Rotational Prior Knowledge for SVMs

Arkady Epshteyn and Gerald DeJong

University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
{aepshtey, dejong}@uiuc.edu

Abstract. Incorporation of prior knowledge into the learning process can significantly improve low-sample classification accuracy. We show how to introduce prior knowledge into linear support vector machines in form of constraints on the rotation of the normal to the separating hyperplane. Such knowledge frequently arises naturally, e.g., as inhibitory and excitatory influences of input variables. We demonstrate that the generalization ability of rotationally-constrained classifiers is improved by analyzing their VC and fat-shattering dimensions. Interestingly, the analysis shows that large-margin classification framework justifies the use of stronger prior knowledge than the traditional VC framework. Empirical experiments with text categorization and political party affiliation prediction confirm the usefulness of rotational prior knowledge.

1 Introduction

Support vector machines (SVMs) have outperformed competing classifiers on many classification tasks [1,2,3]. However, the amount of labeled data needed for SVM training can be prohibitively large for some domains. Intelligent user interfaces, for example, must adopt to the behavior of an individual user after a limited amount of interaction in order to be useful. Medical systems diagnosing rare diseases have to generalize well after seeing very few examples. Natural language processing systems learning to identify infrequent social events (e.g., revolutions, wars, etc.) from news articles have access to very few training examples. Moreover, they rely on manually labeled data for training, and such data is often expensive to obtain. Various techniques have been proposed specifically to deal with the problem of learning from very small datasets. These include active learning [4], hybrid generative-discriminative classification [5], learning-to-learn by extracting common information from related learning tasks [6], and using prior knowledge.

In this work, we focus on the problem of using prior knowledge to increase the accuracy of a large margin classifier at low sample sizes. Several studies have shown the efficacy of this method. Scholkopf et. al. [7] demonstrate how to integrate prior knowledge about invariance under transformations and importance of local structure into the kernel function. Fung et. al. [8] use domain knowledge in form of labeled polyhedral sets to augment the training data. Wu and Srihari [9] allow human users to specify their confidence in the example's label, varying the effect of each example on the separating hyperplane proportionately to its confidence. Mangasarian et. al. [10] introduce prior knowledge into the large-margin regression framework.

J. Gama et al. (Eds.): ECML 2005, LNAI 3720, pp. 108–119, 2005.

While the ability of prior knowledge to improve any classifier's generalization performance is well-known, the properties of large margin classifiers with prior knowledge are not well understood. In order to study this problem, we introduce a new form of prior knowledge for SVMs (rotational constraints) and prove that it is possible to obtain stronger guarantees for the generalization ability of constrained classifiers in the large-margin framework than in the classical VC framework. Specifically, we show that the VC dimension of our classifier remains large even when its hypothesis space is severely constrained by prior knowledge. The fat-shattering dimension, however, continues to decrease with decreasing hypothesis space, justifying the use of stronger domain knowledge. We conduct experiments to demonstrate improvements in performance due to rotational prior knowledge and compare them with improvements achievable by active learning.

2 Preliminaries

The SVM classifier with a linear kernel learns a function of the form

$$sign(f(x;\omega,\theta) = \omega^T x + \theta = \sum_{i=1}^{n} \omega_i x_i + \theta)^1 \tag{1}$$

that maps $(x;\omega,\theta) \in \mathbb{R}^n \mathrm{x} W, \Theta$ to one of the two possible output labels $\{1,-1\}$. Given a training sample of m points $(x_1,y_1)...(x_m,y_m)$, SVM seeks to maximize the margin between the separating hyperplane and the points closest to it [1]. For canonical hyperplanes (i.e., hyperplanes with unit margins), the maximum-margin hyperplane minimizes the regularized risk functional

$$R_{reg}[f,l] = \frac{1}{m} \sum_{i=1}^{m} l(y_i, f(x_i;\omega,\theta)) + \frac{C_1}{2} \|\omega\|_2^2 \tag{2}$$

with hard margin 0-1 loss given by $l(y_i, f(x_i;\omega,\theta)) = I_{\{-y_i f(x_i;\omega,\theta)>0\}}$.

The soft margin formulation allows for deviation from the objective of maximizing the margin in order to better fit the data. This is done by substituting the hinge loss function $l(y_i, f(x_i;\omega,\theta)) = max(1 - y_i f(x_i;\omega,\theta), 0)$ into (2).

Minimizing the regularized risk (2) in the soft margin case is equivalent to solving the following (primal) optimization problem:

$$\underset{\omega,\theta,\xi}{minimize}\ \tfrac{1}{2}\|\omega\|_2^2 + \tfrac{1}{C_1 m}\sum_{i=1}^{m}\xi_i\ \ subj.to\ \ y_i(\omega^T x + \theta) \geq 1 - \xi_i,\ i = 1...m \tag{3}$$

[1] $sign(y) = 1$ if $y \geq 0$, -1 otherwise

Calculating the Wolfe dual from (3) and solving the resulting maximization problem:

$$\underset{\alpha}{maximize} \sum_{i=1}^{m} \alpha_i - \tfrac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j) \tag{4}$$

$$subject\,to \quad \tfrac{1}{Cm} \geq \alpha_i \geq 0,\ i = 1...m \qquad and \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$\text{yields the solution } \omega = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{5}$$

Setting $\xi_i = 0$, $i = 1...m$, (3), (4), and (5) can be used to define and solve the original hard margin optimization problem.

The generalization error of a classifier is governed by its VC dimension [1]:

Definition 1. *A set of points $S = \{x^1...x^m\}$ is shattered by a set of functions F mapping from a domain X to $\{-1, 1\}$ if, for each $b \in \{-1, 1\}^m$, there is a function f_b in F with $bf_b(x^i) = 1$, $i = 1..m$. The VC-dimension of F is the cardinality of the largest shattered set S.*

Alternatively, the fat-shattering dimension can be used to bound the generalization error of a large margin classifier[11]:

Definition 2. *A set of points $S = \{x^1...x^m\}$ is γ-shattered by a set of functions F mapping from a domain X to $\mathbb{R}$ if there are real numbers $r^1, ..., r^m$ such that, for each $b \in \{-1, 1\}^m$, there is a function f_b in F with $b(f_b(x^i) - r^i) \geq \gamma$, $i = 1..m$. We say that $r^1, ..., r^m$ witness the shattering. Then the fat-shattering dimension of F is a function $fat_F(\gamma)$ that maps γ to the cardinality of the largest γ-shattered set S.*

3 Problem Formulation and Generalization Error

In this work, we introduce prior knowledge which has not been previously applied in the SVM framework. This prior is specified in terms of explicit constraints placed on the normal vector of the separating hyperplane. For example, consider the task of determining whether a posting came from the newsgroup alt.atheism or talk.politics.guns, based on the presence of the words "gun" and "atheism" in the posting. Consider the unthresholded perceptron $f(posting; \omega_{atheism}, \omega_{gun}, \theta)$ $= \omega_{atheism} * I_{\{atheism\,present\}} + \omega_{gun} * I_{\{gun\,present\}} + \theta$ ($I_{\{x\,present\}}$ is the indicator function that is 1 when the word x is present in the *posting* and 0 otherwise). A positive value of $\omega_{atheism}$ captures excitatory influence of the word "atheism" on the outcome of classification by ensuring that the value of $f(posting; \omega_{atheism}, \omega_{gun}, \theta)$ increases when the word "atheism" is encountered

in the posting, all other things being equal. Similarly, constraining ω_{gun} to be negative captures an inhibitory influence. Note that such constraints restrict the rotation of the hyperplane, but not its translation offset θ. Thus, prior knowledge by itself does not determine the decision boundary. However, it does restrict the hypothesis space.

We are interested in imposing constraints on the parameters of the family F of functions $sign(f(x;\omega,\theta))$ defined by (1). Constraints of the form $\omega^T c > 0$ generalize excitatory and inhibitory sign constraints[2] (e.g., $\omega_i > 0$ is given by $c = [c_1 = 0, ..., c_i = 1, ..., c_n = 0]^T$). In addition, sometimes it is possible to determine the approximate orientation of the hyperplane a-priori. Normalizing all the coefficients ω_i in the range $[-1, 1]$ enables the domain expert to specify the strength of the contribution of ω_{gun} and $\omega_{atheism}$ in addition to to the signs of their influence. When prior knowledge is specified in terms of an orientation vector v, the conic constraint $\frac{\omega^T v}{\|\omega\|\|v\|} > \rho$ ($\rho \in [-1, 1)$) prevents the normal ω from deviating too far from v.

It is well-known that the VC-dimension of F in $\mathbb{R}^n$ is $n + 1$ (see, e.g., [12]). Interestingly, the VC-dimension of *constrained* F is at least n with any number of constraints imposed on $\omega \in W$ as long as there is an open subset of W that satisfies the constraints (this result follows from [13]). This means that any value of ρ in the conic constraint cannot result in significant improvement in the classifier's generalization ability as measured by its VC-dimension. Similarly, sign constraints placed on all the input variables cannot decrease the classifier's VC-dimension by more than 1. The following theorem shows that the VC-dimension of a relatively weakly constrained classifier achieves this lower bound of n:

Theorem 1. *For the class* $F_C = \{x \rightarrow sign(\sum_{i=1}^{n} \omega_i x_i + \theta) : \omega_1 > 0\}$, *VC-dimension of* $F_C = n$.

Proof. The proof uses techniques from [12]. Let $F_C = \{x \rightarrow sign(\omega_1 x_1 + \overline{\omega}^T \overline{x} + \theta) : \omega_1 > 0\}$, where $\overline{x} = [x_2, ..., x_n]^T$ is the projection of x into the hyperplane $\{\omega_1 = 0\}$ and $\overline{\omega} = [\omega_2, ...\omega_n]^T$.

First, observe that $\{\omega_1 > 0\}$ defines an open subset of W. Hence, the VC-dimension of F_C is at least n. Now, we show by contradiction that a set of $n + 1$ points cannot be shattered by F_C. Assume that some set of points $x^1, ..., x^{n+1} \in \mathbb{R}^n$ can be shattered. Let $\overline{x^1}, ..., \overline{x^{n+1}} \in \mathbb{R}^{n-1}$ be their projections into the hyperplane $\{\omega_1 = 0\}$. There are two cases: Case 1: $\overline{x^1}, ..., \overline{x^{n+1}}$ are distinct. Since these are $n + 1$ points in an $(n - 1)$-dimensional hyperplane, by Radon's Theorem [14] they can be divided into two sets S_1 and S_2 whose convex hulls intersect. Thus, $\exists \lambda_i, \lambda_j (0 \leq \lambda_i, \lambda_j \leq 1)$

[2] In the rest of the paper, we refer to excitatory and inhibitory constraints of the form $\omega_i > 0$ ($\omega_i < 0$) as sign constraints because they constrain the sign of w_i.

$$\text{such that} \sum_{i\,:\,\overline{x^i} \in S_1} \lambda_i \overline{x^i} = \sum_{j\,:\,\overline{x^j} \in S_2} \lambda_j \overline{x^j} \tag{6}$$

$$\text{and} \sum_{i\,:\,\overline{x^i} \in S_1} \lambda_i = \sum_{j\,:\,\overline{x^j} \in S_2} \lambda_j = 1 \tag{7}$$

Since $x^1, ..., x^{n+1}$ are shattered in $\mathbb{R}^n$, $\exists \omega_1, \overline{\omega}, \theta$ such that $\omega_1 x_1^i + \overline{\omega}^T \overline{x^i} \geq \theta$ for all $\overline{x^i} \in S_1$. Multiplying by λ_i and summing over i, we get (after applying (7))

$$\overline{\omega}^T \sum_{i\,:\,\overline{x^i} \in S_1} \lambda_i \overline{x^i} \geq \theta - \omega_1 \sum_{i\,:\,\overline{x^i} \in S_1} \lambda_i x_1^i \tag{8}$$

Similarly, for all $\overline{x^j} \in S_2$, $\omega_1 x_1^j + \overline{\omega}^T \overline{x^j} < \theta \Rightarrow$

$$\overline{\omega}^T \sum_{j\,:\,\overline{x^j} \in S_2} \lambda_j \overline{x^j} < \theta - \omega_1 \sum_{j\,:\,\overline{x^j} \in S_2} \lambda_j x_1^j \tag{9}$$

$$\text{Combining (8), (9), and (6) yields } \omega_1 \Big(\sum_{j\,:\,\overline{x^j} \in S_2} \lambda_j x_1^j - \sum_{i\,:\,\overline{x^i} \in S_1} \lambda_i x_1^i \Big) < 0 \tag{10}$$

$$\text{Since } \omega_1 > 0, \Big(\sum_{j\,:\,\overline{x^j} \in S_2} \lambda_j x_1^j - \sum_{i\,:\,\overline{x^i} \in S_1} \lambda_i x_1^i \Big) < 0 \tag{11}$$

Now, shattering the same set of points, but reversing the labels of S_1 and S_2 implies that $\exists \omega_1', \overline{\omega'}, \theta'$ such that $\omega_1' x_1^i + \overline{\omega'}^T \overline{x^i} < \theta'$ for all $\overline{x^i} \in S_1$ and $\omega_1' x_1^j + \overline{\omega'}^T \overline{x^j} \geq \theta'$ for all $\overline{x^j} \in S_2$. An argument identical to the one above shows that

$$\omega_1' \Big(\sum_{j\,:\,\overline{x^j} \in S_2} \lambda_j x_1^j - \sum_{i\,:\,\overline{x^i} \in S_1} \lambda_i x_1^i \Big) > 0 \tag{12}$$

$$\text{Since } \omega_1' > 0, \Big(\sum_{j\,:\,\overline{x^j} \in S_2} \lambda_j x_1^j - \sum_{i\,:\,\overline{x^i} \in S_1} \lambda_i x_1^i \Big) > 0, \text{ which contradicts (11)}$$

Case 2: Two distinct points x^1 and x^2 project to the same point $\overline{x^1} = \overline{x^2}$ (13) on the hyperplane $\{\omega_1 = 0\}$. Assume, wlog, that $x_1^1 < x_1^2$ (14). Since x^1 and x^2 are shattered, $\exists \omega_1, \overline{\omega}, \theta$ such that $\omega_1 x_1^1 + \overline{\omega}^T \overline{x^1} \geq \theta > \omega_1 x_1^2 + \overline{\omega}^T \overline{x^2}$, which, together with (13) and (14), implies that $\omega_1 < 0$, a contradiction. □

This result means that imposing a sign constraint on a single input variable or using $\rho = 0$ in the conic constraint is sufficient to achieve the maximum theoretical improvement within the VC framework[3]. However, it is unsatisfactory in a sense that it contradicts our intuition (and empirical results) which suggests that stronger prior knowledge should help the classifier reduce its generalization error faster. The following theorem shows that the fat-shattering dimension decreases continuously with increasing ρ in the conic constraint, giving us the desired guarantee. Technically, the fat-shattering dimension is a function of the margin γ, so we use the following definition of function domination to specify what we mean by decreasing fat-shattering dimension:

Definition 3. *A function $f_1(x)$ is dominated by a function $f_2(x)$ if, for all x, $f_1(x) \leq f_2(x)$ and, at least for one a, $f_1(a) < f_2(a)$. When we say that $f_\rho(x)$ decreases with increasing ρ, we mean that $\rho_1 < \rho_2$ implies that $f_{\rho_2}(x)$ is dominated by $f_{\rho_1}(x)$.*

Theorem 2. *For the class $F_{v,\rho} = \{x \to \omega^T x + \theta : \|\omega\|_2 = 1, \|v\|_2 = 1, \|x\|_2 \leq R, \omega^T v > \rho \geq 0\}$, $fat_{F_{v,\rho}}(\gamma)$ decreases with increasing ρ.* [4]

Proof. The fat-shattering dimension obviously cannot increase with increasing ρ, so we only need to find a value of γ where it decreases. We show that this happens at $\gamma' = R\sqrt{1-\rho_2^2}$. First, we upper bound $\mathrm{fat}_{F_{v,\rho_2}}(\gamma')$ by showing that, in order to γ'-shatter two points, the separating hyperplane must be able to rotate through a larger angle than that allowed by the constraint ${\omega^1}^T v > \rho_2$. Assume that two points x^1, x^2 can be γ'-shattered by F_{v,ρ_2}. Then $\exists \omega^1, \omega^2, \theta^1, \theta^2, r^1, r^2$ such that ${\omega^1}^T x^1 + \theta^1 - r^1 \geq \gamma'$, ${\omega^1}^T x^2 + \theta^1 - r^2 \leq -\gamma'$, ${\omega^2}^T x^1 + \theta^2 - r^1 \leq -\gamma'$, ${\omega^2}^T x^2 + \theta^2 - r^2 \geq \gamma'$. Combining the terms and applying the Cauchy-Schwartz inequality, we get $\left\|\omega^1 - \omega^2\right\| \geq \frac{2\gamma'}{R}$. Squaring both sides, expanding $\left\|\omega^1 - \omega^2\right\|^2$ as $\left\|\omega^1\right\|^2 + \left\|\omega^2\right\|^2 - 2{\omega^1}^T\omega^2$, and using the fact that $\left\|\omega^1\right\| = \left\|\omega^2\right\| = 1$ yields

$$ {\omega^1}^T\omega^2 \leq 1 - \frac{2\gamma'^2}{R^2} = 2\rho_2^2 - 1 \tag{15} $$

Since the angle between ω^1 and ω^2 cannot exceed the sum of the angle between ω^1 and the prior v and the angle between v and ω^2, both of which are bounded above by $\arccos(\rho_2)$, we get (after some algebra) ${\omega^1}^T\omega^2 > 2\rho_2^2 - 1$, which contradicts (15).

$$ \text{Thus, } fat_{F_{v,\rho_2}}(R\sqrt{1-\rho_2^2}) < 2 \tag{16} $$

[3] The constraint $\{w_1 > 0\}$ is weak since it only cuts the volume of the hypothesis space by $\frac{1}{2}$.

[4] Note that the statement of this theorem deals with hyperplanes with unit normals, not canonical hyperplanes. The margin of a unit-normal hyperplane is given by $min_{i=1..m}|\omega x^i + \theta|$.

Now, we lower bound $\text{fat}_{F_{v,\rho_1}}(\gamma')$ by exhibiting two points γ'-shattered by F_{v,ρ_1}. Wlog, let $v = [0, 1, 0, ..0]^T$. It is easy to verify that $x^1 = [R, 0, ..0]^T$ and $x^2 = [-R, 0, ..0]^T$ can be $R\sqrt{1-\rho_2^2}$-shattered by F_{v,ρ_1}, witnessed by $r^1 = r^2 = 0$.

$$\text{Hence, } fat_{F_{v,\rho_1}}(R\sqrt{1-\rho_2^2}) \geq 2 \tag{17}$$

which, combined with (16), completes the argument. □

The result of Theorem 1 is important because it shows that even weak prior knowledge improves the classifier's generalization performance in the VC framework which makes less assumptions about the data than the fat-shattering framework. However, it is the result of Theorem 2 within the fat-shattering framework which justifies the use of stronger prior knowledge.

4 Implementation

The quadratic optimization problem for finding the maximum margin separating hyperplane (2) can be easily modified to take into account linear rotational constraints of the form $\omega^T c_j > 0$, $j = 1...l$. The soft margin/soft constraint formulation that allows for possibility of violating both the margin maximization objective and the rotational constraints minimizes the following regularization functional:

$$R_{reg}[f, l, l'] = \frac{1}{m}\sum_{i=1}^{m} l(y_i, f(x_i; \omega, \theta)) + \frac{C_1}{C_2 l}\sum_{j=1}^{l} l'(\omega, c_j) + \frac{C_1}{2}\|\omega\|_2^2 \tag{18}$$

with 0-1 losses for the data and the prior: $l(y_i, f(x_i; \omega, \theta)) = I_{\{-y_i f(x_i;\omega,\theta)>0\}}$ and $l'(\omega, c_j) = I_{\{-\omega^T c_j > 0\}}$ in the hard margin/hard rotational constraints case and hinge losses: $l(y_i, f(x_i; \omega, \theta)) = max(1 - y_i f(x_i; \omega, \theta), 0)$,$l'(\omega, c_j) = max(-\omega^T c_j, 0)$ in the soft margin/soft rotational constraints case. The regularization functional above is the same as in (2) with an additional loss function which penalizes the hyperplanes that violate the prior. Minimizing (18) with hinge loss functions is equivalent to solving:

$$\underset{\omega,\theta,\xi,\nu}{minimize}\ \tfrac{1}{2}\|\omega\|_2^2 + \tfrac{1}{C_1 m}\sum_{i=1}^{m}\xi_i + \tfrac{1}{C_2 l}\sum_{j=1}^{l}\nu_j \tag{19}$$

$$subject\ to\ y_i(\omega^T x + \theta) \geq 1 - \xi_i, \xi_i \geq 0,\ i = 1...m,$$
$$\omega^T c_j \geq 0 - \nu_j, \nu_j \geq 0,\ j = 1...l.$$

Constructing the Lagrangian from (19) and calculating the Wolfe dual results in the following maximization problem:

$$\underset{\alpha,\beta}{maximize}\ \sum_{i=1}^{m}\alpha_i - \tfrac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j (x_i^T x_j) -$$
$$\sum_{i=1}^{m}\sum_{j=1}^{l}\alpha_i\beta_j y_i (x_i^T c_j) - \frac{1}{2}\sum_{i,j=1}^{l}\beta_i\beta_j (x_i^T c_j) \tag{20}$$

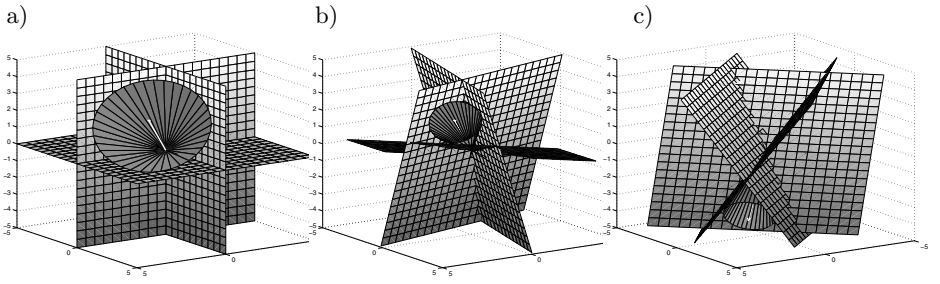

Fig. 1. Approximating a conic constraint:

a) Start with the known constraints $\omega_1 \geq 0$, $\omega_2 \geq 0$, and $\omega_3 \geq 0$ around $v^1 = [\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}]^T$. The figure shows linear constraints around v^1 (white vector) and the cone $\frac{\omega^T v^1}{\|\omega\|} > \frac{1}{\sqrt{3}}$ approximated by these constraints.
b) Rotate the bounding hyperplanes $\{\omega_1 = 0\}$, $\{\omega_2 = 0\}$, $\{\omega_3 = 0\}$ into v^1, approximating a cone with the required angle ρ' around v^1
c) Rotate the whole boundary from v^1 (white vector in (a),(b)) to the required orientation v' (white vector in (c)).

$$subj.\,to\ \tfrac{1}{C_1 m} \geq \alpha_i \geq 0,\ i = 1...m,\ \ \tfrac{1}{C_2 l} \geq \beta_j \geq 0,\ j = 1...l,\ and\ \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$\text{The solution to (20) is given by } \omega = \sum_{i=1}^{m} \alpha_i y_i x_i + \sum_{j=1}^{l} \beta_j c_j. \tag{21}$$

As before, setting $\xi_i = 0$, $\nu_j = 0$, $i = 1...m$, $j = 1...l$, (19), (20), and (21) can be used to solve the hard margin/hard rotational constraints optimization problem. Note that in the soft-margin formulation, constants C_1 and C_2 define a trade-off between fitting the data, maximizing the margin, and respecting the rotational constraints.

The above calculation can impose linear constraints on the orientation of the large margin separating hyperplane when such constraints are given. This is the case with sign-constrained prior knowledge. However, domain knowledge in form of a cone centered around an arbitrary rotational vector v' cannot be represented as a linear constraint in the quadratic optimization problem given by (19). The approach taken in this work is to approximate an n-dimensional cone with n hyperplanes. For example, sign constraints $\omega_1 \geq 0$, $\omega_2 \geq 0$, and $\omega_3 \geq 0$ approximate a cone of angle $\rho^1 = \frac{1}{\sqrt{3}}$ around $v^1 = [\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}]^T$ (see Figure 1-(a)). To approximate a cone of arbitrary angle ρ' around an arbitrary orientation vector v', 1) the normal ω_i' of each bounding hyperplane $\{\omega_i = 0\}$ (as defined by the sign constraints above) is rotated in the plane spanned by $\{\omega_i', v^1\}$ by an angle $acos(\omega_i'^T v^1) - \rho'$, and 2) a solid body rotation that transforms v^1 into v' is subsequently applied to all the bounding hyperplanes, as illustrated in Figure 1. This construction generalizes in a straightforward way from $\mathbb{R}^3$ to $\mathbb{R}^n$.

5 Experiments

Experiments were performed on two distinct real-world domains:

Voting Records. This is a UCI database [15] of congressional voting records. The vote of each representative is recorded on the 16 key issues. The task is to predict the representative's political party (Democrat or Republican) based on his/her votes. The domain theory was specified in form of inhibitory/excitatory sign constraints. An excitatory constraint means that the vote of "yea" correlates with the Democratic position on the issue, an inhibitory constraint means that Republicans favor the proposal. The complete domain theory is specified in Figure 2-(a). Note that sign constraints are imposed on relatively few features (7 out of 16). Since this type of domain knowledge is weak, a hard rotational constraint SVM was used. Only representatives whose positions are known on all the 16 issues were used in this experiment. The results shown in Figure 2-(b) demonstrate that sign constraints decrease the generalization error of the classifier. As expected, prior knowledge helps more when the data is scarce.

Text classification. The task is to determine the newsgroup that a posting was taken from based on the posting's content. We used the 20-newsgroups dataset [16]. Each posting was treated as a bag-of-words, with each binary feature encoding whether or not the word is present in the posting. Stemming was used in the preprocessing stage to reduce the number of features. Feature selection based on mutual information between each individual feature and the label was employed (300 maximally informative features were chosen). Since SVMs are best suited for binary classification tasks, all of our experiments involve pairwise newsgroup classification. The problem of applying SVMs to multicategory classification has been researched extensively([2,3]), and is orthogonal to our work.

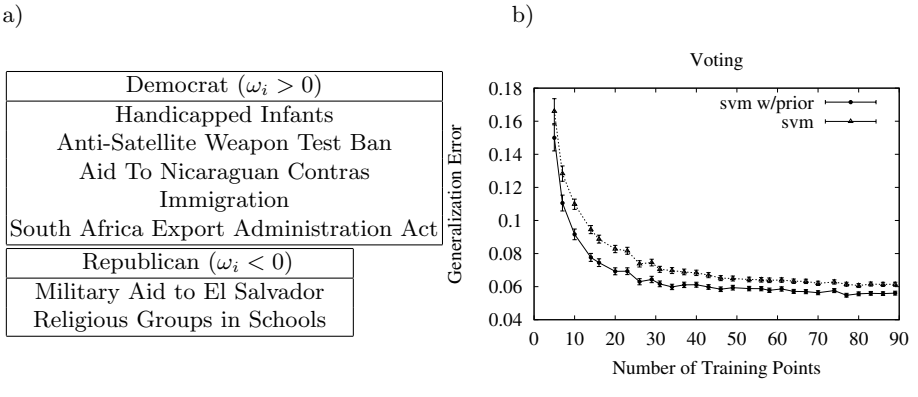

Fig. 2. a) Prior Knowledge for voting b)Generalization error as a percentage versus number of training points for voting classification. For each classification task, the data set is split randomly into training and test sets in 1000 different ways. SVM classifier is trained on the training set with and without prior knowledge, and its average error on the test set is plotted, along with error bars showing 95% confidence intervals.

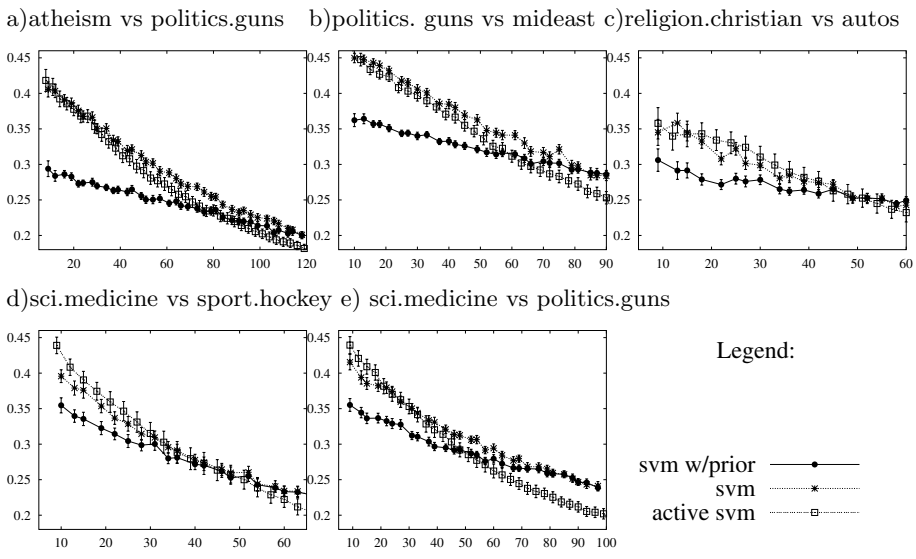

Fig. 3. Generalization error as a percentage versus number of training points for 5 different classification experiments. For each random sample selection, the data set is split randomly into training and test sets in 100 different ways. For active learning experiments, the data set is split randomly into two equal-sized sets in 100 different ways, with one set used as the unlabeled pool for query selection, and the other set - for testing. All error bars are based on 95% confidence intervals.

Prior knowledge in this experiment is represented by a conic constraint around a specific orientation vector v'. While it may be hard for human experts to supply such a prior, there are readily available sources of domain knowledge that were not developed specifically for the classification task at hand. In order to be able to utilize them, it is essential to decode the information into a form usable by the learning algorithm. This is the virtue of rotational constraints: they are directly usable by SVMs and they can approximate more sophisticated pre-existing forms of information. In our experiments, domain knwoledge from Wordnet, a lexical system which encodes semantic relations between words [17], is automatically converted into v'. The coefficient v'_x of each word x is calculated from the relative proximity of x to each category label in the hypernym (is-a) hierarchy of Wordnet (measured in hops). A natural approximation of v'_x is given by $\frac{hops(x,label_+)}{hops(x,label_-)+hops(x,label_+)}$, normalized by a linear mapping to the required range $[-1, 1]$, where $label_+$ and $label_-$ are the names of the two newsgroups. Performance of the following three classifiers on this task was evaluated:

1. A soft rotational constraint SVM ($C_1 = C_2 = 10^{-5}$) with Wordnet prior ($\rho = 0.99$) (reasonable values of constants were picked based on the alt.atheism vs. politics.guns classification task, with no attempt to optimize them for other tasks).

2. An SVM which actively selects the points to be labeled out of a pool of unlabeled newsgroup postings. We implemented a strategy suggested in [4] which always queries the point closest to the separating hyperplane.
3. Traditional SVM ($C_1 = 10^{-5}$) trained on a randomly selected sample.

Typical results of this experiment for a few different pairwise classification tasks appear in Figure 3. For small data samples, the prior consistently decreases generalization error by up to 25%, showing that even a very approximate prior orientation vector v' can result in significant performance improvement. Since prior knowledge is imposed with soft constraints, the data overwhelms the prior with increasing sample size. Figure 3 also compares the effect of introducing rotational constraints with the effect of active learning. It has been shown theoretically that active learning can improve the convergence rate of the classification error under a favorable distribution of the input data [18], although no such guarantees exist for general distributions. In our experiments, active learning begins to improve performance only after enough data is collected. Active learning does not help when the sample size is very small, probably due to the fact that the separating hyperplane of the classifier cannot be approximated well, resulting in uninformative choices of query points. Rotational prior knowledge, on the other hand, is more helpful for lowest sample sizes and ceases to be useful in the region where active learning helps. Thus, the strengths of prior knowledge and active learning are complementary. Combining them is a direction for future research.

6 Conclusions

We presented a simple framework for incorporating rotational prior knowledge into support vector machines. This framework has proven not only practically useful, but also useful for gaining insight into generalization ability of a-priori constrained large-margin classifiers.

Related work includes using Wordnet for feature creation for text categorization ([19]) and introducing sign constraints into the perceptron learning algorithm [20,21]. These studies do not provide generalization error guarantees for classification.

Acknowledgement. We thank Ilya Shpitser and anonymous reviewers for helpful suggestions on improving this paper. This material is based upon work supported in part by the National Science Foundation under Award NSF CCR 01-21401 ITR and in part by the Information Processing Technology Office of the Defense Advanced Research Projects Agency under award HR0011-05-1-0040. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Defense Advanced Research Projects Agency.

References

1. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag (1995)
2. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the Tenth European Conference on Machine Learning. Number 1398 (1998)
3. Dumas, S., Platt, J., Heckerman, D., Sahami, M.: Inductive learning algorithms and representations for text categorization. Proceedings of the Seventh International Conference on Information and Knowledge Management (1998)
4. Campbell, C., Cristianini, N., Smola, A.: Query learning with large margin classifiers. Proceedings of The Seventeenth International Conference on Machine Learning (2000) 111–118
5. Raina, R., Shen, Y., Ng, A., McCallum, A.: Classification with hybrid generative/discriminative models. Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems (2003)
6. Fink, M.: Object classification from a single example utilizing class relevance metrics. Proceedings of the Eighteenth Annual Conference on Neural Information Processing Systems (2004)
7. Scholkopf, B., Simard, P., Vapnik, V., Smola, A.: Prior knowledge in support vector kernels. Advances in kernel methods - support vector learning (2002)
8. Fung, G., Mangasarian, O., Shavlik, J.: Knowledge-based support vector machine classifiers. Proceedings of the Sixteenth Annual Conference on Neural Information Processing Systems (2002)
9. Wu, X., Srihari, R.: Incorporating prior knowledge with weighted margin support vector machines. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2004)
10. Mangasarian, O., Shavlik, J., Wild, E.: Knowledge-based kernel approximation. Journal of Machine Learning Research (2004)
11. Shawe-Taylor, J., Bartlett, P.L., Williamson, R.C., Anthony, M.: Structural risk minimization over data-dependent hierarchies. IEEE Transactions on Information Theory **44** (1998)
12. Anthony, M., Biggs, N.: PAC learning and artificial neural networks. Technical report (2000)
13. Erlich, Y., Chazan, D., Petrack, S., Levy, A.: Lower bound on VC-dimension by local shattering. Neural Computation **9** (1997)
14. Grunbaum, B.: Convex Polytopes. John Wiley (1967)
15. Blake, C., Merz, C.: UCI repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/mlrepository.html (1998)
16. Blake, C., Merz, C.: 20 newsgroups database, http://people.csail.mit.edu/people/jrennie/ 20newsgroups/ (1998)
17. Miller, G.: WordNet: an online lexical database. International Journal of Lexicography **3** (1990)
18. Dasgupta, S., Kalai, A.T., Monteleoni, C.: Analysis of perceptron-based active learning. Eighteenth Annual Conference on Learning Theory (2005)
19. Gabrilovich, E., Markovitch, S.: Text categorization with many redundant features: Using aggressive feature selection to make svms competitive with c4.5. Proceedings of The Twenty-First International Conference on Machine Learning (2004)
20. Amit, D., Campbell, C., Wong, K.: The interaction space of neural networks with sign-constrained weights. Journal of Physics (1989)
21. Barber, D., Saad, D.: Does extra knowledge necessarily improve generalization? Neural Computation **8** (1996)