Ensemble Manifold Regularization

Bo Geng, Dacheng Tao, *Senior Member*, *IEEE*, Chao Xu, Linjun Yang, and Xian-Sheng Hua, *Member*, *IEEE*

Abstract—We propose an automatic approximation of the intrinsic manifold for general semi-supervised learning (SSL) problems. Unfortunately, it is not trivial to define an optimization function to obtain optimal hyperparameters. Usually, cross validation is applied, but it does not necessarily scale up. Other problems derive from the suboptimality incurred by discrete grid search and the overfitting. Therefore, we develop an ensemble manifold regularization (EMR) framework to approximate the intrinsic manifold by combining several initial guesses. Algorithmically, we designed EMR carefully so it 1) learns both the composite manifold and the semi-supervised learner jointly, 2) is fully automatic for learning the intrinsic manifold hyperparameters implicitly, 3) is conditionally optimal for intrinsic manifold approximation under a mild and reasonable assumption, and 4) is scalable for a large number of candidate manifold hyperparameters, from both time and space perspectives. Furthermore, we prove the convergence property of EMR to the deterministic matrix at rate root-n. Extensive experiments over both synthetic and real data sets demonstrate the effectiveness of the proposed framework.

Index Terms—Manifold learning, semi-supervised learning, ensemble manifold regularization.

1 Introduction

In practical applications, e.g., handwritten digit recognition, video scene classification, and document categorization, the effort of labeling examples is generally laborious, though vast amounts of unlabeled samples are readily available and provide auxiliary information. *Semi-supervised learning* (SSL) under such scenarios is specifically designed to improve the generalization ability of the supervised learning by the leverage of unlabeled samples.

The common motivation of the SSL algorithms [5], [7], [12], [13], [14], [17], [21], [24], [26], [28], [29], [30], [31] is trying to exploit the intrinsic geometry of the probability distribution of unlabeled samples by restricting the inductive or transductive prediction to comply with this geometry. The *manifold regularization* framework [7], one of the most representative works, assumes that the geometry of the intrinsic data probability distribution is supported on the low-dimensional manifold. To approximate the manifold, the Laplacian of the adjacency graph is computed in an unsupervised manner from samples by using the Laplacian Eigenmap in the feature space [6]. The manifold approximation and the learning model are combined together under the conventional regularization framework [15], which smooths the model output along the manifold. The conventional regularization framework [15] shows that the solution of an ill-posed problem can be approximated by the variational principle, which contains both samples and the prior smoothness information. The *manifold regularization* utilizes the manifold to replace the smoothness assumption in [15], where the manifold is determined by the graph Laplacian with the predefined hyperparameters.

However, in general there are no explicit rules to choose graph hyperparameters for intrinsic manifold estimation because it is nontrivial to define an objective function to obtain these hyperparameters. Usually the cross validation [20] is utilized for parameter selection. However, this grid-search technique tries to select parameters from discrete states in the parameter space, and lacks the ability to approximate the optimal solution. Furthermore, it does not scale up well for a huge number of possible parameters. Moreover, performance measurements of the learned model, e.g., the classification accuracy, are weakly relevant to the difference between the approximated and intrinsic manifolds. Finally, the pure cross validation-based parameter selection inevitably drives the model to overfit the training and the validation set, and thus the learner cannot generalize well on the test set. As a consequence, an automatic and data-driven manifold approximation will be valuable for the *manifold regularization*-based SSL.

In this paper, to tackle the aforementioned problems, we propose an *ensemble manifold regularization* (EMR) framework, which combines the automatic intrinsic manifold approximation and SSL. By providing a series of initial guesses of the graph Laplacian, the framework learns to combine them to approximate the intrinsic manifold in a conditionally optimal way. Meanwhile, the semi-supervised model is learned and restricted to be smooth along the estimated manifold. We designed the EMR framework carefully so it

1. learns both a composite manifold and a semi-supervised learner jointly, leading to a unified framework;
2. is fully automatic for learning hyperparameters of the intrinsic manifold implicitly and avoids problems caused by the pure cross validation;
3. is conditionally optimal for the intrinsic manifold approximation under a mild and reasonable assumption, i.e., the optimal manifold lies in the convex hull of the initially guessed manifolds; and
4. is scalable for a large number of candidate manifold hyperparameters, from both time and space perspectives, because cross validation is not required for the hyperparameter selection and the induced graph structure is sparse.

2 Related Works

In recent years, two groups of SSL algorithms have shown themselves superior to classical supervised inductive learners.

The first group of methods assume that data in the same cluster share similar labels. Based on the *cluster assumption*, Nigam et al. [21] applied the EM algorithm on a mixture of multinomials for text classification, and showed that the accuracy of learned text classifiers can be improved by augmenting a small number of labeled training documents with a large number of unlabeled documents. Vapnik [28] proposed the *transductive support vector machine* (TSVM), which maximizes the margin in the presence of the unlabeled data and learns a decision boundary that traverses through low data-density regions. Joachims [17] implemented semi-supervised SVM based on a local combinational search strategy, after which various techniques had been applied to solve the nonconvex optimization problem associated with semi-supervised SVM [8], [24]. A comprehensive review can be found in [13]. Self-Training [29] and Cotraining [10] iteratively label some unlabeled examples according to the predictions of the current classifier, and retrain a new classifier with the additional labeled examples.

The second category of SSL algorithms are graph-based. They define a similarity graph over labeled and unlabeled examples and

- *B. Geng and C. Xu are with the Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China. E-mail: bogeng@pku.edu.cn, xuchao@cis.pku.edu.cn.*
- *D. Tao is with the Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney, Broadway, NSW 2007, Australia. E-mail: dacheng.tao@uts.edu.au.*
- *L. Yang is with Microsoft Research Asia, Beijing 100190, China. E-mail: linjuny@microsoft.com.*
- *X.-S. Hua is with Microsoft, One Microsoft Way, Redmond, WA 98052. E-mail: xshua@microsoft.com.*

Manuscript received 6 June 2010; revised 26 Apr. 2011; accepted 31 Jan. 2012; published online 21 Feb. 2012.
Recommended for acceptance by D.D. Lee.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-06-0423.
Digital Object Identifier no. 10.1109/TPAMI.2012.57.

constrain the prediction to be smooth over the graph. Zhu et al. [31] adopted the Gaussian fields and characterized the mean of the field in terms of harmonic functions. Zhou et al. [30] proposed to iteratively spread the label information of each example to its neighbors, so the classifier boundary is sufficiently smooth w.r.t. the intrinsic structure. Belkin and Niyogi [5] built a classifier over the samples represented by the eigenfunctions revealed from both the labeled and unlabeled data. Chapelle et al. [14] and Smola and Kondor [26] found that the spectral transformation of a Laplacian results in kernels suitable for semi-supervised learning. Belkin et al. [7] modeled the manifold structure under the regularization framework [15], which regularizes the solution of the SSL problem.

3 Ensemble Manifold Regularization

Consider the *semi-supervised learning* setting, where two sets of samples $x \in \mathbb{R}^d$ are available, i.e., l labeled samples, $L = \{(x_i, y_i)\}_{i=1}^{l}$, and u unlabeled samples, $U = \{x_i\}_{i=l+1}^{u+l}$, with $y_i \in \mathbb{R}$ as the label of x_i, for a total of $n = l + u$ samples. Suppose, labeled samples are $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ pairs drawn from a probability distribution P, and unlabeled samples are $x \in \mathbb{R}^d$ simply drawn according to the marginal distribution P_X of P.

To utilize P_X induced by unlabeled samples for SSL, the well-known *manifold regularization* framework is proposed. It assumes that the support of P_X is a compact manifold, and incorporates an additional regularization term to minimize the function complexity along the manifold [7]. The problem takes the following form:

$$\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(f, x_i, y_i) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2, \tag{1}$$

where $\mathcal{H}_K$ is the *reproducing kernel hilbert space* (RKHS), V is a general loss function, e.g., the least square error or the hinge loss, $\|f\|_K^2$ penalizes the classifier complexities measured in an appropriately chosen RKHS and is similar to that in SVM [28], and $\|f\|_I^2$ is the smooth penalty term to reflect the smoothness along the manifold supporting P_X. Parameters γ_A and γ_I balance between the loss function V and regularizations $\|f\|_K^2$ and $\|f\|_I^2$. The manifold regularization term $\|f\|_I^2$ plays a key role to SSL and models the classifier smoothness along the manifold estimated from the unlabeled samples.

It turns out that in an appropriate exponential coordinate system, which to the first order coincides with the local coordinate system given by a tangent plane in $\mathbb{R}^d$, $\|f\|_I^2$ is approximated by the graph Laplacian L and the function prediction $\mathbf{f} = [f(x_1), \ldots, f(x_n)]^T$, i.e., $\|f\|_I^2 = \frac{1}{n(n-1)} \mathbf{f}^T L \mathbf{f}$ [22]. In the above setting, the graph Laplacian is defined as $L = W - D$ or $L = D^{-\frac{1}{2}}(W - D)D^{-\frac{1}{2}}$ if normalized. The matrix $W \in \mathbb{R}^n \times \mathbb{R}^n$ is the data adjacency graph, wherein each element W_{ij} is an edge weight between two samples x_i and x_j. In the diagonal matrix $D \in \mathbb{R}^n \times \mathbb{R}^n$, the ith entry $D_{ii} = \sum_{j=1}^{n} W_{ij}$. Various extensions of the graph Laplacian have been proposed [16].

The construction of the graph Laplacian involves setting hyperparameters for creating the data adjacency graph, which is data dependent and generally performed by cross validation. Our framework is designed to automatically and effectively approximate the optimal graph Laplacian.

3.1 The General Framework

It is nontrivial to directly obtain the optimal graph Laplacian hyperparameters according to (1). Therefore, we propose an alternative approach by assuming that the intrinsic manifold lies in the convex hull of the pregiven manifold candidates. Because the optimal graph Laplacian is the discrete approximation to the manifold, the above assumption is equivalent to constraining the search space of possible graph Laplacians, i.e.,

$$L = \sum_{k=1}^{m} \mu_k L_k, \quad \text{s.t.} \sum_{k=1}^{m} \mu_k = 1, \ \mu_k \geq 0, \ \text{for } k = 1, \ldots, m, \tag{2}$$

where we define a set of candidate graph Laplacians $C = \{L_1, \ldots, L_m\}$ and denote the convex hull of set A as: $\mathbf{conv}A = \{\theta_1 x_1 + \cdots + \theta_m x_m | \theta_1 + \cdots + \theta_m = 1, \ x_i \in A, \theta_i \geq 0, i = 1, \ldots, m\}$. Therefore, we have $L \in \mathbf{conv}C$, which is also a graph Laplacian.

Under this constraint, the optimal graph Laplacian hyperparameter estimation is turned into the problem of learning the optimal linear combination of the pregiven candidates. Because each candidate L_i represents a certain manifold of the given samples, the EMR framework can be understood geometrically as follows: First, compute all possible approximated manifolds, each of which corresponds to a "guess" at the intrinsic data distribution, and then learn to linearly combine them for an optimal composite. To minimize the classifier complexity over the composite manifold, we introduce a new manifold regularization term, i.e., $\|f\|_I^2 = \frac{1}{n(n-1)} \mathbf{f}^T (\sum_{k=1}^{m} \mu_k L_k) \mathbf{f} = \sum_{k=1}^{m} \mu_k \|f\|_{I(k)}^2$. Then, we obtain the EMR framework:

$$\begin{aligned} \min_{f \in \mathcal{H}_K, \mu \in \mathbb{R}^m} & \frac{1}{l} \sum_{i=1}^{l} V(f, x_i, y_i) + \gamma_A \|f\|_K^2 + \gamma_I \sum_{k=1}^{m} \mu_k \|f\|_{I(k)}^2 + \gamma_R \|\mu\|^2 \\ \text{s.t.} & \sum_{k=1}^{m} \mu_k = 1, \ \mu_k \geq 0, \ k = 1, \ldots, m, \end{aligned} \tag{3}$$

where the regularization term $\|\mu\|^2$ is introduced to avoid the parameter μ overfitting to one manifold and $\gamma_R \in \mathbb{R}^+$ is the tradeoff parameter for controlling the contribution of the regularization term $\|\mu\|^2$. Because (3) contains a weighted combination of multiple manifold regularization terms, we name the new regularization framework *ensemble manifold regularization*. It is worth emphasizing that EMR is different from the work in [4], which adopts a multiple kernel learning approach to combine the kernels obtained from the inverse of graph Laplacian instead of learning an optimal graph combination to inference the data manifold.

For a fixed μ, (3) degenerates to (1), with $L = \sum_{k=1}^{m} \mu_k L_k$ for $\|f\|_I^2$. On the other hand, for a fixed f, (3) is simplified to

$$\min_{\mu \in \mathbb{R}^m} \sum_{k=1}^{m} \mu_k s_k + \gamma_R \|\mu\|^2, \quad \text{s.t.} \sum_{k=1}^{m} \mu_k = 1, \ \mu_k \geq 0, \ k = 1, \ldots, m, \tag{4}$$

where $s_k = \frac{\gamma_I}{n(n-1)} \mathbf{f}^T L_k \mathbf{f}$. Under this condition, if $\gamma_R = 0$, the solution of (4) will be: $\mu_k = 1$ if $s_k = \min_{j=1,\ldots,m} s_j$ and $\mu_k = 0$ otherwise. This trivial case will assign all the weight to one manifold, which is undesirable for learning a composite manifold. If $\gamma_R \to +\infty$, the solution tends to give the same weight to all graph Laplacians.

3.2 Theoretical Analysis

This section presents some theoretical analysis of EMR, and all the proofs are presented in our supplementary materials, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.57. Because $L \in \mathbf{conv}C$ is a graph Laplacian, according to [7], the representor theorem follows for a fixed μ.

Theorem 1. *For an $L \in \mathbf{conv}C$, the minimization of (3) w.r.t. $f \in \mathcal{H}_K$ with a fixed μ exists and admits the following representation:*

$$f^\star(x) = \sum_{i=1}^{n} \alpha_i^\star K(x_i, x), \tag{5}$$

which is an expansion in terms of the labeled and unlabeled examples.

The representor theorem presents us with the existence and the general form of the solution of (3) under a fixed μ. However, EMR is

motivated to learn both the SSL classifier f and the linear combination coefficients μ. Fortunately, we can adopt the alternating optimization technique [9] to solve (3) in an iterative manner, i.e., first solving (3) w.r.t. f with μ fixed, resulting in the solution represented by (5); then optimizing (3) w.r.t. μ, with f taking the value solved in the last iteration round; and alternatively iterating the above two steps, until the decrement of the objective function is zero. It is necessary to prove the convergence of the above alternating iteration rounds, and the convergence theorem is given below.

Theorem 2. *For a convex loss function $V \in \mathcal{H}_K \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^+$, the alternating optimization of (3) iteratively between the parameters f and μ converges.*

Besides the convergence of the learning algorithm, generally it is also important to verify the consistency of the regularization, which provides a theoretical guarantee on whether the empirical estimation can be deemed as a good candidate of its expected counterpart as long as we have sufficient training data. In all, consistency proof, complementary to empirical studies, shows the confidence that we can trust the estimate.

Specially, for a linear model $f = w^T x$, EMR reduces to $\|f\|_I^2 = \frac{1}{n(n-1)} \sum_{k=1}^m \mu_k w^T X L_k X^T w$, where $R = \frac{1}{n(n-1)} \sum_{k=1}^m \mu_k X L_k X^T$ is a random variable that involves the data dependent term composed of various manifold estimation candidates. We can prove the convergence property of EMR in this condition. More precisely, for $X = \{x_i\}_{i=1}^n$, we prove the data set dependent variable R converges to the deterministic matrix $E(R)$ at root-n for a fixed μ, with $n \to \infty$. We denote the distribution of samples $X = \{x_i\}_{i=1}^n$ as $P(X)$, and its fourth order moment as $E(\|\mathrm{vec}(xx^T)\mathrm{vec}^T(xx^T)\|)$, with $\mathrm{vec}(A)$ representing the vectorization of a matrix A into a column vector. For general choice of kernel function $\varphi_k(\cdot)$, we also abbreviate the conditional expectation as $\varphi_k(x) = E(\phi_k(z-x)|x)$ and $\psi_k(x) = E(\phi_k(z-x)z|x)$, respectively, for samples x and z drawn from $P(X)$.

Theorem 3. *Under the conditions that*

1. *the examples $X = \{x_i\}_{i=1}^n$ are sampled independently,*
2. *the kernel function $\phi_k(x)$ satisfies $\phi_k(0) = 1$ and $|\phi_k(x)| \le 1$ for $\forall k$,*
3. *$E(\|(\mathrm{vec}(xx^T)\mathrm{vec}^T(xx^T))\|) < \infty$, and*
4. *μ is fixed,*

the sample-based estimation

$$R = \frac{1}{n(n-1)} \sum_{k=1}^m \mu_k X L_k X^T$$

converges almost surely to a deterministic matrix. $E(R) = \sum_{k=1}^m \mu_k (E(\varphi_k(x) xx^T) - E(x\psi_k(x)^T))$ at the rate $n^{-1/2}$, i.e., $R \xrightarrow{a.s.} E(R) + O(n^{-1/2})$.

3.3 Discussions

The proof of Theorem 2, demonstrates that for any convex loss function, (3) is convex not only w.r.t. f for fixed μ but also w.r.t. μ for fixed f. However, (3) is not always convex for (f, μ) jointly. We can briefly prove this as follows: Suppose, that the Hessian matrix of the objective function is $H = \left[\begin{smallmatrix} A & B \\ B^T & C \end{smallmatrix}\right]$, wherein $A = \frac{\partial^2 F(f,\mu)}{\partial f^2}$, $B = \frac{\partial^2 F(f,\mu)}{\partial f \partial \mu}$, and $C = \frac{\partial^2 F(f,\mu)}{\partial \mu^2}$. The convexity of (3) w.r.t. f under a fixed μ ensures that $A \succeq 0$. For a large enough γ_A, we have $A \succ 0$. For the matrix B, its ith column is $B_i = L_i f$. The matrix C takes the form $C = \gamma_R I$, wherein I is an $m \times m$-dimensional identity matrix. According to the property of Schur Complement [11], i.e., if $A \succ 0$, $H \succeq 0$ iff $C - B^T A^{-1} B \succeq 0$. This is not always true because B contains f, which could be arbitrary.

However, we can solve the problem based on two strategies: 1) set a large value for γ_R so that (3) is convex w.r.t. (f, μ); 2) initialize $\mu = \frac{1}{m}$. The later strategy initializes as the mean of the graph Laplacians in the candidate set, which usually leads to a satisfied solution. In this paper, we adopt the second strategy for all experiments and show its effectiveness, leaving the discussion about the effects of γ_R independently to Section 4.

The theoretical analyses shown above present the theoretical properties of EMR and ensure that the framework can be implemented into numerous algorithms for various machine learning applications, e.g., classification and regression.

4 EMR Algorithms

We consider the general machine learning applications, e.g., classification and regression, in this section, and show how to implement EMR framework for these applications.

4.1 EMR Support Vector Machine

In support vector machine (SVM) [28], the hinge loss is adopted, i.e., $V(f, x_i, y_i) = (1 - y_i f(x_i))_+ = \max(0, 1 - y_i f(x_i))$. The Laplacian Support Vector Machine (LapSVM) is formulated to optimize the following problem:

$$\min_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^l (1 - y_i f(x_i))_+ + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2.$$

By incorporating EMR, we can extend LapSVM to EMR-SVM as the following:

$$\min_{f \in \mathcal{H}_K, \mu \in \mathbb{R}^m} \frac{1}{l} \sum_{i=1}^l (1 - y_i f(x_i))_+ + \gamma_A \|f\|_K^2 + \gamma_I \sum_{k=1}^m \mu_k \|f\|_{I(k)}^2 + \gamma_R \|\mu\|^2.$$

For a fixed μ, we can resort to Theorem 1 and the solution is given by (5). Substituting (5) into the framework (3), we can obtain the following optimization problem:

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^n, \xi \in \mathbb{R}^l, \mu \in \mathbb{R}^m} & \frac{1}{l} \sum_{i=1}^l \xi_i + \gamma_A \alpha^T K \alpha \\ & + \frac{\gamma_I}{n(n-1)} \sum_{k=1}^m \mu_k \alpha^T K L_k K \alpha + \gamma_R \|\mu\|^2 \\ \text{s.t. } & y_i \left(\sum_{j=1}^n \alpha_j K(x_i, x_j) + b \right) \ge 1 - \xi_i, \ i = 1, \ldots, l, \\ & \xi_i \ge 0, \ i = 1, \ldots, l, \quad \sum_{k=1}^m \mu_k = 1, \ \mu_k \ge 0, \ k = 1, \ldots, m, \end{aligned} \tag{6}$$

where $K \in \mathbb{R}^n \times \mathbb{R}^n$ is the gram matrix and its entry is $K_{ij} = K(x_i, x_j)$.

To adopt the alternating optimization for obtaining the solution of (6), we need to get the solution of f (represented by α according to (5)) with a fixed μ, as well as the solution μ of with a fixed f.

For a fixed μ, we can introduce nonnegative Lagrange multipliers β_i and ς_i for the inequality constraints in (6), which leads to

$$\begin{aligned} L(\alpha, \xi, b, \beta, \varsigma) = & \frac{1}{2} \alpha^T \left(2\gamma_A K + \frac{2\gamma_A}{n(n-1)} K \left(\sum_{k=1}^m \mu_k L_k \right) K \right) \alpha \\ & + \frac{1}{l} \sum_{i=1}^l \xi_i - \sum_{i=1}^l \beta_i \left(y_i \left(\sum_{j=1}^n \alpha_j K(x_i, x_j) + b \right) - 1 + \xi_i \right) - \sum_{i=1}^l \varsigma_i \xi_i. \end{aligned} \tag{7}$$

By taking the partial derivative of (7) w.r.t. α, β, ς and letting them be zero, we can derive the solutions of each primal variable represented by dual variables, which can be further substituted back into (7) to obtain the dual form:

$$\beta^\star = \max_{\beta \in \mathbb{R}^l} \sum_{i=1}^{l} \beta_i - \frac{1}{2}\beta^T Q \beta, \quad \text{s.t. } \sum_{i=1}^{l} \beta_i y_i = 0,\ 0 \le \beta_i \le \frac{1}{l}, i = 1, \ldots, l, \tag{8}$$

where $Q = YJK(2\gamma_A I_n + \frac{2\gamma_I}{n(n-1)}\sum_{k=1}^{m}\mu_k L_k K)^{-1} J^T Y \in \mathbb{R}^l \times \mathbb{R}^l$, $Y = \mathrm{diag}(y_1, y_2, \ldots, y_l) \in \mathbb{R}^l \times \mathbb{R}^l$, $I_n \in \mathbb{R}^n \times \mathbb{R}^n$ denotes an n-dimensional identity matrix, and $J = [I_l, 0] \in \mathbb{R}^l \times \mathbb{R}^n$.

Finally, the classifier parameter α can be obtained as

$$\alpha^\star = \left(2\gamma_A I + \frac{2\gamma_I}{n(n-1)}\sum_{k=1}^{m}\mu_k L_k K\right)^{-1} J^T Y \beta^\star. \tag{9}$$

The learning procedure combines different graph Laplacians into Q, and the optimization of (8) is approximately independent of the number of graph Laplacians m. Therefore, with a fixed μ, we do not incorporate additional computational costs compared against solving LapSVM, except for some sparse matrix additions.

On the other hand, for learning μ with a fixed f, (6) degenerates to (5), and we can adopt the coordinate descent-based algorithm. In each iteration round, we select two elements in μ for updating while the others are fixed. Suppose at an iteration round, the ith and jth elements of μ are selected. Due to the constraint $\sum_{k=1}^{m}\mu_k = 1$, the summation of μ_i and μ_j will not change after this iteration round. Therefore, we have the solution

$$\begin{cases} \mu_i^\star = 0,\ \mu_j^\star = \mu_i + \mu_j, \text{ if } 2\gamma_R(\mu_i + \mu_j) + (s_j - s_i) \le 0 \\ \mu_i^\star = \mu_i + \mu_j,\ \mu_j^\star = 0, \text{ if } 2\gamma_R(\mu_i + \mu_j) + (s_i - s_j) \le 0 \\ \mu_i^\star = \frac{2\gamma_R(\mu_i+\mu_j)+(s_j-s_i)}{4\gamma_R}, \mu_j^\star = \mu_i + \mu_j - \mu_i^\star, \text{ else.} \end{cases} \tag{10}$$

We iteratively traverse over all pairs of elements in μ and adopt (10) to optimize the two elements until the objective function in (6) does not decrease. Intuitively, the update criteria in (10) tends to assign larger value to μ_k with smaller s_k. Because $s_k = \frac{\gamma_I}{n(n-1)}\mathbf{f}^T L_k \mathbf{f}$ measures the smoothness of the function f over the ith manifold approximated by the graph Laplacian L_k, the algorithm will prefer the pregiven manifold that coincides with current iteration round SSL classifier f better.

4.2 EMR Regularized Least Squares (EMR-RLSs)

Besides classification problems, regularized least squares (RLSs) are also widely adopted for various regression problems. In RLSs, the squared loss is adopted as the loss function, i.e., $V(f, x_i, y_i) = (y_i - f(x_i))^2$. The Laplacian regularized least squares (LapRLSs) is proposed to optimize the following problem:

$$\min_{f \in \mathcal{H}_K} \frac{1}{l}\sum_{i=1}^{l}(y_i - f(x_i))^2 + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2.$$

We can extend LapRLSs into the following EMR regularized least squares:

$$\min_{f \in \mathcal{H}_K, \mu \in \mathbb{R}^m} \frac{1}{l}\sum_{i=1}^{l}(y_i - f(x_i))^2 + \gamma_A \|f\|_K^2 + \gamma_I \sum_{k=1}^{m}\mu_k \|f\|_{I(k)}^2 + \gamma_R \|\mu\|^2. \tag{11}$$

According to Theorem 1, for a fixed μ, we can substitute the solution (5) $f^*(x) = \sum_{i=1}^{n}\alpha_i^* K(x, x_i)$ back into (11), resulting in the optimization problem as follows:

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{l}\sum_{i=1}^{l}(Y - JK\alpha)^T(Y - JK\alpha) + \gamma_A \alpha^T K \alpha + \frac{\gamma_I}{n(n-1)}\sum_{k=1}^{m}\mu_k \alpha^T K L_k K \alpha + \gamma_R \|\mu\|^2, \tag{12}$$

where $K \in \mathbb{R}^n \times \mathbb{R}^n$ is the gram matrix over labeled and unlabeled samples, with each entry defined as $K_{ij} = K(x_i, x_j)$, Y is an n-dimensional labeled vector given by $Y = [y_1, y_2, \ldots, y_l, 0, \ldots, 0] \in \mathbb{R}^n$, and $J = \mathrm{diag}(y_1, y_2, \ldots, y_l, 0, \ldots, 0) \in \mathbb{R}^n \times \mathbb{R}^n$ is a diagonal matrix with the first l diagonal element as one and the rest 0.

Taking the derivatives of (12) w.r.t. α and setting it to zero, we can obtain the solution

$$\alpha* = \left(JK + \gamma_A l I_n + \frac{\gamma_I l}{n(n-1)}\sum_{k=1}^{m}\mu_k L_k K\right)^{-1} Y. \tag{13}$$

For learning μ with a fixed f, we still need to solve problem (4), similar to that used in EMR-SVM, by iteratively adopting (10) to obtain the solution.

5 Experiments

In this section, experiments were conducted extensively over synthetic (*two moons*) [7], [31], USPS handwritten digits recognition [3], Biomedical [1], and USC scene classification [23] to demonstrate the effectiveness of EMR. The proposed EMR-based algorithms were compared with conventional SVM [28], Regularized Least Square, transductive SVM [17], LapSVM [7], and LapRLS [7]. More results of the Heart [3], Isolet spoken letter recognition [3], and Newsgroups Text Categorization [27] data sets, as well as the efficiency analysis, can be found in the online supplementary materials.

The RBF kernel $K(x_i, x_j) = \exp(-\gamma_K \|x_i - x_j\|^2)$ was applied for all experiments except the text categorization. The heat Kernel [22] was adopted to compute the edge weights of graph, i.e., if $x_i \in N(x_j)$ or $x_j \in N(x_i)$, $W_{ij} = \exp(-t\|x_i - x_j\|^2)$ and 0 otherwise.

There are three hyperparameters in the graph Laplacian, i.e., the heat Kernel parameter t, the number of nearest neighbors k, and the degree of the graph Laplacian p. For EMR, we created two graph Laplacian sets for different purposes. For the first set, we chose and fixed $t = \{\frac{\tau}{50}, \frac{\tau}{45}, \ldots, \frac{\tau}{5}, \tau, 5\tau, \ldots, 60\tau, 65\tau\}$, $k = 10$, and $p = 2$, which led to 24 graphs. This simplified version focuses on the variation of hyperparameter t, which is easy for the algorithm analysis. For another one, the candidate hyperparameters were chosen as $t = \{\frac{\tau}{15}, \frac{\tau}{10}, \frac{\tau}{5}, \tau, 5\tau, 10\tau, 15\tau, 20\tau\}$, $k \in \{5, 10, 15\}$, and $p \in \{1, 2, 3\}$, where we got 72 graphs. This comparatively larger graph Laplacian set varies all hyperparameters, and is intended to prove that EMR can automatically estimate all hyperparameters introduced by graph Laplacian. Here, without explicit mentions, τ was empirically set as the inverse of $\frac{1}{n^2}\sum_{i,j=1}^{n}\|x_i - x_j\|^2$.

The other parameters of all algorithms were determined by the twofold cross validation over the training set. For EMR-based algorithms, we adaptively set $\gamma_R = \frac{\gamma_I}{m}\sum_{k=1}^{m}\|f\|_{I(k)}^2$ in a data driven manner, and left the discussion about its insensitive property later on.

5.1 Two Moons

In this section, we utilized the original *two moons*, obtained from [7], to justify the significance of EMR for automatically approximating the intrinsic manifold. For the binary classification of data belonging to different moons, the data generator randomly drew 200 samples for both moons, each of which contains only one labeled sample (therefore, two training samples and 398 testing samples in all). The 24 graph set mentioned above was adopted for EMR-SVM. Here, to show the impact of Laplacian parameters to the performance we change the graph of LapSVM from its original binary graph edge weights to the Heat kernel edge weights. Except for t, all the other parameters were set according to the code downloaded from [2]. For EMR-SVM-24G, we vary the parameter τ correspondingly to see how the proposed method can select effective graphs for different

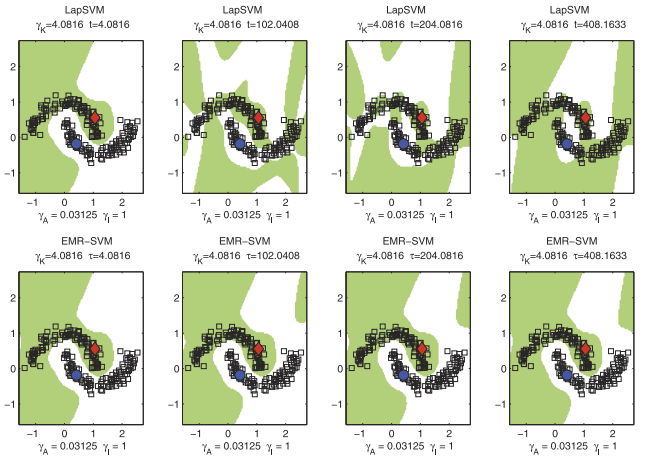

Fig. 1. Two moons data set: the classifiers for LapSVM and EMR-SVM-24G for different parameters of t and τ, respectively. Labeled samples are highlighted.

graph sets. For both t and τ, we gradually change their values from 4.0816 to $25 * 4.0816$, $50 * 4.0816$, and $100 * 4.0816$.

The classifier boundaries of LapSVM and EMR-SVM-24G were plotted in Fig. 1. It can be found that the performance of LapSVM is sensitive to the graph parameter t. Except for the initial value, the performance of all three of the other ones is undesirable. However, our EMR-SVM creates a nearly perfect classifier boundary by learning to combine different manifolds, and the performance is quite stable for various graph candidate sets. This suggests that the composite manifold learned by EMR can select the most effective graphs and combine them to synthesize a global optimal solution.

5.2 Handwritten Digits Recognition

The USPST set is the test data part of the USPS data set in the UCI repository [3]. It contains 2,007 samples with 10 classes, each of which corresponds to a handwritten digit. The data set is widely used to evaluate SSL algorithms [7], [25], where it is randomly divided into 10 splits, with 50 labeled and 1,957 unlabeled samples in each split. We first used the 24 graph set to conduct one-versus-rest multiclass classification experiments.

To study the effectiveness of EMR, we selected one split and present the manifold composite weight μ learned by EMR-SVM in each one-versus-rest classification task in Fig. 2, together with the classification error rate of each graph Laplacian using LapSVM. Obviously, the graph Laplacian selected by EMR-SVM are generally consistent with the corresponding effective graph Laplacians.

It is also important to investigate the sensitivity of the parameter γ_R for EMR because this investigation is helpful to deeply understand how and why the regularization term $\gamma_R\|\mu\|^2$ affects the whole framework. We compared how the multiclass

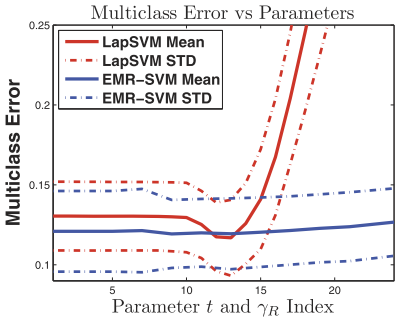

Fig. 3. Parameter sensitivity comparison of LapSVM t and EMR-SVM γ_R.

error rate varies over different γ_R for EMR-SVM and different t for LapSVM. In our setting, the parameter γ_R varied from $10^{-4}\gamma_I$ to γ_I with 12 parameters. We chose the 24 graph set, where t changed from 1.66×10^{-4} to 5.4×10^{-1} with 24 hyperparameters. Note that the domain of γ_R was comparatively larger than that of t. The experimental results averaged over 10 splits are shown in Fig. 3. We can find that the average multiclass error rate is not sensitive to the parameter γ_R in EMR-SVM, while the performance is much more sensitive to the hyperparameter t in LapSVM. According to the analysis in Section 2.1, the best composite graph corresponds to γ_R in the middle range value, i.e., a balance between unanimous weighting and a single manifold.

The one-versus-rest multiclass error rates are shown in Table 1. Besides the results of the 24 graph set we also showed the results using 72 graph set. From the results we can observe that SVM, RLS, and TSVM do not perform as well as manifold regularization-based methods. On the other hand, the performances of EMR-SVM and EMR-RLS are consistently better than LapSVM, LapRLS, and EMR-RLS with 72 graph Laplacians (EMR-RLS-72G) yields the best performance.

5.3 Biomedical Data Set

For biomedical data, such as lung cancer and Colon Tumor, very few labeled training data are available. Therefore, semi-supervised learning is very helpful for biomedical data mining applications. The data sets were all downloaded from the Kent Ridge Biomedical Data Set [1]. We chose seven of them for the experiments to testify to the effectiveness of the proposed EMR framework, namely, Leukemia-ALLAML WhiteHead MIT (Leukemia), Breast Cancer, Colon Tumor, Diffuse Large B-Cell Lymphoma of Harvard (DLBCL), Lung Cancer Michigan, Central Nervous System, Ovarian Cancer-NCI-PBSII-061902. The features of each data set are first normalized to zero mean and unit variance, and subsequently reduced to 100 dimensions by using Principle Component Analysis [18]. We randomly split the data sets by 20 percent for the training examples and 80 percent for the

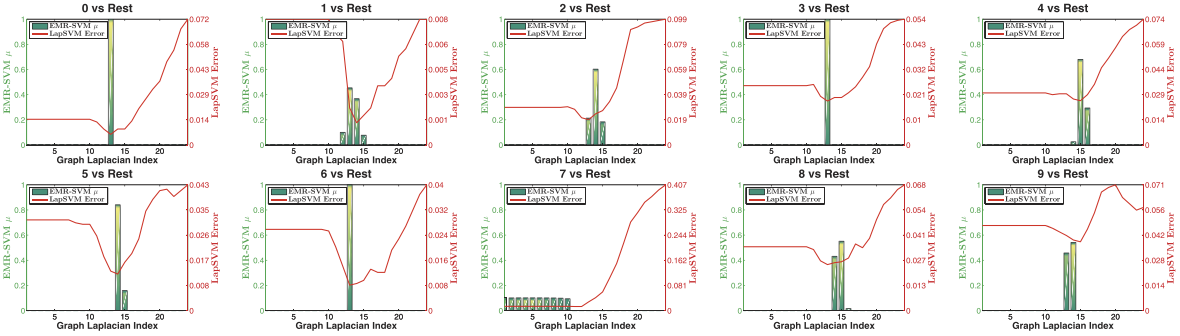

Fig. 2. The manifold combination coefficient μ learned by EMR-SVM over 10 one-versus-rest classification tasks, and the classification error rate of LapSVM using each graph Laplacian over each task.

TABLE 1
Performance Comparisons of Different Algorithms over the USPST Data Set

	SVM	RLS	TSVM	LapSVM	LapRLS	EMR-SVM		EMR-RLS	
						24G	72G	24G	72G
USPST	22.94	23.60	26.28	12.38	12.41	11.64	11.75	11.26	**11.13**

TABLE 2
Performance Comparisons of Different Algorithms over Biomedical Data Sets

	SVM	TSVM	LapSVM	LapRLS	EMR-SVM		EMR-RLS		p-value
					24G	72G	24G	72G	
Leukemia	17.59	17.93	14.17	13.53	13.10	12.76	12.84	**12.67**	2.17e-02
Breast Cancer	38.59	**35.19**	36.17	36.46	35.41	35.90	35.86	35.92	8.84e-01
Colon Tumor	32.70	30.90	26.09	26.17	25.11	24.44	**24.11**	**24.11**	5.65e-03
DLBCL	23.56	22.90	18.83	21.46	19.95	**17.74**	20.54	20.54	6.73e-03
Lung Cancer	0.84	0.96	**0**	**0**	**0**	**0**	**0**	0.19	-
Central Nervous System	50.83	42.92	43.71	44.38	42.81	**41.88**	44.73	44.52	2.73e-03
Ovarian Cancer	1.53	2.65	0.92	0.87	0.40	**0.30**	0.40	0.60	3.18e-04

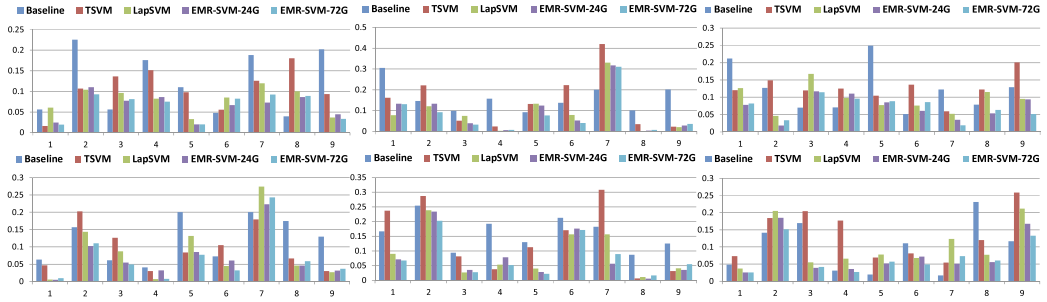

Fig. 4. False negative and false positive rates (subfigures from top to bottom) of different classification methods: baseline [23], TSVM, LapSVM, EMR-SVM-24G, and EMR-SVM-72G over each segment of ACB, AnF, and FDF sites (subfigures from left to right), respectively.

testing ones, with 20 different random splits. The reported results are the average error rates over 20 splits and are shown in Table 2.

We can derive that EMR-based methods show the best performance over all except for Breast Cancer, where TSVM is best performed. In addition, EMR 24G methods (EMR-SVM-24G and EMR-RLS-24G) are comparable to their corresponding single graph methods (LapSVM and LapRLS), respectively. In addition, the parameter spaces of EMR 24G methods are much smaller than those of single manifold ones. For EMR 72G methods, the performance improves a lot and takes the best performance in six of the seven data sets. This justifies that EMR-based algorithms can effectively select graphs for better classification performance. The p-value of the t-test for the best EMR result is better than the best baseline result, showing that the improvements for five out of seven data sets are significant.

5.4 Scene Classification

Experiments were also performed on the USC scene data set [23] for the scene classification task. The data set contains 375 video clips from three sites, namely the Ahamanson Center for Biology (ACB), Associate and Founders Park (AnF), and Frederick D. Fagg Park (FDF). Each site is divided into nine segments, each of which is taken as a scene in a site. The gist feature [23] was adopted for the experiments, which takes color, intensity, and orientation information into consideration for 544 dimensions. The results of neural network classifier in [23] is reported as the baseline. As a classification task, we reported the performances of EMR-SVM-24G and EMR-SVM-72G in comparing with TSVM and LapSVM.

The false negative and false positive rates of different classification methods over each segment are reported, respectively, in Fig. 4, where EMR-based methods consistently achieved better performance at each scene. Table 3 shows the multiclass error rates of different methods. As before, SSL methods dramatically outperform the baseline. Especially, EMR-SVM-based algorithms show the best results at all sites. This again demonstrates the effectiveness and applicability of EMR for video scene classification task. The t-test for the performance of EMR-based methods is better than the best baseline result show that all the improvements are significant under the alpha level of 0.05.

TABLE 3
The Multiclass Error Rates of Different Classification Algorithms on the USC Scene Data Set

	Baseline	TSVM	LapSVM	EMR-SVM		p-value
				24G	72G	
ACB	12.04	9.03	8.39	6.81	**6.20**	3.29e-04
AnF	15.79	10.85	7.26	**6.99**	7.06	3.31e-03
FDF	11.38	11.09	8.28	7.04	**6.63**	7.85e-04

6 Conclusion

Numerous applications can be handled well by the *manifold regularization*-based *semi-supervised learning* algorithms. However, they all fall short because the estimation of hyperparameters in the manifold regularization is not convenient. In this paper, we propose an *ensemble manifold regularization* framework to automatically and implicitly estimate the hyperparameters of the manifold regularization. By providing some initial guesses of manifolds, EMR learns to combine them for a conditionally optimal estimation of the intrinsic manifold. The alternating optimization technique is utilized to unify the learning of the semi-supervised classifier and the manifolds combination coefficients together.

Based on extensive experiments over the two moons test, handwritten digits recognition (USPS), spoken letter recognition (Isolet), heart, USC scene classification, and 20 Newsgroups text categorization data sets, we have the following observations:

- The graph selection over two moons and USPS shows that EMR is effective for approximating the intrinsic manifold by combining some initial candidates.
- The results over the spoken letter recognition data set prove that EMR generalizes well over not only the unlabeled data whose intrinsic distribution is observable, but also the test data whose intrinsic distribution is unknown.
- The varying of graph Laplacian sets demonstrates that EMR is automatically learning the hyperparameters of intrinsic manifold, and is scalable to a large number of graphs.
- The success of both EMR-SVM and EMR-RLS over various data sets is interpreted as EMR possesses the ability to be generalized to different learning problems.
- Extensive statistical results over these six data sets from various sources sufficiently demonstrate that EMR-based methods are superior to discovering the optimal manifold and improving the learner's generalization ability.
- Our EMR approach provides an alternative approach to approximate the parameters of manifold regularization. We will consider applying the work in [19] to optimize other hyperparameters of EMR in the future work.

Acknowledgments

This work is partially supported by the Australian ARC discovery project (ARC DP-120103730), Chinese NBRPC 2011CB302400, NSFC 60975014, 61121002, and NSB 4102024.

References

[1] http://datam.i2r.a-star.edu.sg/data sets/krbd/, 2012.
[2] http://manifold.cs.uchicago.edu/manifold_regularization/manifold.html, 2012.
[3] http://www.ics.uci.edu/mlearn/, 2012.
[4] A. Argyriou, M. Herbster, and M . Pontil, "Combining Graph Laplacians for Semi-Supervised Learning," *Proc. Advances in Neural Information Processing Systems 18*, pp. 67-74, 2005.
[5] M. Belkin and P. Niyogi, "Using Manifold Structure for Partially Labelled Classification," *Proc. Neural Information Processing System Conf.*, 2002.
[6] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Computation*, vol. 15, pp. 1373-1396, 2003.
[7] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *J. Machine Learning Research*, vol. 7, pp. 2399-2434, 2006.
[8] K.P. Bennett and A. Demiriz, "Semi-Supervised Support Vector Machines," *Advances in Neural Information Processing Systems*, vol. 12, pp. 368-374, 1998.
[9] J.C. Bezdek and R.J. Hathaway, "Convergence of Alternating Optimization," *Neural, Parallel and Scientific Computations*, vol. 11, pp. 351-368, 2003.
[10] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," *Proc. 11th Ann. Conf. Computational Learning Theory*, 1998.
[11] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ., 2004.
[12] *Semi-Supervised Learning*, O. Chapelle B. Schölkopf and A. Zien, eds. MIT Press, 2006.
[13] O. Chapelle, V. Sindhwani, and S.S. Keerthi, "Optimization Techniques for Semi-Supervised Support Vector Machines," *J. Machine Learning Research*, vol. 9, pp. 203-233, 2008.
[14] O. Chapelle, J. Weston, and B. Schölkopf, "Cluster Kernels for Semi-Supervised Learning," *Proc. Advances in Neural Information Processing Systems 15*, 2001.
[15] F. Girosi, M. Jones, and T. Poggio, "Regularization Theory and Neural Networks Architectures," *Neural Computation*, vol. 7, pp. 219-269, 1995.
[16] X. He and P. Niyogi, "Locality Preserving Projections," *Proc. Advances in Neural Information Processing Systems 18*, 2004.
[17] T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," *Proc. 16th Int'l Conf. Machine Learning*, 1999.
[18] I. Jolliffe, *Principal Component Analysis*. Springer 1986.
[19] S. Keerthi, V. Sindhwani, and O. Chapelle, "An Efficient Method for Gradient-Based Adaptation of Hyperparameters in SVM Models," *Proc. Advances in Neural Information Processing Systems 19*, 2007.
[20] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proc. 14th Int'l Joint Conf. Artificial Intelligence*, pp. 1137-1145, 1995.
[21] K. Nigam, A.K. McCallum, S. Thrun, and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents Using Em," *Machine Learning*, vol. 39, nos. 2/3, pp. 103-134, 2000.
[22] S. Rosenberg, *The Laplacian on a Riemmannian Manifolds*. Cambridge Univ., 1997.
[23] C. Siagian and L. Itti, "Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 300-312, Feb. 2007.
[24] V. Sindhwani, S.S. Keerthi, and O. Chapelle, "Deterministic Annealing for Semi-Supervised Kernel Machines," *Proc. 23rd Int'l Conf. Machine Learning*, pp. 841-848, 2006.
[25] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the Point Cloud: From Transductive to Semi-Supervised Learning," *Proc. 22nd Int'l Conf. Machine Learning*, 2005.
[26] A. Smola and R. Kondor, "Kernels and Regularization on Graphs," *Proc. Conf. Learning Theory and Kernel Machines*, 2003.
[27] S. Tong and D. Koller, "Support Vector Machine Active Learning with Applications to Text Classification," *J. Machine Learning Research*, vol. 2, pp. 999-1006, 2000.
[28] V.N. Vapnik, *Statistical Learning Theory*. Wiley, 1998.
[29] D. Yarowsky, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods," *Proc. 33rd Ann. Meeting Assoc. for Computational Linguistics*, pp. 189-196, 1995.
[30] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Scholkopf, "Learning with Local and Global Consistency," *Advances in Neural Information Processing Systems*, vol. 16, pp. 321-328, 2004.
[31] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-Supervised Learning Using Gaussian Fields and Hamonic Functions," *Proc. 20th Int'l Conf. Machine Learning*, 2003.