Ensemble Manifold Regularization

Bo Geng
Key Laboratory of Machine Perception
Peking University, Beijing 100871
bogeng@pku.edu.cn

Dacheng Tao
School of Computer Engineering
Nanyang Technological University, Singapore 639798
dacheng.tao@gmail.com

Chao Xu
Key Laboratory of Machine Perception
Peking University, Beijing 100871
xuchao@cis.pku.edu.cn

Linjun Yang
Microsoft Research Asia
Beijing 100190
linjuny@microsoft.com

Xian-Sheng Hua
Microsoft Research Asia
Beijing 100190
xshua@microsoft.com

Abstract

We propose an automatic approximation of the intrinsic manifold for general semi-supervised learning problems. Unfortunately, it is not trivial to define an optimization function to obtain optimal hyperparameters. Usually, pure cross-validation is considered but it does not necessarily scale up. A second problem derives from the suboptimality incurred by discrete grid search and overfitting problems. As a consequence, we developed an ensemble manifold regularization (EMR) framework to approximate the intrinsic manifold by combining several initial guesses. Algorithmically, we designed EMR very carefully so that it (a) learns both the composite manifold and the semi-supervised classifier jointly; (b) is fully automatic for learning the intrinsic manifold hyperparameters implicitly; (c) is conditionally optimal for intrinsic manifold approximation under a mild and reasonable assumption; and (d) is scalable for a large number of candidate manifold hyperparameters, from both time and space perspectives. Extensive experiments over both synthetic and real datasets show the effectiveness of the proposed framework.

1. Introduction

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

The common motivation of the SSL algorithms [3, 4, 5, 7, 12, 13, 14, 15] 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 [3], 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 the data points by using the Laplacian Eigenmap [2] in the feature space. The manifold approximation and the classifier learning is combined together under the conventional regularization framework [6], which smoothes the classifier along the manifold. The conventional regularization framework [6] shows that the solution of an ill-posed problem can be approximated by the variational principle, which contains both the data and prior smoothness information. The *manifold regularization* utilizes the manifold to replace the smoothness assumption in [6], where the manifold is determined by the

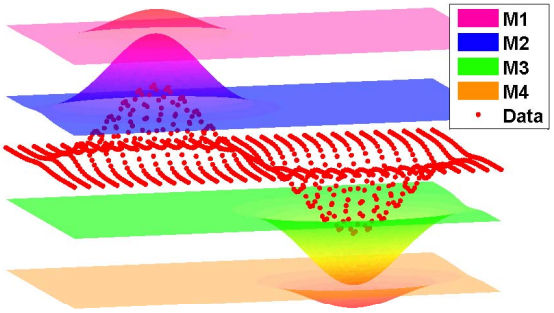

Figure 1. Illustration of EMR. From top to bottom, the IDs of four estimated manifolds are M1, M2, M3, and M4. The intrinsic data manifold is a linear combination of M2 and M3.

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

graph Laplacian with 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, cross-validation 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 (e.g., in Fig. 1, none of the selected parameters could approximate the intrinsic manifold well). Furthermore, it does not scale 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, pure cross-validation based parameter selection inevitably drives the model to overfit the training or the validation set, while the learner loses the generalization ability. As a consequence, automatic and data-driven manifold approximation is invaluable 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 the semi-supervised classifier learning. By providing a series of initial guesses of graph Laplacian, the framework learns to combine them to approximate the intrinsic manifold, e.g., Fig. 1, in a conditionally optimal way. Meanwhile, the semi-supervised classifier is learned and restricted to be smooth along the estimated manifold. We designed the EMR framework carefully so that it: (a) learns both the composite manifold and the semi-supervised classifier jointly, leading to a unified framework; (b) is fully automatic for learning hyperparameters of the intrinsic manifold implicitly and avoids problems caused by pure cross-validation; (c) 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 (e.g., in Fig. 1, the intrinsic manifold is the linear combination of manifolds M2 and M3); and (d) is scalable for a large number of candidate manifold hyperparameters, from both time and space perspectives, due to the avoidance of the cross-validation and the sparsity of the graph structure.

2. Ensemble Manifold Regularization

Consider the *semi-supervised learning* (SSL) 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. Suppose labeled samples are $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ pairs drawn from a probability distribution P, and unlabeled samples are 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 [3]. The problem takes the following form:

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

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 [12];$||f||_I^2$ 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 and regularizations $||f||_K^2$ and $||f||_I^2$. The manifold regularization term $||f||_I^2$ is the key to SSL and models the classifier smoothness along the manifold estimated from the unlabeled samples.

It turns out that in an appropriate coordinate system (exponential, which to the first order coincides with the local coordinate system given by a tangent plan in $\mathbb{R}^d$) [9], $||f||_I^2$ is approximated by the graph Laplacian L and the function prediction $\mathbf{f} = [f(x_1), \ldots, f(x_{l+u})]^T$, i.e.,$||f||_I^2 = \frac{1}{(u+l)^2}\mathbf{f}^T L \mathbf{f}$. 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 $W \in \mathbb{R}^{l+u} \times \mathbb{R}^{l+u}$ is the data adjacency graph, wherein each element W_{ij} is the edge weight between two samples x_i and x_j. In the diagonal matrix $D \in \mathbb{R}^{l+u} \times \mathbb{R}^{l+u}$, $D_{ii} = \sum_{i=1}^{l+u} W_{ij}$.

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

It is nontrivial to directly obtain the optimal graph Laplacian hyperparameters according to (1). As a consequence, we propose an alternative approach to learn the optimal hyperparameters implicitly, by assuming that the intrinsic manifold lies in the convex hull of some pre-given manifold candidates. Since 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.,

$$L = \sum_{j=1}^{m} \mu_j L_j, \qquad (2)$$
$$\text{s.t.} \sum_{j=1}^{m} \mu_j = 1,\ \mu_j \geq 0,\ j = 1, \ldots, m$$

where we define a set of candidate graph Laplacians $C = \{L_1, \ldots, L_m\}$ and denote the convex hull of set A as:

$$\mathbf{conv}A = \{\theta_1 x_1 + \cdots + \theta_k x_k | \theta_1 + \cdots + \theta_k = 1, \quad x_i \in A,\ \theta_i \geq 0, i = 1, \ldots, k\} \tag{3}$$

Therefore, we have $L \in \mathbf{conv}C$. In Section 2.2, we will prove that L 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 some pre-given candidates. Because each candidate graph Laplacian 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.,

$$||f||_I^2 = \frac{1}{(u+l)^2}\mathbf{f}^T L \mathbf{f} = \frac{1}{(u+l)^2}\mathbf{f}^T(\sum_{j=1}^{m}\mu_j L_j)\mathbf{f} \tag{4}$$
$$= \sum_{j=1}^{m}\mu_j ||f||_{I(j)}^2.$$

Then, we obtain the EMR framework as,

$$\min_{f\in\mathcal{H}_K, \mu\in\mathbb{R}^m} \frac{1}{l}\sum_{i=1}^{l} V(f, x_i, y_i) + \gamma_A ||f||_K^2 + \gamma_I \sum_{j=1}^{m}\mu_j ||f||_{I(j)}^2 + \gamma_R ||\mu||_2^2, \tag{5}$$
$$\text{s.t.} \sum_{j=1}^{m}\mu_j = 1,\ \mu_j \geq 0,\ j = 1, \ldots, m$$

where the regularization term $||\mu||_2^2$ is introduced to avoid the parameter overfitting to only one manifold; and $\gamma_R \in \mathbb{R}^+$ is the trade-off parameter to control the contribution of the regularization term $||\mu||_2^2$. Because (5) contains a weighted combination of multiple manifold regularization terms, we name the new regularization framework as *ensemble manifold regularization* (EMR).

For a fixed μ, (5) degenerates to (1), with $L = \sum_{j=1}^{m}\mu_j L_j$ for $||f||_I^2$. On the other hand, for a fixed f, (5) is simplified to:

$$\min_{\mu\in\mathbb{R}^m} \sum_{j=1}^{m}\mu_j s_j + \gamma_R ||\mu||^2, \tag{6}$$
$$\text{s.t.} \sum_{j=1}^{m}\mu_j = 1,\ \mu_j \geq 0,\ j = 1, \ldots, m$$

where $s_j = \frac{\gamma_I}{(u+l)^2}\mathbf{f}^T L_j \mathbf{f}$. Under this condition, if $\gamma_R = 0$, the solution of (6) will be: $\mu_j = 1$ if $s_j = \min_{k=1,\ldots,n} s_k$ and $\mu_j = 0$ otherwise. Such a trivial case will assign all the weight to one manifold, which is extremely sparse and not desirable for learning a composite manifold. If $\gamma_R \to +\infty$, the solution tends to give identical weights to all the graph Laplacians.

We present some theoretical analysis of EMR here. Because the Laplacian matrix for each graph satisfies the semidefinite postive property, i.e., $L_i \in S_{l+u}^+$, their convex combination satisfies $L \in S_{l+u}^+$. Consequently, $L \in \mathbf{conv}C$ is a graph Laplacian. According to [3], the representer theorem follows for a fixed μ.

Theorem 1 *For a $L \in \mathbf{conv}C$, the minimization of (5) w.r.t. f with a fixed μ, exists and has the representation*

$$f^\star(x) = \sum_{i=1}^{l+u}\alpha_i^\star K(x_i, x), \tag{7}$$

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

The representer theorem presents us with the existence and general form of the solution of (5) under any 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 to solve (5) in an iterative manner, i.e., first solving (5) w.r.t. f with a fixed μ, resulting in the solution represented by (7); then optimizing (5) w.r.t. μ, with f taking the value solved in the last iteration; and alternatively iterating the above two steps, until the decrement of the objective function is zero. For any convex loss function $V \in \mathcal{H}_K \times \mathbb{R}^d \times \mathbb{R}$, (5) is convex not only w.r.t. f for fixed μ but also w.r.t. μ for fixed f. Consequently, the alternating optimization of (5), iteratively over parameters f and μ converges.

However, (5) is not convex w.r.t. (f, μ) jointly. We can solve the problem based on two strategies: (a) set a large value for γ_R so that (5) is convex w.r.t. (f, μ); (b) initialize $\mu = \frac{1}{m}$. The later strategy initializes L 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, and leave the discussion about the effects of γ_R independently in Section 4.

The theoretical analysis shown above presents the basic properties of the proposed EMR, and ensures that the framework can be implemented into numerous algorithms for various machine learning applications, e.g., data classification and regression.

3. Algorithms

We consider the general-purpose classification task in this section, and show how to implement EMR framework based on SVM. The regression task is another direct extension with different forms of loss functions, which is not extensively discussed here.

3.1. EMR Support Vector Machines

In SVM, the hinge loss is adopted, i.e., $V(f,x,y) = (1-yf(x))_+ = \max(0, 1-yf(x))$. For a fixed μ, we can resort to the theorem 1, and the solution is given by (7). Substituting (7) into the framework (5), we can obtain the EMR Support Vector Machine (EMR-SVM):

$$\begin{aligned}\min_{\alpha\in\mathbb{R}^{l+u},\xi\in\mathbb{R}^l,\mu\in\mathbb{R}^m} & \frac{1}{l}\sum_{i=1}^{l}\xi_i + \gamma_A\alpha^T K\alpha \\ & + \frac{\gamma_I}{(l+u)^2}\sum_{j=1}^{m}\mu_j\alpha^T K L_j K\alpha + \gamma_R||\mu||_2^2, \quad (8)\\ \text{s.t. } & y_i(\sum_{j=1}^{l+u}\alpha_i K(x_i,x_j)+b) \geq 1-\xi_i,\ i=1,\ldots,l \\ & \xi_i \geq 0,\ i=1,\ldots,l \\ & \sum_{j=1}^{m}\mu_j = 1,\ \mu_j \geq 0,\ j=1,\ldots,m \quad (9)\end{aligned}$$

where $K \in \mathbb{R}^{l+u}\times\mathbb{R}^{l+u}$ 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 (8), we need to get the solution of f (represented by α according to (7)) with a fixed μ, as well as the solution μ of with a fixed f.

3.2. Learning by Alternating Optimization

For a fixed μ, we can introduce non-negative Lagrange multipliers β and ς for the inequality constraints in (8), which leads to:

$$\begin{aligned}& L(\alpha,\xi,b,\beta,\varsigma) \quad (10)\\ =&\frac{1}{2}\alpha^T(2\gamma_A K + \frac{2\gamma_A}{(l+u)^2}K(\sum_{j=1}^{m}\mu_j L_j)K)\alpha + \frac{1}{l}\sum_{i=1}^{l}\xi_i \\ &-\sum_{i=1}^{l}\beta_i(y_i(\sum_{j=1}^{l+u}\alpha_i K(x_i,x_j)+b)-1+\xi_i) - \sum_{i=1}^{l}\varsigma_i\xi_i.\end{aligned}$$

By taking the partial derivative w.r.t. α, β, ς, and requiring them to be zero, we can eliminate those variables by substituting their solutions back into (10) for the dual format:

$$\begin{aligned}\beta^\star = &\max_{\beta\in\mathbb{R}^l}\sum_{i=1}^{l}\beta_i - \frac{1}{2}\beta^T Q\beta, \quad (11)\\ &\text{s.t. } \sum_{i=1}^{l}\beta_i y_i = 0,\ 0\leq\beta_i\leq\frac{1}{l},\ i=1,\ldots,l\end{aligned}$$

where $Q = YJK(2\gamma_A I_{l+u} + \frac{2\gamma_I}{(l+u)^2}\sum_{j=1}^{m}\mu_j L_j K)^{-1}J^T Y$, $Y = \text{diag}(y_1,\ldots,y_l)$, $I \in \mathbb{R}^d\times\mathbb{R}^d$ denotes a d-dimensional identity matrix and $J = [I_l\ \ 0] \in \mathbb{R}^l\times\mathbb{R}^{l+u}$. The solution of (8) is given by

$$\alpha^\star = (2\gamma_A I + \frac{2\gamma_I}{(l+u)^2}\sum_{j=1}^{m}\mu_j L_j K)^{-1}J^T Y\beta^\star. \quad (12)$$

The learning procedure combines different graph Laplacians into Q, and the optimization of (11) is approximately independent of the number of graph Laplacians m. Therefore, with a fixed μ, we do not incorporate additional computational costs, except for some sparse matrix additions, which is negligible.

On the other hand, for learning μ with a fixed f, (8) degenerates to (6), and we can adopt coordinate descent algorithm. In each iteration, we select two elements in for updating while the others are fixed. Suppose at an iteration, the ith and jth elements of are selected. Due to the constraint $\sum_{j=1}^{m}\mu_j = 1$, the summation of μ_i and μ_j will not change after this iteration. Therefore, we have the solution of this iteration:

$$\begin{cases}\mu_i^\star = 0,\ \mu_j^\star = \mu_i+\mu_j, \text{ if } 2\gamma_R(\mu_i+\mu_j)+(s_j-s_i)\leq 0\\ \mu_i^\star = \mu_i+\mu_j,\ \mu_j^\star = 0, \text{ if } 2\gamma_R(\mu_i+\mu_j)+(s_i-s_j)\leq 0\\ \mu_i^\star = \dfrac{2\gamma_R(\mu_i+\mu_j)+(s_j-s_i)}{4\gamma_R}, \mu_j^\star = \mu_i+\mu_j-\mu_i^\star, \text{ else.}\end{cases} \quad (13)$$

We iteratively traversal over all pairs of elements in μ and adopt the solution in (13), until the objective function in (6) does not decrease. Intuitively, the update criteria in (13) tends to assign larger value μ_j to smaller s_j. Because $s_j = \frac{\gamma_I}{(l+u)^2}\mathbf{f}^T L_j\mathbf{f}$ measures the smoothness of the function f over the jth manifold approximated by the graph Laplacian L_j, the algorithm will prefer the pre-given manifold that coincides with current iteration SSL classifier f.

3.3. Learning Complexity Analysis

Suppose we are given n samples, m candidate graph Laplacians, each of which corresponds to a specific manifold hyperparameter setting, and adopt v-fold cross-validation technique for parameter selection. For EMR-SVM, the computation of matrix Q in (11) involves an inversion and several multiplications of $n\times n$ matrix, with

the time complexity $O(n^{2.8})$ using Strassen algorithm [11]; solving (11) is a standard QP problem, where we utilize SVM based solver with the time complexity $O(n^{2.3})$ [8]; for (6), we traversal over all pairs of variables, and the time cost is approximately $O(m^2)$. Denoting the number of alternating iteration as η, and the number of candidate parameters that need the cross-validation as r. Therefore, the total cost of EMR-SVM is $O(v\eta r(n^{2.8} + n^{2.3} + m^2))$. Since the graph number m is generally much smaller than the sample number n, we can approximate its time cost as $O(v\eta r n^{2.8})$. On the contrary, for the single manifold regularization based algorithm LapSVM [3], we need to do v times cross-validation over all graphs, with time complexity $O(vmn^{2.8})$. In section 4, we'll experimentally demonstrate that the iteration number η is usually quite small, even for a large number of graphs. The number r is small since γ_R is comparatively insensitive to the performance of the classifier. However, m is linearly dependent on the number of graphs. As a consequence, EMR-SVM is much more efficient than LapSVM for a large number of graph Laplacian hyperparameters.

For the space cost, since the graph Laplacian is a sparse matrix structure, EMR-SVM can easily store all the graphs into the memory, with the space cost $O(mkn)$, where k is quite small for general problem with $mk < n$. However, for matrix Q , we have to allocate $O(n^2)$ space to store it. As a consequence, the space cost of EMR-SVM is comparable to that of LapSVM, and is around $O(n^2)$.

4. Experiments

In this section, experiments are conducted extensively over synthetic (Two Moons) [10, 3] and UCI Machine Learning Repository (USPS test set, Heart) [1], to demonstrate the effectiveness and efficiency of EMR. The proposed classification algorithm EMR-SVM is compared with conventional SVM [12], transductive SVM (TSVM) [7] and LapSVM [3].

The RBF kernel $K(x_i, x_j) = \exp(-\gamma_K||x_i - x_j||^2)$ is utilized for all experiments. For graph construction, the heat Kernel [9] is adopted to compute the edge weights, i.e., if $x_i \in N(x_j)$ or $x_j \in N(x_i)$ we have $W_{ij} = \exp(-t||x_i - x_j||^2)$, 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 create two graph Laplacian sets for different purposes. For the first set, we choose $t = \{\frac{\tau}{50}, \frac{\tau}{45}, \ldots, \frac{\tau}{5}, \tau, 5\tau, \ldots, 60\tau, 65\tau\}$ and fix $k = 10$ and $p = 2$, which leads to 24 graphs. This simplified version focuses on the variation of hyperparameter t, which is suitable for algorithm analysis. For another one, the candidate hyperparameters are chosen as $t = \{\frac{\tau}{15}, \frac{\tau}{10}, \frac{\tau}{5}, \tau, 5\tau, \ldots, 10\tau, 15\tau, 20\tau\}$, $k = \{5, 10, 15\}$ and $p = \{1, 2, 3\}$, where we get 72 graphs. This comparatively larger graph Laplacian set varies all hyperparameters, and is intended to evaluate the performance of EMR-SVM and to prove that EMR can automatically estimate all hyperparameters introduced by graph Laplacian. Here, τ is empirically set as the inverse of the mean square of Euclidian distances between all pairs of samples in a specific dataset, i.e.,$\tau = (\frac{1}{(l+u)^2}\sum_{i,j=1}^{l+u}||x_i - x_j||^2)^{-1}$.

The general parameters of all the algorithms are determined by the two-fold cross-validation over the training set. For LapSVM, the hyperparameters t, k and p are cross-validated carefully. For EMR-SVM, we empirically set $\gamma_R = 0.1\gamma_I$, and leave the discussion about its insensitive property later on.

4.1. Two Moons

In this section, we utilize the well known synthetic dataset *two moons*, to justify the significance of the proposed EMR for automatically approximating the intrinsic manifold. For the binary classification of data belonging to different moons, the data generator randomly draws 150 samples for both moons, each of which contains only one labeled sample (therefore, 2 training samples and 298 testing samples in all). The 24 graph set mentioned above is adopted for EMR-SVM.

The classifier boundaries of different algorithms are plotted in Fig. 2. Obviously, for SVM without unlabeled data considered, a direct max-margin classifier is built between the two labeled samples. Although TSVM employ the information of unlabeled samples, it cannot avoid getting stuck to a suboptimal solution. For LapSVM, we present the one with the best classification accuracy in our graph Laplacian set. Unfortunately, it does not perform as well as expected, because all graph Laplacian hyperparameters are suboptimal for approximating the intrinsic manifold. Our EMR-SVM, by learning to combine different manifolds, creates a nearly perfect classifier boundary. This demonstrates that the composite manifold learned by EMR is much better than grid search based hyperparameter selection method.

Fig. 3 presents us some detailed results of LapSVM over each graph Laplacian hyperparameter t. Due to the space limitation, we only present the top four results. It proves that the hyperparameter t is essential for the manifold approximation, and its estimation is nontrivial.

4.2. UCI Machine Learning Repository

The USPST set is the test data part of the famous handwritten digits USPS dataset, in the UCI Machine Learning Repository [1]. It contains 2007 samples with 10 classes, each of which corresponds to a handwritten digit. The dataset is widely used for evaluating the effectiveness of

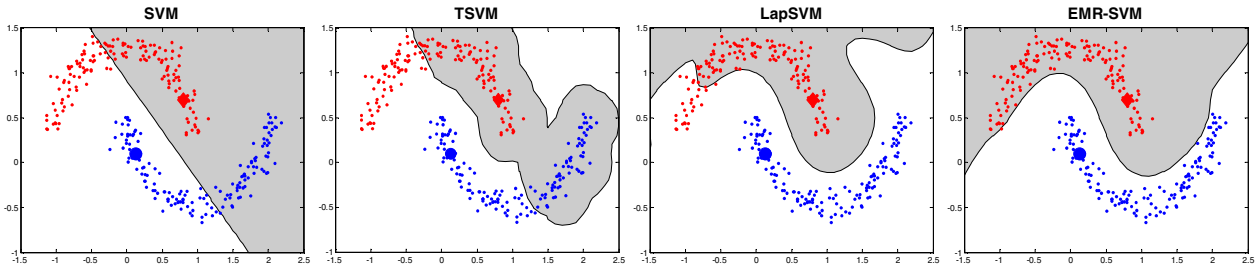

Figure 2. Two Moons Dataset: the best classifiers for SVM, TSVM, LapSVM and EMR-SVM. Labeled samples are high-lighted.

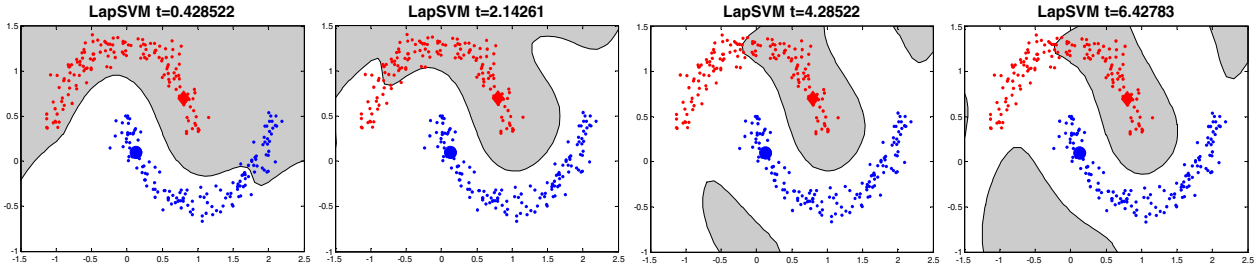

Figure 3. Two Moons Dataset: The best four results of LapSVM over different graph Laplacians. Labeled samples are high-lighted.

SSL algorithms [3, 10], where the dataset is randomly divided into 10 splits, with 50 labeled and 1957 unlabeled samples in each split. We use the 24 graph set to conduct one-vs-rest multi-class classification experiments.

To study the effectiveness of EMR, we select one split and present the manifold composite weight μ learned by EMR-SVM in each one-vs-rest classification task in Fig. 4, together with the classification error rate of each graph Laplacian using LapSVM. 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 it is helpful to deeply understand how and why the regularization term $\gamma_R||\mu||^2$ affects the whole framework. We compare how the multiclass error rate varies over different γ_R for EMR-SVM, and different t for LapSVM. Here, the parameter γ_R varies from $10^{-4}\gamma_I$ to γ_I with 12 parameters. For hyperparameter t, we choose the 24 graph set, in which t changes from 1.66×10^{-4} to 5.4×10^{-1}. Note that the domain of γ_R is comparatively larger than that of t. The experimental results averaged over ten splits are shown in Fig. 5. We can find that the average multiclass error rate is insensitive to the parameter in EMR-SVM, while the performance is much more sensitive to the hyperparameter t in LapSVM. The right figure in Fig. 5 shows a detailed results of γ_R. Experimentally, the best results is achieved when $\gamma_R = 0.01\gamma_I$. Analog to the conclusion from *two moons* dataset and the analyses in Section 2, the best composite graph is a balance between unanimous weighting and a single manifold.

The one-vs-rest multiclass error rates are shown in Table 1, where the presented results of LapSVM are from [3].

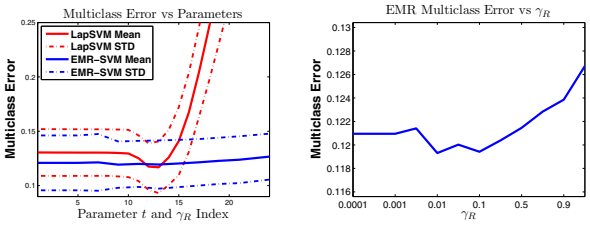

Figure 5. Parameter sensitivity comparison of LapSVM t and EMR-SVM γ_R (Left) and details for EMR-SVM γ_R(Right)

Besides the results of 24 graph set (abbreviated as EMR-24G), we also show the results using 72 graph set (abbreviated as EMR-72G), to demonstrate that EMR could be extended to learn other graph Laplacian hyperparameters, including p and k. As the results of *two moons* dataset, SVM and TSVM do not perform as well as manifold regularization based methods. On the other hand, the performance of EMR-SVM is consistently better than LapSVM, and EMR-72G yields the best performance. Furthermore, it is worth emphasizing that EMR-SVM approximates the intrinsic graph Laplacian automatically without performing cross-validations over the graph hyperparameters.

In Table 1, we also show experimental results of the Heart dataset [1] for binary classification. The dataset is randomly divided by the ratio 1:9 between training and test sets with 10 different splits. Based on this experiment, we can draw the same conclusion as above.

5. Conclusion

Numerous practical applications can be handled by the *semi-supervised learning* algorithms based on *manifold reg-*

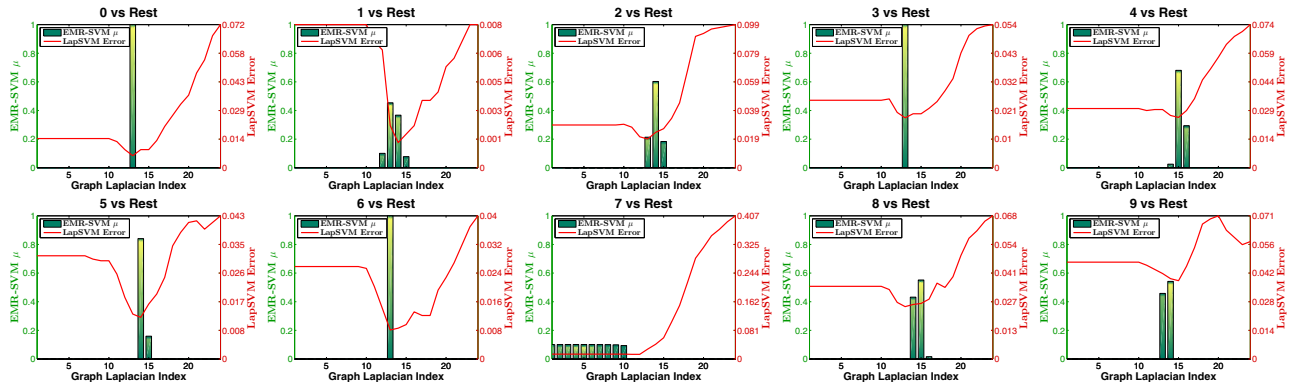

Figure 4. USPST Dataset: The manifold combination coefficient μ learned by EMR-SVM over ten one-vs-rest classification tasks, as well as the classification error rate of LapSVM using each graph Laplacian over each task.

Table 1. Performance comparisons of different algorithms over USPST and Heart dataset.

	SVM	TSVM	LapSVM	EMR-SVM	
				24G	72G
USPST	22.94	26.28	12.67	11.94	11.53
Heart	20.07	20.01	18.55	18.18	17.59

ularization, or its variants and extensions. However, they all fall short, because the estimation of the hyperparameters incorporated in the manifold regularization is not trivial. In this paper, we propose an *ensemble manifold regularization* (EMR) framework for automatically and implicitly estimating the hyperparameters of 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. Extensive experiments over synthetic and real dataset demonstrate that EMR is superior to LapSVM, for effectively approximating the intrinsic manifold and improving the classification performance.

6. Acknowledgements

This project was supported by the Microsoft Operations PTE LTD-NTU Joint R&D (under project number M48020065), Nanyang Technological University Start-Up Grant (under project number M58020010), China 973 Research Program under Grant 2004CB318005 and 2009CB320900.

References

[1] http://www.ics.uci.edu/ mlearn/.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 2003.

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 2006.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. *COLT*, 1998.

[5] O. Chapelle and A. Zien. Semi-supervised classification by low density separation. *AISTATS*, 2005.

[6] F. Girosim, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 1995.

[7] T. Joachims. Transductive inference for text classification using support vector machines. *ICML*, 1999.

[8] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. *In Advances in Kernel Methods - Support Vector Learning*, 1999.

[9] S. Rosenberg. The laplacian on a riemmannian manifolds. *Cambridge University*, 1997.

[10] V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. *ICML*, 2005.

[11] V. Strassen. Gaussian elimination is not optimal. *Numer. Math.*, 1969.

[12] V. N. Vapnik. Statistical learning theory. *Wiley*, 1998.

[13] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Semi-supervised adapted hmms for unusual event detection. *CVPR*, 2005.

[14] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. *NIPS*, 2004.

[15] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. *ICML*, 2003.