Neurocomputing 147 (2015) 299–306

Contents lists available at ScienceDirect

Neurocomputing

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

Semi-supervised classification learning by discrimination-aware manifold regularization

Yunyun Wang [a,b], Songcan Chen [b,*], Hui Xue [c], Zhenyong Fu [a]

[a] Department of Computer Science and Engineering, Nanjing University of Posts & Telecommunications, Nanjing 210046, PR China
[b] Department of Computer Science and Engineering, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, PR China
[c] School of Computer Science and Engineering, Southeast University, Nanjing 210096, PR China

ARTICLE INFO

Article history:
Received 13 December 2013
Received in revised form
20 April 2014
Accepted 23 June 2014
Communicated by Feiping Nie
Available online 30 June 2014

Keywords:
Semi-supervised classification
Manifold regularization
Discrimination
Unsupervised clustering

ABSTRACT

Manifold regularization (MR) provides a powerful framework for semi-supervised classification (SSC) using both the labeled and unlabeled data. It first constructs a single Laplacian graph over the whole dataset for representing the manifold structure, and then enforces the smoothness constraint over such graph by a Laplacian regularizer in learning. However, the smoothness over such a single Laplacian graph may take the risk of ignoring the discrimination among boundary instances, which are very likely from different classes though highly close to each other on the manifold. To compensate for such deficiency, researches have already been devoted by taking into account the discrimination together with the smoothness in learning. However, those works are only confined to the discrimination of the labeled instances, thus rather limited in boosting the semi-supervised learning. To mitigate such an unfavorable situation, we attempt to discover the possible discrimination in the available instances first by performing some unsupervised clustering over the whole dataset, and then incorporate it into MR to develop a novel discrimination-aware manifold regularization (DAMR) framework. In DAMR, instances with high similarity on the manifold will be restricted to share the same class label if belonging to the same cluster, or to have different class labels, otherwise. Our empirical results show the competitiveness of DAMR compared to MR and its variants likewise incorporating the discrimination in learning.

1. Introduction

In many real applications, the unlabeled data can be easily and cheaply collected, while the acquisition of labeled data is usually quite expensive and time-consuming, especially involving manual effort. For instance, in web page recommendation, huge amounts of web pages are available, but few users are willing to spend time marking which web pages they are interested in. In spam email detection, a large number of emails can be automatically collected, yet few of them have been labeled spam or not by users. Consequently, semi-supervised learning, which exploits a large amount of unlabeled data jointly with the limited labeled data for learning, has attracted intensive attention during the past decades. In this paper, we focus on semi-supervised classification, and so far, lots of semi-supervised classification methods have been developed [1–4].

Generally, semi-supervised classification methods attempt to exploit the intrinsic data distribution information disclosed by the unlabeled data in learning, and the information is usually considered to be helpful for learning. To exploit the unlabeled data, some assumption should be adopted for learning. Two common assumptions in semi-supervised classification are the cluster assumption and the manifold assumption [3–5]. The former assumes that similar instances are likely to share the same class label, thus guides the classification boundary passing through the low density region between clusters. The latter assumes that the data are resided on some low dimensional manifold represented by a Laplacian graph, and similar instances should share similar classification outputs according to the graph. Almost all off-the-shelf semi-supervised classification methods adopt one or both of those assumptions explicitly or implicitly [1,4]. For instance, the large margin semi-supervised classification methods, such as transductive Support Vector Machine (TSVM) [6], semi-supervised SVM (S3VM) [7] and their variants [8,9], adopt the cluster assumption. The graph-based semi-supervised classification methods, such as label propagation [10,11], graph cuts [12] and manifold regularization (MR) [13], adopt the manifold assumption. Furthermore, there are also methods combining both assumptions for better performances, such as RegBoost [14] and SemiBoost [15], etc.

In this paper, we concentrate on the MR framework [13], which provides an effective way for semi-supervised classification [16], and has been applied in diverse applications such as image

* Corresponding author. Tel.: +86 25 84892956; fax: +86 25 84892811.
E-mail address: s.chen@nuaa.edu.cn (S. Chen).

http://dx.doi.org/10.1016/j.neucom.2014.06.059
0925-2312/

retrieval [17] and web spam identification [18], etc. At the same time, the manifold learning concept has also successfully applied in many other learning tasks including clustering [19], dimensionality reduction [20], and non-negative matrix factorization [21,22], etc.

MR for semi-supervised classification represents the manifold structure for the whole dataset by a single Laplacian graph, which is different from MR for supervised classification constructing the respective Laplacian graphs for individual classes, and then imposes the smoothness constraint over such a representation by a Laplacian regularizer in learning. However, the smoothness constraint imposed over a single Laplacian graph may take the risk of ignoring the discrimination among the boundary instances, which are very likely to belong to different classes though close over the manifold, consequently, MR may misclassify the boundary instances between clusters [16].

In fact, many researches have already been devoted to compensating for this deficiency by utilizing the dissimilarity or discrimination in the learning of MR. In [23], Andrew et al. considered both the label similarity and dissimilarity in learning, and developed a new dissimilarity encoded MR framework based on mixed graph. However, the dissimilarity should be given beforehand. In [24], Wang and Zhang constructed an unsupervised discriminative kernel based on discriminant analysis, and then used it to derive specific algorithms, including semi-supervised discriminative regularization (SSDR) and semi-parametric discriminative semi-supervised classification (SDSC). However, the derived methods do not fall into the methodology of manifold regularization. Recently in [16], Wu et al. incorporated linear discriminant analysis (LDA) and MR into a coherent framework and developed a semi-supervised discriminative regularization (SSDR). Specifically, the intra-class and inter-class graphs are constructed first in SSDR based on the labeled data, and then the corresponding intra-class compactness and inter-class separation are optimized simultaneously in the learning of MR. However, SSDR in [16] only utilize the discrimination of the labeled data, while the label information is usually rather limited in semi-supervised learning, consequently, its improvement over MR is not so distinct in the experiments.

In this paper, we attempt to incorporate the discrimination of both the labeled and unlabeled data into MR so as to develop a discrimination-aware MR framework for semi-supervised classification. In fact, due to the lack of label information in semi-supervised learning and thus the difficulty for formulating the discrimination of the whole data, SSDR in [16] only uses the discrimination of the labeled data, while the label instances are usually scarce in semi-supervised classification. For discovering the discrimination of all given data, we adopt the strategy of a pre-performed unsupervised clustering method as an example. Specifically, by performing some unsupervised clustering method such as FCM beforehand, we can get the within/between-cluster information of all instance pairs, which is much analog to the must/cannot-link information in semi-supervised clustering. Then we incorporate such information into MR such that for instances with high similarity over the manifold structure, they are restricted to share the same class label if belonging to the same cluster, or to have different class labels, otherwise. In this way, DAMR actually utilizes both the cluster and manifold assumptions in learning. It has been demonstrated by previous work that methods working on multiple data distribution assumptions can achieve better classification than those working on a single one [14,15], thus DAMR is able to be expected to perform better than MR.

The rest of this paper is organized as follows. Section 2 introduces the related works, Section 3 presents the proposed discrimination-aware manifold regularization framework, Section 4 presents a specific algorithm DA_LapRLSC through adopting the square loss function, Section 5 gives the empirical results, and some conclusions are drawn in Section 6.

2. Related works

2.1. Manifold regularization

Manifold assumption is one of the most commonly-used data distribution assumptions in semi-supervised learning [2,4]. Generally, the manifold structure is captured by an undirected graph according to some similarity measure, in which the vertices represent the instances and the edge-weights represent the similarities between instance pairs, and the manifold assumption assumes that similar instances over the manifold structure should share similar classification outputs. Lots of semi-supervised classification methods have been proposed based on the manifold assumption, mainly including the graph-based methods such as label propagation, graph cuts and manifold regularization, etc. Most graph-based methods, including label propagation and graph cuts, aim to learn only the class labels for the available unlabeled instances, thus learn in the transductive learning style [4]. However, many real applications actually need inductive methods for predicting unseen instances [15], and manifold regularization (MR) is exactly an inductive learning framework for semi-supervised classification based on the manifold assumption, which has been applied in diverse applications during the recent years.

Given labeled data $X_l=\{x_i\}_{i=1}^{l}$ with the corresponding labels $Y=\{y_i\}_{i=1}^{l}$, and unlabeled data $X_u=\{x_j\}_{j=l+1}^{n}$, where each $x_i\in R^d$ and $u=n-l$. $G=\{W_{ij}\}_{i,j=1}^{n}$ is a Laplacian graph over the whole dataset, where each weight W_{ij} represents the similarity between instances x_i and x_j. The Laplacian graph can be defined by many strategies such as the 0–1 weighting, i.e., $W_{ij}=1$ if and only if x_i and x_j are connected by an edge over the graph, the heat kernel weighting with $W_{ij}=e^{-||x_i-x_j||^2/\sigma}$ if x_i and x_j are connected, or the dot-product weighting with $W_{ij}=x_i^Tx_j$ if x_i and x_j are connected.

Then with a decision function $f(x)$, the framework of MR can be formulated as

$$\min_f \frac{1}{l}\sum_{i=1}^{l} V(x_i,y_i,f)+\gamma_A\|f\|_K^2+\frac{\gamma_I}{2(l+u)^2}\sum_{i,j=1}^{l+u} W_{ij}(f(x_i)-f(x_j))^2 \tag{1}$$

where $V(x_i,y_i,f)$ is some loss function, such as the hinge loss $\max[0,1-y_if(x_i)]$ for support vector machine (SVM) or the square loss $(y_i-f(x_i))^2$ for regularized least square classifier (RLSC), in this way, the MR framework naturally embodies the specific algorithms LapSVM and LapRLSC [13]. $\|f\|_K^2$ is a regularization term for smoothness in the Reproducing Kernel Hilbert Space (RKHS). The third term guarantees the prediction smoothness over the graph, which can be further written as

$$\frac{1}{2}\sum_{i,j=1}^{l+u} W_{ij}(f(x_i)-f(x_j))^2=\mathbf{f}^T\mathbf{L}\mathbf{f} \tag{2}$$

where $\mathbf{f}=[f(x_1),\ldots,f(x_{l+u})]^{\mathrm{T}}$, and $\mathbf{L}$ is the graph Laplacian given by $\mathbf{L}=\mathbf{D}-\mathbf{W}$, $\mathbf{W}$ is the weight matrix of graph G and $\mathbf{D}$ is a diagonal matrix with the diagonal component given by $\mathbf{D}_{ii}=\sum_{j=1}^{n}\mathbf{W}_{ij}$. According to the Representer theorem [13], the minimizer of problem (1) has the form

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

where K: $X\times X\to R$ is a Mercer kernel (the bias of the decision function can be omitted by augmenting each instance with an 1-valued element).

It is clear that in MR, if instances x_i and x_j are similar in terms of W_{ij}, then it is restricted that their class labels are similar as well. *Such a smoothness restriction is also imposed on the boundary instance pairs, however, instance pairs in the boundary area are very*

likely to belong to different classes. Thus in this paper, we attempt to utilize the discrimination among instances together with the smoothness over the manifold.

2.2. Semi-supervised discriminative regularization (SSDR) [16]

The performance of LDA usually deteriorates when the label information is insufficient, and at the same time, MR tends to misclassify instances near the boundary between clusters during classification [16]. With those in mind, Wu et al. [16] incorporated LDA and MR into a coherent framework for semi-supervised classification, and proposed a novel semi-supervised discriminative regularization (SSDR) method. SSDR exploits both the discrimination (label information of the labeled data) and the data distribution for learning. Specifically, besides the graph G over the $(l+u)$ instances to model the intrinsic geometrical structure of data manifold, a intra-class graph G_w (with weight matrix $\mathbf{W}_w$) and a inter-class graph G_b (with weight matrix $\mathbf{W}_b$) are also constructed respectively by

$$W_{w,ij}=\begin{cases}1/l & \text{if both } x_i \text{ and} x_j \text{ are labeled and } y_i=y_j\\ 0 & \text{otherwise}\end{cases} \tag{4}$$

$$W_{b,ij}=\begin{cases}1/l & \text{if both } x_i \text{ and } x_j \text{ are labeled and} y_i\neq y_j\\ 0 & \text{otherwise}\end{cases} \tag{5}$$

Then the optimization problem of SSDR is formulated as

$$\begin{aligned}\min_f \frac{1}{l}\sum_{i=1}^{l}V(x_i,y_i,f)+\gamma_A\|f\|_K^2+\frac{\gamma_I}{(l+u)^2}\mathbf{f}^T\mathbf{L}\mathbf{f}\\ +\frac{\gamma_D}{2}\sum_{i,j=1}^{l+u}W_{w,ij}(f(x_i)-f(x_j))^2+\frac{1-\gamma_D}{2}\sum_{i,j=1}^{l+u}W_{b,ij}(f(x_i)-f(x_j))^2\end{aligned} \tag{6}$$

The last two terms in (6) can be further written as

$$\begin{aligned}&\frac{\gamma_D}{2}\sum_{i,j=1}^{l+u}W_{w,ij}(f(x_i)-f(x_j))^2+\frac{1-\gamma_D}{2}\sum_{i,j=1}^{l+u}W_{b,ij}(f(x_i)-f(x_j))^2\\ &=\mathbf{f}^T(\gamma_D\mathbf{L}_w+(1-\gamma_D)\mathbf{L}_d)\mathbf{f}\end{aligned} \tag{7}$$

where $\mathbf{L}_w$ and $\mathbf{L}_d$ are the graph Laplacian over the intra-class graph G_w and the inter-class graph G_b, respectively.

SSDR actually aims to exploit the discrimination in the learning of MR, and is able to partly reduce the risk of misclassifying the boundary instances. However, from the definition of G_w and G_b in (4) and (5) respectively, *SSDR exploits only the discrimination from the labeled data, while in semi-supervised learning, the labeled instances are usually quite limited*. Thus in this paper, we attempt to utilize the discrimination from all available data, including both the given labeled and unlabeled instances.

3. Discrimination-aware manifold regularization (DAMR) for semi-supervised learning

3.1. Motivation

MR imposes the smoothness constraints over instances from both classes, thus may ignore the discrimination among instances in the boundary areas, and consequently misclassify those boundary instances, since they are very likely to belong to different classes though highly similar over the manifold. Fig. 1 gives an illustration over a linear manifold dataset consisting of two classes, each class containing 200 instances with only 10 labeled. From Fig. 1, concentrating on the smoothness only, MR (more specifically LapRLSC) derives a classification boundary (denoted by the dotted line) deviating from the real one (denoted by the dot–dash line), and thus misclassifies several boundary instances. To compensate for it, SSDR [16] utilizes the discrimination of the labeled instances together with the smoothness in MR. However, the labeled instances are usually scarce in semi-supervised classification, thus for each labeled instance, labeled instances in the opposite classes may not fall into its neighbor in the construction of the Laplacian graph, consequently, the discrimination from those labeled instances may have little effect on classification. As can be seen in Fig. 1, SSDR [16] derives the same classification boundary as MR. While considering the discrimination over the whole dataset together with the smoothness in MR, our proposed DAMR (more specifically DA_LapRLSC, detailed later) derives a classification boundary (denoted by the solid line) closer to the real one. Correspondingly, the classification accuracies for the unlabeled instances by MR, SSDR [16] and DAMR are 0.8974, 0.8974 and 0.9132, respectively. As a result, it is reasonable to consider the discrimination of all available instances in the learning of MR, which is exactly the motivation of our DAMR.

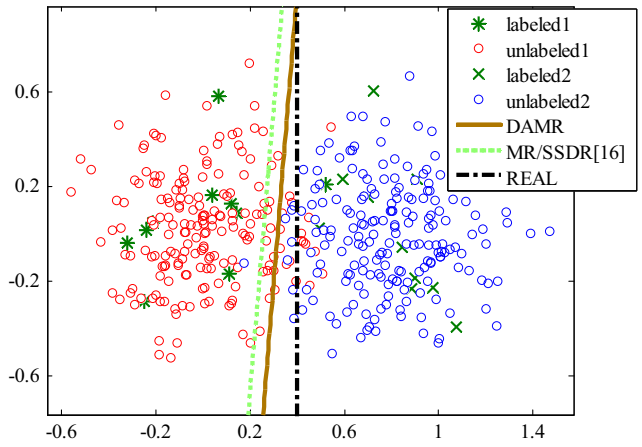

Fig. 1. The toy dataset and the corresponding classification boundaries of MR (LapRLSC), SSDR [16] and DAMR (DA_LapRLSC), together with the real one, labeled1 and labeled2 denote the labeled instances in individual classes, unlabeled1 and unlabeled2 denote the unlabeled instances in individual classes, MR and SSDR in [16] derive the same decision boundary.

3.2. DAMR framework

In order to exploit the discrimination of all available data, we first perform some unsupervised clustering method, such as FCM, to discover the intrinsic cluster structure of the data distribution. Suppose we get a cluster indicator vector for instances from the unsupervised clustering method denoted by $\mathbf{Y}_c=[y_1^c,\ldots,y_{l+u}^c]\in R^{(l+u)}$, where each $y_j^c\in\{-1,1\}$ represents the cluster label for the jth instance, then we define a discriminative matrix by $\mathbf{S}_c=\mathbf{Y}_c\mathbf{Y}_c^T\in\mathbf{R}^{(l+u)\times(l+u)}$, where each $S_{ij}^c=y_i^cy_j^c\in\{-1,1\}$. $S_{ij}^c=1$ indicates that instances x_i and x_j belong to the same cluster, and $S_{ij}^c=-1$ indicates that x_i and x_j belong to different clusters. For the labeled instances, we keep the discrimination from the given labels, and define a discriminative matrix $\mathbf{S}^l\in\mathbf{R}^{(l+u)\times(l+u)}$, in which each

$$S_{ij}^l=\begin{cases}1, \text{ if } x_i \text{ and } x_j \text{ are both labeled, and belong to the same class}\\ -1, \text{ if } x_i \text{ and } x_j \text{ are both labeled, and belong to different classes}\\ 0, \text{ otherwise (either or both of } x_i \text{ and } x_j \text{ are unlabeled)}\end{cases} \tag{8}$$

Finally, a discriminative matrix $\mathbf{S_0}$ combing the given label information and discrimination from the unsupervised clustering

method can be defined by

$$S_{ij}^0 = \begin{cases} S_{ij}^l, & \text{if} x_i \text{ and } x_j \text{ are both labeled} \\ S_{ij}^c, & \text{otherwise} \end{cases} \tag{9}$$

where each $S_{ij}^0 \in \{-1, 1\}$.

Further, we introduce a weight vector for instances denoted by $\mathbf{w}=[w_1,\ldots w_{l+u}]^T$, where $w_i=1$ for the labeled instance and $w_j=\gamma(0 \le \gamma \le 1)$ for the unlabeled instance, and set $\mathbf{T}=\mathbf{w}\mathbf{w}^T$, then we renew the discriminative matrix by

$$S_{ij} = \begin{cases} 1, \text{ if } S_{ij}^0 = 1, \text{ or } T_{ij} \le 0.5 \\ -1, \text{ otherwise (if } S_{ij}^0 = -1, \text{ and } T_{ij} > 0.5) \end{cases} \tag{10}$$

From (10), there are two cases in which the S_{ij} is set to 1, i.e., (1) if instances x_i and x_j belong to the same class by the class label, or the same cluster by the clustering results. (2) The confidence relating to the unlabeled instance is low, or the clustering result is unreliable such that $T_{ij} \le 0.5$. In those two specific cases, we set S_{ij} to 1 such that the discriminative (smoothness) constraints in DAMR mainly degenerates to the smoothness constraints in MR. Otherwise, if x_i and x_j belong to different classes or clusters, and at the same time, the clustering result is reliable such that $T_{ij} > 0.5$, most of the discrimination constraints are kept in DAMR.

Further, we renew the weight matrix $\mathbf{W}$ over G as

$$W_{ij} = \begin{cases} W_{ij}^0, & \text{if } S_{ij}^0 = 1 \\ T_{ij} W_{ij}^0, & \text{otherwise} \end{cases} \tag{11}$$

where W_{ij}^0 represents the similarity between instances x_i and x_j over the manifold by some distance metric. Note that the weight matrix $\mathbf{W}$ here differs from that in MR in that the component S_{ij} are weighted by T_{ij} when x_i and x_j belong to different classes or clusters.

In order to incorporate such discrimination, we assume that "with high similarity over the manifold, instances in the same cluster should share the same class label, and have different class labels, otherwise", then we formulate the discrimination-aware manifold regularization framework as follows,

$$\min_f J = \frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \frac{\gamma_D}{2(l+u)^2}\sum_{i,j=1}^{l+u} W_{ij}(f(x_i) - S_{ij} f(x_j))^2 \tag{12}$$

where $V(x_i, y_i, f)$ is some loss function. The last term can be further written as

$$\frac{1}{2}\sum_{i,j=1}^{l+u} W_{ij}(f(x_i) - S_{ij} f(x_j))^2 = \mathbf{f}^T(\mathbf{D} - \mathbf{W} \circ \mathbf{S})\mathbf{f} \triangleq \mathbf{f}^T \mathbf{L}_D \mathbf{f} \tag{13}$$

where $\mathbf{L}_D = \mathbf{D} - \mathbf{W} \circ \mathbf{S}$ is actually a new Laplacian matrix incorporated with the discrimination over all available instances. When $S_{ij}=1$, i.e., x_i and x_j belong to the same class or cluster, then x_i and x_j are restricted to share the same class label when they are similar over the manifold structure, otherwise, when $S_{ij}=-1$, i.e., x_i and x_j belong to different classes or clusters, then x_i and x_j are restricted to belong to different classes though they are similar over the manifold structure. Note that we do not simply add the discriminative information to w_{ij}, i.e., let $w_{ij} < 0$ when x_i and x_j belong to different classes or clusters, since it would lead to several problems such as non-convex [23]. On the other hand, the parameter γ actually adjusts the reliability or importance of the discrimination from the unsupervised clustering method. If such information is reliable or important, one can attach importance to it by setting a large γ value, otherwise, one can accordingly set a small γ value. When γ approaches 0, DAMR actually degenerates to MR concerning on the smoothness alone. Moreover, through exploiting the cluster discrimination in MR, the proposed DAMR framework utilizes both the cluster and manifold assumptions. Researchers have demonstrated that methods working on multiple data distribution assumptions can achiever better classification than those just working on a single one [14,15]. As a result, DAMR is able to be expected to perform better than MR.

As in MR, the minimizer of the optimization problem (13) can also be formulated as $f^*(x) = \sum_{i=1}^{l+u} \alpha_i K(x_i, x)$ according to the Representation Theorem.

4. Discrimination-aware Lap_RLSC (DA_LapRLSC)

4.1. Algorithm

In terms of different loss functions, we can develop different algorithms for DAMR. We adopt the square loss in this section as an example, and derive an algorithm called Discrimination-aware Lap_RLSC (DA_LapRLSC). The optimization problem of DA_LapRLSC can be written as,

$$\min_f J = \frac{1}{l}\sum_{i=1}^{l} (y_i - f(x_i))^2 + \gamma_A \|f\|_K^2 + \frac{\gamma_D}{2(l+u)^2}\sum_{i,j=1}^{l+u} W_{ij}(f(x_i) - S_{ij} f(x_j))^2 \tag{14}$$

As in MR, the minimizer of (14) can also be formulated as $f^*(x) = \sum_{i=1}^{l+u} \alpha_i K(x_i, x)$ according to the Representation Theorem. Then (14) can be further written as

$$\min_f J = \frac{1}{l}(\mathbf{K}_l\boldsymbol{\alpha} - \mathbf{Y})^T(\mathbf{K}_l\boldsymbol{\alpha} - \mathbf{Y}) + \gamma_A \boldsymbol{\alpha}^T \mathbf{K}\boldsymbol{\alpha} + \frac{\gamma_D}{(l+u)^2}\boldsymbol{\alpha}^T \mathbf{K}\mathbf{L}_D\mathbf{K}\boldsymbol{\alpha} \tag{15}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots \alpha_{l+u}]^T$ is the vector of Lagrange multipliers. $\mathbf{K}_l = (\mathbf{X}_l, \mathbf{X})_H \in R^{l \times (l+u)}$ and $\mathbf{K} = (\mathbf{X}, \mathbf{X})_H \in R^{(l+u)\times(l+u)}$ are the kernel matrices, where $\mathbf{X}_l$ and $\mathbf{X}$ denote the labeled and the whole datasets, respectively. $\mathbf{Y} = [y_1, \ldots y_l]^T$ is the vector of class labels for the labeled data. Zeroing the derivation of (15) w.r.t. $\boldsymbol{\alpha}$, we have

$$\frac{\partial J}{\partial \boldsymbol{\alpha}} = \frac{1}{l}\mathbf{K}_l^T(\mathbf{K}_l\boldsymbol{\alpha} - \mathbf{Y}) + \gamma_A \mathbf{K}\boldsymbol{\alpha} + \frac{\gamma_D}{(l+u)^2}\mathbf{K}\mathbf{L}_D\mathbf{K}\boldsymbol{\alpha} \tag{16}$$

Finally, we have

$$\boldsymbol{\alpha} = \left(\frac{1}{l}\mathbf{K}_l^T\mathbf{K}_l + \gamma_A \mathbf{K} + \frac{\gamma_D}{(l+u)^2}\mathbf{K}\mathbf{L}_D\mathbf{K}\right)^{-1}\frac{1}{l}\mathbf{K}_l^T\mathbf{Y} \tag{17}$$

The concrete algorithm description of DA_LapRLSC is summarized in Table 1 below.

Note that though the square loss function is adopted here as an example, one can also adopt other loss functions to develop specific algorithms within the framework of DAMR.

4.2. Computation complexity

In Lap_RLSC, the construction of the weight matrix needs a computation complexity of $O(n^2)$, where n is the size of the whole dataset, and the solution involves the inverse of the kernel matrix over the whole dataset with a computation complexity of $O(n^3)$. As a result, the computation complexity of Lap_RLSC is $O(n^3)$. In SSDR in ref. [20], there is the inverse of the kernel matrix in both the construction of the discriminative kernel and the solution process, thus its computation complexity is also $O(n^3)$. In SSDR in ref. [16], the complexities of constructing the intra-class and inter-class graphs are both $O(n^2)$, and the complexity of the solution is $O(n^3)$ as in Lap_RLSC, thus the overall computation complexity is still O (n^3). In our DA_LapRLSC, the complexity of the clustering method FCM is $O(c^2ndi)$ [25], where c is the number of clusters, d is the dimension of the dataset, and i is the number of iterations in FCM. The complexity of constructing the discriminative matrix is $O(n^2)$, and the complexity of the solution is $O(n^3)$. As a result, the total

Table 1
The algorithm description of DA_LapRLSC.

Input	$\mathbf{X}_l$, $\mathbf{X}_u$ – the labeled and unlabeled dataset $\mathbf{Y}_l$ – the label set of $\mathbf{X}_l$ γ_A, γ_D, γ – the regularization parameters
Output	$f(x)$ – the decision function
Procedure Get the initial cluster indicator vector $\mathbf{Y}_c$ by some unsupervised clustering method such as FCM; Construct the discriminative matrix $\mathbf{S}$; Construct the new Laplacian matrix $\mathbf{L}_D$ with the discrimination; Get $\boldsymbol{\alpha}^*$ for DA_LapRLSC by (17); Predict any instance x by decision function $f^*(x)=\sum_{i=1}^{l+u}\alpha_i K(x_i,x)$;	

Table 2
The description of the 14 UCI datasets.

Dataset	Size	Dimension	Dataset	Size	Dimension
Automobile	159	25	Horse	366	27
Austra	690	15	Ionosphere	351	34
Biomed	194	5	Isolet	600	51
Bupa	345	6	Pima	768	8
Ethn	2630	30	Sonar	208	60
German	1000	24	Vehicle	435	16
Heart	270	9	Wdbc	569	14

computation complexity of DA_LapRLSC is $O(c^2ndi+n^3)$, or $O(4ndi+n^3)$ with the cluster number set to the class number 2 here. Since both the data dimension d and iteration number i are commonly much smaller than the instance number n, the computation complexity of DA_LapRLSC can also be considered to be comparable to $O(n^3)$.

5. Experiments

In this section, we evaluate the performance of the developed DAMR framework (specifically, DA_LapRLSC) over 14 UCI datasets compared with Lap_RLSC, and two off-the-shelf improvements of MR considering the discrimination, i.e., SSDR in [24] and [16], respectively.

The descriptions of the 14 UCI datasets are given in Table 2. Each UCI dataset is randomly split into two halves, one for training and the other for testing, and the training set contains 10 and 100 labeled instances, respectively. The FCM method and linear kernel are adopted here, and the neighbor graph is used with the neighbor number k fixed to 10, and the cluster number is set to the class number.

5.1. Performance comparison

The heat kernel weighting strategy is used in the graph construction. When 10 instances are labeled, γ_A and γ_I in Lap_RLSC are fixed to 1 and 1, respectively, γ_1, γ_2 and λ in SSDR [24] are fixed to 1, 1 and 1, respectively, γ_A, γ_I and γ_D in SSDR [16] are fixed to 1, 1, and 0.6, respectively, and γ_A, γ_D and γ in DA_LapRLSC are fixed to 1, 1 and 1, respectively. When 100 instances are labeled, γ in DA_LapRLSC and γ_D in SSDR [16] are both selected by 5-cross validation from [0, 0.2, 0.4, 0.6, 0.8, 1], and the other parameters are all selected from the range [0.01, 0.1, 1, 10, 100]. This training process is repeated 20 times and the average accuracy and variance are reported in Tables 3 and 4, respectively, in which the best performance over each dataset is highlighted in bold in each row.

Table 3
The comparative results with 10 labeled instances.

Dataset	Lap_RLSC	SSDR [24]	SSDR [16]	DA_LapRLSC
Automobile	78.22 ± 0.69	**80.34 ± 0.84**	78.69 ± 0.69	76.61 ± 0.45
Austra	59.34 ± 0.44	59.22 ± 0.46	59.89 ± 0.44	**60.38 ± 0.41**
Biomed	55.22 ± 0.65	60.82 ± 0.67	55.22 ± 0.65	**64.62 ± 0.58**
Bupa	**49.67 ± 0.03**	48.37 ± 0.10	49.57 ± 0.03	49.03 ± 0.03
Ethn	59.30 ± 1.29	59.41 ± 1.24	59.86 ± 1.25	**61.25 ± 1.25**
German	62.54 ± 0.25	60.80 ± 0.22	62.54 ± 0.25	**64.20 ± 0.07**
Heart	51.56 ± 0.31	52.67 ± 0.35	51.56 ± 0.31	**64.81 ± 1.01**
Horse	53.72 ± 0.24	50.87 ± 0.35	53.80 ± 0.26	53.72 ± 0.32
Ionosphere	69.52 ± 0.25	**72.33 ± 0.15**	69.55 ± 0.26	71.39 ± 0.43
Isolet	80.93 ± 2.07	92.27 ± 0.09	80.93 ± 2.07	**94.13 ± 0.10**
Pima	47.73 ± 0.39	**54.37 ± 0.11**	47.75 ± 0.40	51.50 ± 0.32
Sonar	47.75 ± 0.04	47.60 ± 0.05	47.75 ± 0.04	**50.45 ± 0.15**
Vehicle	67.75 ± 0.48	**75.41 ± 0.49**	68.29 ± 0.42	70.18 ± 0.76
Wdbc	62.81 ± 0.00	62.81 ± 0.00	62.89 ± 0.00	**67.33 ± 0.04**

Table 4
The comparative results with 100 labeled instances.

Dataset	Lap_RLSC	SSDR [24]	SSDR [16]	DA_LapRLSC
Automobile	90.17 ± 0.05	90.76 ± 0.10	**91.19 ± 0.05**	90.25 ± 0.06
Austra	63.56 ± 0.02	63.32 ± 0.03	63.56 ± 0.02	**67.85 ± 0.02**
Biomed	77.39 ± 0.15	77.71 ± 0.16	78.44 ± 0.15	**88.62 ± 0.01**
Bupa	56.03 ± 0.03	58.62 ± 0.07	56.99 ± 0.05	**64.88 ± 0.09**
Ethn	92.37 ± 0.03	92.45 ± 0.02	**92.59 ± 0.02**	92.40 ± 0.02
German	64.68 ± 0.37	64.93 ± 0.32	64.92 ± 0.37	**68.44 ± 0.04**
Heart	82.17 ± 0.3	82.27 ± 0.04	82.96 ± 0.02	**83.41 ± 0.01**
Horse	62.86 ± 0.14	**66.43 ± 0.17**	63.17 ± 0.14	63.59 ± 0.23
Ionosphere	88.61 ± 0.01	88.98 ± 0.01	88.98 ± 0.01	**89.22 ± 0.00**
Isolet	98.34 ± 0.00	**98.64 ± 0.01**	98.57 ± 0.00	98.57 ± 0.00
Pima	57.10 ± 0.02	58.70 ± 0.02	57.97 ± 0.01	**60.40 ± 0.02**
Sonar	77.13 ± 0.09	**78.52 ± 0.09**	77.36 ± 0.08	77.73 ± 0.08
Vehicle	79.87 ± 0.14	77.61 ± 0.02	**81.11 ± 0.09**	76.67 ± 0.02
Wdbc	91.07 ± 0.01	**95.10 ± 0.01**	91.27 ± 0.01	91.57 ± 0.01

From Tables 3 and 4, we can get several observations as follows,

- SSDR in [24] defines a new discriminative kernel for exploiting the discrimination in semi-supervised classification learning, and usually performs better than Lap_RLSC. Specifically, SSDR in [24] performs better than Lap_RLSC over 9 out of the 14 datasets when 10 instances are labeled, and outperforms Lap_RLSC over 12 datasets when 100 instances are labeled. As a result, utilizing the discrimination can help improving the performance of semi-supervised classification.
- Utilizing the discrimination of the labeled data, SSDR in [16] also performs better than Lap_RLSC in most cases, indicating again that the discrimination can boost learning. Specifically, SSDR in [16] performs better than Lap_RLSC over 8 datasets when 10 instances are labeled, and outperforms Lap_RLSC over

13 datasets when 100 instances are labeled. However, the improvement is not so distinct when the labeled data are scarce. As can be seen, the performance improvement of SSDR in [16] is much more distinct in the case of 100 labeled instances than that of 10 labeled instances. As a result, we attempt to exploit the discrimination of all available data in this paper.

- Through using the discrimination of all given data, DA_LapRLSC performs better than Lap_RLSC over 11 datasets when 10 instances are labeled, and outperforms Lap_RLSC over 13 datasets when 100 instances are labeled. Moreover, it performs the best over 8 datasets when 10 instances are labeled, and outperforms the other methods over 7 datasets when 100

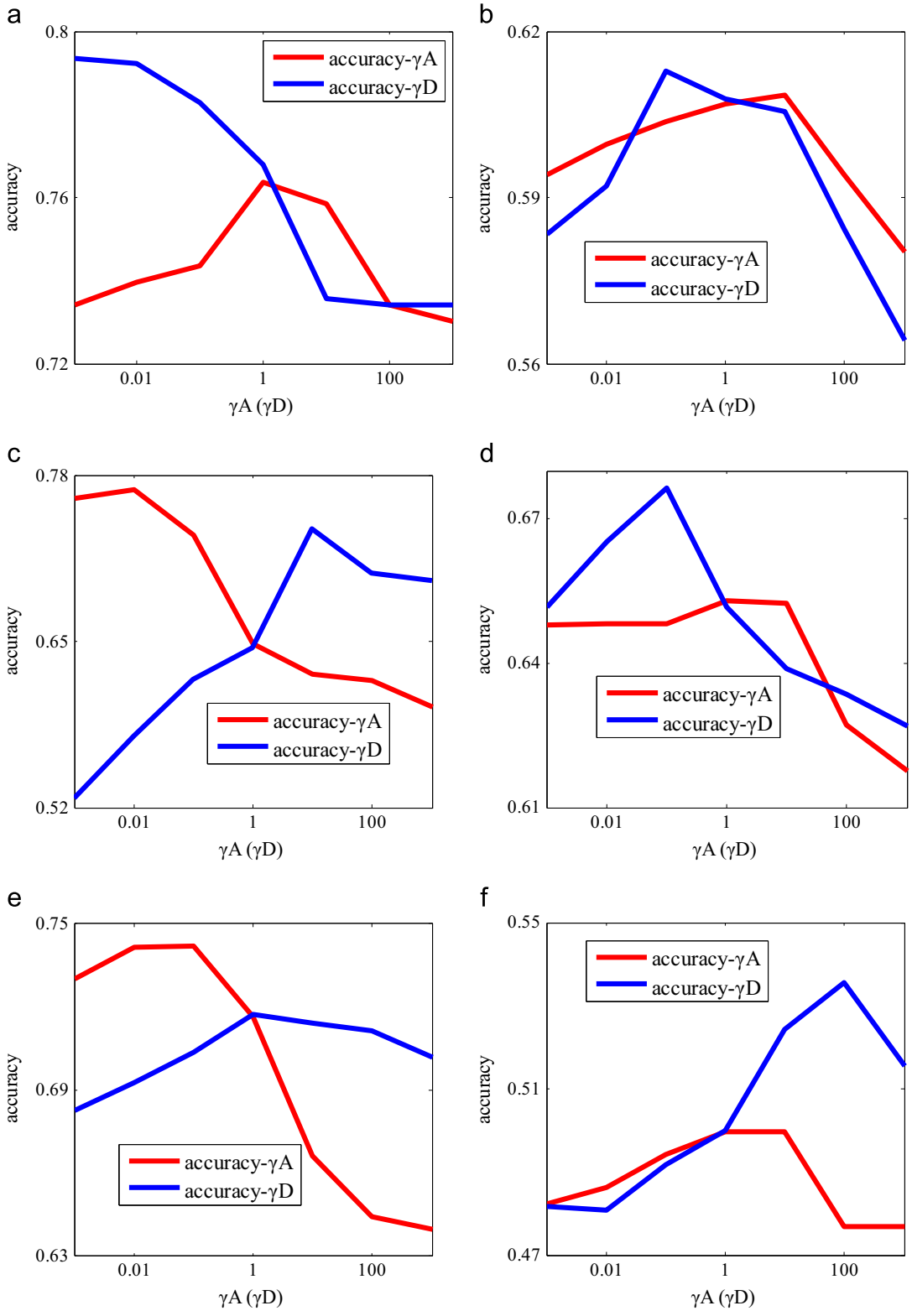

Fig. 2. The performances of DA_LapRLSC w.r.t. different values of γ_A (γ_D) from {0.001, 0.01, 0.1, 1, 10, 100, 1000} over (a) automobile (b) austra (c) biomed (d) heart (e) ionosphere and (f) sonar, accuracy-γ_A indicates the performance curve w.r.t different values of γ_A, and accuracy-γ_D is the performance curve w.r.t. different values of γ_D.

instances are labeled, indicating the effectiveness of the proposed DA_LapRLSC compared with the other methods.

5.2. Parameter analysis

We show the performance of DA_LapRLSC w.r.t different values of γ_A and γ_D, respectively, from {0.001, 0.01, 0.1, 1, 10, 100, 1000} over 6 datasets with 10 labeled instances in Fig. 2, with γ fixed to 1. Moreover, we also give the performance of DA_LapRLSC w.r.t different values of γ from {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1} in Fig. 3, with γ_A and γ_D both fixed to 1.

In Fig. 2, the performance of DA_LapRLSC (accuracy-γ_A) tends to increase and then decrease with the increase of γ_A. Since when γ_A is small, DA_LapRLSC considers little about the smoothness of the decision function in the RKHS space, and when γ_A is large, the classification of DA_LapRLSC will be dominated by such smoothness.

At the same time, the performance of DA_LapRLSC (accuracy-γ_D) in Fig. 2 tends to increase and then decrease with the increase of γ_D over most datasets. The reason is that when γ_D is small, DA_LapRLSC considers little about the discriminative smoothness constraints over the whole dataset, and when γ_D is large, the classification of DA_LapRLSC will be dominated by those constraints. However, it is not the case over automobile. Specifically, the performance of DA_LapRLSC tends to decrease with the increase of γ_D over automobile, indicating that the discrimination smoothness constraints may not be helpful for learning in this case.

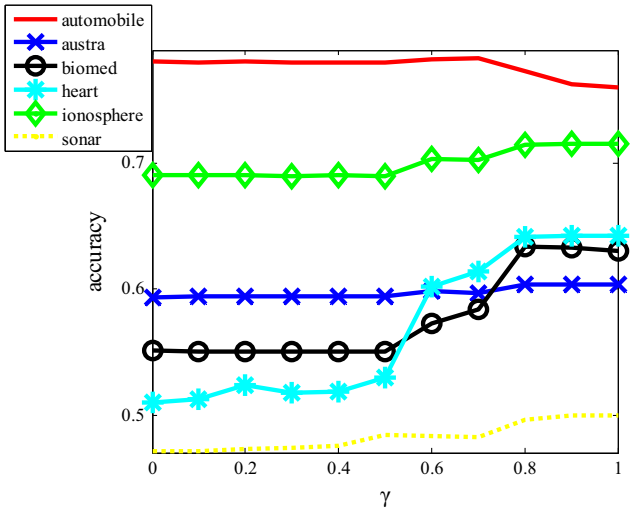

Fig. 3. The performances of DA_LapRLSC w.r.t. different values of γ from {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1} over (a) automobile, (b) austra, (c) biomed, (d) heart, (e) ionosphere and (f) sonar.

Further, when γ is small, the instance pairs satisfying $T_{ij} > 0.5$ are scarce, and the discriminative smoothness constraints in DA_LapRLSC mainly degenerate to the smoothness constraints in MR, otherwise, most of the discriminative smoothness constraints are kept in DA_LapRLSC. From Fig. 3, we can see that the performance of DA_LapRLSC tends to descend with the increase of γ with more discriminative smoothness constraints over automobile, indicating that the discrimination smoothness constraints may not be helpful for learning here. At the same time, the performance tends to ascend with the increase of γ over the other 5 datasets, indicating that the discrimination smoothness constraints are indeed helpful for learning in those cases.

5.3. Time complexity

We show the computation time of the compared methods with 10 instance labels in Table 5 below. From Table 5, we can see that though our DA_LapRLSC is not so competitive in efficiency, its computation time is still acceptable trading for better classification performances.

6. Conclusion

Considering that MR imposes the smoothness constraint over each instance pairs, including the instances in the boundary area, thus may misclassify those boundary instances, we develop a discrimination-aware MR (DAMR) framework in this paper through incorporating the discrimination of all available data from some unsupervised learning method. In the learning of DAMR, with high similarity over the manifold structure, instances in the same cluster are restricted to share the same class output, while instances in different clusters are restricted to have different class outputs. In this way, DAMR actually utilizes both the manifold and cluster assumptions. In the implementation, we simply use the square loss function and FCM to develop a specific DA_LapRLSC algorithm as an example. Empirical results show the competitive results of DA_LapRLSC compared with Lap_RLSC, as well as SSDR in [24] and [16], respectively.

Though we adopt the FCM method in the experiment as an example, other unsupervised learning methods or even some semi-supervised clustering method can also be adopted for exploiting the data structure with both labeled and unlabeled data, moreover, the clustering method cannot avoid grouping instances in different classes into the same cluster, thus some other strategy is also worth

Table 5
The computation times(s) of individual methods with 10 instances labeled.

Dataset	Lap_RLSC	SSDR [24]	SSDR [16]	DA_LapRLSC
Automobile	0.2402 ± 0.01	0.0468 ± 0.002	0.2246 ± 0.0001	0.3368 ± 0.0003
Austra	3.1824 ± 0.0268	7.1783 ± 0.0752	3.2649 ± 0.0211	5.0879 ± 0.0296
Biomed	0.2465 ± 0.0053	0.1622 ± 0.0007	0.2496 ± 0.0005	0.2680 ± 0.0064
Bupa	0.5054 ± 0.0047	0.6583 ± 0.0052	0.5460 ± 0.0021	0.7421 ± 0.0179
Ethn	193.6908 ± 12.5806	541.8539 ± 21.3753	202.0057 ± 15.2876	218.6510 ± 11.6523
German	11.0969 ± 0.1583	27.0818 ± 0.1557	10.6705 ± 0.1003	14.6953 ± 0.1221
Heart	0.3713 ± 0.0024	0.4649 ± 0.0000	0.4368 ± 0.0013	0.6268 ± 0.0002
Horse	0.5554 ± 0.0020	0.6802 ± 0.0167	0.5834 ± 0.0105	0.6670 ± 0.0133
Ionosphere	0.4867 ± 0.0120	0.3463 ± 0.0005	0.5023 ± 0.0024	0.7951 ± 0.0039
Isolet	1.8252 ± 0.0090	3.7544 ± 0.0619	1.8824 ± 0.0079	3.3072 ± 0.0068
Pima	5.0420 ± 0.4375	10.0433 ± 0.1282	4.7612 ± 0.0080	6.8422 ± 0.0030
Sonar	0.2527 ± 0.0031	0.0874 ± 0.0002	0.2527 ± 0.0003	0.3585 ± 0.0026
Vehicle	0.6833 ± 0.0077	1.1731 ± 0.0170	0.7706 ± 0.0115	0.9038 ± 0.0255
Wdbc	1.7316 ± 0.0032	3.3759 ± 0.0849	1.8096 ± 0.0375	2.9578 ± 0.0192

investigating, which is exactly one of our future works. At the same time, seeking for strategies for selecting the parameters in DAMR, and further improving the efficiency of DAMR are both learning problems worth investigating in our future work.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant nos. 61300165, 61272422, 61375057 and 61300164, the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant no. 20133223120009, the Introduction of Talent Research Foundation of Nanjing University of Posts and Telecommunications under Grant nos. NY213033 and NY213031, the Natural Science Foundation of Jiangsu Province of China under Grant no. BK20131298, and sponsored by Jiangsu Qinglan project.

References

[1] Z.-H. Zhou, M. Li, Semi-supervised learning by disagreement, Knowl. Inf. Syst. 24 (2010) 415–439.
[2] X. Zhu, A.B. Goldberg, Introduction to Semi-Supervised Learning, Morgan & Claypool, San Rafael, 2009.
[3] X. Zhu, Semi-supervised Learning Literature Survey, Technical Report, Computer Sciences, University of Wisconsin-Madison, MA, 2008.
[4] O. Chapelle, B. Scholkopf, A. Zien, Semi-supervised Learning, MIT Press, Cambridge, Massachusetts, USA, 2006.
[5] P.K. Mallapragada, R. Jin, A.K. Jain, Y. Liu, Semi-Boost: boosting for semi-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 2000–2014.
[6] T. Joachims, Transductive inference for text classification using support vector machines, in: Proceedings of the 16th International Conference on Machine Learning, Bled, Slovenia, 1999, pp. 200–209.
[7] G. Fung, O.L. Mangasarian, Semi-supervised support vector machine for unlabeled data classification, Optim. Methods Softw. 15 (2001) 99–105.
[8] R. Collobert, F. Sinz, J. Weston, L. Bottou, Large scale transductive SVMs, J. Mach. Learn. Res. 7 (2006) 1687–1712.
[9] Y.-F. Li, J.T. Kwok, Z.-H. Zhou, Semi-supervised learning using label mean, in: Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009, pp. 633–640.
[10] Y. Bengio, O. Delalleau, N.L. Roux, Label Propagation and Quadratic Criterion, in Semi-Supervised Learning, MIT Press, Cambridge, Massachusetts, USA, 2006.
[11] X. Zhu, Z. Ghahramani, Learning from Labeled and Unlabeled Data with Label Propagation, Technical Report, Carnegie Mellon University, 2002.
[12] A. Blum, S. Chawla, Learning from labeled and unlabeled data using graph mincuts, in: Proceedings of the 18th International Conference on Machine Learning, Williamstown, MA, USA, 2001, pp. 19–26.
[13] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (2006) 2399–2434.
[14] K. Chen, S. Wang, Semi-supervised learning via regularized boosting working on multiple semi-supervised assumptions, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 129–143.
[15] P.K. Mallapragada, R. Jin, A.K. Jain, Y. Liu, SemiBoost: boosting for semi-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 2000–2014.
[16] F. Wu, W. Wang, Y. Yang, Y. Zhuang, F. Nie, Classification by semi-supervised discriminative regularization, Neurocomputing 73 (2010) 1641–1651.
[17] X. He, Laplacian regularized D-optimal design for active learning and its application to image retrieval, IEEE Trans. Image Process. 19 (2010) 254–263.
[18] J. Abernethy, O. Chapelle, C. Castillo, Web spam identification through content and hyperlinks, in: Proceedings of the 4th International Workshop on Adversarial information retrieval on the web, Beijing, China, 2008, pp. 41–44.
19 M Belkin, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Neural Information Processing Systems, Vancouver, British Columbia, Canada, 2001, pp. 585–591.
[20] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Am. Assoc. Adv. Sci. 290 (2000) 2323–2326.
[21] D. Cai, X. He, J. Han, T. Huang, Graph regularized non-negative matrix factorization for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 1548–1560.
[22] J. Huang, F. Nie, H. Huang, C. Ding, Robust manifold non-negative matrix factorization, ACM Trans. Knowl. Data Discov. 8 (2014) 11.
[23] A.B. Goldberg, X. Zhu, S. Wright, Dissimilarity in graph-based semi-supervised classification, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, Monte Carlo Resort, Las Vegas, Nevada, USA, 2007, pp. 155–162.
[24] F. Wang, C. Zhang, On discriminative semi-supervied classificaiton, in: Proceedings of the 23nd AAAI Conference on Artificial Intelligence, Chicago, Illinois, 2008, pp. 720–725.
[25] S. Ghosh, S.K. Dubey, Comparative analysis of K-means and fuzzy C-means algorithms, Int. J. Adv. Comput. Sci. Appl. 4 (2013) 35–39.

Yunyun Wang received the B.S degree in Computer Science and Engineering from Anhui Normal University in 2002, and the Ph.D. degree in Computer Science and Technology from Nanjing University of Aeronautics and Astronautics in 2012. She is currently with the School of Computer Science and Technology in Nanjing University of Posts and Telecommunications. Her current research interests include pattern recognition, machine learning and neural computing.

Songcan Chen received his B.S. degree in mathematics from Hangzhou University (now merged into Zhejiang University) in 1983. In 1985, he completed his M.S. degree in computer applications at Shanghai Jiaotong University and then worked at NUAA in January 1986. There he received a Ph.D. degree in communication and information systems in 1997. Since 1998, as a full-time professor, he has been with the Department of Computer Science & Engineering at NUAA. His research interests include pattern recognition, machine learning and neural computing. In these fields, he has authored or coauthored over 130 scientific peer-reviewed papers.

Hui Xue received her B.S. degree in mathematics from Nanjing Normal University in 2002. In 2005, she received her M.S. degree in mathematics from Nanjing University of Aeronautics & Astronautics (NUAA). And she also received her Ph.D. degree in computer application technology at NUAA in 2008. Since 2009, she has been with the School of Computer Science & Engineering at Southeast University. Her research interests include pattern recognition, machine learning and neural computing.

Zhenyong Fu received the B.S degree in Mathematics from the Kunming University of Science and Technology in 2002, the M.E. degree in Computer Science from the Fudan University in 2005, and the Ph.D. degree in the Department of Computer Science and Engineering from Shanghai Jiao Tong University in 2013. He is currently with the School of Computer Science and Technology in Nanjing University of Posts and Telecommunications. His research interest includes machine learning, computer vision and pattern recognition.