Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007

Maximum Margin based Semi-supervised Spectral Kernel Learning

Zenglin Xu, Jianke Zhu, Michael R. Lyu, *Fellow, IEEE,* and Irwin King, *Member, IEEE*

***Abstract*— Semi-supervised kernel learning is attracting increasing research interests recently. It works by learning an embedding of data from the input space to a Hilbert space using both labeled data and unlabeled data, and then searching for relations among the embedded data points. One of the most well-known semi-supervised kernel learning approaches is the spectral kernel learning methodology which usually tunes the spectral empirically or through optimizing some generalized performance measures. However, the kernel designing process does not involve the bias of a kernel-based learning algorithm, the deduced kernel matrix cannot necessarily facilitate a specific learning algorithm. To supplement the spectral kernel learning methods, this paper proposes a novel approach, which not only learns a kernel matrix by maximizing another generalized performance measure, the margin between two classes of data, but also leads directly to a convex optimization method for learning the margin parameters in support vector machines. Moreover, experimental results demonstrate that our proposed spectral kernel learning method achieves promising results against other spectral kernel learning methods.**

I. Introduction

Kernel methods provide a new learning framework in machine learning for their conceptual simplicity and good performance on many tasks [20], [21]. They work by embedding data from the input space to a Hilbert space, and then searching for relations among the embedded data points. Semi-supervised learning has been actively studied in the machine learning communities [6], which can take advantage of the unlabeled data.

The graph Laplacians method [29], [27], [26] is one of the most well-known kernel-based semi-supervised learning approach. In this family of semi-supervised kernel learning methods, kernels are usually constructed by transforming the spectrum of a "local similarity" graph over both labeled and unlabeled data. During learning such a kernel, Zhu et al. [29] propose to learn coefficients corresponding to smooth eigenvectors of a spectral graph [8] via maximizing the kernel target alignment [9] which measures the similarity between the feature space induced by a kernel matrix and the feature space induced by labels. Later, Hoi et. al [10] extend the work in [29] through equipping the kernel matrix with a faster spectral decay rate.

However, the kernel designing process does not involve the bias of a kernel-based learning algorithm, the deduced kernel matrix cannot necessarily facilitate a specific learning algorithm. It is known that different kernel methods try to utilize different prior knowledge in order to derive the separating hyperplane. For example, SVM maximizes the boundary between two classes of data in the kernel-induced feature space; Kernel Fisher Discriminant Analysis (KFDA) [17] maximizes the between-class covariance while minimizes the within covariance; and Minimax Probability Machine (MPM) [15], [11] finds a hyperplane in the feature space, which minimizes the maximum Mahalanobis distances to two classes. Therefore, it is necessary to incorporate the bias or the prior of a learning algorithm into the kernel designing process in order to make a classifier adequately utilize the prior information underlying the labeled data and the unlabeled data.

To supplement the spectral kernel learning methods, this paper proposes a novel approach, which not only learns a kernel matrix by maximizing another generalized performance measure, the margin between two classes of data, but also leads directly to a convex optimization method for learning the margin parameters in SVMs. More specifically, our semi-supervised spectral kernel learning approach learns a kernel matrix with a fast spectral decay rate, which utilizes the labels of the training set as well as the underlying distribution of the whole data to maximize the soft margin between different classes.

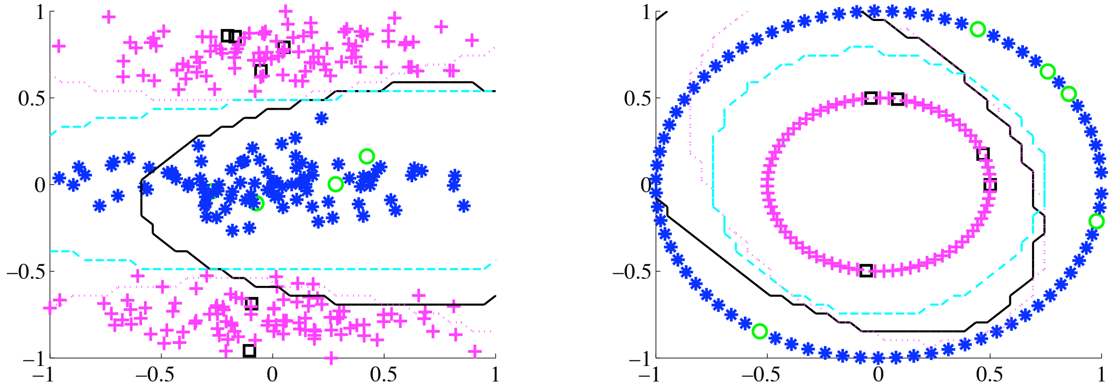
To understand the characteristics of the proposed spectral kernel learning method, we employ two synthetic data sets with a cluster structure as examples. **Relevance** is a data set where only one dimension of the data is relevant to separate the data. **Twocircles** is composed by two circles with the same center. Figure 1 draws the decision boundaries of different algorithms. The unlabeled data including the data over a grid, are utilized to draw the decision boundaries. It indicates that generalization ability can be strengthened through utilizing the information of unlabeled data to learn a kernel matrix. Therefore, the performance of the classifier can be improved, and it is especially significant in the case that the kernel matrix is learned by maximizing the margins.

The rest of this paper is organized as follows. Section 2 reviews the related work in kernel learning. In Section 3, we derives the proposed kernel learning approach which maximizes a generalized performance measure and optimizes the margin parameters in SVM. Section 4 describes the experimental results of the proposed kernel learning approach as well the baseline methods. Section 5 sets out our conclusion and discusses future work.

We use the following notations. Let $\mathcal{X}$ denotes the original input space, which is an arbitrary subset of $\mathcal{R}^d$ where d is a positive number. Let $\mathcal{C} = \{1, 2, ..., m\}$ be the set of labels where m is the number of classes. Let l be the number of labeled data points and n be the amount of labeled data and unlabeled data. A kernel function is defined as a symmetric function κ, such that $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$

Zenglin Xu, Jianke Zhu, Irwin King, and Michael R. Lyu are with the Department of Computer Science and Engineering, the Chinese University of Hong Kong, Shatin, N.T., Hong Kong. E-mail: {zlxu,jkzhu,king,lyu}@cse.cuhk.edu.hk.

1-4244-1380-X/07/$25.00 ©2007 IEEE



(a) Relevance

(b) Twocircles

Fig. 1. The decision boundaries on **Relevance** and **Twocircles**. The points represented by squares (in black) and circles (in green) are the labeled data. Those represented by pluses (in magenta) and asterisks (in blue) are unlabeled data. SVMs equipped with RBF kernels are used as the classifiers. The separating lines were obtained by projecting test data over a grid. The lines in black (dark), magenta (doted), and cyan (dashed) represent decision boundaries of kernel SVM with a regular RBF kernel, a fast-decay spectral kernel attained by maximizing the kernel target alignment, a fast-decay spectral kernel attained by maximizing the margin, respectively.

for all $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$, where Φ is a mapping from $\mathcal{X}$ to a feature space $\mathcal{H}$. The form of kernel function κ could be a linear kernel function, $\kappa(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i} \cdot \mathbf{x_j}$, or a Gaussian RBF kernel function, $\kappa(\mathbf{x_i}, \mathbf{x_j}) = \exp(-|\mathbf{x_i} - \mathbf{x_j}\|_2^2/\sigma)$, or a polynomial kernel function, $\kappa(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + 1)^p$, for some σ and p respectively. A standard kernel matrix or Gram matrix $K \in \mathcal{R}^{n \times n}$ is a positive semidefinite matrix such that $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ for any $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$. We denote the eigenvalues and eigenvectors of a kernel matrix as λ and ϕ, such that $K = \sum_{i=1}^{n} \lambda_i \phi_i \phi_i^T$. Note that except that it is specified clearly, the eigenvectors are sorted according to the decreasing order of eigenvalues.

II. Related Work

Kernel-based learning algorithms have been widely studied in machine learning (see, for example, [20], [21]). They work by embedding the data from the input space to a Hilbert space, and then searching for relations among the embedded data points. The embedding implicitly defines the geometry of the feature space and induces a notion of similarity in the input space. According to Mercer's Theorem [20], any kernel function κ implicitly maps data in the input space to a high dimensional Hilbert space $\mathcal{H}$ through the mapping function $\Phi : \mathcal{X} \rightarrow \mathcal{H}$. Therefore, it is important to learn a kernel matrix corresponding to the entire data set. Lanckriet et. al proposes Semi-definite Programming (SDP) algorithms to learn a combination of different kernel matrices [14]. Note that in [14], two performance measures, the kernel target alignment and the margin between two classes of data are utilized. Other optimal kernel learning algorithms in machine learning can be found in [5], [12], [16], [2], [9].

On the other hand, spectral graph theory [8] has attracted the focus of semi-supervised kernel learning [27], [22]. Several semi-supervised learning algorithms have been proposed based on Spectral Graph Theory, for example, diffusion kernels [13], Gaussian fields [28], and the order-constrained spectral kernel [29]. Typically, a graph is constructed where the nodes are the data instances and the edges define the "local similarity" measures among data points. For example, the local similarity measure can be the Euclidean distance and the edge can be constructed by the node's k nearest neighbors. The edge between two data points suggests that they may share the same label. In general, it is believed that smaller eigenvalues correspond to smoother eigenvectors over the graph. Thus smaller eigenvalues and corresponding eigenvectors are used to compose the initial graph Laplacian which is further employed to maximizes the alignment between the learned kernel matrix and the target kernel in order to learn a new kernel matrix. In [29], the experimental results imply that the order-constrained spectral kernel achieves better performance than the diffusion kernel and the Gaussian field kernel. Moreover, Hoi et. al [10] still optimize the kernel target alignment, and extend the spectral kernel learning method by specifying a fast spectral decay rate.

Some recent theoretical work builds the connection between spectral graph theory and kernel learning. Smola and Condor [22] show some theoretical understanding between kernel and regularization based on the graph theory. In addition, Berkin et al. develop a regularization framework for regularization on graphs [1]. In most recent, Zhang et al. provide a theoretical framework for semi-supervised learning based on unsupervised kernel design and derive a generalization error bound [26]. It demonstrates that a kernel with a fast decay rate is useful for the classification task [25], [26]. All of the above work build the solid foundation of this paper.

III. Spectral Kernel Learning

In this section, we first describe the theoretical foundation of spectral kernel learning, and then present the maximum margin based spectral kernel learning approach.

A. Theoretical Foundation

We review the theoretical foundation from the perspectives of unsupervised kernel design rule and the optimization criteria for a good kernel matrix. Then we summarize the foundation into a semi-supervised spectral kernel learning rule.

1) Unsupervised Kernel Design Rule: From the perspective of standard supervised learning, the objective is to learn a function f so that the empirical loss is as small as possible [24]. To avoid overfitting, one needs to restrict the hypothesis function family size. Thus, we consider the following regularized linear prediction method on the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$:

$$\hat{f} = \arg\inf_{f\in\mathcal{H}} \frac{1}{l}\sum_{i=1}^{l} L(h(\mathbf{x}_i), y_i) + r||h||_{\mathcal{H}}^2, \tag{1}$$

where r is a regularization coefficient. According to the Representer Theorem [24], the solution of Eq. (1) can be represented as $\hat{f}(\mathbf{x}) = \sum_{i=1}^{l} \hat{\alpha}_i \kappa(\mathbf{x})$, where $\alpha = (\hat{\alpha}_1, \dots, \hat{\alpha}_l)$ is given by

$$\alpha = \arg\min_{\alpha\in\mathcal{R}_l} \frac{1}{l} L(\sum_{j=1}^{l} \alpha_j \kappa(\mathbf{x}_i, \mathbf{x}_j), y_j) + r\sum_{j,k=1}^{l} \alpha_j\alpha_k\kappa(\mathbf{x}_j, \mathbf{x}_k). \tag{2}$$

Consider a semi-supervised setting, we try to learn the real-valued vectors $f \in \mathcal{R}^n$, such that

$$\hat{f} = \arg\inf_{f\in\mathcal{R}^n} \frac{1}{l}\sum_{i=1}^{l} L(f_i, y_i) + rf^T K^{-1} f. \tag{3}$$

It is proven that the solution of the above semi-supervised learning is equivalent to the solution of supervised learning in Eq. (1), such that

$$\hat{f}_j = \hat{h}(\mathbf{x}_j), \quad j = 1, \dots, n. \tag{4}$$

Therefor, it provides a way of unsupervised kernel design by replacing the kernel function κ with $\bar{\kappa}$, or replacing the kernel matrix K with $\bar{K}$, i.e.,

$$\bar{K} = \sum_{i=1}^{n} g(\lambda_i)\phi_i\phi_i^T, \tag{5}$$

where $g(\cdot)$ is a transformation function of the spectra of a kernel matrix and λ_i is sorted in a decreasing order. This is also consistent with the general principle for creating a semi-supervised kernel from the graph Laplacian as suggested in [7], [22]. Depending on different forms of $g(\cdot)$, different kernel matrices can be learned. We summarize the settings of $g(\cdot)$ as well as their corresponding kernels in Table I. Note that for these spectral kernels, parameters σ, ϵ, q, w are tuned using cross-validation. μ is the optimization variable in [29] and [10] to optimize the alignment between the learned kernel matrix and the target kernel.

2) Optimization Criteria: Kernel methods choose a function that is linear in the feature space by optimizing some criterion over the samples. More specifically, the optimization criteria include the kernel alignment, the margin between different classes, and the Fisher discriminant ratio [18], [9], [14], [20]. We focus our attention on the kernel alignment and the margin between different classes, because they can be conveniently used in kernel learning.

Definition 1 Kernel Alignment. The empirical alignment of a kernel κ_1 with a kernel κ_2 with respect to the sample $\mathcal{X}$ is the quantity:

$$\omega_A(\mathcal{X}, \kappa_1, \kappa_2) = \frac{\langle K_1, K_2\rangle_F}{\sqrt{\langle K_1, K_1\rangle_F \langle K_2, K_2\rangle_F}}, \tag{6}$$

where K_i is the kernel matrix for the sample $\mathcal{X}$ using the kernel function κ_i and $\langle\cdot,\cdot\rangle_F$ is the Frobenius inner product between two matrices, i.e., $\langle K_1, K_2\rangle_F = \sum_{i,j=1}^{n} \kappa_1(\mathbf{x_1}, \mathbf{x}_2)\kappa_2(\mathbf{x_1}, \mathbf{x}_2)$.

This offers a principle to learn a kernel matrix through assessing the relationship between a given kernel and the target kernel induced by the given labels. When the vector $\mathbf{y}$ of $\{\pm 1\}$ is known, we can consider $T = \mathbf{y}\mathbf{y}^T$ as the target kernel. Let K_{tr} as the "training-block" of the kernel matrix, which are composed by data with known labels. Then the alignment of the training-block of the kernel matrix and the target kernel matrix can be formulated as follows:

$$\omega_A(\mathcal{X}, K_{tr}, T) = \frac{\langle K_{tr}, \mathbf{y}\mathbf{y}^T\rangle_F}{\sqrt{\langle K_{tr}, K_{tr}\rangle_F \langle \mathbf{y}\mathbf{y}^T, \mathbf{y}\mathbf{y}^T\rangle_F}}. \tag{7}$$

Since $\langle \mathbf{y}\mathbf{y}^T, \mathbf{y}\mathbf{y}^T\rangle_F = l^2$, the above equation is equivalent to

$$\omega_A(\mathcal{X}, K_{tr}, T) = \frac{\langle K_{tr}, \mathbf{y}\mathbf{y}^T\rangle_F}{l\sqrt{\langle K_{tr}, K_{tr}\rangle_F}}. \tag{8}$$

Definition 2 Soft Margin. Given a labeled sample $\mathcal{X}_l$, the hyperplane $(\mathbf{w}_*, b_*)$ that solves the optimization problem

$$\begin{aligned} &\min_{\mathbf{w},b} \quad \langle \mathbf{w}, \mathbf{w}\rangle + C\sum_{i=1}^{l}\xi_i \\ &s.t. \quad y_i(\langle \mathbf{w}, \Phi(\mathbf{x}_i) + b\rangle) \geq 1 - \xi_i, i = 1, \dots, l, \\ &\xi_i \geq 0, \end{aligned} \tag{9}$$

realizes the maximal margin classifier with geometric margin $\gamma = 1/||\mathbf{w}_*||_2$, assuming it exists.

By formulating Eq. (9) into its corresponding Lagrangian dual problem, the solution can be derived as below:

$$\begin{aligned} \omega_M(K^{tr}) &= \langle \mathbf{w}_*, \mathbf{w}_*\rangle + C\sum_{i=1}^{l}\xi_{i*} \\ &= \max_{\alpha} 2\alpha^T\mathbf{e} - \alpha^T G(K^{tr})\alpha : \quad \alpha^T\mathbf{y} = 0, \end{aligned} \tag{10}$$

where $\mathbf{e}$ is the l-dimensional vector of ones, $C \geq \alpha \geq 0$, $\alpha \in \mathcal{R}^l$, $G(K^{tr})$ is defined by $G_{ij}(K^{tr}) = [K^{tr}]_{ij}y_iy_j = \kappa(\mathbf{x}_i, \mathbf{x}_j)y_iy_j$, and $\alpha \geq 0$ means $\alpha_i \geq 0$, $i = 1, \dots, l$.

TABLE I

SEMI-SUPERVISED KERNELS ACHIEVED BY DIFFERENT SPECTRAL TRANSFORMATION.

$g(\lambda)$	**Parameter(s)**	**Kernels**	**Refernces**
$g(\lambda) = \exp(-\frac{\sigma^2}{2}\lambda)$	σ	the diffusion kernel	[13]
$g(\lambda) = \frac{1}{\lambda+\epsilon}$	ϵ	the Gaussian filed kernel	[28]
$g(\lambda) = \mu_i, \mu_i \leq \mu_{i+1}, i = 1, \ldots, n-1$	μ	the order-constrained spectral kernel	[29]
$g(\lambda) = \mu_i, \mu_i \geq w\mu_{i+1}, i = 1, \ldots, q-1$	μ,$w \geq 1$	the fast-decay spectral kernel	[10]

B. Semi-supervised Spectral Kernel Learning Framework

Based on the unsupervised kernel design rule and the optimization criteria, we are able to formulate a semi-supervised kernel learning framework as follows:

$$\begin{aligned} \max_{g(\lambda)} \quad & \omega(\bar{K}) \\ s.t. \quad & \bar{K} = \sum_{i=1}^{n} g(\lambda_i)\phi_i\phi_i^T, \end{aligned} \tag{11}$$

where $\omega(\bar{K})$ is either the kernel target alignment or the soft margin. Theoretically, $g(\cdot)$ can be any function listed in Table I. In addition, it is convenient to obtain a a global optimum solution when the optimization problem is a convex programming. Especially, it is desirable that the learned kernel matrix has a fast spectral decay rate. Therefore, the fast-decay spectral kernel is considered in this framework, this leads to the following optimization problem:

$$\begin{aligned} \max_{\mu} \quad & \omega(\bar{K}) \\ s.t. \quad & \bar{K} = \sum_{i=1}^{q} \mu_i\phi_i\phi_i^T, \\ & trace(\bar{K}) = \delta, \\ & \mu_i \geq 0, \\ & \mu_i \geq w\mu_{i+1}, i = 1, \ldots, q-1, \end{aligned} \tag{12}$$

where δ is a constant, w is a pre-defined spectral decay factor that satisfies $w \geq 1$, the eigenvectors are sorted in the decreasing order of the eigenvalues and only eigenvectors corresponding to q largest eigenvalues are selected. In the case of selecting the kernel target alignment ω_A as the optimization criterion, the optimization problem is reduced to that proposed in [10].

However, the kernel designing process does not consider the bias of a kernel-based learning algorithm, the deduced kernel matrix cannot necessarily facilitate a specific learning algorithm. It is meaningful to incorporate the bias or the prior of a learning algorithm into the kernel learning process. To supplement the spectral kernel learning methods, this paper proposes to employ the margin between two classes of data, ω_M, as the optimization criterion. The resulted approach not only learns a kernel matrix, but also leads directly to a convex method for learning the margin parameters in SVMs.

C. Maximum Margin Based Spectral Kernel Learning

By maximizing the margin (Eq. (11)) between two class of data along with the above semi-supervised learning framework (Eq. (12)), we have the following semi-supervised learning problem:

$$\begin{aligned} \max_{\mu,\alpha} \quad & 2\alpha^T\mathbf{e} - \alpha^T G(\bar{K}^{tr})\alpha \\ s.t. \quad & \bar{K} = \sum_{i=1}^{d} \mu_i\phi_i\phi_i^T, \\ & trace(\bar{K}) = \delta, \\ & \alpha^T\mathbf{y} = 0, \\ & 0 \leq \alpha_j \leq C, j = 1, \ldots, n, \\ & \mu_i \geq 0, \\ & \mu_i \geq w\mu_{i+1}, i = 1, \ldots, q-1, \end{aligned} \tag{13}$$

where $G(\bar{K}^{tr}) = D(\mathbf{y})\bar{K}^{tr}D(\mathbf{y})$, $D(\mathbf{y})$ is the diagonal matrix of the label vector $\mathbf{y}$.

We note each rank-one kernel matrix $\bar{K}_i = \phi_i\phi_i^T$, then $\bar{K} = \sum_{i=1}^{q} \mu_i\bar{K}_i$. Following [14], it can be proven that the above optimization problem is able to further formulated as below:

$$\begin{aligned} \max_{\alpha,\mu} \quad & 2\alpha^T\mathbf{e} - \delta\rho \\ s.t. \quad & \delta = \mu^T\mathbf{t}, \\ & \rho \geq \frac{1}{t_i}\alpha^T G(\bar{K}_i^{tr})\alpha, \quad 1 \leq i \leq q, \\ & \mu \geq \mathbf{0}, \\ & \alpha^T\mathbf{y} = 0, \\ & 0 \leq \alpha_j \leq C, j = 1, \ldots, n, \\ & \mu_i \geq w\mu_{i+1}, i = 1, \ldots, q-1, \end{aligned} \tag{14}$$

where $G(\bar{K}_i^{tr}) = D(\mathbf{y})\bar{K}_i^{tr}D(\mathbf{y})$, and $\mathbf{t} = \{t_1, t_2, \ldots, t_q\}$ is the trace vector of K_i, i.e., $trace(\bar{K}_i) = t_i$. This is a Quadratically Constrained Quadratic Program (QCQP), which is regarded as a special form of Second Order Cone Program (SOCP) [4]. Typically, SOCP problem can be efficiently solved by interior point method [19], which is implemented in SeDumi [23].

According to Karush-Kuhn-Tucker conditions [20], [21], the discriminant function of SVM in the kernel-induced feature space is represented by the linear span of the support vectors, i.e., $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \Phi(\mathbf{x}_i)$. Thus α is sparse and only positive for the support vectors. Let the threshold b is set to 0, and then the discriminant function can be directly written as:

$$f(\mathbf{z}) = \sum_{i=1}^{n} y_i\alpha_i K(\mathbf{x}_i, \mathbf{z}), \tag{15}$$

where $\mathbf{z}$ is a test data point.

Remark. The above spectral kernel learning method not only optimizes the margin between different classes, but also solves the margin parameter α of SVM. To differentiate from the spectral kernel maximizing the kernel target alignment, we name the proposed spectral kernel as the fast-decay spectral kernel with maximum margin (abbreviated as “MM”).

IV. Experimental Results

In this section, we report the experimental results on several benchmark data sets. For performance comparison, we also implemented three competitive methods. These methods include, the standard linear kernel and RBF kernel, the order constrained spectral kernel (abbreviated as “order”) [29], and the fast-decay spectral kernel optimizing the kernel alignment (noted as “KA”) [10].

Experimental Data Sets. To make evaluations comprehensive, we have collected both the synthetic data sets and the UCI data sets [3] as our experimental test beds. Table II summarizes the information of the benchmark data sets. Two synthetic data sets described in Section 1 and four benchmark data sets from the UCI machine learning repository are employed to evaluate the performance of the proposed kernel learning algorithm.

TABLE II

Data information

Data set	# Samples	# Features	# Classes
Ionosphere	351	34	2
Banana	400	2	2
Sonar	208	60	2
Solar-flare	666	9	2

Experimental Setup. The parameters of different algorithms are set in the following. SVM is used as the classifier for evaluating all kernel matrices. To facilitate a fair comparison, we select the top 20 smallest eigenvalues and eigenvectors of the graph Laplacian, which is constructed with 10-NN unweighed graphs. Moreover, both the linear kernel and the RBF kernel are used to construct the input kernel matrix for KA and MM. The parameter $\mathbf{C}$ of SVM is fixed to 100 in the experiments. The parameter γ in RBF kernel is tuned by 10-cross validation for data sets. For synthetic data sets, the training size is set to 10. For benchmark data sets, the training size is set from 10 to 30. For each training set size, we conduct 20 random trials and each trial is conducted according to a modified 10-fold cross-validation. In each trial, the training set contains each class of data, and the remaining data are then used as the unlabeled (test) data. The spectral decay rate w and the number of eigenvectors q used in KA and MM, are selected from the range $[8, 20]$ and $[1.1, 2.0]$, respectively.

Table III reports the prediction accuracy and the standard error of classifiers for four kernel matrices on two synthetic data sets. As illustrated in Fig. 1, the spectral kernel learning method which maximizes the margin between different classes of data gets the best performance for SVM.

TABLE III

Experimental results on two synthetic data sets (%).

Algorithm	Relevance	Twocircles
RBF	81.52±4.63	78.74±5.02
Order	62.41±3.32	51.14±1.71
KA	91.27±4.57	84.10±4.44
MM	**93.15**±3.49	**94.98**±3.13

The prediction accuracy and standard errors on the benchmark data sets can be observed from Table IV, where two standard kernels and five semi-supervised kernels are compared based on SVM classifiers with different sizes of labeled data. For KA and MM, the words in the parenthesis specify the input kernel type. From the experimental results, it can be concluded that the order-constrained kernel performs slightly worse than standard kernels, though its advantage is no parameter required to be chosen. We observe that for most of the data sets, our proposed spectral kernel learning method performs better than other semi-supervised kernels and the standard kernels. Especially, for the banana data set, the improvement is larger than 10% in prediction accuracy.

V. Conclusions

In this paper, we discuss a semi-supervised spectral kernel learning framework, where previous methods do not incorporate the classifier bias into the spectral kernel learning. To supplement this framework, we have proposed a novel approach, which not only learns a kernel matrix by maximizing the margin between two classes of data, but also leads directly to a convex optimization method for learning the margin parameters in support vector machines. Experimental results on four UCI data sets have demonstrated that our proposed spectral kernel learning method achieves promising results against other spectral kernel learning methods.

One of the future work of this paper is to extend the semi-supervised kernel learning to multiway classification. Another is to apply the proposed method in large-scale text categorization and other applications, where the data sets have a cluster structure.

Acknowledgment

The work described in this paper is fully supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4205/04E and Project No. CUHK4235/04E). The authors would like to thank Dr. Steven Hoi for his fruitful discussion and providing the code of the Unified Kernel Machines.

References

[1] M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. In *COLT*, pages 624–638, 2004.

[2] J. Bi, T. Zhang, and K. P. Bennett. Column-generation boosting methods for mixture of kernels. In *KDD ’04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 521–526, New York, NY, USA, 2004. ACM Press.

[3] C. L. Blake and C. J. Merz. Repository of machine learning databases, University of California, Irvine, http://www.ics.uci.edu/∼mlearn/mlrepository.html, 1998.

TABLE IV

CLASSIFICATION PERFORMANCE OF DIFFERENT KERNELS.

Training	Standard Kernels		Semi-supervised Kernels				
Size	Linear	RBF	Order	KA (Linear)	KA (RBF)	MM (Linear)	MM (RBF)
Ionosphere (%)							
10	71.51±2.12	66.56±2.04	62.31±3.92	74.36±2.47	70.24±4.99	**74.45**±2.54	69.56±2.26
20	77.50±1.20	71.37±2.48	63.64±2.71	78.75±1.89	76.62±3.12	**78.83**±1.74	77.55±3.04
30	80.23±0.90	77.82±2.52	63.52±2.44	81.21±1.17	80.51±2.80	81.47±1.08	**82.5**9±0.96
Banana (%)							
10	53.69±1.69	55.63±2.07	50.22±0.94	53.87±1.34	62.68±2.18	53.95±1.54	**64.92**±2.26
20	55.30±1.86	58.73±2.39	50.44±0.93	54.74±1.63	66.18±2.46	55.14±1.76	**69.88**±1.87
30	56.07±2.43	60.48±1.57	50.73±0.93	55.72±1.55	69.33±1.96	56.24±2.07	**74.87**±1.33
Sonar (%)							
10	63.89±2.25	57.52±1.70	49.96±1.16	**64.30**±1.88	60.92±2.22	64.14±1.77	61.95±2.44
20	68.72±1.50	65.73±1.71	49.80±0.62	69.17±1.64	67.91±1.87	68.94±1.49	**69.18**±1.73
30	71.98±1.20	71.20±1.32	49.73±1.09	72.31±1.86	70.90±1.34	**73.22**±1.61	71.32±1.60
Solar-flare (%)							
10	55.92±1.78	56.58±2.53	51.45±1.83	57.75±2.08	57.88±2.23	**58.11**±1.92	57.95±1.93
20	59.73±1.97	60.44±2.27	51.14±1.56	60.64±1.84	60.87±1.96	60.60±1.68	**61.08**±1.77
30	61.77±1.44	61.67±1.53	50.85±2.06	**62.19**±1.01	62.14±1.42	61.95±1.21	61.75±1.11

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[5] T. P. Centeno and N. D. Lawrence. Optimising kernel parameters and regularisation coefficients for non-linear discriminant analysis. *Journal of Machine Learning Research*, 7(2):455–491, 2006.

[6] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[7] O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. volume 15 of *NIPS*, 2002.

[8] F. R. Chung. *Spectral graph theory*. the American Mathematical Society, 1997.

[9] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola. On kernel-target alignment. In *Neural Information Processing Systems (NIPS 13)*, pages 367–373, 2001.

[10] S. C. H. Hoi, M. R. Lyu, and E. Y. Chang. Learning the unified kernel machines for classification. In *Proceedings of Twentith International Conference on Knowledge Discovery and Data Mining (KDD-2006)*, pages 187–196, New York, NY, USA, 2006. ACM Press.

[11] K. Huang, H. Yang, I. King, M. R. Lyu, and L. Chan. Minimum error minimax probability machine. *Journal of Machine Learning Research*, 5:1253–1286, 2004.

[12] S.-J. Kim, A. Magnani, and S. Boyd. Optimal kernel selection in kernel Fisher discriminant analysis. In *Proceedings of the 23rd international conference on Machine learning (ICML-2006)*, pages 465–472, New York, NY, USA, 2006. ACM Press.

[13] R. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces, 2002.

[14] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[15] G. R. G. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2002.

[16] C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.

[17] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller. Fisher discriminant analysis with kernels. In *Proceedings of IEEE Neural Network for Signal Processing Workshop*, pages 41–48, 1999.

[18] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K. Müller. Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):623–633, 2003.

[19] Y. Nesterov and A. Nemirovsky. *Interior point polynomial methods in convex programming: Theory and applications*. Studies in Applied Mathematics. Philadelphia, 1994.

[20] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[21] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, UK, 2004.

[22] A. Smola and R. Kondor. Kernels and regularization on graphs, 2003.

[23] J. F. Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11:625–653, 1999.

[24] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[25] C. K. I. Williams and M. Seeger. The effect of the input density distribution on kernel-based classifiers. In *Proceedings of 17th International Conf. on Machine Learning (ICML-2000)*, pages 1159 – 1166. Morgan Kaufmann, San Francisco, CA, 2000.

[26] T. Zhang and R. Ando. Analysis of spectral kernel design based semi-supervised learning. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems (NIPS 18)*, pages 1601–1608. MIT Press, Cambridge, MA, 2006.

[27] X. Zhu. Semi-supervised learning literature survey. Technical report, Computer Sciences, University of Wisconsin-Madison, 2005.

[28] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of Twentith International Conference on Machine Learning (ICML-2003)*, pages 912–919, 2003.

[29] X. Zhu, J. Kandola, Z. Ghahramani, and J. Lafferty. Nonparametric transforms of graph kernels for semi-supervised learning. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems (NIPS 17)*, pages 1641–1648. MIT Press, Cambridge, MA, 2005.