Semi-Supervised Learning with Mixed Knowledge Information

Fanhua Shang, L. C. Jiao
Key Laboratory of Intelligent Perception and Image
Understanding of Ministry of Education of China
Xidian University, Xi'an 710071, PR China
shangfanhua@hotmail.com
jlcxidian@163.com

Fei Wang
Healthcare Analytics Research Group,
IBM T. J. Watson Research Center
Hawthorne, NY, USA
feiwang03@gmail.com

ABSTRACT

Integrating new knowledge sources into various learning tasks to improve their performance has recently become an interesting topic. In this paper we propose a novel semi-supervised learning (SSL) approach, called semi-supervised learning with *Mixed Knowledge Information* (SSL-MKI) which can simultaneously handle both sparse labeled data and additional pairwise constraints together with unlabeled data. Specifically, we first construct a unified SSL framework to combine the manifold assumption and the pairwise constraints assumption for classification tasks. Then we present a *Modified Fixed Point Continuation* (MFPC) algorithm with an eigenvalue thresholding (EVT) operator to learn the enhanced kernel matrix. Finally, we develop a two-stage optimization strategy and provide an efficient SSL approach that takes advantage of Laplacian spectral regularization: semi-supervised learning with *Enhanced Spectral Kernel* (ESK). Experimental results on a variety of synthetic and real-world datasets demonstrate the effectiveness of the proposed ESK approach.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms

Algorithms, Performance, Experimentation

Keywords

Semi-supervised learning (SSL), Kernel learning, Graph Laplacian, Nuclear norm regularization, Pairwise constraints

1. INTRODUCTION

Semi-Supervised Learning (SSL) has recently received a significant amount of attention in the machine learning and data mining communities [1, 2]. A large amount of SSL approaches have been proposed, including EM with generative mixture models, self-training, co-training, transductive support vector machines (TSVMs), and graph-based methods. Among them, a family of graph-based SSL methods is label propagation techniques [3, 4] or provides graph-based regularization frameworks for learning from labeled and unlabeled data [5, 6]. Although graph-based SSL has been studied extensively, so far there are few comprehensive techniques to integrate weakly labeled data and pairwise constraints together for classification tasks.

In SSL, the class labels of data are the most widely used supervisory information [7]. But labeled instances are often difficult, expensive, or time consuming to obtain, as they require the efforts of the domain experts. Integrating new knowledge sources such as side information into classification tasks with insufficient training data has recently become an interesting topic [8]. In this paper we are particularly interested in how to incorporate additional pairwise constraints to improve classification accuracy. *Pairwise constraints* may be relatively easy to collect, and indicate that some pairs of instances are in the same class and some are not, known as the *Must-Link* (ML) and the *Cannot-Link* (CL), respectively, [1, 9, 10]. Pairwise constraints can also be obtained from data labels where objects with the same label are must-link while objects with different labels are cannot-link. Pairwise constraints have been widely used in various tasks such as clustering [9-12] and distance metric learning [13-15] where it has been shown that the presence of appropriate pairwise constraints can achieve reasonable performance improvement. However, there are relatively fewer works using additional pairwise constraints to support semi-supervised classification tasks [8].

Generally, SSL methods are derived based on two fundamental assumptions: the *cluster assumption* (also called label consistency) and the *manifold assumption*. In the cluster assumption, decision boundaries should not cross high density regions, but instead lie in low density regions. Typical methods include TSVMs [16, 17] and some convex relaxation methods. In the manifold assumption, data are assumed to be sampled from a low-dimensional manifold that is embedded inside of a higher dimensional input space. Many SSL approaches implement such an assumption by using the graph Laplacian of a neighborhood graph which can characterize the underlying manifold structure. Typical methods include Zhu et al.'s Gaussian random fields (GRF) method [3], Zhou et al.'s learning with local and global consistency (LGC) [4], manifold regularization [5, 6], etc. More recently, Li et al. [7] presented a *pairwise constraint assumption* that is very effective for classification tasks together with the cluster assumption. In the pairwise constraint assumption, those unlabeled data points involved in any ML constraint are classified into the same class

KDD'12, August 12 - 16 2012, Beijing, China

and those involved in any CL constraint are classified into different classes.

Despite many successes, most graph-based SSL methods mentioned above have a limitation in the difficulty of tuning optimal graph parameters. Specially, with limited labeled data, it may be ineffective to learn kernel parameters by cross-validation from labeled data only [18]. To address this limitation, various kernel learning approaches [7, 18-23] (also called non-parametric kernel learning) are proposed to learn a positive semidefinite (PSD) kernel matrix directly from the data incorporating label or side information. Several existing studies [7, 21] have shown that time complexity of standard interior-point semidefinite programming (SDP) solvers to learn the entire kernel matrix could be as high as $O(n^{6.5})$, where n is the number of data points, which prohibits those approaches for practical applications [18], whereas there are several efficient approaches derived from the spectral decomposition of graph Laplacians, such as the *Order-constrained Spectral Kernel* (OSK) [19, 20], the *Transductive Spectral Kernel* (TSK) [22], and the graph *Laplacian Regularized Kernel* (LRK) [24].

In this paper, we consider a more general problem of semi-supervised classification which can handle sparse labeled data and additional pairwise constraints together with abundant unlabeled data, and propose a novel SSL approach, also called semi-supervised learning with *Mixed Knowledge Information* (SSL-MKI). We first construct a unified SSL-MKI framework to implement both the manifold assumption and the pairwise constraint assumption. Under the SSL-MKI framework, we also present a semi-supervised low-rank kernel learning model with nuclear norm regularization. Then we develop a two-stage optimization strategy and provide an efficient SSL-MKI algorithm that takes advantage of Laplacian spectral regularization: semi-supervised learning with *Enhanced Spectral Kernel* (ESK).

2. NOTATIONS AND BACKGROUND

Given a data set of n instances $X = \{x_1, x_2, \cdots, x_l, x_{l+1}, \cdots, x_n\}$, the first small number l points $X_L = \{x_i\}_{i=1}^{l}$ belonging to c classes are labeled, and the remaining points $X_U = \{x_i\}_{i=l+1}^{n}$ are unlabeled, and two sets of pairwise must-link and cannot-link constraints are denoted respectively by $\mathrm{ML} = \{(x_i, x_j)\}$ where x_i and x_j should be in the same class, and $\mathrm{CL} = \{(x_i, x_j)\}$ where x_i and x_j should be in different classes.

Before we go into the details of our method, we will briefly review some of the related works in this section. As mentioned in the previous section, the common denominator of graph-based methods is to model the whole data set as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the vertex set $\mathcal{V} = X$, and the weight w_{ij} on the edge $e_{ij} \in \mathcal{E}$ representing the similarity between x_i and x_j. For convenience, we adopt the local scaling parameter trick proposed in [25] to construct $\mathcal{G}$ as a T-nearest neighbor (T-NN) graph. Let us define a selecting function of the local scale,

$$h(x) = \left\| x - x^{(T)} \right\|, \tag{1}$$

where $x^{(T)}$ is the T-th nearest neighbor of x in X. The weight matrix $W \in \mathbb{R}^{n \times n}$ associated with $\mathcal{G}$ is formed subsequently as

$$W_{ij} = \begin{cases} \exp\left(-\dfrac{\left\| x_i - x_j \right\|^2}{h(x_i)h(x_j)} \right), & \text{if } x_j \in N(x_i) \text{ or } x_i \in N(x_j); \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Note that we set $W_{ii} = 0$ to avoid self-loops. We further denote the diagonal degree matrix $D \in \mathbb{R}^{n \times n}$ whose entries are given by $D_{ii} = \sum_j W_{ij}$, and the normalized *graph Laplacian* $L = I - D^{-1/2} W D^{-1/2}$.

The regularization framework proposed by Zhou et al. [4] can implement a global classification task as follows:

$$\begin{aligned} \mathcal{Q}(F) &= \mathcal{S}(F) + \mathcal{L}(F) \\ &= \frac{1}{2} \sum_{i,j=1}^{n} W_{ij} \left\| \frac{F_i}{\sqrt{D_{ii}}} - \frac{F_j}{\sqrt{D_{jj}}} \right\|^2 + \mu \sum_{i=1}^{n} \left\| F_i - Y_i \right\|^2, \end{aligned} \tag{3}$$

where F is a vectorial function, $F: X \to \mathbb{R}^c$, to assign a label $f_i = \arg\max_{j \le c} F_{ij}$ to each instance x_i, $\mu > 0$ is the regularization parameter, and $Y \in \mathbb{R}^{n \times c}$ is a class indicator matrix, where $Y_{ik} = 1$ if x_i is labeled as k, and $Y_{ik} = 0$ otherwise. The $\mathcal{S}(\cdot)$ is a regularizer to penalize the smoothness of the classifying function F over the graph, and the $\mathcal{L}(\cdot)$ is a loss function measuring the inconsistency between the predicted and initial label assignment. Then the classification function is

$$F^* = \arg\min_F \mathcal{Q}(F) = (1-\alpha)(I - \alpha S)^{-1} Y, \tag{4}$$

where α ($0 < \alpha = 1/(1+\mu) < 1$) is the regularization parameter and $S = D^{-1/2} W D^{-1/2}$ symmetrically normalizes W. The calculated matrix F^* stacks the final class assignment. But this method is an unreliable approach for model selection if only very few labeled instances are available [26].

In recent years, a large amount of low-rank matrix recovery methods have been proposed for matrix completion [27]-[30] problems. Among of them, there are several representative methods such as SVT [27], FPCA [28], and ALM [29]. And some works also provide theoretical guarantee that the task of the *rank minimization* problem can be accomplished via solving the *nuclear norm* (also known as the trace norm) minimization under some reasonable conditions. Specifically, Candès and Recht [30] proved that a given incomplete low-rank matrix (but unknown) $Z \in \mathbb{R}^{n \times n}$ satisfying certain incoherence conditions can be exactly recovered by the following model (5) with probability at least $1 - B_1 n^{-3}$ from a subset Ω of uniformly sampled entries $\{Z_{ij} : (i,j) \in \Omega\}$ whose cardinality $|\Omega|$ is of the form $B_2 r n^{5/4} \log n$, where r is the rank of the desired matrix, and B_1 and B_2 are two positive constants.

$$\begin{aligned} &\min_K \left\| K \right\|_*, \\ &\text{s.t., } M \odot (K - Z) = 0, \end{aligned} \tag{5}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix, i.e., the sum of its singular values, the operator $\odot$ denotes element-wise multiplication, and the weight matrix M is defined as

$$M_{ij}=\begin{cases}1, & \text{if } (i,j)\in\Omega,\\ 0, & \text{otherwise.}\end{cases} \tag{6}$$

3. FRAMEWORK OF SSL-MKI

3.1 A General Framework

As mentioned above, our goal is to use the given labeled data and pairwise constraints together with unlabeled instances for classification tasks. We propose a general SSL-MKI framework as follows:

$$\mathcal{Q}(F,K)=\mathcal{S}(F,K)+\mu_1\mathcal{L}_1(F)+\mu_2\mathcal{L}_2(K), \tag{7}$$

where K is a desired "ideal" kernel matrix, $\mathcal{S}(\cdot,\cdot)$ is a regularizer to penalize the smoothness of the classifying function F, $\mathcal{L}_1(F)$ is a loss function penalizing the inconsistency between the predicted and initial label assignment, and $\mathcal{L}_2(K)$ is a loss function to measure the change between the predicted and "ideal" kernels corresponding to the given pairwise constraints, such as the squared loss or hinge loss functions. $\mu_1>0$ and $\mu_2>0$ are the regularization parameters for $\mathcal{L}_1(F)$ and $\mathcal{L}_2(K)$, respectively. In the unified framework, we can implement both the *manifold assumption* and the *pairwise constraint assumption*.

3.2 Nuclear Norm Regularized Model

Under the squared loss function, the model (7) is formulated as follows:

$$\begin{aligned}\mathcal{Q}(F,K)=&\frac{1}{2}\sum_{i,j=1}^{n}\tilde{K}_{ij}\left\|\frac{F_i}{\sqrt{D_{ii}}}-\frac{F_j}{\sqrt{D_{jj}}}\right\|^2+\mu_1\sum_{i=1}^{n}\left\|F_i-Y_i\right\|^2\\ &+\mu_2\left(\sum_{i=1}^{n}(K_{ii}-1)^2+\sum_{(x_i,x_j)\in\mathrm{ML}}^{n}(K_{ij}-1)^2+\sum_{(x_i,x_j)\in\mathrm{CL}}^{n}(K_{ij}-0)^2\right),\\ =&\frac{1}{2}\sum_{i,j=1}^{n}\tilde{K}_{ij}\left\|\frac{F_i}{\sqrt{D_{ii}}}-\frac{F_j}{\sqrt{D_{jj}}}\right\|^2+\mu_1\sum_{i=1}^{n}\left\|F_i-Y_i\right\|^2+\mu_2\sum_{(x_i,x_j,t_{ij})\in S}(K_{ij}-t_{ij})^2,\end{aligned} \tag{8}$$

where $S=\{(x_i,x_j,t_{ij})\}$ is the set of pairwise constraints, and t_{ij} is a binary variable that takes 1 or 0 to denote x_i and x_j belonging to the same class or not, and

$$\tilde{K}_{ij}=\begin{cases}K_{ij}, & K_{ij}\geq 0;\\ 0, & \text{otherwise.}\end{cases} \tag{9}$$

The time complexity of learning the entire kernel matrix in [7, 21] could be as high as $O(n^{6.5})$. An effective heuristic of speedup is to perform low-rank kernel matrix approximation and matrix factorization [19, 22, 24, 31, 32]. Moreover, the number of the given pairwise constraints is far less than the one which is sufficient to complete the low-rank kernel matrix with high probability. Thus, we incorporate the Laplacian spectral regularization into the above model (8). In other words, we set $K=QUQ^T$, where $Q=(q_1,\ldots,q_n)^T\in\mathbb{R}^{n\times m}$ consists of the m smoothest eigenvectors of the graph Laplacian L, and U is of size $m\times m$. The completion problem of the desired kernel K is then converted into the learning problem of a much smaller matrix U ($m\ll n$), subject to the constraint that $U\geq 0$.

Considering that the desired kernel matrix K is low-rank, and $\|K\|_*=\|QUQ^T\|_*=\|U\|_*$, we present the following nuclear norm regularized problem,

$$\begin{aligned}\mathcal{Q}(F,U)=&\frac{1}{2}\sum_{i,j=1}^{n}\tilde{K}_{ij}\left\|\frac{F_i}{\sqrt{D_{ii}}}-\frac{F_j}{\sqrt{D_{jj}}}\right\|^2+\mu_1\sum_{i=1}^{n}\left\|F_i-Y_i\right\|^2+\\ &\mu_2\left(\mu\|U\|_*+\frac{1}{2}\|M\odot(QUQ^T-Z)\|_F^2\right),\end{aligned} \tag{10}$$

where μ is a positive regularization parameter, Ω is the set of indices of known entries of Z, which is defined as

$$Z_{ij}=\begin{cases}1, & i=j;\\ 1, & (i,j)\in\mathrm{ML};\\ 0, & (i,j)\in\mathrm{CL}.\end{cases}$$

It is generally difficult that the optimization problem (10) is minimized with respect to both variables simultaneously. Thus, we employ an alternating optimization idea to solve the above problem. In the next section, we present a two-stage optimization strategy and provide an efficient *modified fixed point continuous* algorithm to learn the enhanced matrix U.

4. THE ALGORITHM

In this section, we present an effective two-stage optimization strategy for the SSL-MKI model (10). Furthermore, we should first choose to respect the *pairwise constraints assumption* and then the *manifold assumption*, considering that the given pairwise constraints are from reliable knowledge. Then we design an efficient *modified fixed point continuous* algorithm to learn the enhanced matrix U, and present a semi-supervised classification (SSC) algorithm, which aims to solve transduction classification tasks.

The objective function (10) can be approximated by a two-stage optimization strategy as follows:

$$\begin{cases}\min\limits_{U}\ \mu\|U\|_*+\frac{1}{2}\left\|M\odot(QUQ^T-Z)\right\|_F^2,\\ \min\limits_{F}\ \frac{1}{2}\sum_{i,j=1}^{n}\tilde{K}_{ij}\left\|\frac{F_i}{\sqrt{D_{ii}}}-\frac{F_j}{\sqrt{D_{jj}}}\right\|^2+\mu_1\sum_{i=1}^{n}\left\|F_i-Y_i\right\|^2.\end{cases} \tag{11}$$

In other words, the problem (10) can be efficiently solved using the following two stages: the first stage involving only one variable U is to compute the desired enhanced matrix U with the given pairwise constraints; and the second stage is to achieve the classification assignment based on the learned similarity matrix $\tilde{K}$ from the first stage.

4.1 Modified Fixed Point Algorithm

In the first stage, the optimization problem is formulated as follows:

$$\min_{U} \mu\left\|U\right\|_{*} + \frac{1}{2}\left\|M \odot (QUQ^{T} - Z)\right\|_{F}^{2}, \qquad (12)$$
$$\text{s.t., } U \geq 0.$$

While the nuclear norm minimization problem (12) can be converted into a semidefinite programming (SDP) problem, the time complexity of each iteration of standard SDP solvers based on the interior-point method could be at least as $O(m^6)$ [33]. To overcome this issue, many first-order algorithms have been developed to solve those problems, such as FPCA [28], which is a fixed point continuation algorithm. Furthermore, the FPCA method provably converges to the globally optimal solution and has been shown to outperform SDP solvers in terms of matrix recoverability. More recently, Ni et al. [34] proposed an *Augmented Lagrange Multiplier* (ALM) method for solving the low-rank representation problem with a PSD constraint. Our model (12) is also a nuclear norm minimization problem with a PSD constraint. In this part, we propose a *Modified Fixed Point Continuation* (MFPC) algorithm with an eigenvalue thresholding (EVT) operator to learn the enhanced matrix U, also called MFPC. The proposed MFPC algorithm can reduce the number of the auxiliary variables used in the ALM method [29] to accelerate its convergence. In the following subsections, we describe the proposed MFPC algorithm, and discuss the stopping criteria for iterations to acquire the optimal solution.

Inspired by the fixed point continuation algorithm proposed by Ma et al. [28], which has been used to multi-label transductive learning [35], we develop a modified fixed point iterative algorithm with an EVT operator to solve the proposed nuclear norm minimization problem (12).

Let $g(U) := \mu \| U \|_{*} + \frac{1}{2} \| M \odot (QUQ^{T} - Z) \|_{F}^{2}$, the derivative of the function $g(\cdot)$ with respect to U is given by

$$\partial g = \mu \partial \| U \|_{*} + H,$$

where $\partial \| U \|_{*}$ is the set of the subgradients of the nuclear norm, and $H = h(U) := Q^{T}\left(M \odot (QUQ^{T} - Z)\right)Q.$

Following [36], an explicit expression of the subdifferential of the nuclear norm at a symmetric matrix is given by the following lemma.

Lemma 1. Let $U \in \mathbb{R}^{m \times m}$ be a real symmetric matrix, then

$$\partial \| U \|_{*} = \{V^{(1)}[V^{(1)}]^{T} - V^{(2)}[V^{(2)}]^{T} + S : [V^{(1)}, V^{(2)}]^{T} S = 0,\ \| S \|_{2} \leq 1\},$$

where $V^{(1)}$ and $V^{(2)}$ are orthogonal eigenvectors associated with the positive and negative eigenvalues of U respectively, and $\| \cdot \|_{2}$ denotes the spectral norm of a matrix.

In addition, the following optimality condition in [37] can be adopted for the proposed nuclear norm minimization problem (12).

Theorem 1. Let $g(\cdot)$ be a convex function. Then U^{*} is an optimal solution to the problem (12), if and only if $U^{*} \geq 0$, and there exists a matrix $E \in \partial g(U^{*})$ such that

$$\left\langle E, F - U^{*} \right\rangle \geq 0, \quad \text{for all } F \geq 0.$$

Based on the above theorem, we can develop a modified fixed point iterative scheme for solving the problem (12) by adopting the operator splitting technique.

The operator $T(\cdot)$ is defined as

$$T(\cdot) := \tau\mu\partial \| \cdot \|_{*} + \tau h(\cdot),$$

where $\tau > 0$. And $T(\cdot)$ can be split into two parts:

$$T(\cdot) = T_{1}(\cdot) - T_{2}(\cdot),$$

where $T_{1}(\cdot) = \tau\mu\partial \| \cdot \|_{*} + I(\cdot)$, $T_{2}(\cdot) = I(\cdot) - \tau h(\cdot)$, and $I(\cdot)$ is an identity operator.

Let $Y = T_{2}(U)$, then $T(U) = \tau\mu\partial \| U \|_{*} + U - Y$, and $U \geq 0$. For tackling the proposed model (12), we need to solve the following nuclear norm minimization problem,

$$\min_{U \geq 0} \tau\mu \| U \|_{*} + \frac{1}{2} \| U - Y \|_{F}^{2}. \qquad (13)$$

The convex optimization problem (13) has a closed-form optimal solution [34], and the optimal solution is given by the eigenvalue thresholding (EVT) operator which will be defined later:

$$U^{*} = \mathrm{EVT}_{\tau\mu}\left(Y\right).$$

Thus, our modified fixed point scheme for solving the problem (12) can be expressed by the following two-step iteration as follows:

$$\begin{cases} Y^{k} = U^{k} - \tau h(U^{k}), \\ U^{k+1} = \mathrm{EVT}_{\tau\mu}(Y^{k}). \end{cases} \qquad (14)$$

Definition 1 (Eigenvalue thresholding (EVT) operator) Assume $U = U^{T} \geq 0$, and its eigenvalue decomposition is given by $U = V\mathrm{diag}(\lambda)V^{T}$, where $V \in \mathbb{R}^{m \times r}$, and $\lambda \in \mathbb{R}_{+}^{r}$. Given $\nu > 0$, $\mathrm{EVT}_{\nu}(\cdot)$ is defined as:

$$\mathrm{EVT}_{\nu}(U) := V\mathrm{diag}(\max\{\lambda - \nu, 0\})V^{T}, \qquad (15)$$

where $\max\{\cdot,\cdot\}$ should be understood element-wise.

Theorem 2. Suppose a symmetric matrix $U^{*} \geq 0$ satisfies:

1. $\| M \odot (QU^{*}Q^{T} - Z) \|_{F}^{2} < \mu / m$ for a small positive constant μ.

2. $U^{*} = \mathrm{EVT}_{\tau\mu}(U^{*} + \tau h(U^{*}))$. (16)

Then U^{*} is the unique optimal solution of the problem (12).

Proof. Please refer to [28, 37]. □

4.2 Implementation

We develop a modified fixed point iterative scheme to learn the enhanced matrix U with a PSD constraint. As suggested in [28, 37], the continuation technique can accelerate the convergence of the fixed point iterative method, and the parameter β determines the rate of reduction of consecutive μ_{k},

$$\mu_{k+1} = \max\{\mu_{k}\beta, \bar{\mu}\}. \qquad (17)$$

where $\bar{\mu}$ is a moderately small constant. Thus, the continuation strategy is also adopted by our modified fixed point algorithm, which solves a sequence of the problem (12), easy to difficult, corresponding to a sequence of large to small values of μ_k.

In the implementation of [28], the parameter τ is always set to 1, in contrast, it is set to $\tau \in (0, 2/\|\Psi\|_2)$ for the proposed fixed point continuation algorithm so that our algorithm's convergence is guaranteed, where $\Psi := (Q^T \otimes Q^T) I_\Omega (Q \otimes Q)$, $\otimes$ denotes the Kronecker product of two matrices, and $I_\Omega \in \mathbb{R}^{n^2 \times n^2}$ is a diagonal matrix which entries associated with Ω are set to 1, and 0 otherwise. There are many ways to select the parameter τ to accelerate the convergence of gradient algorithms for compressing sensing tasks. We now specify a strategy, which is based on the *Barzilai-Borwein* (BB) method [38] for choosing the parameter τ_k. The shrinkage iteration (14) first takes a gradient descent step with the step size τ_k along the negative gradient direction h^k of the smooth function $\| M \odot (QUQ^T - Z) \|_F^2 / 2$, and then applies the EVT operator $\mathrm{EVT}_v(\cdot)$ to accommodate the non-smooth term $\|U\|_*$. Therefore, it is natural to choose the parameter τ_k based on the function $\| M \odot (QUQ^T - Z) \|_F^2 / 2$ alone.

Let $H^k = Q^T \left(M \odot (QUQ^T - Z) \right) Q$, $\Delta U = U^k - U^{k-1}$, and $\Delta h = H^k - H^{k-1}$, then the BB step is defined by

$$\tau_k = \frac{\langle \Delta U, \Delta h \rangle}{\langle \Delta h, \Delta h \rangle}, \text{ or } \tau_k = \frac{\langle \Delta U, \Delta U \rangle}{\langle \Delta U, \Delta h \rangle}.$$

To avoid the BB step size τ_k being either too small or too large, we take

$$\tau_k = \max\left\{ \tau_{\min}, \min\left\{ \tau_k, \tau_{\max} \right\} \right\}, \tag{18}$$

where $0 < \tau_{\min} < \tau_{\max} < \infty$ are two constants.

Because our ultimate goal is to learn the enhanced matrix U, the accurate solution of the problem (12) is not required. Therefore, we use the following criterion as a stopping rule,

$$\frac{\|U^{k+1} - U^k\|_F}{\max\{1, \|U^k\|_F\}} < tol, \tag{19}$$

where tol is a small positive number. Experiments shows that $tol = 10^{-4}$ is good enough for obtaining the optimal matrix U^*.

Based on the previous analysis, we develop a *Modified Fixed Point Continuation* (MFPC) algorithm to learn the enhanced matrix U, as listed in **Algorithm 1**.

Theorem 3. The sequence $\{U^k\}$ generated by our modified fixed point iterations with $\tau \in (0, 2/\|\Psi\|_2)$ converges to some $\bar{U} \in \Gamma$, where Γ is the set of optimal solutions of the problem (12).

Proof. Please refer to [28, 37]. □

We now claim that our modified fixed-point continuation algorithm converges to an optimal solution of the problem (12).

```
Algorithm 1: MFPC algorithm
Input: A data set of $n$ instances $X = \{x_1, x_2, \cdots, x_l, x_{l+1}, \cdots, x_n\}$,
$X_L = \{x_i\}_{i=1}^{l}$ are labeled, and $X_U = \{x_i\}_{i=l+1}^{n}$ are unlabeled.
$\mathrm{ML} = \{(x_i, x_j)\}$ is the set of must-link constraints, and
$\mathrm{CL} = \{(x_i, x_j)\}$ is the set of cannot-link constraints. The
number of nearest neighbors $T$ and the constant $m$.
Output: The enhanced matrix $U$.
Initialize: Given $M$, $U^0$, $\bar{\mu}$, $\beta$, and $tol$.
And select $\mu_1 > \mu_2 > \cdots > \mu_L = \bar{\mu} > 0$.
1. Construct the T-NN graph and compute the normalized
   graph Laplacian $L = I - D^{-1/2} W D^{-1/2}$.
2. Compute the $m$ eigenvectors $\phi_1, \ldots, \phi_m$ of $L$ associated
   with the first $m$ smallest eigenvalues, and form the spectral
   representation matrix $Q = [\phi_1, \ldots, \phi_m] \in \mathbb{R}^{n \times m}$.

for $\mu_k = \mu_1, \mu_2, \cdots, \mu_L$, do
  while not converged do
    1. Choose the BB step size $\tau_k$ by Eq. (18).
    2. Update $Y^k$
         $H^k = Q^T \left( M \odot (QUQ^T - Z) \right) Q$, $Y^k = U^k - \tau_k H^k$.
    3. Update $U^{k+1}$
         $U^{k+1} = \mathrm{EVT}_{\tau_k \mu_k}\left( Y^k \right)$.
    stop condition: $\frac{\|U^{k+1} - U^k\|_F}{\max\{1, \|U^k\|_F\}} < tol$.
  end while
end for
```

4.3 Label Propagation

The enhanced spectral kernel K has been constructed using the above proposed MFPC algorithm, and we would have to take advantage of it to predict the labels of the unlabeled instances. We also present a semi-supervised learning method with *enhanced spectral kernel*, also called ESK, as shown in **Algorithm 2**. Here, our iteration equation can be written as follows:

$$F^{t+1} = \alpha P F^t + (1 - \alpha) Y, \tag{20}$$

where $P = D^{-1/2} \tilde{K} D^{-1/2}$ [1]. We will use the equation (20) to update the labels of each data point until convergence.

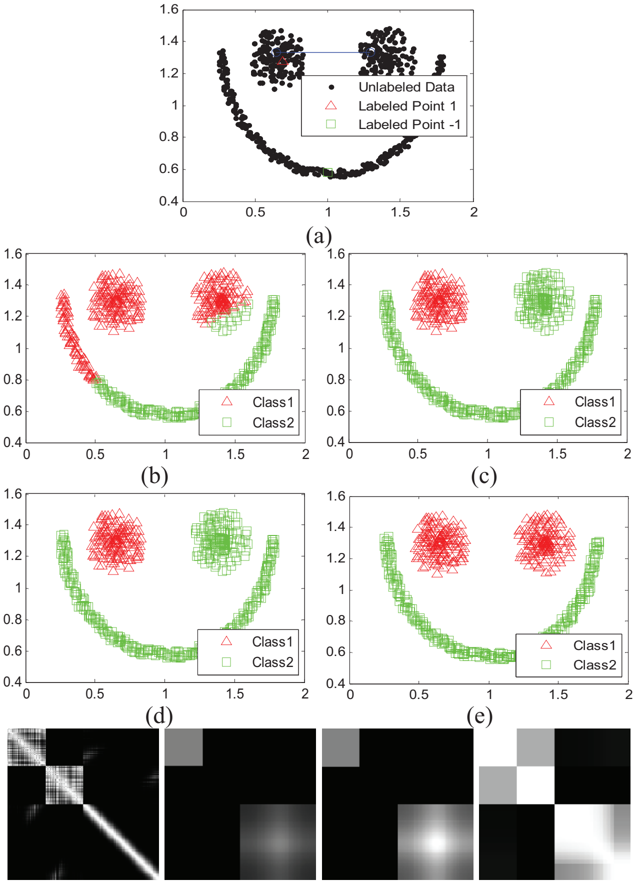
We give a toy example to illustrate how our ESK algorithm works, as shown in Figure 1. At first glance, the toy data consists of three separate groups, and is composed of a mixture of Gaussian-like and curve-like groups, as shown in Figure 1(a). Moreover, we also present the comparison between the similarity matrix in the input space and the kernel matrices learned by OSK, TSK, and MFPC, where the data are ordered such that all the instances in two Gaussian-like groups appear first; all the instances in the curve-like group appear second. It can be clearly observed that the enhanced kernel matrix learned by our MFPC algorithm exhibits two clear block structures so that the two classes are well-separated groups. In addition, we can draw a similar

[1] We set the enhanced similarity matrix $\tilde{K}$ by Eq. (9) for the proposed ESK algorithm. However, there is no need to change the enhanced spectral kernel as described above when it is used in traditional kernel machines such as SVMs.

conclusion as [23] that the kernel matrices learned by OSK and TSK have some uninformative eigenvectors even though which are optimally combined according to their own optimization criteria, and they fail to classify data points into the proper class.

```
Algorithm 2: ESK Algorithm
Input: The enhanced matrix $U$ and the constant $\alpha$.
Output: The assigned labels of all the data points.

1. Obtain the enhanced matrix $U$ by solving the problem (12)
   via the proposed MFPC algorithm.
2. Construct the enhanced spectral kernel matrix $K = QUQ^T$,
   and iterate Eq. (20) until convergence.
3. Let $F^*$ be the limit of the sequence $\{F^t\}$, and assign the
   labels of each data point $x_i$ by $y(x_i) = \mathrm{argmax}_{k \le c} F^*_{ik}$.
```



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

Figure 1: Classification results on the toy data set. (a) Toy data set with two labeled points and one ML constraint. (b)–(e) Classification results using LGC with $\sigma = 0.2$, OSK, TSK, and the proposed ESK algorithm with only one iteration. (f) Similarity matrix for the toy data set in the input space. (g)–(i) Learned kernel matrices by OSK, TSK, and MFPC with the $m = 5$ smoothest eigenvectors of graph Laplacians and a neighborhood size $T = 7$. The brighter a pixel, the greater similarity the pixel represents.

4.4 Valid Kernel

Theorem 4. If a normalized graph Laplacian $L \in \mathbb{R}^{n \times n}$ has the first m eigenvectors $\phi_1, \ldots, \phi_m$ corresponding to the m smallest engenvalues, and the enhanced matrix U obtained by solving the problem (12) for ESK is symmetric positive semidefinite. Then the family of matrices $K = QUQ^T$ is a valid kernel matrix.

Proof: Because $U = U^T \ge 0$, $U = U^{1/2}(U^{1/2})^T$, and $K = QUQ^T = QU^{1/2}(QU^{1/2})^T$, the enhanced spectral kernel matrix K is certainly positive semidefinite and thus a valid kernel matrix. □

Remark: Similar to the existing spectral kernel learning approaches such as OSK and TSK, the kernel matrix K learned by the proposed MFPC algorithm is also nonparametric spectral kernels from the graph Laplacian kernel L, and is referred to as the enhanced spectral kernel. Hence, the enhanced spectral kernel can be used in traditional kernel machines such as SVMs.

4.5 Complexity Analysis

The main running time of the proposed ESK algorithm is consumed by constructing the *k*-NN graph, computing the enhanced kernel matrix, and iterating the procedure (20). The time complexity of computing the enhanced matrix U by solving the problem (12) is $O(n^2 + t_1 m^2 n + t_1 m^3)$, where t_1 is the number of iterations. The total time complexity of ESK is $O(n^2 + t_1 m^2 n + t_1 m^3 + n^2 m + t_2 n^2 c)$, where t_2 is the number of iterations in the procedure (20). In the ESK algorithm, computing the m smoothest eigenvectors of the sparse matrix L can be efficiently performed using the Lanczos algorithm [39].

5. EXPERIMENTS

In this section, we present a set of experiments on many data sets, including a synthetic data set and many transductive settings.

5.1 Compared Algorithms

We compared the performance of the proposed ESK approach with the existing state-of-the-art SSL algorithms or related SSL methods, and the results averaged over 50 independent runs are reported.

- We use one-versus-rest SVMs[2] (SVM) [40] as the baseline. The width of the RBF kernel for SVM is set using 5-fold cross validation.

- GRF [3] and LGC [4]. The affinity matrix is constructed by a Gaussian function whose width is set by 5-fold cross validation.

- LapSVM[3] [6]: The base kernel is also selected to be Gaussian whose width is set by 5-fold cross validation, and all of the other hyperparameters are set by grid search as in [6].

- TSK [22] and OSK[4] [19]. For TSK, the decay factor γ is set to 2, and other parameters in TSK are set as in our ESK algorithm. For OSK, all parameters in OSK are set as in our algorithm.

- The proposed ESK algorithm. In all of the experiments, we set the constant $\alpha = 0.01$. And the number of nearest neighbors T is set by grid search as in [6].

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
[3] http://www.dii.unisi.it/~melacci/lapsvmp/.
[4] http://pages.cs.wisc.edu/~jerryzhu/publications.html.

5.2 Real-world Datasets

We use three categories of real-world data sets in our experiments, which are selected to cover a wide range of properties. Specifically, these data sets include:

- *UCI Data*[5]. We perform experiments on five UCI data sets, including Ionosphere, Sonar, Balance, Iris and Glass, and an artificial data set, G50c.
- *Image Data.* We perform experiments on six image data sets: MNIST [41], USPS[6], COIL20 [42], Caltech4[7], ORL[8] and YaleB3 [43].
- *Text Data*. We also perform experiments on two text data sets: 20-newsgroup[9] and WebKB[10].

The basic information of those data sets together with additional randomly chosen pairwise constraints is summarized in Table 1.

Table 1: Descriptions of the data sets.

Category	Data	Class	Feature	Labeled	Size	num_*M*	num_*C*
UCI	G50c	2	50	20	550	10	5
	Ionosphere	2	33	20	351	10	5
	Sonar	2	60	20	208	10	5
	Balance	3	4	20	625	15	15
	Iris	3	4	20	150	15	15
	Glass	6	9	30	214	12	30
Images	MNIST0123	4	784	8	4157	8	12
	USPS0123	4	256	8	3588	8	12
	COIL20	20	1024	40	1440	40	100
	Caltech4	4	4200	20	3479	8	12
	ORL	40	1024	80	400	80	200
	Yale3	3	1200	3	1755	6	6
Text	20-News	4	8014	40	3970	20	30
	Text1	2	7511	20	1946	20	10
	WK-CL	7	4134	70	827	14	72
	WK-TX	7	4029	70	814	14	72
	WK-WT	7	4165	70	1166	14	72
	WK-WC	7	4189	70	1210	14	72

Note that num_*M* and num_*C* denote the numbers of randomly chosen must-link constraints and cannot-link constraints, respectively.

5.3 Transduction Classification Results

The performances of the existing state-of-the-art SSL methods and the proposed ESK algorithm on these real-world data sets and G50c data set are shown in Tables 2, 3, and 4, in which the best performance for each data set is shown in bold. Here, we fairly compare the performance of the proposed ESK algorithm only using the given labeled data (denoted as ESK_L) with four existing state-of-the-art SSL approaches and SVMs. And we also provide the classification results of the proposed ESK algorithm both using the sparse labeled data and additional constraints (denoted as ESK_LC). By applying SVMs as the final classifier, we contrast the enhanced spectral kernel for ESK with two competitive spectral kernels such as OSK and TSK. From these tables, we can observe the following:

- LapSVM usually outperforms SVMs, GRF and LGC, especially on the image data sets, since there are clear nonlinear underlying manifolds behind those data sets, and LapSVM algorithm can make use of both the labeled data and the geometrical structure information contained in data.
- TSK, OSK, and the proposed ESK algorithm are often better than SVMs, GRF, LGC, and LapSVM since the flexible kernel from spectral transforms is more data-driven than the standard kernel, e.g., Gaussian kernel. And TSK has been shown to very effective for text data sets [22].
- The proposed ESK algorithm always performs at least as good as the best of the other algorithms. On the text data sets, ESK usually outperforms all other state-of-the-art SSL algorithms. Additional pairwise constraints have been shown to consistently improve the performance of the proposed ESK algorithm on all data sets since ESK can take advantage of both the given labeled data and pairwise constraints together.
- ESK+SVMs often significantly outperforms the other two learned spectral kernel machines including OSK+SVMs and TSK+SVMs.

In the second part of these experiments, we illustrate classification accuracies using the proposed ESK algorithm on the G50c, USPS0123 and 20-News data sets with the number of randomly labeled points varying from 2 to 20, from 4 to 40, and from 4 to 40, respectively, and against a number of randomly chosen pairwise constraints with only one labeled data point in each class, as shown in Figures 2 and 3. In the figures, the abscissa denotes the number of randomly labeled data or chosen pairwise constraints (we guarantee that there is at least one labeled point in each class), and the ordinate is the classification accuracy value averaged over 50 independent runs. For comparison, the classification results of four state-of-the-art SSL algorithms and SVMs are also plotted in the corresponding figure. It can be clearly observed that the proposed ESK algorithm is very stable, that is, even when we only label a very small fraction of the data, it can still get high classification accuracies and consistently outperforms the other five algorithms with the same amount of labeled data. Moreover, as the number of sparse constraints grows, the classification accuracy of the proposed ESK algorithm can be considerably improved and is better than that of graph Laplacian regularized kernel (LRK) [24]. This confirms that the proposed kernel learning model with nuclear norm regularization can avoid the over-fitting problems of LRK.

6. CONCLUSIONS

In this paper we have proposed a novel semi-supervised learning approach with *Mixed Knowledge Information* (SSL-MKI), which can handle both labeled data and additional pairwise constraints together with unlabeled data. We first constructed a unified SSL-MKI framework that can implement both the manifold assumption and the pairwise constraint assumption. Under the above framework, we also presented a *Modified Fixed Point Continuation* (MFPC) algorithm with an eigenvalue thresholding (EVT) operator to learn the enhanced kernel matrix. Then we developed a two-stage optimization strategy and provided an efficient ESK approach. Unlike the general SSL method, the proposed ESK approach can effectively make use of the given pairwise constraints that can often be obtained with little human effort. Finally, we provided a variety of experiments to show the effectiveness of our ESK approach, from which we found that the proposed ESK algorithm outperforms the state-of-the-art SSL methods.

[5] http://archive.ics.uci.edu/ml/.

[6] http://www.kernel-machines.org/data.html.

[7] http://www.robots.ox.ac.uk/~vgg/data3.html.

[8] http://www.uk.research.att.com/facedatabase.html.

[9] http://people.csail.mit.edu/jrennie/20Newsgroups/.

[10] http://www.cs.cmu.edu/~WebKB/.

Table 2: Classification accuracies (mean and standard deviation, %) on UCI datasets.

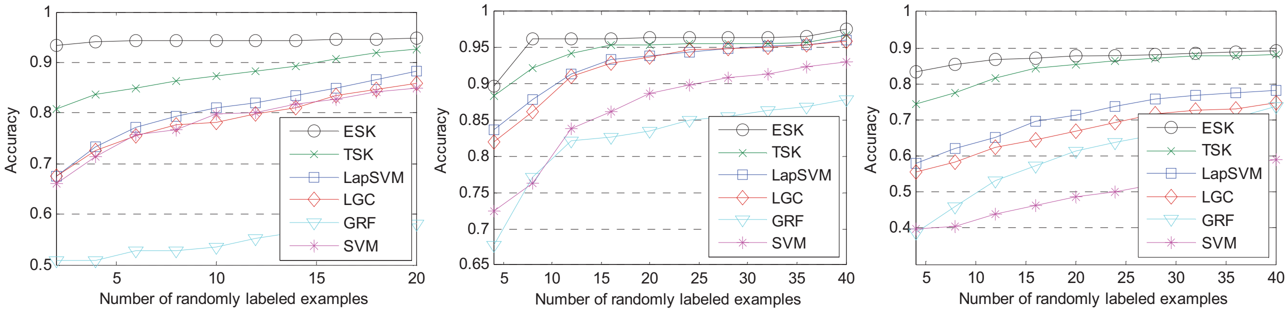
Data	G50c	Ionosphere	Sonar	Balance	Iris	Glass
SVM	85.36±2.46	74.54±6.79	66.24±4.83	71.39±6.14	94.35±2.42	57.38±3.64
GRF	57.36±9.36	78.54±6.84	60.82±6.35	67.46±6.93	93.81±2.46	57.81±5.32
LGC	86.78±2.44	83.10±4.39	62.18±6.03	70.03±8.19	93.11±2.40	56.25±3.92
LapSVM	86.65±3.22	82.95±1.84	**68.24±1.28**	63.86±7.43	95.42±1.83	60.11±4.98
TSK	92.69±2.13	76.10±7.19	64.31±5.02	68.41±4.19	93.89±3.88	57.80±4.02
ESK_L	**94.57±0.25**	**84.72±1.52**	66.86±1.43	**71.70±4.06**	**96.10±1.36**	**60.87±3.63**
ESK_LC	94.64±0.27	85.69±1.33	67.54±1.75	72.55±3.87	96.74±1.40	62.21±3.38
OSK+SVM	91.79±3.25	82.43±3.40	64.57±1.66	67.58±8.64	93.83±4.29	58.70±5.94
TSK+SVM	93.09±4.40	86.59±3.15	69.68±2.97	72.82±4.05	94.16±1.35	61.96±8.05
ESK+SVM	95.06±0.21	87.43±2.01	69.22±2.13	73.64±5.61	96.58±1.45	63.49±4.62

Table 3: Classification accuracies (mean and standard deviation, %) on images datasets.

Data	MNIST0123	USPS0123	COIL20	Caltech4	ORL	YALE3
SVM	74.06±4.20	84.35±4.27	74.96±2.11	59.89±7.40	76.82±2.71	91.00±7.22
GRF	68.94±6.03	77.09±7.54	82.36±2.76	65.84±4.40	76.92±2.77	95.18±6.92
LGC	80.13±3.32	90.14±4.14	80.38±2.10	88.00±5.69	76.40±2.39	94.68±5.47
LapSVM	76.36±5.02	87.88±5.75	86.58±1.53	**90.81±4.98**	77.34±2.60	**96.95±0.51**
TSK	93.49±0.73	95.25±1.79	83.19±1.29	87.81±6.98	76.13±3.02	93.35±3.92
ESK_L	**95.56±2.15**	**95.82±0.93**	**87.75±1.26**	90.74±3.87	**78.91±2.31**	96.87±1.13
ESK_LC	95.90±1.86	96.45±0.74	88.66±2.45	91.03±3.62	83.44±2.27	97.35±1.22
OSK+SVM	81.35±6.72	90.57±4.11	86.52±4.51	88.51±8.23	76.34±4.22	95.14±2.69
TSK+SVM	94.08±4.23	95.79±1.26	90.00±3.93	91.96±6.16	82.22±8.63	95.37±7.75
ESK+SVM	96.17±1.65	96.92±3.08	90.39±2.67	91.58±4.70	83.51±4.82	97.60±1.64

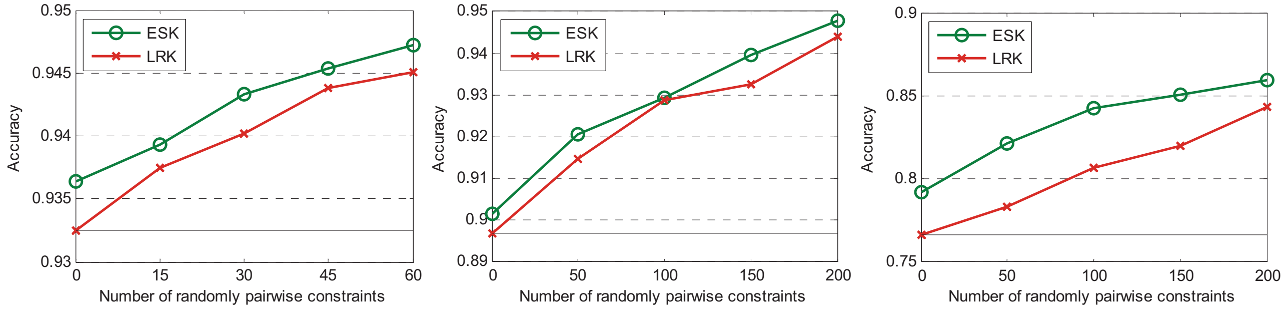
Table 4: Classification accuracies (mean and standard deviation, %) on text datasets.

Data	20-News	Text1	WK-CL	WK-TX	WK-WT	WK-WC
SVM	58.88±8.17	76.05±5.18	73.00±0.46	71.92±0.53	79.48±0.25	75.36±0.22
GRF	71.63±1.83	77.55±9.79	73.26±0.36	72.20±0.41	79.57±0.28	75.56±0.29
LGC	73.99±2.48	74.89±9.91	73.15±0.41	71.86±0.31	79.40±0.26	75.40±0.24
LapSVM	74.36±0.18	80.72±1.51	74.62±0.80	72.50±0.52	80.18±0.23	76.25±0.28
TSK	86.07±3.24	87.91±2.93	75.28±2.62	74.98±3.99	80.39±1.92	75.79±1.70
ESK_L	**89.16±0.74**	**89.65±2.63**	**79.53±3.18**	**78.60±4.21**	**82.93±1.57**	**80.49±1.63**
ESK_LC	89.32±0.67	89.77±2.25	79.74±2.85	79.16±3.42	83.46±1.60	81.35±1.49
OSK+SVM	86.19±4.55	82.74±4.82	76.63±2.32	75.50±3.36	81.48±0.76	77.86±1.68
TSK+SVM	89.30±2.90	85.76±3.63	77.25±1.67	76.77±2.82	82.04±2.63	79.93±0.45
ESK+SVM	91.57±0.76	89.93±1.69	81.76±2.96	79.61±3.69	83.78±1.34	82.26±1.55



(a) The G50c dataset (b) The USPS0123 dataset (c) The 20-News dataset

Figure 2: Classification results of different algorithms against a number of randomly labeled data points.



(a) The G50c dataset (b) The USPS0123 dataset (c) The 20-News dataset

Figure 3: Classification results of LRK and ESK algorithms against a number of randomly chosen pairwise constraints.

7. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China Nos. 61003198, 60970067; the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) No. B07048; the Program for Cheung Kong Scholars and Innovative Research Team in University No. IRT1170.

8. REFERENCES

[1] O. Chapelle, B. Schölkopf, A. Zien. Semi-Supervised Learning. The MIT Press, Cambridge, MA, 2006.

[2] X. Zhu. Semi-supervised learning literature survey. Tech. rep., Computer Sciences, University of Wisconsin-Madison, 2008.

[3] X. Zhu, Z. Ghahramani, J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.

[4] D. Zhou, O. Bousquet, T. Lal, J. Weston, B. Schölkopf. Learning with local and global consistency. In *NIPS*, pages 321–328, 2004.

[5] M. Belkin, P. Niyogi, V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, 2006.

[6] S. Melacci, M. Belkin. Laplacian support vector machines trained in the primal. *J. Mach. Learn. Res.*, 12:1149–1184, 2011.

[7] Z. Li, J. Liu, X. Tang. Pairwise constraint propagation by semidefinite programming for semi-supervised classification. In *ICML*, pages 576–583, 2008.

[8] R. Yan, J. Zhang, J. Yang, A. Hauptmann. A discriminative learning framework with pairwise constraints for video object classification. *IEEE Trans. Pattern Anal. Mach. Intell.,* 28(4):578–593, 2006.

[9] K. Wagstaff, C. Cardie. Clustering with instance-level constraints. In *ICML*, pages 1103–1110, 2000.

[10] D. Klein, S. Kamvar, C. Manning. From instance-level constraints to space-level constraints: making the most of prior knowledge in data clustering. In *ICML*, pages 307–314, 2002.

[11] S. Basu, M. Bilenko, R. Mooney. A probabilistic framework for semi-supervised clustering. In *KDD*, pages 59–68, 2004.

[12] B. Kulis, S. Basu, I. Dhillon, R. Mooney. Semi-supervised graph clustering: A kernel approach. In *ICML*, pages 457–464, 2005.

[13] E. Xing, A. Ng, M. Jordan, S. Russell. Distance metric learning, with application to clustering with side-information. In *NIPS*, pages 505–512, 2003.

[14] M. Bilenko, S. Basu, R. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *ICML*, pages 81–88, 2004.

[15] J. Davis, B. Kulis, P. Jain, S. Sra, I. S. Dhillon. Information theoretic metric learning. In *ICML*, pages 209–216, 2007.

[16] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209, 1999.

[17] O. Chapelle, A. Zien. Semi-supervised classification by low density separation. In *AISTATS*, pages 57–64, 2005.

[18] J. Zhuang, I. Tsang, S.C.H. Hoi. A family of simple non-parametric kernel learning algorithms. *J. Mach. Learn. Res.*, 12:1313–1347, 2011.

[19] X. Zhu, J. S. Kandola, Z. Ghahramani, J. D. Lafferty. Nonparametric transforms of graph kernels for semi-supervised learning. In *NIPS*, pages 1641–1648, 2005.

[20] S. Hoi, M. Lyu, E. Chang. Learning the unified kernel machines for classification. In *KDD*, pages 187–196, 2006.

[21] S. Hoi, R. Jin, M. Lyu. Learning nonparametric kernel matrices from pairwise constraints. In *ICML*, pages 361–368, 2007.

[22] W. Liu, B. Qian, J. Cui, J. Liu. Spectral kernel learning for semi-supervised classification. In *IJCAI*, pages 1150–1155, 2009.

[23] E. Hu, S. Chen, D. Zhang, X. Yin. Semisupervised kernel matrix learning by kernel propagation. *IEEE Trans. Neural Netw.*, 21(11):1831–1841, 2010.

[24] X. -M. Wu, A. So, Z. Li, S. Li. Fast graph Laplacian regularized kernel learning via semidefinite-quadratic-linear programming. In *NIPS*, pages 1964–1972, 2009.

[25] L. Zelnik-Manor, P. Perona. Self-tuning spectral clustering. In *NIPS*, pages 1601–1608, 2004.

[26] F. Wang, C. Zhang. Label propagation through linear neighborhoods. *IEEE Trans. Knowl. Data Eng.*, 20(1):55–67, 2008.

[27] J. Cai, E. J. Candès, Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.*, 20(4):1956–1982, 2010.

[28] S. Ma, D. Goldfarb, L. Chen. Fixed point and Bregman iterative methods for matrix rank minimization. *Math. Program.*, 128(1):321–353, 2011.

[29] Z. Lin, M. Chen, L. Wu. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. *Math. Program.*, submitted, 2009.

[30] E. J. Candès, B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[31] K. Q. Weinberger, F. Sha, Q. Zhu, L. K. Saul. Graph Laplacian regularization for large-scale semidefinite programming. In *NIPS*, pages 1489–1496, 2007.

[32] F. Shang, Y. Liu, F. Wang. Learning spectral embedding for semisupervised clustering. In *ICDM*, pages 597–606, 2011.

[33] Z. Liu, L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM J. Matrix Anal. Appl.*, 31(3):1235–1256, 2010.

[34] Y. Ni, J. Sun, X. Yuan, S. Yan, L. Cheong. Robust low-rank subspace segmentation with semidefinite guarantees. In *ICDM*, pages 1179–1188, 2010.

[35] A. B. Goldberg, X. Zhu, B. Recht, J. Xu, R. Nowak. Transduction with matrix completion: three birds with one stone. In *NIPS*, 2010.

[36] G. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl.*, 170:33–45, 1992.

[37] Y. Ma, L. Zhi. The minimum-rank gram matrix completion via modified fixed point continuation method. In *ISSAC*, pages 241–248, 2011.

[38] J. Barzilai, J. Borwein. Two-point step size gradient methods. *IMA J. Numer. Anal.*, 8:141–148, 1988.

[39] G. H. Golub, C. F. V. Loan. Matrix computations. Third edition, Johns Hopkins University Press, 1996.

[40] C. Chang, C. Lin. LIBSVM: A library for support vector machines, 2001. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[41] Y. LeCun, C. Cortes. The MNIST database of handwritten digits, 2009. Available: http://yann.lecun.com/exdb/mnist/.

[42] S.A. Nene, S.K. Nayar, J. Murase. Columbia object image library (COIL-20). Tech. rep., CUCS-005-96, Columbia Univ., 1996.

[43] 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):643–660, 2001.