (S-RLSC) algorithm.

In order to avoid confusion, we give a list of the main notations used in this paper in Table 1. Throughout this paper, all data points and the corresponding label vectors are in the form of column vectors and denoted by lowercases. All sets are represented by capital curlicue letters. Matrices are denoted by normal capital letters.

3.1. Sparse representation of data

Sparse representation [2] refers to the representation that accounts for most or all information of a signal within a linear combination of a small number of elementary signals called basis.

Table 1
Notations.

$\mathbb{R}^N$	The input N-dimensional Euclidean space
X	$X = [x_1, \ldots, x_l, \ldots, x_{l+u}] \in \mathbb{R}^{N \times (l+u)}$ is the training data matrix. $\{x_i\}_{i=1}^{l}$ are labeled points, and $\{x_i\}_{i=l+1}^{l+u}$ are unlabeled points
C	The number of classes that the samples belong to
Y	$Y = (y_1, \ldots, y_l, \mathbf{0}, \ldots, \mathbf{0}) \in \mathbb{R}^{C \times (l+u)}$ is the 0–1 label matrix. $y_i \in \mathbb{R}^C$ is the label vector of x_i, and all components of y_i are 0 s except one being 1
Z	$Z = (z_1, z_2, \ldots, z_l)^T \in \mathbb{R}^l$ is the label vector with z_i the label of x_i, where $z_i \in \{-1, 1\}$ if $C = 2$
$F(\cdot)$	$F(x) = (f_1(x), \ldots f_C(x))^T$ is the discriminative vector function. The index of the class which x belongs to is that of the component with the maximum value.
$g(\cdot)$	The classifier function with $g(x) \in \{1, 2, \ldots, C\}$ being the index of the class that x comes from
$k(u,v)$	Kernel function of variables u and v
K	Kernel matrix $K = \{k(x_i, x_j)\} \in \mathbb{R}^{(l+u) \times (l+u)}$
B	$B = (b_1, \ldots, b_{l+u}) \in \mathbb{R}^{C \times (l+u)}$. Its columns are the coefficients of the kernel function to represent the discriminative function $F(\cdot)$
$\|\cdot\|_0$	l^0 norm where $\|w\|_0$ counts the number of nonzero entries for a vector $w \in \mathbb{R}^m$
$\|\cdot\|_1$	l^1 norm where $\|w\|_1 = \sum_{i=1}^m \|w_i\|$ for a vector $w \in \mathbb{R}^m$
$\|\cdot\|_2$	l^2 norm where $\|w\|_2 = \sqrt{\sum_{i=1}^m w_i^2}$ for a vector $w \in \mathbb{R}^m$

Although originally proposed for compression and encoding signals, it is naturally discriminative [21] and thus can be used to solve classification tasks.

Given sufficient data points, any data point from the data set approximately lies in the linear span of the other data points, which can be described as

$$x_i = X\alpha_i, \quad i = 1, \ldots, l+u,$$

where $X = (x_1, \ldots, x_l, \ldots, x_{l+u}) \in \mathbb{R}^{N \times (l+u)}$ is the training data matrix, $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ii-1}, 0, \alpha_{ii+1}, \ldots, \alpha_{il+u})^T \in \mathbb{R}^{l+u}$ is the reconstruction coefficients. Then the sparse representation of x_i can be found via the l_0-minimization problem as

$$\alpha_i^* = \operatorname*{argmin}_{\alpha_i} \|\alpha_i\|_0 \quad \text{s.t.} \quad x_i = X\alpha_i, \tag{5}$$

where $\|\cdot\|_0$ denotes the l^0-norm which is defined to be the number of nonzero entries in a vector. The solution α_i^*, which is the sparse representation of x_i, describes how to express x_i with a linear combination of the smallest quantity of the training data points. It has been shown that the training data points of the same class associated with the tested one are preferred in the sparse sense. Therefore, using the training data points as basis elements, the sparse representation of the tested data point has strong discriminative ability.

However, this problem (5) is proved to be an NP-hard problem [1]. Fortunately, theoretical results show that if the solution α_i^* obtained is sparse enough then the solution of the l^1 and l^0 minimization problems are equivalent [12,6,7]. Therefore, the sparse representation problem can also be addressed by the following l^1-optimization problem:

$$\alpha_i^* = \operatorname*{argmin}_{\alpha_i} \|\alpha_i\|_1 \quad \text{s.t. } x_i = X\alpha_i.$$

Taking account of the effect of noise, insufficient training points or outliers, the optimization problem can be formulated as

$$\alpha_i^* = \operatorname*{argmin}_{\alpha_i} (\|\alpha_i\|_1 + \lambda \|e_i\|_p^q) \quad \text{s.t. } x_i = X\alpha_i + e_i, \tag{6}$$

where $e_i = (e_{i1}, \ldots, e_{iN})^T \in \mathbb{R}^N$ is the error vector, $\|e_i\|_p^q = (\sum_{j=1}^N |e_{ij}|^{1/p})^q$, p and q are positive integers.

The problem (6) is a general approach of constructing the sparse representation of data. If $p = 1/2$ and $q = 2$, then the traditional lasso method can be obtained, while if $p = q = 1$ and $\lambda = 1$ then the l^1-graph method of Wright et al. [22] can be obtained.

For convenience, in this paper we set $p = q = 1$ and $\lambda = 1$. This gives the l^1-graph method:

$$\alpha_i^* = \operatorname*{argmin}_{\alpha_i} (\|\alpha_i\|_1 + \|e_i\|_1) \quad \text{s.t } .x_i = X\alpha_i + e_i. \tag{7}$$

Let $\theta_i = (\alpha_i^T e_i^T)^T$ and $\hat{X} = (XI)$, where I is an $N \times N$ identity matrix. Then the proposed l^1-minimization problem (7) can be transformed to

$$\theta_i^* = \operatorname*{argmin}_{\theta_i} \|\theta_i\|_1 \quad \text{s.t. } x_i = \hat{X}\theta_i. \tag{8}$$

This minimization problem can be easily solved using convex programming [5,8].

3.2. Sparse regularization and S-RLSC algorithm

In our method, y_i is a C-dimensional label vector with the elements 0 or 1, where C is the number of classes. If x_i belongs to the k-th class, then the k-th component of y_i takes the value 1 and the rest components take the value 0. Our discriminative vector function $F(\cdot)$ is defined as

$$F(x) = (f_1(x), f_2(x), \ldots f_C(x))^T,$$

where

$$f_s(x)=\sum_{i=1}^{u+l} b_{si}k(x_i,x),\quad s=1,\ldots,C.$$

The class of x is determined as that of the component which takes the maximum value of $f_{s(x)}$, $s=1,2,\ldots,C$.

Many existing semi-supervised learning algorithms are based on the manifold assumption that the data points distribute on an intrinsic low-dimensional manifold and that if two data points x_i, x_j are close on the manifold then their labels are the same similar, that is, the label function varies smoothly on the manifold. However, this assumption is very restriction and brings an extra input parameter, the neighborhood size.

On the other hand, it is clear that the sparse representing coefficients α_{ij} of the data point x_i characterizes how the rest points contribute to the sparse representation of x_i and, therefore, is naturally discriminative. Thus, it is meaningful to require that the discriminative function on the data points keeps the sparse representing coefficients in the square loss sense. Hence, the label of a data point will be reconstructed by the labels of the other data points using the sparse representing coefficients. Based on this, we define the penalty term in this section to measure the error of the discriminative function F in preserving the sparse representing coefficients in the average sense as follows:

$$\|F\|_I^2=\frac{1}{(l+u)^2}\sum_{i=1}^{l+u}\|F(x_i)-\sum_{j=1}^{l+u}\alpha_{ij}F(x_j)\|_2^2, \tag{9}$$

where α_{ij} is the j-th element of α_i.

For a given Mercer kernel function $k(u,v)$, there is an associated reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ of functions. It can be constructed by considering the space of finite linear combinations $\sum_i\eta_i k(u_i,\cdot)$ of kernels and taking completion with respect to the inner product given by $\langle k(u,\cdot),k(v,\cdot)\rangle_K=k(u,v)$. Thus, for any functions $f,g\in\mathcal{H}_K$, if $f=\sum_i\eta_i k(u_i,\cdot), g=\sum_i\rho_i k(u_i,\cdot)$ then the inner product of f,g and the square of the norm of f can be expressed, respectively, as

$$\langle f,g\rangle_K=\sum_{i,j}\eta_i\rho_j k(u_i,v_j),$$

$$\|f\|_K^2=\sum_{i,j}\eta_i\eta_j k(u_i,u_j).$$

In the regularization theory, $\|f\|_K^2$ can be used as a measure of the complexity in RKHS of the function.

For vector functions

$$F(x)=(f_1(x),f_2(x),\ldots,f_C(x))^T,$$

$$G(x)=(g_1(x),g_2(x),\ldots,g_C(x))^T,$$

their inner product can be defined as

$$\langle F,G\rangle_K=\sum_{i=1}^{C}\langle f_i,g_i\rangle_K.$$

Then the regularization term of F, which measures the complexity of the classifier in the ambient space, can be expressed as

$$\|F\|_K^2=\sum_{i=1}^{C}\|f_i\|_K^2. \tag{10}$$

Similarly as in the MR framework (1), making use of (9) and (10) to be the penalty term and the regularization term for the discriminative function F, respectively, we have the sparse regularization method as follows:

$$F^*=\underset{f_s\in\mathcal{H}_K,s=1,\ldots,C}{\operatorname{argmin}}\left\{\frac{1}{l}\sum_{i=1}^{l}V(x_i,y_i,F)+\gamma_A\|F\|_K^2+\gamma_I\|F\|_I^2\right\}, \tag{11}$$

where γ_A and γ_I are the given regularization parameters and V is some loss function.

For convenience, here we set $V(x,y,f)=\|y-F(x)\|_2^2$. Then the S-RLSC method is given as follows:

$$F^*=\underset{f_s\in\mathcal{H}_K,s=1,\ldots,C}{\operatorname{argmin}}\left\{\frac{1}{l}\sum_{i=1}^{l}\|y_i-F(x_i)\|_2^2+\gamma_A\|F\|_K^2+\gamma_I\|F\|_I^2\right\}. \tag{12}$$

By matrix computation, the Eq. (9) can be expressed in the matrix form

$$\|F\|_I^2=\frac{1}{(l+u)^2}\operatorname{tr}(H(I-A)^T(I-A)H^T), \tag{13}$$

where

$$A=(\alpha_1,\ldots,\alpha_{l+u})^T\in\mathbb{R}^{(l+u)\times(l+u)},$$

$$H=(F(x_1),\ldots,F(x_{l+u}))\in\mathbb{R}^{C\times(l+u)},$$

and $\operatorname{tr}(M)$ denotes the trace of the matrix M, that is, the sum of the diagonal elements of the matrix M. On the other hand, the discriminative vector function can be written as

$$F^*(x)=\sum_{i=1}^{l+u}b_i k(x_i,x), \tag{14}$$

where $b_i=(b_{1i},\ldots,b_{ci})^T$. Let $B=(b_1,\ldots,b_{l+u})\in\mathbb{R}^{C\times(l+u)}$ be the matrix of coefficients b_i and let $K=(k(x_i,x_j))\in\mathbb{R}^{(l+u)\times(l+u)}$ be the kernel matrix. Then we have

$$\|F\|_K^2=\sum_{i=1}^{C}\|f_i\|_K^2=\sum_{i=1}^{C}b_i^TKb_i=\operatorname{tr}(BKB^T). \tag{15}$$

Assume that $Y=(y_1,\ldots,y_l,\mathbf{0},\ldots,\mathbf{0})\in\mathbb{R}^{C\times(l+u)}$ is the label matrix with elements 0 or 1, where $\mathbf{0}\in\mathbb{R}^C$ is the zero vector, and $J\in\mathbb{R}^{(l+u)\times(l+u)}$ is a diagonal matrix with the first l diagonal elements being 1 and the rest being 0. By substituting (14), (13) and (15) into (12), the following convex optimization problem can be obtained:

$$B^*=\underset{B\in\mathbb{R}^{(l+u)\times C}}{\operatorname{argmin}}\left\{\frac{1}{l}\operatorname{tr}((Y-BKJ)(Y-BKJ^T))+\gamma_A\operatorname{tr}(BKB^T)+\frac{\gamma_I}{(l+u)^2}\operatorname{tr}(BK(I-A)^T(I-A)KB^T)\right\}. \tag{16}$$

The solution of the optimization problem (16) is given by

$$B^*=Y\left(KJ+\gamma_A ll+\frac{\gamma_I l}{(l+u)^2}K(I-A)^T(I-A)\right)^{-1}, \tag{17}$$

where I is the identity matrix. Then the solution of the problem (12) is

$$F^*(x)=(f_1^*(x),f_2^*(x),\ldots f_C^*(x))^T=\sum_{i=1}^{l+u}b_i^*k(x_i,x) \tag{18}$$

with $B^*=(b_1^*,\ldots,b_{l+u}^*)$. Thus, the S-RLSC classifier is obtained as

$$g^*(x)=i^*=\underset{i=1,2,\ldots,C}{\operatorname{argmax}}\{f_i^*(x)\}.$$

Based on the above, the corresponding S-RLSC algorithm can be summarized as follows.

Algorithm 1. The S-RLSC algorithm

Input: Data set $\{(x_i,y_i),x_{l+j},i=1,\ldots,l,j=1,\ldots,u\}$, regularization parameters γ_A, γ_I and the kernel parameter σ.

1: Normalize the columns of X to have the unit l^2-norm.

2: Compute the best sparse representing coefficients for each point in $\{x_i\}_{i=1}^{l+u}$ by solving the l^1-norm minimization problem (8).

3: Compute the kernel matrix $K=(k(x_i,x_j))$.

4: Compute the coefficient matrix B^* of the kernel matrix K by the Eq. (17).
5: Compute the discriminative function $F^*(x)$ by (18).
Output: The final classifier $g^*(x) = i^* = \text{argmax}_{i=1,2,\ldots,C}\{f_i^*(x)\}$.

3.3. Contributions

The proposed SR approach and the derived S-RLSC algorithm have three desirable features: less parameters, highly discriminative ability and a natural out-of-sample extension for new points, as discussed as follows:

(1) *Adaptively establishing the relationship between data points.* Based on the sparse representation theory, the relationship between each data point and the rest data points can be automatically obtained by solving an l^1-norm minimization problem, while the traditional methods, which always use a fixed neighborhood size (e.g. the k-nearest-neighbor or the ε-nearest-neighbor) to construct the relationship, cannot be adaptive to uneven data.
(2) *Highly discriminative ability.* The l^1-minimization problem finds a sparse representation for each data point. The represented data point and the representing data points are naturally in the same class. This desirability works well in a high-dimensional space and is related to the number of classes but not the number of data points.
(3) *A natural out-of-sample extension for new data points.* Our algorithm not only inherits the highly discriminative ability of sparse representation of data but also has a natural out-of-sample extension for newly coming data points. This property allows much faster classification speed compared with the sparse classification and has a better performance than the MR algorithms.

4. Experiments

To evaluate the performance of the S-RLSC algorithm, we perform experiments on four real-world data sets: the MIT CBCL data set [11], the Intelligent Traffic System (ITS) data set [9], the Extended Yale B data set [13,16] and the USPS handwritten digit data set [15]. Comparison is made with four important classification methods: LapRLSC method [4], Gaussian fields and harmonic functions (GFHF) method [24], the sparse representation-based classification (SRC) [21] and the 1-nearest-neighbor (1-NN) algorithm.

4.1. Data set description

The MIT CBCL database contains 2429 face images and 4548 non-face images. Each image has 19×19 pixels and is transformed into a 361-dimensional vector. This database contains two classes of data points, face and non-face. In the experiment, we used a subset of this database which consists of 1000 face and 1000 non-face images. Fig. 1 shows some face and non-face images from this data set.

The ITS data set contains 2000 images of two classes: 1000 images of humans walking or running and 1000 images of a common scene on road without human activity. Each data point is a cropped 32×16 gray-scale image and is transformed into a 512-dimensional vector. Some samples of both human walking images and the non-human images are presented in Fig. 2.

The USPS database contains 8-bit gray-scale images of classes '0'–'9' with 1100 data points of each class. Each data point of this data set is a 16×16 image of a handwritten digit and is transformed into a 256-dimensional vector. For each class, we randomly selected 250 data points for our experiments. Therefore, our USPS data set contains 2500 data points. Some sample data points of class '1' from this data set are shown in Fig. 3.

The Extended Yale B face database contains 2114 frontal-face images of 38 individuals. Each data point is a cropped 64×64 gray-scale face image and was captured under various lighting conditions. For each class, there are about 60 images, and each image is stacked to a 1024-dimensional data vector. We use the whole database as one of our experimental data sets. Fig. 4 presents some sample face images of two individuals from the data set.

4.2. Parameter selection and experimental settings

LapRLSC and GFHF methods need the neighborhood size k as a key parameter for implementation. Experiments show that, for each data set, the results are stable with a wide range of neighborhood sizes. In our experiments, the neighborhood size k is set to be 7 for all the four data sets.

Fig. 1. Some image samples from the CBCL data set. The first two rows show some face images and the last two rows show some non-face images.

Fig. 2. Some image samples from the ITS data set. The first two rows show some human walking images and the last two rows show some non-human images.

Fig. 3. Some image samples from the USPS handwritten data set of class '1'.

Fig. 4. Some face image samples of two individuals from the Extended Yale B data set.

The LapRLSC and S-RLSC methods have regularization parameters γ_A and γ_I. Let $CA=\gamma_A l$, $CI=\gamma_I l/(l+u)^2$ and let the kernel function $k(x,y)=\exp(-\|x-y\|^2/\sigma^2)$. In the experiments it was found that both algorithms perform well with a wide range of regularization parameters. In our experiments, we set $CA=0.005$, $CI=0.01$ and $\sigma=0.5$ for all four data sets. The SRC method needs only one parameter, error tolerance ε. In [21], this value was set to be 0.05 throughout all their experiments. Therefore, we also used this value for the error tolerance ε in our experiments.

For each data set $\mathcal{X}$, we first rearrange the order of data points randomly. Then, in each class of $\mathcal{X}$, 15% of the data points are left for out-of-sample extension experiment. Denote by $\mathcal{X}_1$ the rest data points of the data set $\mathcal{X}$. In each class of $\mathcal{X}_1$, we randomly label m data points to train the algorithms. For the SRC and 1-NN classifiers, the training set comprises only the labeled data points from $\mathcal{X}_1$. For the GFHF, LapRLSC and S-RLSC classifiers, the training set consists of the whole $\mathcal{X}_1$, including the labeled and the unlabeled data points. After obtaining the classifiers, classification was performed first on the unlabeled data points in $\mathcal{X}_1$ and then on the out-of-sample extension data points. It should be noted that in all the experiment results the classification accuracy is lower for a small number of labeled training data samples (e.g. $m=5,10$) for all algorithms, which should be natural due to the small number of training samples.

4.3. Experimental results

For the CBCL and ITS data sets, which have data points of two-class, the number m of labeled data points (in each class) changes from 5 to 650. The recognition accuracy of the algorithms on the unlabeled data points of $\mathcal{X}_1$ are shown in Figs. 5 and 6, respectively. As can be seen from Figs. 5 and 6, the S-RLSC algorithm clearly outperforms the other four methods. In particular, for the CBCL data set, in the case when $m \geq 500$, the recognition rate of the S-RLSC algorithm is over 99%. Note that the LapRLSC algorithm has also better recognition results than the GFHF, SRC and 1-NN algorithms. This may be because the LapRLSC algorithm takes account of the smoothness of the classifier function on the manifold. It should also be noted that the recognition result of the SRC algorithm is not very satisfactory when m is relatively small. The reason may be that SRC only uses the labeled points for training, but it is based on sparse representation which may become invalid for insufficient training samples.

The recognition result of the algorithms is shown in Figs. 7 and 8, respectively, on the unlabeled data points of $\mathcal{X}_1$ for the USPS and

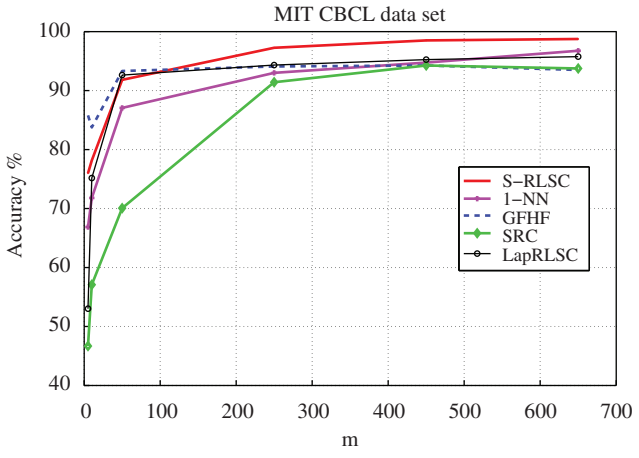

Fig. 5. Classification results of the unlabeled data points in $\mathcal{X}_1$ for the CBCL data set, where m is the number of labeled data points in each class.

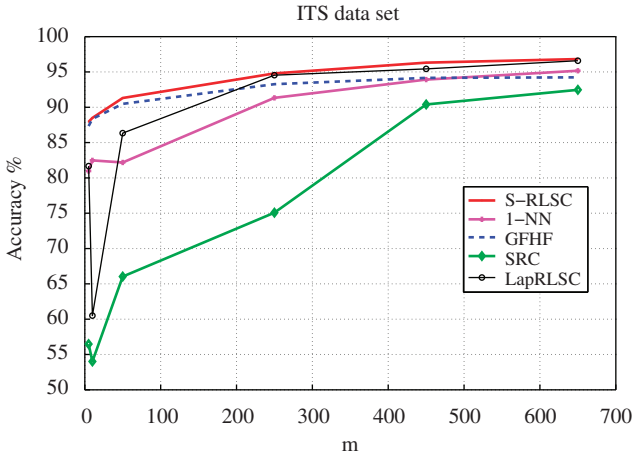

Fig. 6. Classification results of the unlabeled data points in $\mathcal{X}_1$ for the ITS data set, where m is the number of labeled data points in each class.

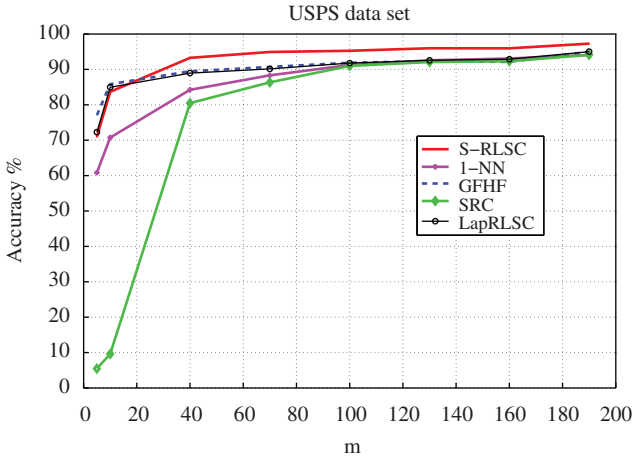

Fig. 7. Classification results of the unlabeled data points in $\mathcal{X}_1$ for the USPS data set, where m is the number of labeled data points in each class.

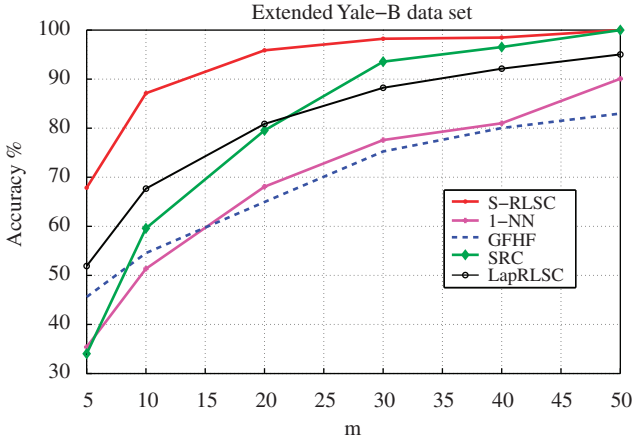

Fig. 8. Classification results of the unlabeled data points in $\mathcal{X}_1$ for the Extended Yale B data set, where m is the number of labeled data points in each class.

Extended Yale B data sets. Since the data points of these two data sets belong to more than two classes, the multi-class LapRLSC method is implemented on them, where the label of x_i is changed to the vector form. The number m (in each class) of the labeled data points changes from 5 to 190 for the USPS data set and from 5 to 50

Table 2
Out-of-sample extension results on the CBCL data set.

Method	S-RLSC (%)	LapRLSC (%)	SRC (%)	1-NN (%)	GFHF (%)
$m=5$	74.00	52.33	44.33	65.67	**86.33**
$m=10$	75.67	74.66	58.33	70.67	**81.00**
$m=50$	**95.00**	94.67	67.33	88.67	92.00
$m=250$	**99.00**	98.00	93.33	94.67	94.67
$m=450$	**99.33**	98.67	95.33	96.00	94.33
$m=650$	**98.67**	**98.67**	96.00	96.67	94.67

Table 3
Out-of-sample extension results on the ITS data set.

Method	S-RLSC (%)	LapRLSC (%)	SRC (%)	1-NN (%)	GFHF (%)
$m=5$	**87.94**	74.21	57.89	80.62	74.47
$m=10$	**88.49**	64.47	53.16	80.52	77.63
$m=50$	**91.31**	82.89	65.26	78.42	78.94
$m=250$	**94.78**	91.56	72.37	88.68	87.11
$m=450$	**96.31**	93.68	87.11	90.26	88.68
$m=650$	**96.82**	94.74	90.26	91.84	91.05

Table 4
Out-of-sample extension results on the USPS data set.

Method	S-RLSC (%)	LapRLSC (%)	SRC (%)	1-NN (%)	GFHF (%)
$m=5$	**72.63**	**72.63**	5.26	61.84	68.94
$m=10$	**86.58**	85.00	9.47	73.16	79.74
$m=40$	**95.00**	93.68	82.11	87.11	88.41
$m=100$	**98.16**	97.36	92.11	91.01	93.74
$m=160$	98.68	**98.94**	93.68	94.74	94.74

Table 5
Out-of-sample extension results on the Yale B data set.

Method	S-RLSC (%)	LapRLSC (%)	SRC (%)	1-NN (%)	GFHF (%)
$m=5$	**65.41**	48.25	36.46	33.78	21.18
$m=10$	**84.18**	71.05	58.17	54.69	35.65
$m=20$	**94.10**	85.52	76.94	68.90	55.22
$m=30$	**96.24**	91.42	90.61	77.21	71.58
$m=40$	**98.12**	95.44	96.51	79.89	80.70
$m=50$	98.12	97.32	**98.39**	83.38	74.72

for the Extended Yale B data set. It can be seen from Figs. 7 and 8 that the S-RLSC algorithm has the best recognition result among all the algorithms except for the USPS data set in the case when $m=5$ and 10 and the Extended Yale B data set in the case when $m=50$. In addition, for relatively large m the result of the SRC algorithm is satisfactory on the two data sets.

The out-of-sample extension result of the algorithms on the four data sets is shown in Tables 2–5 for several values of m, where the best classification results are in boldface for each fixed value of m. As can be seen from the tables, the discriminative ability of the S-RLSC algorithm is much better than the other algorithms.

4.4. Data sets with noise

In order to evaluate the performance of the S-RLSC algorithm on noise data, the five algorithms are implemented on the Extended Yale-B data set which is corrupted by noise of three different levels. At each level, each data point is corrupted by adding i.i.d. noise from a uniform distribution with zero mean to the pixels of the data point. The variance of the added noise varies from 0, $0.01\times\sigma$ to $0.1\times\sigma$, where $\sigma=2630$ is the mean of the pairwise Euclidean distances of the data set (before normalization). Three sample images of a data point are presented in Fig. 9, which are corrupted by noise of three different levels as described above.

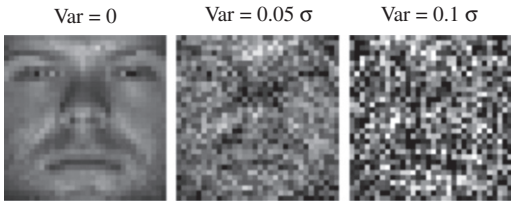

Fig. 9. Some samples of the noised Extended Yale-B images.

Table 6
Classification results on the noised Yale B data set.

	S-RLSC (%)	LapRLSC (%)	SRC (%)	1-NN (%)	GFHF (%)
$Var=0$					
$m=5$	**72.21**	55.75	44.08	39.29	40.69
$m=10$	**84.86**	72.79	62.53	55.00	52.60
$m=25$	**95.45**	93.21	84.04	73.70	72.70
$Var=0.05\sigma$					
$m=5$	**55.96**	24.34	3.55	23.05	26.90
$m=10$	**56.97**	40.12	2.18	34.55	39.88
$m=25$	**84.31**	52.94	7.45	46.67	50.59
$Var=0.1\sigma$					
$m=5$	**12.49**	6.45	4.88	9.76	11.91
$m=10$	**38.71**	21.92	2.48	12.82	15.05
$m=25$	**58.89**	58.73	2.23	16.63	20.01

Table 7
Out-of-sample extension results on the noised Yale B data set.

	S-RLSC (%)	LapRLSC (%)	SRC (%)	1-NN (%)	GFHF (%)
$Var=0$					
$m=5$	**71.43**	48.07	40.09	39.01	37.54
$m=10$	**85.09**	62.54	57.81	52.24	44.73
$m=25$	**95.29**	83.13	81.57	74.90	59.60
$Var=0.05\sigma$					
$m=5$	**52.44**	35.48	3.23	22.25	27.63
$m=10$	56.99	**58.64**	2.98	32.75	39.62
$m=25$	**85.36**	84.28	8.35	46.48	56.41
$Var=0.1\sigma$					
$m=5$	8.28	7.88	4.53	8.47	**11.52**
$m=10$	**29.81**	12.36	1.09	12.48	15.63
$m=25$	**48.23**	22.75	3.13	15.29	20.02

In these experiments, we use 50 percents of the whole data set as the training set and the rest 50 percents as the test set. For each test, we label $m=5$, 10 and 25 data points in the training set to train the algorithms. The setting of the parameters is the same as in the previous experiments.

As seen from Tables 6 and 7, the performance of the five algorithms deteriorates rapidly with the level of noise. The S-RLSC algorithms has the best performance than the other four algorithms. It should be noted that the SRC algorithm has the worst performance on the noised Yale-B data set. There may be two reasons for this phenomenon. First, the noise corrupts all the pixels of each image which is not sparse; this is different from [21], where the noise is sparse and corrupts only a portion of certain randomly chosen pixels for each image. Secondly, the SRC algorithm is a supervised classification method, which does not use the structure information of unlabeled data points to improve its classification performance.

In summary, the above experimental results show the advantage of the proposed S-RLSC algorithm over the other four algorithms.

5. Conclusion

In this paper, we have proposed a novel semi-supervised classification algorithm called the S-RLSC algorithm. The S-RLSC algorithm assumes that the discriminative function which contains the label information the sparse representing coefficients of data points. Comparison has been made of the S-RLSC algorithm with four important classification algorithms: the LapRLSC, GFHF, SRC and 1-NN algorithms through experiments on four real-world data sets, and the results have shown that the S-RLSC algorithm has a better recognition result and stable performance with respect to the parameters compared with the other four algorithms.

Acknowledgments

This work was partly supported by the NNSF of China Grant no. 90820007, the Outstanding Youth Fund of the NNSF of China Grant no. 60725310, the 863 Program of China Grant no. 2007AA04Z228 and the 973 Program of China Grant no. 2007CB311002. The authors thank the referees for their invaluable comments and suggestions which helped improve the paper greatly.

References

[1] E. Amaldi, V. Kann, On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems, Theor. Comput. Sci. 209 (1998) 237–260.
[2] A. Barron, J. Rissanen, B. Yu, The minimum description length principle in coding and modeling, IEEE Trans. Inf. Theory 44 (1998) 2743–2760.
[3] M. Belkin, P. Niyogi, Semi-supervised learning on Riemannian manifolds, Mach. Learn. 56 (1–3) (2004) 209–239.
[4] M. Belkin, V. Sindhwani, P. Niyogi, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (2006) 2399–2434.
[5] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, Cambridge, UK, 2004.
[6] E. Candes, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Comm. Pure Appl. Math. 59 (8) (2006) 1207–1223.
[7] E. Candes, T. Tao, Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory 52 (12) (2006) 5406–5425.
[8] E. Candes, J. Romberg, l^1-magic: recovery of sparse signals via convex programming, 2005, 〈http://www.acm.caltech.edu/l1magic/downloads/l1magic.pdf〉.
[9] X. Cao, H. Qiao, J. Keane, A low-cost pedestrian-detection system with a single optical camera, IEEE Trans. Intelligent Transport. Syst. 9 (1) (2008) 58–67.
[10] Z. Chen, S. Haykin, On different facets of regularization theory, Neural Comput. 14 (12) (2002) 2791–2846.
[11] CBCL Face Database 1, MIT Center For Biological and Computation Learning, 〈 http://www.ai.mit.edu/projects/cbcl 〉.
[12] D. Donoho, For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution, Comm. Pure Appl. Math. 59 (6) (2006) 797–829.
[13] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 643–660.
[14] F. Girosi, M. Jones, T. Poggio, Regularization theory and neural networks architectures, Neural Comput. 7 (2) (1995) 219–269.
[15] J.J. Hull, A database for handwritten text recognition research, IEEE Trans. Pattern Anal. Mach. Intell. 16 (5) (1998) 550–554.
[16] K.C. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 684–698.
[17] T. Poggio, S. Smale, The mathematics of learning: dealing with data, Not. Am. Math. Soc. 50 (5) (2003) 537–544.
[18] A.N. Tikhonov, Regularization of incorrectly posed problems, Sov. Math. Dokl. 4 (1963) 1624–1627.
[19] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[20] F. Wang, C. Zhang, Robust self-tuning semi-supervised learning, Neurocomputing 70 (16–18) (2007) 2931–2939.
[21] J. Wright, A. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 210–227.
[22] J. Wright, Y. Ma, J. Mairal, G. Spairo, T. Huang, S. Yan, Sparse representation for computer vision and pattern recognition, Proceedings of the IEEE 98 (6) (2010) 1031–1044.
[23] H. Xue, S. Chen, Q. Yang, Discriminatively regularized least-squares classification, Pattern Recognition 42 (1) (2009) 93–104.
[24] X. Zhu, Z. Ghahramani, J. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, in: The 20th International Conference on Machine Learning (ICML), 2003, pp. 912–919.

Mingyu Fan received the B.Sc. degree from the Central University for Nationalities, Beijing, China, in 2006. He is currently working toward the Ph.D. degree in applied mathematics in the Institution of Applied Mathematics, Academy of Mathematics and System Science, Chinese Academy of Science, Beijing, China. His current research interests are theory and application of manifold learning and nonlinear dimensionality reduction.

Nan-Nan Gu received the B.Sc. degree in information and computing science from Xi'an Jiaotong University, Xi'an, China in 2006, and the M.Sc. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China in 2009. Currently, she is working toward the Ph.D. degree in pattern recognition in the Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her research interests include theory and application of manifold learning, nonlinear dimensionality reduction and classification.

Hong Qiao (SM'06) received the B.E. degree in hydraulics and control and the M.E. degree in robotics from Xian Jiaotong University, Xian, China, the M.Phil. degree in robotics control from the Industrial Control Center, University of Strathclyde, Glasgow, U.K., and the Ph.D. degree in robotics and artificial intelligence from De Montfort University, Leicester, U.K., in 1995.

She was a University Research Fellow with De Montfort University from 1995 to 1997. She was a Research Assistant Professor from 1997 to 2000 and an Assistant Professor from 2000 to 2002 with the Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong. Since January 2002, she has been a Lecturer with the School of Informatics, University of Manchester, Manchester, U.K. Currently, she is also a Professor with the Laboratory of Complex Systems and Intelligent Science, Institute of Automation, Chinese Academy of Sciences, Beijing, China. She first proposed the concept of "the attractive region in strategy investigation," which has successfully been applied by herself in robot assembly, robot grasping, and part recognition. The work has been reported in Advanced Manufacturing Alert (Wiley, 1999). Her current research interests include information-based strategy investigation, robotics and intelligent agents, animation, machine learning (neural networks and support vector machines), and pattern recognition.

Dr. Qiao is a member of the Program Committee of the IEEE International Conference on Robotics and Automation from 2001 to 2004. She is currently the Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART B and the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING.

Bo Zhang received the B.Sc. degree in mathematics from Shandong University, Jinan, China, the M.Sc. degree in mathematics from Xi'an Jiaotong University, Xi'an, China, and the Ph.D. degree in applied mathematics from the University of Strathclyde, Glasgow, U.K., in 1983, 1985, and 1992, respectively.

From 1985 to 1988, he was a Lecturer with the Department of Mathematics, Xi'an Jiaotong University, China. He was a Postdoctoral Research Fellow with the Department of Mathematics, Keele University, Keele, U.K. from January 1992 to December 1994, and with the Department of Mathematical Sciences, Brunel University, Uxbridge, U.K. from January 1995 to October 1997. In November 1997, he joined the School of Mathematical and Informational Sciences, Coventry University, Coventry, U.K., as a Senior Lecturer, where he was promoted to Reader in Applied Mathematics in September 2000 and to Professor of Applied Mathematics in September 2003. Currently, he is a Professor with the Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China. His current research interests include direct and inverse scattering problems, computational electromagnetics, partial differential equations, and machine learning.