Generalizing the Bias Term of Support Vector Machines

Wenye Li and **Kwong-Sak Leung** and **Kin-Hong Lee**

Department of Computer Science and Engineering

The Chinese University of Hong Kong

{wyli, ksleung, khlee}@cse.cuhk.edu.hk

Abstract

Based on the study of a generalized form of representer theorem and a specific trick in constructing kernels, a generic learning model is proposed and applied to support vector machines. An algorithm is obtained which naturally generalizes the bias term of SVM. Unlike the solution of standard SVM which consists of a linear expansion of kernel functions and a bias term, the generalized algorithm maps predefined features onto a Hilbert space as well and takes them into special consideration by leaving part of the space unregularized when seeking a solution in the space. Empirical evaluations have confirmed the effectiveness from the generalization in classification tasks.

1 Introduction

Support vector machines (SVM) have been shown to perform well in many machine learning applications. Based on Vapnik's seminal work in statistical learning theory[Vapnik, 1998], the algorithm starts with the ideas of separating hyperplanes and margins. Given the data $(\mathbf{x}_i; y_i)_{i=1}^m$ where $\mathbf{x}_i \in \mathcal{R}^d$ and $y_i \in \{+1, -1\}$, one searches for the linear hyperplane that separates the positive and negative samples with the largest margin (the distance from the hyperplane to the nearest data point). In the nonseparable case, a soft margin method is used to choose a hyperplane that splits the examples as cleanly as possible. Using a kernel $K_{\mathbf{x}}(\mathbf{x}') = K(\mathbf{x}, \mathbf{x}')$ (a positive definite function) to extend the algorithm to the nonlinear case, an optimal hyperplane is found which has a simple representation as a linear expansion of kernel functions and a constant

$$f^* = \sum_{i=1}^m c_i K_{\mathbf{x}_i} + b \qquad (1)$$

where each c_i is a real number and b is called a bias term.

In the case of nonseparability, some meanings of the margin motivation are lost. [Girosi, 1998; Evgeniou *et al.*, 2000; Poggio and Smale, 2003] suggest a different view of SVM based on a framework championed by Poggio and other researchers[Poggio and Girosi, 1990a; 1990b] which implicitly treats learning as an approximation problem and tries to

find a solution to a regularized minimization problem

$$\min_{f \in \mathcal{H}_K} \frac{1}{m} \sum_{i=1}^m V(y_i, f(\mathbf{x}_i)) + \gamma \|f\|_K^2 \qquad (2)$$

in $\mathcal{H}_K$ for some $\gamma > 0$. This expression represents a tradeoff between the empirical error as calculated by the loss function V, and "smoothness" of the solution as represented by the norm of f in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ which is induced by a kernel K. $\gamma \|f\|_K^2$ is also called a regularizer. SVM can be partially derived from this framework with the choice of a hinge loss function $V(y, f(\mathbf{x})) \equiv \max(1 - yf(\mathbf{x}), 0)$. The only *incompatibility* comes from the bias term b, which makes the combined model (1) generally not in $\mathcal{H}_K$ any more.

In this paper, we follow the regularization point of view. Based on the study of a generalized regularizer which leaves part of the hypothesis space unregularized and a specific trick in constructing kernels, we propose a new learning scheme which allows user to predefine features, and uses these features and kernel expansions to derive a solution in the hypothesis space, instead of considering kernel expansions only as in existing kernel-based methods. When the idea is applied to SVM, we obtain an algorithm which naturally generalizes the bias term of SVM. The generalized term linearly combines the predefined features instead of being a constant only. Different from the empirical results that the existence of the bias term does not make much significance in practice[Rifkin, 2002], we will show the term is not trivial any more when it is generalized, and appropriate choices of the term will improve the applicability of the algorithm.

The paper is organized as follows. In section 2, we investigate a generalized regularizer in the regularized learning framework and study its associated solution when it is applied to SVM. In section 3, after introducing a kernel construction trick, with which predefined features are mapped onto an RKHS, we further clarify the mathematical details in deriving SVM. A new algorithm is proposed in section 4 based on the previous discussions. Empirical results are demonstrated in section 5. Finally, section 6 presents our discussions and conclusions.

2 Generalized Regularized Learning

Suppose $\mathcal{H}_K$ is the direct sum of two subspaces: $\mathcal{H}_K = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0 = span(\varphi_1, \cdots, \varphi_\ell)$ is spanned by $\ell (\leq m)$ linearly independent features. We consider the generalized regularized learning by minimizing

$$\min_{f \in \mathcal{H}_K} \frac{1}{m} \sum_{i=1}^{m} V\left(y_i, f\left(\mathbf{x}_i\right)\right) + \gamma \left\|P_1 f\right\|_K^2 \qquad (3)$$

where $P_1 f$ is the orthogonal projection of f onto $\mathcal{H}_1$ and $\gamma \left\|P_1 f\right\|_K^2$ is called a generalized regularizer. When the model is applied to SVM, we consider V as the hinge loss function mentioned above. By introducing slake variables ξ_i corresponding to the empirical error at point $\mathbf{x}_i$, our problem becomes

$$\min_{f \in \mathcal{H}_K} \frac{1}{m} \sum_{i=1}^{m} \xi_i + \gamma \left\langle P_1 f, P_1 f \right\rangle_K$$

satisfying

$$\begin{aligned} y_i f\left(\mathbf{x}_i\right) &\geq & 1 - \xi_i \\ \xi_i &\geq & 0 \end{aligned}$$

where $\langle \cdot, \cdot \rangle_K$ denotes the inner product in $\mathcal{H}_K$. We derive the dual quadratic program by using the technique of Lagrange multipliers:

$$\begin{aligned} L &=& \frac{1}{m} \sum_{i=1}^{m} \xi_i + \gamma \left\langle P_1 f, P_1 f \right\rangle_K \qquad (4) \\ && - \sum_{i=1}^{m} \alpha_i \left(y_i f\left(\mathbf{x}_i\right) - 1 + \xi_i\right) - \sum_{i=1}^{m} \zeta_i \xi_i. \end{aligned}$$

We want to minimize L with respect to f and ξ_i and maximize w.r.t. α_i and ζ_i, subject to the constraints of the primal problem and nonnegativity constraints on α_i and ζ_i. Taking derivative w.r.t. ξ_i and setting it to zero, we have

$$\frac{\partial L}{\partial \xi_i} = \frac{1}{m} - \alpha_i - \zeta_i = 0. \qquad (5)$$

Substituting (5) into (4), we have

$$L = \gamma \left\langle P_1 f, P_1 f \right\rangle_K - \sum_{i=1}^{m} \alpha_i y_i f\left(\mathbf{x}_i\right) + \sum_{i=1}^{m} \alpha_i. \qquad (6)$$

Suppose f^* is the minimizer to (6). For any $f \in \mathcal{H}_K$, let $f = f^* + \delta g$ where $\delta \in \mathcal{R}$ and $g \in \mathcal{H}_K$. We have

$$\begin{aligned} L &=& \gamma \left\langle P_1 f^* + \delta P_1 g, P_1 f^* + \delta P_1 g \right\rangle_K \\ && - \sum_{i=1}^{m} \alpha_i y_i \left(f^* + \delta g\right)\left(\mathbf{x}_i\right) + \sum_{i=1}^{m} \alpha_i \end{aligned}$$

By taking derivative w.r.t. δ, we have

$$\frac{\partial L}{\partial \delta} = 2\gamma \left\langle P_1 f^* + \delta P_1 g, P_1 g \right\rangle_K - \sum_{i=1}^{m} \alpha_i y_i g\left(\mathbf{x}_i\right).$$

Since f^* is the minimizer, we have $\frac{\partial L}{\partial \delta}|_{\delta=0} = 0$. Then the problem becomes

$$2\gamma \left\langle P_1 f^*, P_1 g \right\rangle_K - \sum_{i=1}^{m} \alpha_i y_i g\left(\mathbf{x}_i\right) = 0.$$

The equation holds for all $g \in \mathcal{H}_K$. Specifically, letting $g = K_{\mathbf{x}}(\cdot)$ gives

$$P_1 f^* = \frac{1}{2\gamma} \sum_{i=1}^{m} \alpha_i y_i K_{\mathbf{x}_i}$$

by the reproducing property of kernels. Or,

$$f^* = (f^* - P_1 f^*) + \frac{1}{2\gamma} \sum_{i=1}^{m} \alpha_i y_i K_{\mathbf{x}_i}.$$

$f^* - P_1 f^*$ is the orthogonal projection of f^* onto H_0, and hence it can be represented as $\sum_{p=1}^{\ell} \lambda_p \varphi_p$. So we have,

$$f^* = \sum_{p=1}^{\ell} \lambda_p \varphi_p + \sum_{i=1}^{m} c_i K_{\mathbf{x}_i} \qquad (7)$$

where $\lambda_p \in \mathcal{R}$ and $c_i = \frac{1}{2\gamma} \alpha_i y_i$.

The derived minimizer in (7) satisfies the *reproduction* property. Suppose $(\mathbf{x}_i; y_i)_{i=1}^{m}$ comes from a model that is perfectly linearly related to $\{\varphi_1, \cdots, \varphi_\ell\}$, it is desirable to get back a solution independent of other features. As an evident result of (3), the property is satisfied. The parameters $c_1, \cdots, c_m$ in (7) will all be zero, which makes the regularizer in (3) equal to zero. As will be shown in the experiments, this property often has the effect of *stabilizing* the results from different choices of kernels and regularization parameters practically.

3 Mapping Predefined Features onto RKHS

By decomposing a hypothesis space $\mathcal{H}_K$ and studying a generalized regularizer, we have proposed a generalized regularized learning model and derived its associated solution which consists of a linear expansion of kernel functions and predefined features. In this section, we will introduce a kernel construction trick which maps kernel functions and predefined features simultaneously onto a Hilbert space. Then we will show the mathematical details in deriving SVM with a generalized bias term.

3.1 A Kernel Construction Trick

Given features $\{\varphi_1, \cdots, \varphi_\ell\}$ and a strictly positive definite function Φ, let us consider the following reproducing kernel

$$K\left(\mathbf{x}, \mathbf{y}\right) = H\left(\mathbf{x}, \mathbf{y}\right) + \sum_{p=1}^{\ell} \varphi_p'\left(\mathbf{x}\right) \varphi_p'\left(\mathbf{y}\right) \qquad (8)$$

where

$$H\left(\mathbf{x}, \mathbf{y}\right) \qquad (9)$$

$$\begin{aligned} =& \Phi\left(\mathbf{x}, \mathbf{y}\right) + \sum_{p=1}^{\ell} \sum_{q=1}^{\ell} \varphi_p'\left(\mathbf{x}\right) \varphi_q'\left(\mathbf{y}\right) \Phi\left(\mathbf{x}_p, \mathbf{x}_q\right) \\ & - \sum_{p=1}^{\ell} \varphi_p'\left(\mathbf{x}\right) \Phi\left(\mathbf{x}_p, \mathbf{y}\right) - \sum_{q=1}^{\ell} \varphi_q'\left(\mathbf{y}\right) \Phi\left(\mathbf{x}, \mathbf{x}_q\right), \end{aligned}$$

$\{\varphi_1', \cdots, \varphi_\ell'\}$ defines a linear transformation of the predefined features $\{\varphi_1, \cdots, \varphi_\ell\}$ w.r.t. $\{\mathbf{x}_1, \cdots \mathbf{x}_\ell\}$:

$$
\begin{bmatrix} \varphi_1'(\mathbf{x}) \\ \cdots \\ \varphi_\ell'(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \varphi_1(\mathbf{x}_1) & \cdots & \varphi_1(\mathbf{x}_\ell) \\ \cdots & & \cdots \\ \varphi_\ell(\mathbf{x}_1) & \cdots & \varphi_\ell(\mathbf{x}_\ell) \end{bmatrix}^{-1} \begin{bmatrix} \varphi_1(\mathbf{x}) \\ \cdots \\ \varphi_\ell(\mathbf{x}) \end{bmatrix}
\tag{10}
$$

and satisfies

$$
\varphi_q'(\mathbf{x}_p) = \begin{cases} 1 & 1 \le p = q \le \ell \\ 0 & 1 \le p \ne q \le \ell \end{cases}.
\tag{11}
$$

This trick was studied in [Light and Wayne, 1999] to provide an alternative basis for radial basis functions and first used in a fast RBF interpolation algorithm[Beatson *et al.*, 2000]. A sketch of properties peripheral to our concerns includes

$$
K_{\mathbf{x}_p} = \varphi_p'
\tag{12}
$$

$$
\langle \varphi_p', \varphi_q' \rangle_K = \begin{cases} 1 & p = q \\ 0 & p \ne q \end{cases}
\tag{13}
$$

$$
H_{\mathbf{x}_p} = 0
\tag{14}
$$

$$
\langle H_{\mathbf{x}_i}, \varphi_p' \rangle_K = 0
\tag{15}
$$

$$
\langle H_{\mathbf{x}_i}, H_{\mathbf{x}_j} \rangle_K = H(\mathbf{x}_i, \mathbf{x}_j)
\tag{16}
$$

where $1 \le p, q \le \ell$, and $\ell + 1 \le i, j \le m$. Another useful property is that the matrix $\mathbf{H} = (H(\mathbf{x}_i, \mathbf{x}_j))_{i,j=\ell+1}^{m}$ is strictly positive definite, which will be used in the following computations.

By property (12), we can see that predefined features $\{\varphi_1, \cdots, \varphi_\ell\}$ are explicitly mapped onto $\mathcal{H}_K$, which has a subspace $\mathcal{H}_0 = span(\varphi_1', \cdots, \varphi_\ell') = span(\varphi_1, \cdots, \varphi_\ell)$. By property (13), $\{\varphi_1', \cdots, \varphi_\ell'\}$ also forms an orthonormal basis of $\mathcal{H}_0$.

3.2 Computation

Using the kernel defined in (8) and (9) and its properties in (12) and (14), the minimizer in (7) can be rewritten as:

$$
\begin{aligned}
f^* &= \sum_{p=1}^{\ell} \lambda_p \varphi_p + \sum_{i=1}^{m} c_i K_{\mathbf{x}_i} \\
&= \cdots \\
&= \sum_{p=1}^{\ell} \tilde{\lambda}_p \varphi_p' + \sum_{i=\ell+1}^{m} \tilde{c}_i H_{\mathbf{x}_i}
\end{aligned}
\tag{17}
$$

where $\tilde{\lambda}_1, \cdots, \tilde{\lambda}_\ell, \tilde{c}_{\ell+1}, \cdots, \tilde{c}_m$ are m parameters to be determined. Furthermore, from the orthogonal property (15) between φ_p' and $H_{\mathbf{x}_i}$, we have

$$
P_1 f^* = \sum_{i=\ell+1}^{m} \tilde{c}_i H_{\mathbf{x}_i}.
\tag{18}
$$

By property (16), we have

$$
\langle P_1 f^*, P_1 f^* \rangle_K = \tilde{\mathbf{c}}^T \mathbf{H} \tilde{\mathbf{c}}
$$

where $\tilde{\mathbf{c}} = (\tilde{c}_{\ell+1}, \cdots, \tilde{c}_m)^T$. Substituting it into (6), we obtain in a matrix representation

$$
L = \gamma \tilde{\mathbf{c}}^T \mathbf{H} \tilde{\mathbf{c}} - \alpha_1^T \mathbf{Y}_1 \mathbf{E}_1^T \tilde{\lambda} - \alpha_2^T \mathbf{Y}_2 \left(\mathbf{E}_2^T \tilde{\lambda} + \mathbf{H} \tilde{\mathbf{c}} \right) + \mathbf{1}^T \alpha
\tag{19}
$$

where $\tilde{\lambda} = \left(\tilde{\lambda}_1, \cdots, \tilde{\lambda}_\ell \right)^T$, $\alpha = (\alpha_1, \cdots, \alpha_m)^T$, $\alpha_1 = (\alpha_1, \cdots, \alpha_\ell)^T$, $\alpha_2 = (\alpha_{\ell+1}, \cdots, \alpha_m)^T$, $\mathbf{Y}_1 = diag(y_1, \cdots, y_\ell)$, $\mathbf{Y}_2 = diag(y_{\ell+1}, \cdots, y_m)$, $\mathbf{E}_1 = \left(\varphi_p'(\mathbf{x}_i) \right)_{p=1, i=1}^{\ell, \ell}$, $\mathbf{E}_2 = \left(\varphi_p'(\mathbf{x}_i) \right)_{p=1, i=\ell+1}^{\ell, m}$ and $\mathbf{1}$ is a vector of ones with appropriate size. Taking derivative w.r.t. $\tilde{\lambda}$, we obtain

$$
\mathbf{E}_1 \mathbf{Y}_1 \alpha_1 + \mathbf{E}_2 \mathbf{Y}_2 \alpha_2 = \mathbf{0}.
\tag{20}
$$

Then, we have

$$
L = \gamma \tilde{\mathbf{c}}^T \mathbf{H} \tilde{\mathbf{c}} - \alpha_2^T \mathbf{Y}_2 \mathbf{H} \tilde{\mathbf{c}} + \mathbf{1}^T \alpha
\tag{21}
$$

Taking derivative w.r.t. $\tilde{\mathbf{c}}$ and setting it to zero, we have

$$
2\gamma \mathbf{H} \tilde{\mathbf{c}} - \mathbf{H} \mathbf{Y}_2 \alpha_2 = \mathbf{0}.
$$

$\mathbf{H}$ is strictly positive definite and hence invertible. So we have

$$
\tilde{\mathbf{c}} = \frac{\mathbf{Y}_2 \alpha_2}{2\gamma}.
\tag{22}
$$

Substituting it back into (21), we have

$$
\begin{aligned}
L &= \frac{1}{4\gamma} \alpha_2^T \mathbf{Y}_2^T \mathbf{H} \mathbf{Y}_2 \alpha_2 - \frac{1}{2\gamma} \alpha_2^T \mathbf{Y}_2^T \mathbf{H} \mathbf{Y}_2 \alpha_2 + \mathbf{1}^T \alpha \\
&= -\frac{1}{2\gamma} \left(\frac{1}{2} \alpha_2^T \mathbf{Y}_2^T \mathbf{H} \mathbf{Y}_2 \alpha_2 - 2\gamma \mathbf{1}^T \alpha \right)
\end{aligned}
$$

So, the problem becomes

$$
\min_{\alpha} \frac{1}{2} \alpha_2^T \mathbf{Y}_2^T \mathbf{H} \mathbf{Y}_2 \alpha_2 - 2\gamma \mathbf{1}^T \alpha
\tag{23}
$$

satisfying

$$
\mathbf{0} \le \alpha \le \frac{1}{\mathbf{m}}, \ and \ \mathbf{E}_1 \mathbf{Y}_1 \alpha_1 + \mathbf{E}_2 \mathbf{Y}_2 \alpha_2 = \mathbf{0}.
\tag{24}
$$

The first box constraint comes from (5) and the nonnegativity of α, requiring $0 \le \alpha_i \le \frac{1}{m}$ $(1 \le i \le m)$. The second equality constraint comes from (20). $\mathbf{Y}_2^T \mathbf{H} \mathbf{Y}_2$ is a strictly positive definite matrix, and the problem becomes a standard quadratic program(QP). Comparing with SVM's QP[Vapnik, 1998], the new problem requires ℓ equality constraints instead of one equality constraint. This seems to burden the computations. Actually, because the major computations in solving the QP come from the box constraint, these equality constraints will not affect the computations much.

The solution of $\tilde{\mathbf{c}}$ is obtained after α by (22). Then $\tilde{\lambda}$ comes from a linear program,

$$
\min_{\tilde{\lambda}} \frac{1}{m} \sum_{i=1}^{m} \xi_i
\tag{25}
$$

satisfying

$$
\begin{aligned}
y_i \left(\sum_{p=1}^{\ell} \tilde{\lambda}_p \varphi_p + \sum_{j=\ell+1}^{m} \tilde{c}_j H_{\mathbf{x}_j} \right)(\mathbf{x}_i) &\ge 1 - \xi_i, \\
\xi_i &\ge 0.
\end{aligned}
$$

4 A Generalized Algorithm

4.1 Algorithm

Based on the discussions above, an algorithm which generalizes the bias term of SVM is proposed as follows:

1. Start with data $(\mathbf{x}_i; y_i)_{i=1}^m$.

2. For $\ell \, (\le m)$ predefined linearly independent features $\{\varphi_1, \cdots, \varphi_\ell\}$ of the data, define $\{\varphi_1', \cdots, \varphi_\ell'\}$ according to equation (10).

3. Choose a symmetric, strictly positive definite function $\Phi_{\mathbf{x}}(\mathbf{x}') = \Phi(\mathbf{x}, \mathbf{x}')$ which is continuous on $\mathcal{R}^d \times \mathcal{R}^d$. Get $H_{\mathbf{x}}$ according to equation (9).

4. Define $f : \mathcal{R}^d \to \mathcal{R}$ by

$$f(\mathbf{x}) = \sum_{p=1}^\ell \tilde{\lambda}_p \varphi_p'(\mathbf{x}) + \sum_{i=\ell+1}^m \tilde{c}_i H_{\mathbf{x}_i}(\mathbf{x}), \quad (26)$$

where the solution of $\tilde{c}_{\ell+1}, \cdots, \tilde{c}_m$ comes from a quadratic program (23) and equation (22), and $\tilde{\lambda}_1, \cdots, \tilde{\lambda}_\ell$ is obtained by solving a linear program (25).

4.2 Virtual Samples

The new algorithm needs the construction of an orthonormal basis $\{\varphi_1', \cdots, \varphi_\ell'\}$ from predefined features $\{\varphi_1, \cdots, \varphi_\ell\}$ w.r.t. ℓ samples via a linear transformation. However, algorithms, which search for linear hyperplanes to separate the data, are often not invariant to linear transformations. To provide the invariance, we consider an alternative approach that avoids the transformation.

We introduce ℓ virtual points $\mathbf{x}_{v_1}, \cdots, \mathbf{x}_{v_\ell}$ satisfying

$$\varphi_p(\mathbf{x}_{v_q}) = \begin{cases} 1 & 1 \le p = q \le \ell \\ 0 & 1 \le p \neq q \le \ell \end{cases}.$$

The label y_{v_q} of each virtual point $\mathbf{x}_{v_q}$ is defined to be 0. We study the following problem:

$$\min_{f \in \mathcal{H}_K} L_v(f) \quad (27)$$

where

$$
\begin{aligned}
L_v(f) &= \frac{1}{m} \sum_{q=1}^\ell \left(1 - y_{v_q} f(\mathbf{x}_{v_q})\right)_+ \\
&\quad + \frac{1}{m} \sum_{i=1}^m \left(1 - y_i f(\mathbf{x}_i)\right)_+ + \gamma \|P_1 f\|_K^2 \\
&= \frac{\ell}{m} + \frac{1}{m} \sum_{i=1}^m \left(1 - y_i f(\mathbf{x}_i)\right)_+ + \gamma \|P_1 f\|_K^2.
\end{aligned}
$$

It can be easily seen that $\varphi_1, \cdots, \varphi_\ell$ have formed an orthonormal basis of $\mathcal{H}_0$ w.r.t. the virtual points and there is no need to compute $\varphi_1', \cdots, \varphi_\ell'$ any more. The new regularized minimization problem in (27) is equivalent to the original one in (3) with a hinge loss. The introduction of the virtual points does not change the scope of the classification problem while having avoided the linear transformation to construct an orthonormal basis.
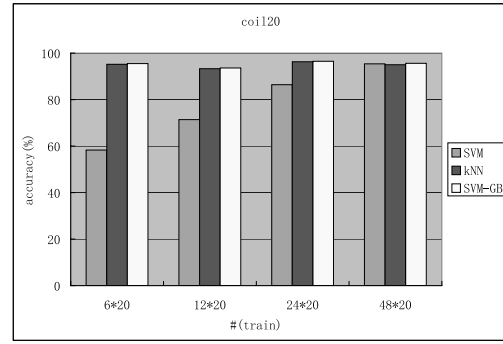

Figure 1: Image classification accuracies using *coil20* dataset with different number of training images. Twenty predefined features along with a Gaussian kernel Φ are used by SVM-GB; Φ is also used by SVM. The kernel and regularization parameters are selected via cross validation. One-versus-one strategy is used for multi-classification.

5 Experiments

To evaluate the effectiveness brought by the generalized bias term, two groups of empirical results are reported. The first group of experiments focused on the *reproduction* property of the generalized bias term. It used Columbia university image library (*coil-20*) dataset[1] for image classification. The dataset has $1,440$ images $\mathbf{x}_1, \cdots, \mathbf{x}_{1440}$ evenly distributed in 20 classes. After preprocessing, each image was represented as a $32 \times 32 = 1024$ dimensional vector with each component having a value between 0 and 255 indicating the gray level at each pixel. Four experiments were performed. Each experiment used the first 6, 12, 24 and 48 images respectively within each class for training and the rest for testing.

Using recently developed nonlinear dimensionality reduction technqiues[Tenenbaum *et al.*, 2000; Roweis and Saul, 2000], we found that these images actually form a manifold with a much lower dimension than the original $1,024$ dimensions. With the prior knowledge, twenty features $\varphi_1, \cdots, \varphi_{20}$ were predefined as follows:

$$\varphi_p(\mathbf{x}_i) = \begin{cases} 1 & 3NN(\mathbf{x}_i) \cap train_p \neq \{\} \\ 0 & otherwise \end{cases}$$

where $1 \le p \le 20$, $1 \le i \le 1440$, $3NN(\mathbf{x})$ defines the three nearest neighbors in geodesic distance to $\mathbf{x}$, and $train_p$ is the set of training images with class label p. The geodesic distance was computed by the graph shortest path algorithm as in [Tenenbaum *et al.*, 2000]. Defining features in this way, we implicitly used a kNN classifier as a generalized bias term. Similar ideas can be generalized to combining different approaches other than kNN.

Figure 1 compares classification accuracies among SVM, kNN, and the new algorithm (denoted by SVM-GB). An insightful study of the results may find that the 20 predefined features have dominated the learned SVM-GB solution due to the *reproduction* property, and made the performance of the algorithm similar to kNN which shows better performance than standard SVM on this dataset.

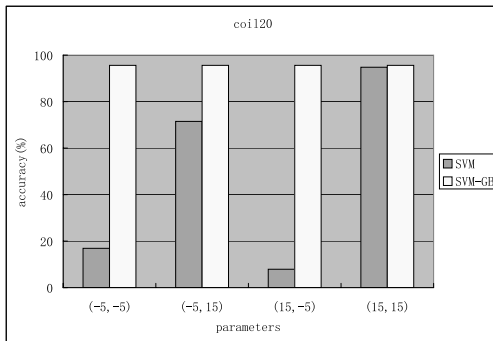[1]http://www1.cs.columbia.edu/CAVE/software/

Figure 2: Image classification accuracies by fixing four groups of kernel and regularization parameters(log-scale).

The *reproduction* property also helps to lessen the sensitivity of the algorithm to the choices of kernels and the regularization parameter γ. To illustrate the effect, another experiment was conducted using 48×20 training images. Instead of using cross validation for parameter selection, four groups of kernel and regularization parameters were fixed. Figure 2 depicts the results. The generalized bias term makes SVM-GB more stable to the variations of kernel and regularization parameters. This property is especially useful concerning the practical situations where *arbitrariness* exists in designing kernels for some applications.

To further study the performance of the new algorithm, the second group of experiments was conducted on text categorization tasks using *20-newsgroups* dataset[2]. The dataset collects UseNet postings into twenty newsgroups and each group has about $1,000$ messages. We experimented with its four major subsets. The first subset has five groups (comp.*), the second four groups (rec.*), the third four groups (sci.*) and the last four groups (talk.*).

For the dataset, we removed all but the $2,000$ words with highest mutual information with the class variable by rainbow package[McCallum, 1996]. Each document was represented by *bag-of-words* (*BoW*) which treats individual words as features. A linear kernel coupled with 10 predefined features $\varphi_1, \cdots, \varphi_{10}$, which were obtained by *probabilistic latent semantic analysis* (*pLSA*) [Hofmann, 1999], was used as input for SVM-GB. Experiments were carried out with different number $(100 \tilde{~} 3, 200)$ of training documents for each subset. One experiment consisted of ten runs and the average accuracy was reported. In each run, the data were separated by the *xval-prep* utility accompanied in *C4.5* package[3].

As a comparison, SVM was tested with a linear kernel, which is a commonly used method for text categorizations. For completeness, both *BoW* and *pLSA* representations of documents were experimented. As can been seen in figure 3, although SVM with *pLSA* representation performs well when the number of training samples is small, the approach loses its advantage when the training set increases to a moderate size. SVM-GB reports the best performance when we have 800 training documents or more. It is also shown that SVM-

GB has better results than SVM in *BoW* representation in all experiments, which gives us the belief that the embedding of *pLSA* features improves the accuracy.

6 Discussion and Conclusion

The idea of regularized learning can be traced back to modern regularization theory[Tikhonov and Arsenin, 1977; Morozov, 1984], which states that the existence of a regularizer helps to provide a unique stable solution to ill-posed problems. In this paper, we have considered the usage of a generalized regularizer in kernel-based learning. One straightforward reason that part of the hypothesis space is left unregularized in the new regularizer is due to a well-accepted fact that finite dimensional problems are often well-posed[Bertero *et al.*, 1988; De Vito *et al.*, 2005] and do not always need regularization, as in the case of learning problems defined in a subspace spanned by a number of predefined features. Similar idea was explored in spline smoothing[Wahba, 1990] that leaves polynomial features unregularized. In kernel-based learning, the most similar to our work is the semiparametric SVM model[Smola *et al.*, 1998], which combines a set of unregularized basis functions with SVM. However, to our knowledge, few algorithms and applications have been investigated for machine learning tasks from a unified RKHS regularization viewpoint in a general way.

With the new regularizer and a specific trick in constructing kernels, predefined features are explicitly mapped onto a Hilbert space and taken into special consideration during the learning, which differentiates our work from standard kernel-based approaches that only consider kernel expansions and a constant in learning. For projecting predefined features onto an RKHS, the idea of a conditionally positive definite function[Micchelli, 1986] is lurking in the background, which goes beyond the discussion of this paper.

When applying the idea to SVM, we have developed a new algorithm which can be regarded as having generalized the bias term of SVM. With domain specific knowledge in predefining features, this generalization makes the bias term not trivial any more. As confirmed by the experimental results, correct choices of features help to improve the accuracy and make the algorithm more *stable* in terms of sensitivity to the selection of kernels and regularization parameters.

Acknowledgments

References

[Beatson *et al.*, 2000] R.K. Beatson, W.A. Light, and S. Billings. Fast solution of the radial basis function interpolation equations: Domain decomposition methods. *SIAM J. Sci. Comput.*, 22:1717–1740, 2000.

[Bertero *et al.*, 1988] M. Bertero, C. De Mol, and E.R. Pike. Linear inverse problems with discrete data: II. stability and regularisation. *Inverse Probl.*, 4(3):573–594, 1988.

---

[2]http://www.cs.cmu.edu/~TextLearning/datasets.html
[3]http://www.rulequest.com/Personal/c4.5r8.tar.gz

[De Vito *et al.*, 2005] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *J. Mach. Learn. Res.*, 6:883–904, 2005.

[Evgeniou *et al.*, 2000] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Adv. Comput. Math.*, 13:1–50, 2000.

[Girosi, 1998] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Comput.*, 10(6):1455–1480, 1998.

[Hofmann, 1999] T. Hofmann. Probabilistic latent semantic analysis. In *UAI'99*, 1999.

[Light and Wayne, 1999] W. Light and H. Wayne. Spaces of distributions, interpolation by translates of a basis function and error estimates. *J. Numer. Math.*, 81:415–450, 1999.

[McCallum, 1996] A.K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/~mccallum/bow, 1996.

[Micchelli, 1986] C.A. Micchelli. Interpolation of scattered data: Distances, matrices, and conditionally positive definite functions. *Constr. Approx.*, 2:11–22, 1986.

[Morozov, 1984] V.A. Morozov. *Methods for Solving Incorrectly Posed Problems*. Springer-Verlag, 1984.

[Poggio and Girosi, 1990a] T. Poggio and F. Girosi. Networks for approximation and learning. *Proc. IEEE*, 78:1481–1497, 1990.

[Poggio and Girosi, 1990b] T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.

[Poggio and Smale, 2003] T. Poggio and S. Smale. The mathematics of learning: Dealing with data. *Not. Am. Math. Soc*, 50:537–544, 2003.

[Rifkin, 2002] R.M. Rifkin. *Everything Old is New Again: A Fresh Look at Historical Approaches in Machine Learning*. PhD thesis, Massachusetts Institute of Technology, 2002.

[Roweis and Saul, 2000] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[Smola *et al.*, 1998] A.J. Smola, T.T. Frieß, and B. Schölkopf. Semiparametric support vector and linear programming machines. In *NIPS 11*, 1998.

[Tenenbaum *et al.*, 2000] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[Tikhonov and Arsenin, 1977] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-Posed Problems*. Winston and Sons, 1977.

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

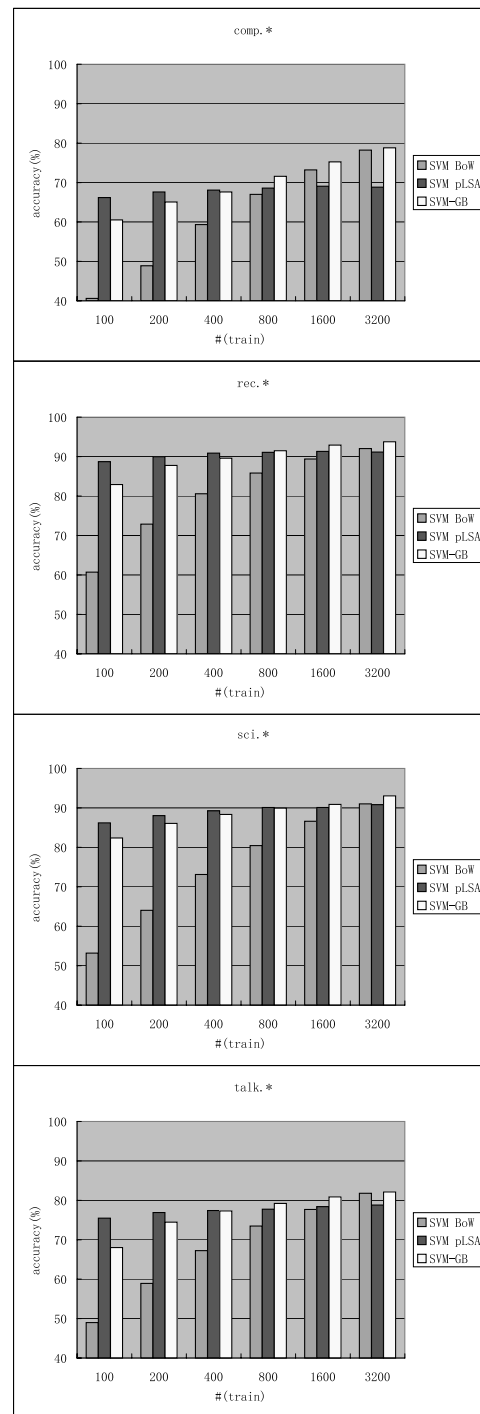[Wahba, 1990] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

Figure 3: Text categorization accuracies on four major subsets of *20-newsgroups* dataset. The classification accuracies are reported by trying training sets with different sizes.