Journal of Machine Learning Research 13 (2012) 3507-3531 Submitted 3/08; Revised 3/12; Published 12/12

PAC-Bayes Bounds with Data Dependent Priors

Emilio Parrado-Hernández EMIPAR@TSC.UC3M.ES
Department of Signal Processing and Communications
University Carlos III of Madrid
Leganés, 28911, Spain

Amiran Ambroladze A.AMBROLADZE@FREEUNI.EDU.GE
Department of Mathematics and Computer Science
Tbilisi Free University
Bedia Street
0182 Tbilisi, Georgia

John Shawe-Taylor J.SHAWE-TAYLOR@CS.UCL.AC.UK
Department of Computer Science
University College London
London, WC1E 6BT, UK

Shiliang Sun SHILIANGSUN@GMAIL.COM
Department of Computer Science and Technology
East China Normal University
500 Dongchuan Road
Shanghai 200241, China

Editor: Gabor Lugosi

Abstract

This paper presents the prior PAC-Bayes bound and explores its capabilities as a tool to provide tight predictions of SVMs' generalization. The computation of the bound involves estimating a prior of the distribution of classifiers from the available data, and then manipulating this prior in the usual PAC-Bayes generalization bound. We explore two alternatives: to learn the prior from a separate data set, or to consider an expectation prior that does not need this separate data set. The prior PAC-Bayes bound motivates two SVM-like classification algorithms, prior SVM and η-prior SVM, whose regularization term pushes towards the minimization of the prior PAC-Bayes bound. The experimental work illustrates that the new bounds can be significantly tighter than the original PAC-Bayes bound when applied to SVMs, and among them the combination of the prior PAC-Bayes bound and the prior SVM algorithm gives the tightest bound.

Keywords: PAC-Bayes bound, support vector machine, generalization capability prediction, classification

1. Introduction

Support vector machines (SVMs) (Boser et al., 1992; Vapnik, 1998; Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002) are accepted among practitioners as one of the most accurate automatic classification techniques. They implement linear classifiers in a high-dimensional feature space using the kernel trick to enable a dual representation and efficient computation. The danger of overfitting in such high-dimensional spaces is conquered by maximizing the margin of the classifier

on the training examples. For this reason there has been considerable interest in bounding the generalization in terms of the margin.

In fact, a main drawback that restrains engineers from using these advanced machine learning techniques is the lack of reliable predictions of generalization, especially in what concerns worst-case performance. In this sense, the widely used cross-validation generalization measures indicate little about the worst-case performance of the algorithms. The error of the classifier on a set of samples follows a binomial distribution whose mean is the true error of the classifier. Cross-validation is a sample mean estimation of the true error, and worst-case performance estimations concern the estimation of the tail of the error distribution. One could then employ statistical learning theory (SLT) tools to bound the tail of the distribution of errors. Early bounds have relied on covering number computations (Shawe-Taylor et al., 1998; Zhang, 2002), while later bounds have considered Rademacher complexity (Bartlett and Mendelson, 2002). The tightest bounds for practical applications appear to be the PAC-Bayes bound (McAllester, 1999; Langford and Shawe-Taylor, 2002; Catoni, 2007) and in particular the form given in Seeger (2002), Langford (2005) and Germain et al. (2009). However, there still exist a remarkable gap between SLT predictions and practitioners' experiences: SLT predictions are too pessimistic when compared to the actual results data analysts get when they apply machine learning algorithms to real-world problems.

Another issue affected by the ability to predict the generalization capability of a classifier is the selection of the hyperparameters that define the training. In the SVM case, these parameters are the trade-off between maximum margin and minimum training error, C, and the kernel parameters. Again, the more standard method of cross-validation has proved to be more reliable in most experiments, despite the fact that it is statistically poorly justified and relatively expensive.

The aim of this paper is to investigate whether the PAC-Bayes bound can be tightened towards less pessimistic predictions of generalization. Another objective is to study the implications of the bound in the training of the classifiers. We specifically address the use of the bound in the model selection stage and in the design of regularization terms other than the maximization of the margin.

The PAC-Bayes bound (retrospected in Section 2) uses a Gaussian prior centered at the origin in the weight space. The key to the new bounds introduced here is to use part of the training set to compute a more informative prior and then compute the bound on the remainder of the examples relative to this prior. This generalisation of the bound, called prior PAC-Bayes bound, is derived in Section 3. The prior PAC-Bayes bound was initially presented by Ambroladze et al. (2007). A slight nuisance of the prior PAC-Bayes bound is that a separate data set should be available in order to fix the prior. In Section 3.2, we further develop the expectation-prior PAC-Bayes bound as an interesting new approach which does not require the existence of the separate data set. We also derive a PAC-Bayes bound with a non-spherical Gaussian prior. To the best of our knowledge this is the first such application for SVMs.

The encouraging results of Ambroladze et al. (2007), motivate a further use of the prior PAC-Bayes bound. Section 4.1 introduces a new classification algorithm, the prior SVM, which replaces the margin maximization in the optimization problem by a regularization term that pushes towards the minimization of the PAC-Bayes bound. The optimization problem that produces the prior SVM is divided into three stages. The first one involves the learning of a prior formed by an ensemble of Gaussian distributions centered at different distances along the same direction. During the second stage, each component of the prior is mapped with a posterior that improves its classification accuracy while tightening the PAC-Bayes bound. In the last stage the prior component/posterior pair that achieves the lowest value of the PAC-Bayes bound is selected as prior SVM classifier. Section

4.2 presents a second algorithm, named η-prior SVM as a variant of prior SVMs where the position of component of the prior that goes into the overall classifier is optimised in a continuous range (not picked from a fixed set). Therefore, η-prior SVMs include a first optimization where the direction of the prior is learnt from a separate set of training patterns, and a second optimization that determines (i) the exact position of the prior along the already learnt direction and (ii) the position of the posterior. Furthermore we show that the performance of the algorithm can be bounded rigorously using PAC-Bayes techniques.

In Section 5 the new bounds and algorithms are evaluated on multiple classification tasks after a parameter selection. The experiments illustrate the capabilities of the prior PAC-Bayes bound to provide tighter predictions of the generalisation of an SVM. Moreover, the combination of the new bounds and the two prior SVM algorithms yields more dramatic tightenings of the bound. Besides, these classifiers achieve good accuracies, comparable to those obtained by an SVM with its parameters fixed with ten fold cross validation. We finish the experimental work showing that the use of a different value of C for the prior and the posterior that form the (η)prior SVM lead to a further tightening of the bound.

Finally, the main conclusions of this work and some related ongoing research are outlined in Section 6.

2. PAC-Bayes Bound for SVMs

This section is devoted to a brief review of the PAC-Bayes bound theorem of Langford (2005). Let us consider a distribution $\mathcal{D}$ of patterns $\mathbf{x}$ lying in a certain input space $\mathcal{X}$ and their corresponding output labels y ($y \in \{-1,1\}$). Suppose Q is a posterior distribution over the classifiers c. For every classifier c, the following two error measures are defined:

Definition 1 (True error) *The true error $c_{\mathcal{D}}$ of a classifier c is defined to be the probability of misclassifying a pattern-label pair $(\mathbf{x}, y)$ selected at random from $\mathcal{D}$*

$$c_{\mathcal{D}} \equiv Pr_{(\mathbf{x},y)\sim\mathcal{D}}(c(\mathbf{x}) \neq y).$$

Definition 2 (Empirical error) *The empirical error $\hat{c}_S$ of a classifier c on a sample S of size m is defined to be the error rate on S*

$$\hat{c}_S \equiv Pr_{(\mathbf{x},y)\sim S}(c(\mathbf{x}) \neq y) = \frac{1}{m}\sum_{i=1}^{m} I(c(\mathbf{x}_i) \neq y_i),$$

where $I(\cdot)$ is an indicator function equal to 1 *if the argument is true and equal to* 0 *if the argument is false.*

Now we define two error measures on the distribution of classifiers: the average true error, $Q_{\mathcal{D}} \equiv \mathbb{E}_{c\sim Q} c_{\mathcal{D}}$, as the probability of misclassifying an instance $\mathbf{x}$ chosen uniformly from $\mathcal{D}$ with a classifier c chosen according to Q; and the average empirical error $\hat{Q}_S \equiv \mathbb{E}_{c\sim Q}\hat{c}_S$, as the probability of classifier c chosen according to Q misclassifying an instance $\mathbf{x}$ chosen from a sample S.

For these two quantities we can derive the PAC-Bayes bound on the true error of the distribution of classifiers:

Theorem 3 (PAC-Bayes bound) *For all prior distributions $P(c)$ over the classifiers c, and for any $\delta \in (0,1]$,*

$$Pr_{S\sim\mathcal{D}^m}\left(\forall Q(c) : KL_+(\hat{Q}_S||Q_\mathcal{D}) \leq \frac{KL(Q(c)||P(c)) + \ln(\frac{m+1}{\delta})}{m}\right) \geq 1-\delta,$$

where $KL(Q(c)||P(c)) = \mathbb{E}_{c\sim Q}\ln\frac{Q(c)}{P(c)}$ is the Kullback-Leibler divergence, and $KL_+(q||p) = q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}$ for $p > q$ and 0 otherwise.

The proof of the theorem can be found in Langford (2005).

This bound can be specialized to the case of linear threshold classifiers. Suppose the m training examples define a linear classifier that can be represented by the following equation:

$$c_\mathbf{u}(\mathbf{x}) = \text{sign}(\mathbf{u}^T\phi(\mathbf{x})), \tag{1}$$

where $\phi(\mathbf{x})$ is a nonlinear projection to a certain feature space[1] where the linear classification actually takes place, and vector $\mathbf{u}$ in the feature space determines the separating hyperplane. Since we are considering only classifiers with threshold set to zero all the classifiers in the paper can be represented with unit vectors ($\|\mathbf{w}\| = 1$). However, as we will be considering distributions of classifiers we use the notation $\mathbf{u}$ to indicate weight vectors that can also be non-unit.

For any unit vector $\mathbf{w}$ we can define a stochastic classifier in the following way: we choose the distribution $Q(c_\mathbf{u}) = Q(c_\mathbf{u}|\mathbf{w},\mu)$, where $\mathbf{u} \sim \mathcal{N}(\mu\mathbf{w}, I)$ is drawn from a spherical Gaussian with identity covariance matrix centered along the direction pointed by $\mathbf{w}$ at a distance μ from the origin. Moreover, we can choose the prior $c_\mathbf{u} : \mathbf{u} \sim \mathcal{N}(\mathbf{0}, I)$ to be a spherical Gaussian with identity covariance matrix centered at the origin. Then, for classifiers of the form in Equation (1) the generalization performance can be bounded as

Corollary 4 (PAC-Bayes bound for SVMs (Langford, 2005)) *For all distributions $\mathcal{D}$, for all $\delta \in (0,1]$, we have*

$$Pr_{S\sim\mathcal{D}^m}\left(\forall \mathbf{w},\mu : KL_+(\hat{Q}_S(\mathbf{w},\mu)||Q_\mathcal{D}(\mathbf{w},\mu)) \leq \frac{\frac{\mu^2}{2} + \ln(\frac{m+1}{\delta})}{m}\right) \geq 1-\delta.$$

It can be shown (see Langford, 2005) that

$$\hat{Q}_S(\mathbf{w},\mu) = \mathbb{E}_m[\tilde{F}(\mu\gamma(\mathbf{x},y))], \tag{2}$$

where $\mathbb{E}_m$ is the average over the m training examples, $\gamma(\mathbf{x},y)$ is the normalized margin of the training examples

$$\gamma(\mathbf{x},y) = \frac{y\mathbf{w}^T\phi(\mathbf{x})}{\|\phi(\mathbf{x})\|}, \tag{3}$$

and $\tilde{F} = 1 - F$ where F is the cumulative normal distribution

$$F(x) = \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\mathrm{d}x. \tag{4}$$

1. This projection is induced by a kernel $\kappa(\cdot)$ satisfying $\kappa(\mathbf{x},\mathbf{y}) = \langle\phi(\mathbf{x}),\phi(\mathbf{y})\rangle$.

Note that the SVM expressed as (1) is computed with a single unit vector $\mathbf{w}$. The generalization error of such a classifier can be bounded by at most twice the average true error $Q_{\mathcal{D}}(\mathbf{w},\mu)$ of the corresponding stochastic classifier involved in Corollary 4 (Langford and Shawe-Taylor, 2002). That is, for all μ we have

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}}\left(\text{sign}(\mathbf{w}^T\phi(\mathbf{x})) \neq y\right) \leq 2Q_{\mathcal{D}}(\mathbf{w},\mu). \tag{5}$$

3. Data Dependent Prior PAC-Bayes Bounds for SVMs

This section presents some versions of the PAC-Bayes bound that aim at yielding a tighter prediction of the true generalization error of the classifier. These new bounds introduce more sophisticated designs for the prior distribution over the classifiers in order to reduce its divergence with the posterior distribution. The first set of bounds learns the prior distribution from a separate training data set that will not be used in the computation of the bound, whilst the second set learns the prior from mathematical expectations, avoiding to leave out a subset of patterns to calculate the bound.

3.1 Bounds Based on a Separate Set of Training Data

This section is a further extension of previous ideas presented by Ambroladze et al. (2007).

Our first contribution is motivated by the fact that the PAC-Bayes bound allows us to choose the prior distribution, $P(c)$. In the standard application of the bound $P(c)$ is chosen to be a spherical Gaussian centered at the origin. We now consider learning a different prior based on training an SVM on a subset T of the training set comprising r training patterns and labels. In the experiments this is taken as a random subset, but for simplicity of the presentation we will assume T comprises the last r examples $\{\mathbf{x}_k, y_k\}_{k=m-r+1}^{m}$.

With these r examples we can learn an (unit and biased) SVM classifier, $\mathbf{w}_r$, and form a prior $P(\mathbf{w}_r,\eta) \sim \mathcal{N}(\eta\mathbf{w}_r, I)$ consisting of a Gaussian distribution with identity covariance matrix centered along $\mathbf{w}_r$ at a distance η from the origin.

The introduction of this prior $P(\mathbf{w}_r,\eta)$ in Theorem 3 results in the following new bound.

Corollary 5 (Single-prior PAC-Bayes bound for SVMs) *Let us consider a prior on the distribution of classifiers consisting of a spherical Gaussian with identity covariance centered along the direction given by $\mathbf{w}_r$ at a distance η from the origin. Classifier $\mathbf{w}_r$ has been learnt from a subset T of r examples a priori separated from a training set S of m samples. Then, for all distributions $\mathcal{D}$, for all $\delta \in (0,1]$, we have*

$$Pr_{S\sim\mathcal{D}^m}\left(\forall \mathbf{w}_m,\mu : KL_+(\hat{Q}_{S\setminus T}||Q_D) \leq \frac{\frac{||\eta\mathbf{w}_r-\mu\mathbf{w}_m||^2}{2} + \ln(\frac{m-r+1}{\delta})}{m-r}\right) \geq 1-\delta,$$

where $\hat{Q}_{S\setminus T}$ is a stochastic measure of the empirical error of the classifier on the $m-r$ samples not used to learn the prior. This stochastic error is computed as in Equation (2) but averaged over $S\setminus T$.

Proof Since we separate r instances to learn the prior, the actual size of the training set to which we apply the bound is $m-r$. In addition, the stochastic error $\hat{Q}$ must be computed only on the instances not used to learn the prior, that is, the subset $S\setminus T$. Note also that the selection of T can not be optimised.

Using a standard expression for the KL divergence between two Gaussians in an N dimensional space,

$$\mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_0,\Sigma_0)\|\mathcal{N}(\boldsymbol{\mu}_1,\Sigma_1)) = \\ \frac{1}{2}\left(\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right)+\mathrm{tr}(\Sigma_1^{-1}\Sigma_0)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\Sigma_1^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)-N\right), \tag{6}$$

the KL divergence between prior and posterior is computed as follows:

$$\mathrm{KL}(Q(\mathbf{w},\mu)||P(\mathbf{w}_r,\eta)) = \mathrm{KL}(\mathcal{N}(\mu\mathbf{w},I)\|\mathcal{N}(\eta\mathbf{w}_r,I)) = \frac{1}{2}||\mu\mathbf{w}-\eta\mathbf{w}_r||^2.$$

■

Intuitively, if the selection of the prior is appropriate, the bound can be tighter than the one given in Corollary 4 when applied to the SVM weight vector on the whole training set. It is worth stressing that the bound holds for all $\mathbf{w}$ and so can be applied to the SVM trained on the whole set. This might at first appear to be 'cheating', but the critical point is that the bound is evaluated on the set $S\backslash T$ not involved in generating the prior. The experimental work illustrates how in fact this bound can be tighter than the standard PAC-Bayes bound.

Moreover, the structure of the prior may be further refined in exchange for a very small increase in the penalty term. This can be achieved with the application of the following result.

Theorem 6 (Mixture prior PAC-Bayes bound) *Let $\mathcal{P}(c)=\sum_{j=1}^{J}\pi_j P_j(c)$ be a prior distribution over classifiers consisting of a mixture of J components $\{P_j(c)\}_{j=1}^{J}$ combined with positive weights $\{\pi_j\}_{j=1}^{J}$ so that $\sum_{j=1}^{J}\pi_j=1$. Then, for all $\delta\in(0,1]$,*

$$Pr_{S\sim\mathcal{D}^m}\left(\forall Q(c): KL_+(\hat{Q}_S||Q_{\mathcal{D}})\leq\min_j\frac{KL(Q(c)||P_j(c))+\ln\frac{m+1}{\delta}+\ln\frac{1}{\pi_j}}{m}\right)\geq 1-\delta.$$

Proof

The bound in Theorem 3 can be instantiated for the ensemble prior $\mathcal{P}(c)$

$$\Pr_{S\sim\mathcal{D}^m}\left(\forall Q(c): \mathrm{KL}_+(\hat{Q}_S||Q_{\mathcal{D}})\leq\frac{\mathrm{KL}(Q(c)||\mathcal{P}(c))+\ln(\frac{m+1}{\delta})}{m}\right)\geq 1-\delta.$$

We now bound the KL divergence between the posterior $Q(c)$ and the ensemble prior $\mathcal{P}(c)$. For any $1\leq i\leq J$:

$$\mathrm{KL}\left(Q(c)\|\mathcal{P}(c)\right)=\int_{c\in C}Q(c)\left(\ln Q(c)-\ln(\sum_{j=1}^{J}\pi_j P_j(c))\right)dc$$

$$\leq\int_{c\in C}Q(c)(\ln Q(c)-\ln(\pi_i P_i(c)))dc=\mathrm{KL}(Q(c)\|P_i(c))-\ln(\pi_i),$$

where the inequality follows from the fact that we have reduced the value inside the $\ln(\cdot)$ term for all c. Finally, the particularisation for the term of minimal $\mathrm{KL}(Q(c)\|P_j(c))-\ln(\pi_j)$ completes the

proof. ■

Note that the inequality in the proof upper bounds the KL divergence to give a bound equivalent to performing a union bound. In particular applications it may be possible to obtain tighter bounds by estimating this KL divergence more closely.

This result can be also specialized for the case of SVM classifiers. The mixture prior is constructed by allocating Gaussian distributions with identity covariance matrix along the direction given by $\mathbf{w}_r$ at distances $\{\eta_j\}_{j=1}^J$ from the origin where $\{\eta_j\}_{j=1}^J$ are positive real numbers. In such a case, we obtain

Corollary 7 (Gaussian Mixture-prior PAC-Bayes bound for SVMs) *Let us consider a prior distribution of classifiers formed by an ensemble of equiprobable spherical Gaussian distributions* $\{P_j(c|\mathbf{w}_r,\eta_j)\}_{j=1}^J$ *with identity covariance and mean* $\eta_j\mathbf{w}_r$*, where* $\{\eta_j\}_{j=1}^J$ *are positive real numbers and* $\mathbf{w}_r$ *is a linear classifier trained using a subset* T *of* r *samples a priori separated from a training set* S *of* m *samples. Then, for all distributions* $\mathcal{D}$*, for all posteriors* $(\mathbf{w},\mu)$ *and for all* $\delta \in (0,1]$*, we have that with probability greater than* $1-\delta$ *over all the training sets* S *of size* m *sampled from* $\mathcal{D}$

$$KL_+(\hat{Q}_{S\setminus T}(\mathbf{w},\mu)||Q_{\mathcal{D}}(\mathbf{w},\mu)) \leq \min_j \frac{\frac{||\eta_j\mathbf{w}_r-\mu\mathbf{w}||^2}{2} + \ln(\frac{m-r+1}{\delta}) + \ln J}{m-r}.$$

Proof The proof is straightforward and can be completed by substituting $1/J$ for all π_j in Theorem 6 and computing the KL divergence between prior and posterior as in the proof of Corollary 5. ■

Note that the $\{\eta_j\}_{j=1}^J$ must be chosen before we actually compute the posterior. A linear search can be implemented for the value of μ that leads to the tightest bound for each particular prior. In the case of a mixture prior, the search is repeated for every member of the ensemble and the reported value of the bound is the tightest one found during the searches.

Moreover, the data distribution can also shape the covariance matrix of the Gaussian prior. Rather than take a spherically symmetric prior distribution we choose the variance in the direction of the prior vector to be $\tau > 1$. As with the prior PAC-Bayes bound the mean of the prior distribution is also shifted from the original in the direction $\mathbf{w}_r$. Seeger (2002) has previously considered non-spherical priors and (different) non-spherical posteriors in bounding Gaussian process classification. Our application to SVMs is not restricted to using specific priors and posteriors so that we have the flexibility to adapt our distributions in order to accommodate the prior derived from the last part of the data.

We introduce notation for the norms of projections for unit vector $\mathbf{u}$, $P_{\mathbf{u}}^{\|}(\mathbf{v}) = \langle \mathbf{u},\mathbf{v}\rangle$ and $P_{\mathbf{u}}^{\perp}(\mathbf{v})^2 = \|\mathbf{v}\|^2 - P_{\mathbf{u}}^{\|}(\mathbf{v})^2$.

Theorem 8 (τ-prior PAC-Bayes bound for linear classifiers) *Let us consider a prior* $P(c|\mathbf{w}_r,\tau,\eta)$ *distribution of classifiers consisting of a Gaussian distribution centred on* $\eta\mathbf{w}_r$*, with identity covariance matrix in all directions except* $\mathbf{w}_r$ *in which the variance is* τ^2*. Then, for all distributions* $\mathcal{D}$*, for all* $\delta \in (0,1]$*, we have that with probability at least* $1-\delta$ *over all the training samples of size* m *drawn from* $\mathcal{D}$*, for all posterior parameters* ($\mathbf{w}$*,* μ)*,*

$$KL(\hat{Q}_{S\setminus T}(\mathbf{w},\mu)\|Q_{\mathcal{D}}(\mathbf{w},\mu)) \leq$$

$$\frac{(\ln(\tau^2)+\tau^{-2}-1+P^{\|}_{\mathbf{w}_r}(\mu\mathbf{w}-\eta\mathbf{w}_r)^2/\tau^2+P^{\perp}_{\mathbf{w}_r}(\mu\mathbf{w})^2)+2\ln(\frac{m-r+1}{\delta})}{2(m-r)}.$$

Proof The application of the PAC-Bayes theorem follows that of Langford (2005) except that we must recompute the KL divergence. Using the expression for the KL divergence between two Gaussian distributions of (6) we obtain

$$\begin{aligned}&\mathrm{KL}(Q(\mathbf{w},\mu)\|P(\mathbf{w}_r,\tau,\eta))=\\&\frac{1}{2}\left(\ln(\tau^2)+\left(\frac{1}{\tau^2}-1\right)+\frac{P^{\|}_{\mathbf{w}_r}(\mu\mathbf{w}-\eta\mathbf{w}_r)^2}{\tau^2}+P^{\perp}_{\mathbf{w}_r}(\mu\mathbf{w})^2\right),\end{aligned}$$

and the result follows. ■

Note that the quantity

$$\hat{Q}_{S\setminus T}(\mathbf{w},\mu)=\mathbb{E}_{m-r}[\tilde{F}(\mu\gamma(\mathbf{x},y))]$$

remains unchanged as the posterior distribution is still a spherical Gaussian centred at $\mathbf{w}$.

3.2 Expectation-Prior PAC-Bayes Bound for SVMs

In this section, we attempt to start an interesting new approach on exploiting priors without the aid of a separate data set. The basic idea is to adopt the mathematical expectation of some quantity and then approximate this expectation by an empirical average computed on the available data.

An expectation that may result in reasonable priors is $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\phi(\mathbf{x})]$, which is used in the derivation of the bound below. Define $\mathbf{w}_p=\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y\phi(\mathbf{x})]$ where $y\in\{+1,-1\}$. A special case of $\mathbf{w}_p$ is $\frac{1}{2}(\mathbf{w}^+-\mathbf{w}^-)$ with $\mathbf{w}^+=\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D},y=+1}[\phi(\mathbf{x})]$, $\mathbf{w}^-=\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D},y=-1}[\phi(\mathbf{x})]$ when each class has the same prior probability. We use its general form in deriving bounds.

Given a sample set S including m examples, the empirical estimate of $\mathbf{w}_p$ would be $\hat{\mathbf{w}}_p=\mathbb{E}_{(\mathbf{x},y)\sim S}[y\phi(\mathbf{x})]=\frac{1}{m}\sum_{i=1}^{m}[y_i\phi(\mathbf{x}_i)]$. We have the following bound.

Theorem 9 (Single-expectation-prior PAC-Bayes bound for SVMs) *For all $\mathcal{D}$, for all Gaussian prior $P\sim\mathcal{N}(\eta\mathbf{w}_p,I)$ over margin classifiers, for all $\delta\in(0,1]$:*

$$\begin{aligned}Pr_{S\sim\mathcal{D}^m}\quad&(\forall\mathbf{w},\mu: KL_+(\hat{Q}_S(\mathbf{w},\mu)||Q_{\mathcal{D}}(\mathbf{w},\mu))\leq\\&\frac{\frac{1}{2}(\|\mu\mathbf{w}-\eta\hat{\mathbf{w}}_p\|+\eta\frac{R}{\sqrt{m}}(2+\sqrt{2\ln\frac{2}{\delta}}))^2+\ln(\frac{2(m+1)}{\delta})}{m})\geq 1-\delta,\end{aligned}$$

where the posterior is $Q\sim\mathcal{N}(\mu\mathbf{w},I)$ with $R=\sup_{\mathbf{x}}\|\phi(\mathbf{x})\|$.

Proof First, we try to bound $KL(Q||P)$. We have

$$\begin{aligned}KL(Q||P)&=\frac{1}{2}\|\mu\mathbf{w}-\eta\mathbf{w}_p\|^2\\&=\frac{1}{2}\|\mu\mathbf{w}-\eta\hat{\mathbf{w}}_p+\eta\hat{\mathbf{w}}_p-\eta\mathbf{w}_p\|^2\\&=\frac{1}{2}\|\mu\mathbf{w}-\eta\hat{\mathbf{w}}_p\|^2+\frac{1}{2}\|\eta\hat{\mathbf{w}}_p-\eta\mathbf{w}_p\|^2+(\mu\mathbf{w}-\eta\hat{\mathbf{w}}_p)^{\top}(\eta\hat{\mathbf{w}}_p-\eta\mathbf{w}_p)\\&\leq\frac{1}{2}\|\mu\mathbf{w}-\eta\hat{\mathbf{w}}_p\|^2+\frac{1}{2}\eta^2\|\hat{\mathbf{w}}_p-\mathbf{w}_p\|^2+\eta\|\mu\mathbf{w}-\eta\hat{\mathbf{w}}_p\|\|\hat{\mathbf{w}}_p-\mathbf{w}_p\|,\end{aligned}\tag{7}$$

where the last inequality uses Cauchy-Schwarz inequality. Now it suffices to bound $\|\hat{\mathbf{w}}_p - \mathbf{w}_p\|$.

Define $R = \sup_{\mathbf{x}} \|\phi(\mathbf{x})\|$. It is simple to show that $\sup_{(\mathbf{x},y)} \|y\phi(\mathbf{x})\| = \sup_{\mathbf{x}} \|\phi(\mathbf{x})\| = R$. With reference to a result on estimating the center of mass (Shawe-Taylor and Cristianini, 2004), we have

$$Pr\left(\|\hat{\mathbf{w}}_p - \mathbf{w}_p\| \geq \frac{2R}{\sqrt{m}} + \varepsilon\right) \leq \exp\left(-\frac{2m\varepsilon^2}{4R^2}\right).$$

Setting the right hand side equal to $\delta/2$, solving for ε shows that with probability at least $1-\delta/2$, we have

$$\|\hat{\mathbf{w}}_p - \mathbf{w}_p\| \leq \frac{R}{\sqrt{m}}\left(2 + \sqrt{2\ln\frac{2}{\delta}}\right). \tag{8}$$

Define $b = \frac{R}{\sqrt{m}}\left(2+\sqrt{2\ln\frac{2}{\delta}}\right)$, we have

$$Pr_{S\sim\mathcal{D}^m}\left(KL(Q||P) \leq \frac{1}{2}\|\mu\mathbf{w} - \eta\hat{\mathbf{w}}_p\|^2 + \frac{1}{2}\eta^2 b^2 + \eta b\|\mu\mathbf{w} - \eta\hat{\mathbf{w}}_p\|\right) \geq 1-\delta/2. \tag{9}$$

Then, according to Theorem 3, we have

$$Pr_{S\sim\mathcal{D}^m}\left(\forall Q(c) : KL_+(\hat{Q}_S||Q_{\mathcal{D}}) \leq \frac{KL(Q||P) + \ln(\frac{2(m+1)}{\delta})}{m}\right) \geq 1-\delta/2. \tag{10}$$

Define $a = \|\mu\mathbf{w} - \eta\hat{\mathbf{w}}_p\|$. Combining (9) and (10), we get

$$Pr_{S\sim\mathcal{D}^m}\left(\forall \mathbf{w},\mu : KL_+(\hat{Q}_S(\mathbf{w},\mu)||Q_{\mathcal{D}}(\mathbf{w},\mu)) \leq \frac{\frac{1}{2}a^2 + \frac{1}{2}\eta^2 b^2 + \eta ab + \ln(\frac{2(m+1)}{\delta})}{m}\right) \geq 1-\delta,$$

where we used $(1-\delta/2)^2 > 1-\delta$. Rewriting the bound as

$$Pr_{S\sim\mathcal{D}^m}\left(\forall \mathbf{w},\mu : KL_+(\hat{Q}_S(\mathbf{w},\mu)||Q_{\mathcal{D}}(\mathbf{w},\mu)) \leq \frac{\frac{1}{2}(a+\eta b)^2 + \ln(\frac{2(m+1)}{\delta})}{m}\right) \geq 1-\delta$$

completes the proof. ■

Considering at the same time Theorem 9 and the mixture-prior PAC-Bayes bound, it is not difficult to reach the following mixture-expectation-prior PAC-Bayes bound for SVMs.

Theorem 10 (Mixture-expectation-prior PAC-Bayes bound for SVMs) *For all $\mathcal{D}$, for all mixtures of Gaussian prior $\mathcal{P}(c) = \sum_{j=1}^{J} \pi_j P_j(c)$ where $P_j \sim \mathcal{N}(\eta_j \mathbf{w}_p, I)$ $(j = 1,\ldots,J)$, $\pi_j \geq 0$ and $\sum_{j=1}^{J}\pi_j = 1$ over margin classifiers, for all $\delta \in (0,1]$:*

$$Pr_{S\sim\mathcal{D}^m} \quad \left(\forall \mathbf{w},\mu : KL_+(\hat{Q}_S(\mathbf{w},\mu)||Q_{\mathcal{D}}(\mathbf{w},\mu)) \leq \min_j \frac{\frac{1}{2}(\|\mu\mathbf{w} - \eta_j\hat{\mathbf{w}}_p\| + \eta_j\frac{R}{\sqrt{m}}(2+\sqrt{2\ln\frac{2}{\delta}}))^2 + \ln(\frac{2(m+1)}{\delta}) + \ln\frac{1}{\pi_j}}{m}\right) \geq 1-\delta,$$

where the posterior is $Q \sim \mathcal{N}(\mu\mathbf{w}, I)$ with $R = \sup_{\mathbf{x}} \|\phi(\mathbf{x})\|$. If we consider equiprobable members in the mixture, then $\ln\frac{1}{\pi_j} = \ln J$.

Moreover, the expectation prior bound can also be extended to the case where the shape of the covariance matrix of the prior is also determined from the training data:

Theorem 11 (τ-Expectation-prior PAC-Bayes bound) *Consider a prior distribution $P \sim \mathcal{N}(\eta \mathbf{w}_p, I, \tau^2)$ of classifiers consisting of a Gaussian distribution centred on $\eta \mathbf{w}_p$, with identity covariance in all directions except $\mathbf{w}_p$ in which the variance is τ^2. Then, for all distributions $\mathcal{D}$, for all $\delta \in (0,1]$, we have*

$$Pr_{S\sim\mathcal{D}^m} \quad (\forall \mathbf{w}, \mu : KL_+(\hat{Q}_S(\mathbf{w},\mu)||Q_{\mathcal{D}}(\mathbf{w},\mu)) \leq \frac{\frac{1}{2}(\ln(\tau^2) + \frac{(\|\mu\mathbf{w} - \eta\hat{\mathbf{w}}_p\| + \eta\frac{R}{\sqrt{m}}(2+\sqrt{2\ln\frac{2}{\delta}}))^2 - \mu^2 + 1}{\tau^2} + \mu^2 - 1) + \ln(\frac{2(m+1)}{\delta})}{m}) \geq 1-\delta,$$

where the posterior is $Q \sim \mathcal{N}(\mu\mathbf{w}, I)$ with $R = \sup_{\mathbf{x}} \|\phi(\mathbf{x})\|$. We can recover Theorem 9 by taking $\tau = 1$.

Proof According to Theorem 8,

$$KL(Q||P) = \frac{1}{2}\left(\ln(\tau^2) + \frac{1}{\tau^2} - 1 + \frac{P^{\|}_{\mathbf{w}_p^*}(\mu\mathbf{w} - \eta\mathbf{w}_p)^2}{\tau^2} + P^{\perp}_{\mathbf{w}_p^*}(\mu\mathbf{w})^2\right),$$

where $\mathbf{w}_p^* = \mathbf{w}_p / \|\mathbf{w}_p\|$. The last two quantities can be rewritten as

$$\begin{aligned}
\frac{P^{\|}_{\mathbf{w}_p^*}(\mu\mathbf{w} - \eta\mathbf{w}_p)^2}{\tau^2} + P^{\perp}_{\mathbf{w}_p^*}(\mu\mathbf{w})^2 &= \frac{1}{\tau^2}(\frac{\mathbf{w}_p^\top}{\|\mathbf{w}_p\|}(\mu\mathbf{w} - \eta\mathbf{w}_p))^2 + \|\mu\mathbf{w}\|^2 - (\frac{\mathbf{w}_p^\top}{\|\mathbf{w}_p\|}\mu\mathbf{w})^2 \\
&= \frac{1}{\tau^2}(\frac{\mathbf{w}_p^\top}{\|\mathbf{w}_p\|}\mu\mathbf{w} - \eta\|\mathbf{w}_p\|)^2 + \|\mu\mathbf{w}\|^2 - (\frac{\mathbf{w}_p^\top}{\|\mathbf{w}_p\|}\mu\mathbf{w})^2 \\
&= \frac{1}{\tau^2}(\eta^2\|\mathbf{w}_p\|^2 - 2\eta\mathbf{w}_p^\top\mu\mathbf{w}) + \|\mu\mathbf{w}\|^2 \\
&= \frac{1}{\tau^2}(\|\mu\mathbf{w} - \eta\mathbf{w}_p\|^2 - \|\mu\mathbf{w}\|^2) + \|\mu\mathbf{w}\|^2 \\
&= \frac{1}{\tau^2}(\|\mu\mathbf{w} - \eta\mathbf{w}_p\|^2 - \mu^2) + \mu^2 .
\end{aligned}$$

By Equation (7), we have

$$\|\mu\mathbf{w} - \eta\mathbf{w}_p\|^2 \leq \|\mu\mathbf{w} - \eta\hat{\mathbf{w}}_p\|^2 + \eta^2\|\hat{\mathbf{w}}_p - \mathbf{w}_p\|^2 + 2\eta\|\mu\mathbf{w} - \eta\hat{\mathbf{w}}_p\|\|\hat{\mathbf{w}}_p - \mathbf{w}_p\| .$$

By Equation (8), we have with probability at least $1 - \delta/2$

$$\|\hat{\mathbf{w}}_p - \mathbf{w}_p\| \leq \frac{R}{\sqrt{m}}\left(2 + \sqrt{2\ln\frac{2}{\delta}}\right).$$

With $a = \|\mu\mathbf{w} - \eta\hat{\mathbf{w}}_p\|$ and $b = \frac{R}{\sqrt{m}}\left(2 + \sqrt{2\ln\frac{2}{\delta}}\right)$, we have

$$Pr_{S\sim\mathcal{D}^m}\left(KL(Q||P) \leq \frac{1}{2}(\ln(\tau^2) + \frac{1}{\tau^2} - 1 + \frac{a^2 + \eta^2 b^2 + 2\eta ab - \mu^2}{\tau^2} + \mu^2)\right) \geq 1 - \delta/2. \qquad (11)$$

Then, according to Theorem 3, we have

$$Pr_{S\sim\mathcal{D}^m}(\forall Q(c) : KL_+(\hat{Q}_S||Q_{\mathcal{D}}) \leq \frac{KL(Q||P)+\ln(\frac{2(m+1)}{\delta})}{m}) \geq 1-\delta/2. \tag{12}$$

Combining (11) and (12) results in

$$Pr_{S\sim\mathcal{D}^m} \quad (\forall \mathbf{w},\mu : KL_+(\hat{Q}_S(\mathbf{w},\mu)||Q_{\mathcal{D}}(\mathbf{w},\mu)) \leq \frac{\frac{1}{2}(\ln(\tau^2)+\frac{(a+\eta b)^2-\mu^2+1}{\tau^2}+\mu^2-1)+\ln(\frac{2(m+1)}{\delta})}{m}) \geq 1-\delta,$$

which completes the proof. ■

4. Optimising the Prior PAC-Bayes Bound in the Design of the Classifier

Up to this point we have introduced the prior PAC-Bayes bounds as a means to tighten the original PAC-Bayes bound (this fact is illustrated in the experiments included in Section 5). The next contribution of this paper consists of the introduction of the optimisation of the prior PAC-Bayes bound into the design of the classifier. The intuition behind this use of the bounds is that classifiers reporting low values for the bound should yield a good generalization capability.

4.1 Prior SVM

The new philosophy is implemented in the prior SVM by replacing the maximization of the margin in the optimization problem defining the original SVM with a term that pushes towards the tightening of the prior PAC-Bayes bound. This subsection introduces the formulation of the new algorithm, a method to determine the classifier by means of off-the-shelf quadratic programming solvers, and a procedure to compute the prior PAC-Bayes bound for these new classifiers.

4.1.1 FORMULATION OF THE PRIOR SVMS

As stated before, the design criterion for the prior SVMs involves the minimization of the prior PAC-Bayes bound. Let us consider the simplest case of the bound, that is, a single prior centered on $\eta\mathbf{w}_r$, where $\mathbf{w}_r$ is the unit vector weight of the SVM constructed with r training samples and η is a scalar fixed a priori. For simplicity, we assume these r samples are the last ones in the training set $\{(\mathbf{x}_l, y_l)\}_{l=m-r+1}^{m}$. Therefore, $\mathbf{w}_r$ can be expressed in terms of these input patterns as:

$$\mathbf{w}_r = \frac{\sum_{l=m-r+1}^{m} y_l \alpha_l \phi(\mathbf{x}_l)}{\left\|\sum_{l=m-r+1}^{m} y_l \alpha_l \phi(\mathbf{x}_l)\right\|}.$$

In such a case, a small bound on the error of the classifier is the result of a small value of $\|\eta\mathbf{w}_r - \mu\mathbf{w}\|^2$, and a large value of the normalized margin of Equation (3) for the remaining training examples $\gamma(\mathbf{x}_i, y_i)$, $i = 1,\ldots,m-r$.

We start by addressing the separable case. Under perfect separability conditions, a good strategy to obtain a classifier of minimal bound is to solve the following optimization problem:

$$\min_{\mathbf{w}} \left[\frac{1}{2}\|\mathbf{w}-\eta\mathbf{w}_r\|^2\right] \tag{13}$$

subject to

$$y_i \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) \geq 1 \quad i = 1, \ldots, m-r. \tag{14}$$

Clearly, the objective function of (13) attempts to reduce the value of the right hand side of the bound, while the constraints in (14) that impose the separability of the classes lead to a small $\hat{Q}_S$.

Once $\mathbf{w}$ is found through the solution of (13) with constraints (14) the proper bound on the average true error of the prior SVM can be obtained by means of a further tuning of μ (that is, using $\mu\mathbf{w}$ instead of $\mathbf{w}$ as mean of the posterior distribution), where this last tuning will not change $\mathbf{w}$.

The extension of the prior SVM to the non-separable case is easily carried out through the introduction of positive slack variables $\{\xi_i\}_{i=1}^{m-r}$. Then the optimization problem becomes

$$\min_{\mathbf{w},\xi_i} \left[\frac{1}{2} \|\mathbf{w} - \mathbf{w}_r\|^2 + C \sum_{i=1}^{m-r} \xi_i \right] \tag{15}$$

subject to

$$y_i \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) \geq 1 - \xi_i \qquad i = 1, \ldots, m-r, \tag{16}$$

$$\xi_i \geq 0 \qquad i = 1, \ldots, m-r. \tag{17}$$

Note that the constraints in (16) also push towards the minimization of the stochastic error $\hat{Q}_S$. In this sense, for a sample $\mathbf{x}$ on the wrong side of the margin we have $\xi = 1 - y\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) > 1$, which leads to a margin $\gamma < 0$ and thus an increase in $\hat{Q}_S$ (see Equations (2) to (4)). Therefore, by penalizing ξ we enforce a small $\hat{Q}_S$.

Furthermore, Corollary 7 allows us to use a mixture of J distributions instead of one at the cheap cost of $\frac{\ln J}{m}$. This can be used to refine the selection of the weight vector of the prior SVMs through the following procedure:

1. First we determine a unit $\mathbf{w}_r$ with samples $\{(\mathbf{x}_l, y_l)\}_{l=m-r+1}^{m}$. Then we construct a mixture prior with J Gaussian components with identity covariance matrices centered at $\eta_j \mathbf{w}_r$, with η_j being J real positive constants.

2. For every element in the mixture we obtain a prior SVM classifier $\mathbf{w}^j$ solving

$$\min_{\mathbf{w}^j,\xi_i} \left[\frac{1}{2} \|\mathbf{w}^j - \eta_j \mathbf{w}_r\|^2 + C \sum_{i=1}^{m-r} \xi_i \right]$$

   subject to

$$y_i \phi(\mathbf{x}_i)^T w^j \geq 1 - \xi_i \quad i = 1, \ldots, m-r,$$

$$\xi_i \geq 0 \quad i = 1, \ldots, m-r.$$

   Afterwards, we obtain the bounds $Q_{\mathcal{D}}^j$ corresponding to the average true error of each one of the J prior SVMs by tuning μ (see Corollary 6).

3. We finally select as the prior SVM the $\mathbf{w}^j$ that reports the lowest bound $Q_{\mathcal{D}}^j$.

It should be pointed out that each prior scaling (η_j) that is tried increases the computational burden of the training of the prior SVMs by an amount corresponding to an SVM problem with $m-r$ data points.

Appendix A details a procedure to determine the solution $\mathbf{w}$ to the optimization problem given by (15) and constraints (16) and (17) based on the usual derivation of the SVM.

4.1.2 COMPUTING THE PAC-BAYES BOUND FOR THE PRIOR SVMS

The remainder of the section presents a method to compute the PAC-Bayes bound for a prior SVM obtained through the procedure described above. To simplify notation we have introduced the nonunit weight vector $\mathbf{w}_{m-r} = \mathbf{w} - \eta \mathbf{w}_r$, that includes the posterior part of the prior SVM. The bound is based on the relationship between two distributions of classifiers: the prior $P(\mathbf{w}_r, \eta) \sim \mathcal{N}(\eta \mathbf{w}_r, I)$ and the posterior $Q(\mathbf{w}, \mu) \sim \mathcal{N}(\mu \mathbf{w}, I)$.

The stochastic error $\hat{Q}_S$ in the left hand side of the bound can be straightforwardly obtained by using a unit $\mathbf{w}$ in (27) in Equations (2) to (4). For the right hand side of the bound, we need to compute $\mathrm{KL}(Q(\mathbf{w}, \mu)||P(\mathbf{w}_r, \eta)) = \frac{\|\eta \mathbf{w}_r - \mu \mathbf{w}\|^2}{2}$ which can be rewritten as

$$\mathrm{KL}(Q(\mathbf{w}, \mu)||P(\mathbf{w}_r, \eta)) = \frac{1}{2}\left(\mu^2 + \eta^2 - 2\mu\eta(\eta + \mathbf{w}_{m-r}^T \mathbf{w}_r)\right).$$

4.2 η-Prior SVM

When the prior SVM is learnt within a mixture priors setting, the last stage of the optimization is the selection of the best prior-component/posterior pair, among the J possibilities. These prior-component/posterior pairs are denoted by $(\eta_j, \mathbf{w}_j)$, where η_j is the jth scaling of the normalized prior $\mathbf{w}_r$. From the point of view of the prior, this selection process can be regarded as a search over the set of scalings using the mixture-prior PAC-Bayes bound as fitness function. Note that the evaluation of such a fitness function involves learning the posterior and the tuning of μ.

The idea presented in this section actually consists of two turns of the screw. First, the search in the discrete set of priors is cast as a linear search for the optimal scaling η in a continuous range of scalings $[\eta_1, \eta_J]$. Second, this linear search is introduced into the optimization of the posterior. Therefore, instead of optimizing a posterior for every scaling of the prior, the optimal scaling and posterior given a normalized prior are the output of the same optimization problem.

The sequel is devoted to the derivation of the resulting algorithm, called the η-prior SVMs, and to its analysis using the prior PAC-Bayes bound framework.

4.2.1 DERIVATION OF THE η-PRIOR SVMS

The η-prior SVM is designed to solve the following problem:

$$\min_{\mathbf{v}, \eta, \xi_i} \left[\frac{1}{2}\|\mathbf{v}\|^2 + C \sum_{i=1}^{m-r} \xi_i \right]$$

subject to

$$\begin{aligned} y_i(\mathbf{v} + \eta \mathbf{w}_r)^T \phi(\mathbf{x}_i) &\geq 1 - \xi_i \qquad i = 1, \ldots, m-r, \\ \xi_i &\geq 0 \qquad i = 1, \ldots, m-r. \end{aligned}$$

The final (unit vector) classifier will be

$$\mathbf{w} = (\mathbf{v} + \eta \mathbf{w}_r)/\|\mathbf{v} + \eta \mathbf{w}_r\|.$$

After a derivation analogous to that presented in Appendix A, we arrive at the following quadratic program

$$\max_{\alpha_i} \sum_{i=1}^{m-r} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m-r} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$$

subject to

$$\sum_{i=1}^{m-r}\sum_{k=m-r+1}^{m}\alpha_i y_i \tilde{\alpha}_k y_k \kappa(\mathbf{x}_i,\mathbf{x}_k) = \sum_{i=1}^{m-r} y_i \alpha_i g_i = 0 \qquad i = 1,\ldots,m-r,$$
$$0 \le \alpha_i \le C \qquad i = 1,\ldots,m-r,$$

where $g_i = \sum_{k=m-r+1}^{m} \tilde{\alpha}_k y_k \kappa(\mathbf{x}_i,\mathbf{x}_k)$ and $\tilde{\alpha}_k$ are the normalized dual variables for the prior learnt from the last r samples, $\{\mathbf{x}_k\}_{k=m-r+1}^{m}$. Once we have solved for α_i, we can compute η by considering some j such that $0 < \alpha_j < C$ and using the equation

$$y_j\left(\sum_{i=1}^{m-r}\alpha_i y_i \kappa(\mathbf{x}_i,\mathbf{x}_j) + \eta g_j\right) = 1.$$

4.2.2 BOUNDS FOR η-PRIOR SVMS

The statistical analysis of the η-prior SVMs can be performed using the τ-prior PAC-Bayes bound of Theorem 8, and τ-expectation prior PAC-Bayes bound. Rather than take a spherically symmetric prior distribution we choose the variance in the direction of the prior vector to be $\tau^2 > 1$. As with the prior SVM analysis the mean of the prior distribution is also shifted from the origin in the direction $\mathbf{w}_r$.

In order to apply the bound we need to consider the range of priors that are needed to cover the data in our application. The experiments conducted in the next section require a range of scalings of $\mathbf{w}_r$ from 1 to 100. For this we can choose $\eta = 50$, $\tau = 50$, and $\mu \le 100$ in all but one of our experiments, giving an increase in the bound over the factor $P_{\mathbf{w}_r}^{\perp}(\mu\mathbf{w})^2$ directly optimized in the algorithm of

$$\frac{\ln(\tau^2) + \tau^{-2} - 1 + P_{\mathbf{w}_r}^{\parallel}(\mu\mathbf{w} - \eta\mathbf{w}_r)^2/\tau^2}{2(m-r)} \le \frac{\ln(\tau) + 0.5\tau^{-2}}{m-r} \approx \frac{3.912}{m-r}. \tag{18}$$

We include Equation (18) to justify that our algorithm optimises a quantity that is very close to the expression in the bound. Note that the evaluation of the bounds presented in the experimental section are computed using the expression from Theorem 8 and not this approximate upper bound. One could envisage making a sequence of applications of the PAC-Bayes bound with spherical priors using the union bound and applying the result with the nearest prior. This strategy leads to a slightly worse bound as it fails to take into account the correlations between the different priors. This fact is illustrated in Section 5.

5. Experiments

This section is devoted to an experimental analysis of the bounds and algorithms introduced in the paper. The comparison of the algorithms is carried out on classification preceded by model selection tasks using some UCI (Blake and Merz, 1998) data sets (see their description in terms of number of instances, input dimensions and numbers of positive/negative examples in Table 1).

5.1 Experimental Setup

For every data set, we prepared 50 different training/test set partitions where 80% of the samples form the training set and the remaining 20% form the test set. From every training set we considered

Problem	# Examples	Input Dim.	Pos/Neg
Handwritten-digits (han)	5620	64	2791 / 2829
Waveform (wav)	5000	21	1647 / 3353
Pima (pim)	768	8	268 / 500
Ringnorm (rin)	7400	20	3664 / 3736
Spam (spa)	4601	57	1813 / 2788

Table 1: Description of data sets in terms of number of examples, number of input variables and number of positive/negative examples.

subsets with 20%, 30%, ..., 100% of the training patterns, in order to analyse the dependence of the bounds with the number of samples used to train the classifier. Note that all the training subsets from the same partition share the same test set.

With each of the training sets we learned a classifier with Gaussian RBF kernels preceded by a model selection. The model selection consists in the determination of an optimal pair of hyperparameters (C, σ). C is the SVM trade-off between the maximization of the margin and the minimization of the number of misclassified training samples; σ is the width of the Gaussian kernel, $\kappa(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/(2\sigma^2))$. The best pair is sought in a 7×5 grid of parameters where $C \in \{0.01, 0.1, 1, 10, 100, 1000, 10000\}$ and $\sigma \in \{\frac{1}{4}\sqrt{d}, \frac{1}{2}\sqrt{d}, \sqrt{d}, 2\sqrt{d}, 4\sqrt{d}\}$, d being the input space dimension.

With respect to the parameters needed by the prior PAC-Bayes bounds, the number of priors J and the amount of patterns separated to learn the prior, the experiments reported by Ambroladze et al. (2007) suggest that $J = 10$ and $r = 50\%$ of the training set size lead to reasonable results.

The setup to calculate the bound values displayed in the next tables was as follows. We trained an instance of the corresponding classifier for each position of the grid of hyperparameters and compute the bound. We selected for that type of classifier the minimum value of the bound found through the whole grid. Then we averaged the 50 values of the bound corresponding to each of the training/testing partitions. We completed the average with the sample standard deviation. Note that proceeding this way we select a (possibly) different pair of hyperparameters for each of the 50 partitions. That is the reason why we name this task model selection plus classification.

The test error rates are computed after the following procedure. For each one of the training/test partitions we carried out the model selection described in the previous paragraph and selected the classifier of minimum bound. We classified the test set with this classifier and obtain the test error rate for those particular classifier and partition. Then we averaged the 50 test error rates to yield the test error rate for those particular data set, model selection method and type of classifier. Note again that the model selection has a significant impact on the reported test error rates.

Moreover, the reported values of the PAC-Bayes and the mixture-prior PAC-Bayes bounds correspond to the mean of the true error over the distribution of classifiers Q_D. The real true error c_D could then be bounded by twice this value (see Equation (5)). In all the experiments the bounds are obtained using a confidence of $\delta = 0.01$.

5.2 Results and Discussion

The section starts presenting an analysis of the performance of SVM with the prior PAC Bayes bounds introduced in this paper. We show how in most cases the use of an informative prior leads to a significant tightening of the bounds on the true error of the classifier. The analysis is then extended towards the new algorithms prior SVM and η-prior SVM. We show how their true error is predicted more accurately by the prior PAC Bayes bound. The observed test errors achieved by these algorithms are comparable to those obtained by SVMs with their hyperparameters fixed through ten fold cross validation. Finally, the prior SVM framework enables the use of a different value of parameter C for prior and posterior, that can be tuned using the prior PAC Bayes bound. The experiments show that the use of different values of C contributes to get even tighter lower bounds.

5.2.1 ANALYSIS OF THE SVM WITH THE PRIOR PAC BAYES BOUNDS

The first set of experiments is devoted to illustrate how tight can be the predictions about the generalisation capabilities of a regular SVM based upon the prior PAC-Bayes bounds. Thus, we have trained SVM using the hyperparameters that arrived at a minimum value of each of the following bounds:

PAC Bayes: the model selection is driven by the PAC Bayes bound of Langford (2005).

Prior PB: model selection driven by the mixture-prior PAC-Bayes bound of Corollary 7 with $J = 10$.

τ-prior PB: τ-prior PAC-Bayes bound of Theorem 8 with $J = 10$ and $\tau = 50$.

$\mathbb{E}$ prior PB: expectation-prior PAC-Bayes bound of Theorem 10.

τ-$\mathbb{E}$ prior PB: τ-expectation prior PAC-Bayes bound of Theorem 11.

Plots in Figure 1 show the performance of the different bounds as a function of the training set size. All the bounds achieve non trivial results even for training set sizes as small as 16% of the complete data set (20% of the training set). In most of the cases, the bounds with an informative prior are tighter than the original PAC Bayes bound with an spherical prior centred on the origin. The expectation prior is significantly better in data sets `wav` and `pim`, whilst the prior PAC Bayes and the τ-prior PAC Bayes are the tighter in problems `rin` and `spa`. Table 2 shows the values of the bounds when the SVM is determined using the 100% of the training set (80% of the data).

Moreover, an examination of the slopes of the plots corresponding to the bounds point out that those that learn the prior from a separate training set do converge faster than the original PAC Bayes and the expectation prior PAC Bayes bounds. Since the former present a $m - r$ in the denominator of the right hand side, one could a priori think that their convergence would be slower than that of the latter, with an m in the denominator. However, the experimental results show that it is better to devote those separate training patterns to acquire a more informative prior than to increase the weight of the denominator in the penalty term.

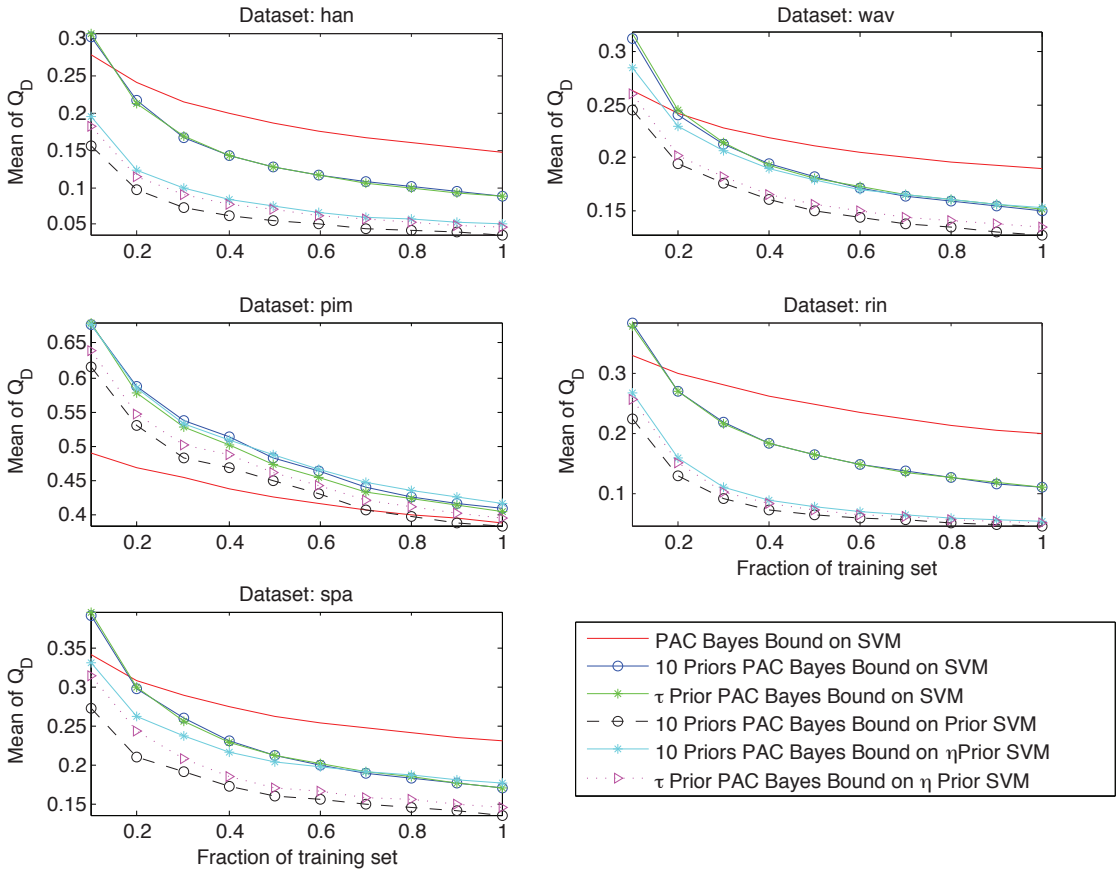

Figure 1: Analysis of SVM with data dependent prior PAC Bayes bounds.

	Data Set				
Bound	han	wav	pim	rin	spa
PAC Bayes	0.148 ± 0.000	0.190 ± 0.000	0.390 ± 0.001	0.198 ± 0.000	0.230 ± 0.000
Prior PB	0.088 ± 0.004	0.151 ± 0.004	0.411 ± 0.015	0.110 ± 0.004	0.171 ± 0.005
τ Prior PB	0.088 ± 0.004	0.152 ± 0.004	0.406 ± 0.013	0.110 ± 0.004	0.172 ± 0.006
$\mathbb{E}$ Prior PB	0.107 ± 0.001	0.133 ± 0.001	0.352 ± 0.004	0.194 ± 0.000	0.221 ± 0.001
τ$\mathbb{E}$ Prior PB	0.149 ± 0.000	0.191 ± 0.000	0.401 ± 0.001	0.199 ± 0.000	0.232 ± 0.000

Table 2: Values of the bounds for SVM.

5.2.2 ANALYSIS OF PRIOR SVM AND η-PRIOR SVM

We repeated the study on the new algorithms, prior SVM and η-prior SVM, which are designed to actually optimise prior PAC-Bayes bounds. The configurations classifier-bound considered for this study were the following:

prior SVM + Prior PB: prior SVM described in page 14 and mixture-prior PAC-Bayes bound of Corollary 7 with $J = 10$ priors .

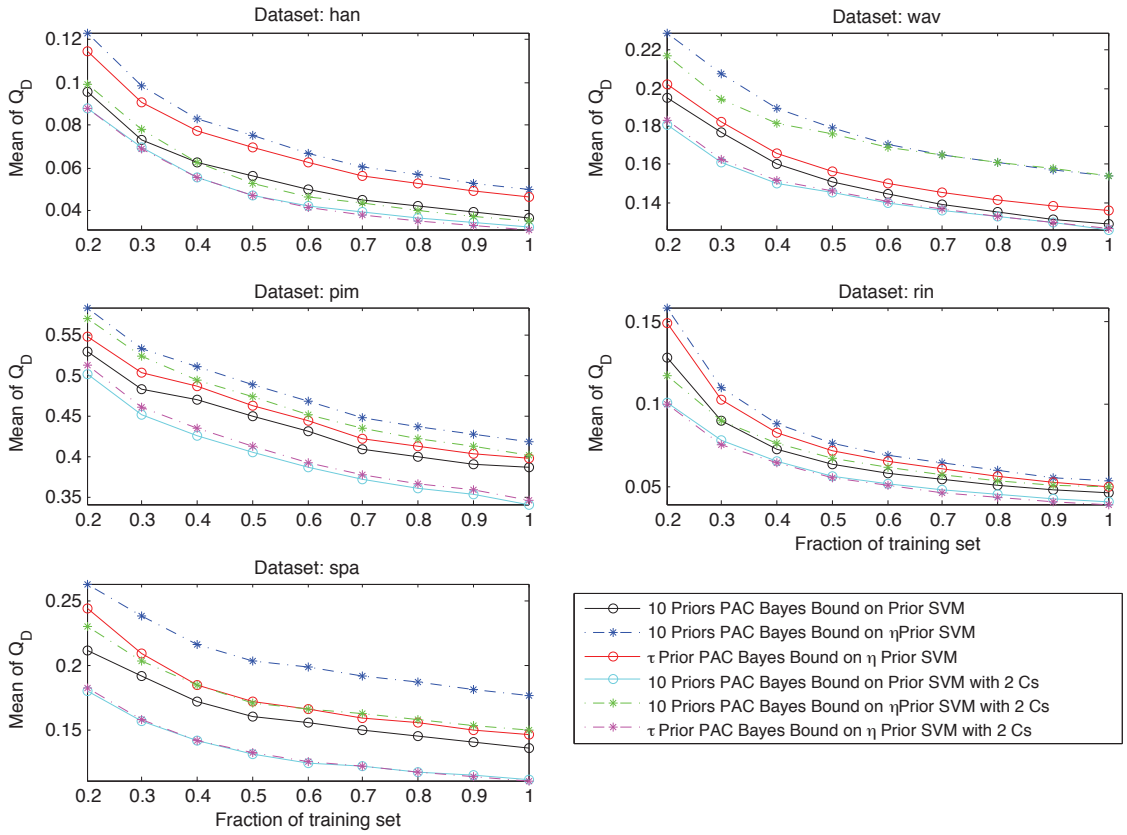

Figure 2: Bounds learning a prior classifier.

η-PSVM + Prior PB: η-prior SVM of Section 4.2.1 and mixture-prior PAC-Bayes bound of Corollary 7 considering η comes from a mixture prior setting of $J = 50$ components $\eta_j \mathbf{w}_r$ with the η_j equally spaced between $\eta_1 = 1$ and $\eta_{50} = 100$. This setting minimizes the penalty term in the prior PAC-Bayes bound as we are not actually using these components to learn the posterior.

η-PSVM + τ-Prior PB: η-prior SVM and the bound in Theorem 8.

As baseline results we include the better bounds found in the analysis of the SVM:

τ-Prior PB: τ prior PAC-Bayes bound of Theorem 8 with $J = 10$ and $\tau = 50$.

$\mathbb{E}$ **Prior PB:** expectation-prior PAC-Bayes bound of Theorem 10.

The plots in Figure 2 show the bounds on the true error, $Q_{\mathcal{D}}$, for the studied configurations bound/classifier as a function of the size of the training set. Table 3 shows these results for a training set of 80% of the complete data. In general, the bounds achieved on prior SVM and η-prior SVM are significantly tighter than the bounds on the SVM, being the mixture-prior PAC Bayes bound on prior SVM the tightest result.

	Data Set				
Bound	han	wav	pim	rin	spa
			Prior SVM		
Prior PB	$\underline{0.037 \pm 0.004}$	$\underline{0.128 \pm 0.004}$	0.386 ± 0.016	$\underline{0.046 \pm 0.003}$	$\underline{0.137 \pm 0.005}$
			η-Prior SVM		
Prior PB	0.050 ± 0.006	0.154 ± 0.004	0.419 ± 0.014	0.053 ± 0.004	0.177 ± 0.006
τ Prior PB	0.047 ± 0.005	0.135 ± 0.004	0.397 ± 0.014	0.050 ± 0.004	0.147 ± 0.006
			SVM		
τ Prior PB	0.088 ± 0.004	0.152 ± 0.004	0.406 ± 0.013	0.110 ± 0.004	0.172 ± 0.006
$\mathbb{E}$ Prior PB	0.107 ± 0.001	0.133 ± 0.001	$\underline{0.352 \pm 0.004}$	0.194 ± 0.000	0.221 ± 0.001

Table 3: Values of the bounds on the prior SVM and η-prior SVM classifiers.

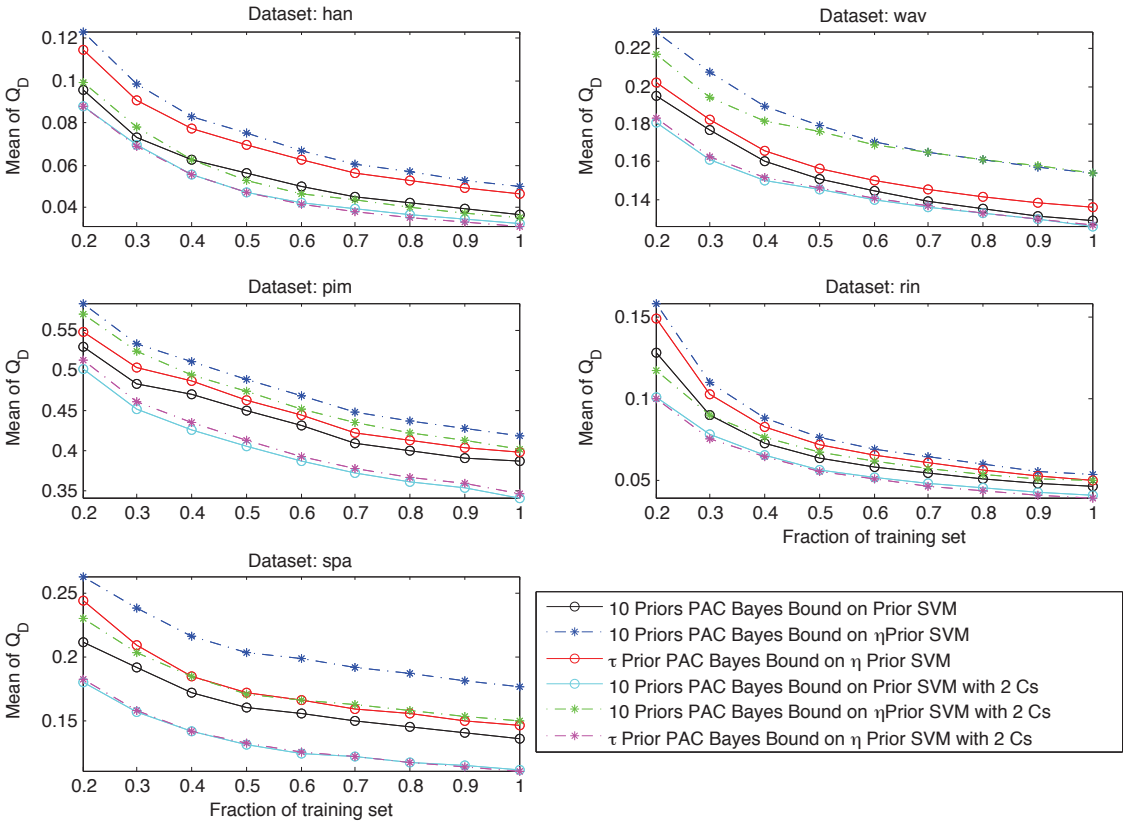

Figure 3: Bounds when prior and posterior have a different value of C.

Notice that in most of the configurations where the prior is learnt from a separate set the new bounds achieve a significant cut in the value of the PAC-Bayes bound, which indicates that learning an informative prior distribution helps to tighten the PAC-Bayes bound.

Furthermore, the two stages training of prior SVM and η-prior SVM enable the use of a different value of C for the prior and posterior classifiers. The intuition behind this proposal is that once the prior is fixed, the posterior could possibly accept a higher value C without overfitting.

	Data Set				
Bound	han	wav	pim	rin	spa
Prior SVM					
Prior PB	0.037 ± 0.004	0.128 ± 0.004	0.386 ± 0.016	0.046 ± 0.003	0.137 ± 0.005
Prior PB 2C	0.033 ± 0.002	0.126 ± 0.004	0.341 ± 0.019	0.041 ± 0.002	0.113 ± 0.004
η-Prior SVM					
Prior PB	0.050 ± 0.006	0.154 ± 0.004	0.419 ± 0.014	0.053 ± 0.004	0.177 ± 0.006
Prior PB 2C	0.035 ± 0.003	0.154 ± 0.004	0.401 ± 0.018	0.049 ± 0.003	0.150 ± 0.005
τ Prior PB	0.047 ± 0.005	0.135 ± 0.004	0.397 ± 0.014	0.050 ± 0.004	0.147 ± 0.006
τ Prior PB 2C	0.031 ± 0.002	0.126 ± 0.004	0.345 ± 0.019	0.039 ± 0.002	0.111 ± 0.005

Table 4: Values of the bounds on the prior SVM and η-prior SVM classifiers when different values of C are used for prior and posterior.

	Data Set				
Bound	han	wav	pim	rin	spa
Prior SVM					
Prior PB	0.010 ± 0.004	0.086 ± 0.007	0.246 ± 0.034	0.016 ± 0.003	0.082 ± 0.009
Prior PB 2C	0.011 ± 0.003	0.091 ± 0.009	0.251 ± 0.038	0.017 ± 0.003	0.069 ± 0.007
η-Prior SVM					
Prior PB	0.010 ± 0.005	0.086 ± 0.006	0.236 ± 0.028	0.016 ± 0.003	0.080 ± 0.009
Prior PB 2C	0.011 ± 0.003	0.087 ± 0.009	0.242 ± 0.039	0.018 ± 0.003	0.068 ± 0.008
τ Prior PB	0.010 ± 0.005	0.085 ± 0.006	0.238 ± 0.028	0.016 ± 0.003	0.080 ± 0.009
τ Prior PB 2C	0.011 ± 0.003	0.092 ± 0.010	0.248 ± 0.042	0.018 ± 0.003	0.070 ± 0.007
SVM					
10 FCV	0.008 ± 0.003	0.087 ± 0.007	0.251 ± 0.023	0.016 ± 0.003	0.067 ± 0.006

Table 5: Test error rates achieved by prior SVM and η-prior SVM classifiers when the hyperparameters are those that minimise a PAC Bayes bound. Prior and posterior are allowed to use a different value of the hyperparameter C.

To evaluate the goodness of this modification, we carried out again the experiments in this subsection but now allowing the prior and posterior to take different values of C from within the range proposed at the beginning of the section. The results displayed in Figure 3 and Table 4 show that the introduction of a different C significantly reduces the value of the bound.

Finally, Table 5 gives some insight about the performance of the new algorithms in terms of observed test error. The joint analysis of the bounds and the error rates on a separate test set shows that the prior PAC Bayes bounds are achieving predictions on the true error very close to the empirical estimations; as an example, for data set `wav` the bound on $Q_{\mathcal{D}}$ is around 13% and the empirical estimation is around 9%. Moreover, the combination of the new classifiers and bounds perform similarly to an SVM plus ten fold cross validation in terms of accuracy.

Figure 4 tries to illustrate qualitatively the discrepancies among the test error rate observed in crossvalidated SVM and that observed in the prior SVM. The figure shows the observed test error and the value of bounds on Q_D as functions of C for data sets `wav` and `pim`. The vertical pink line shows the crossvalidated C. The value of σ was fixed in both cases to the square root of the input data. In both cases, it is very noticeable the dramatic increase in the value of the bound as C increases, compared with a slight increase in the observed test error. A broadly accepted intuition

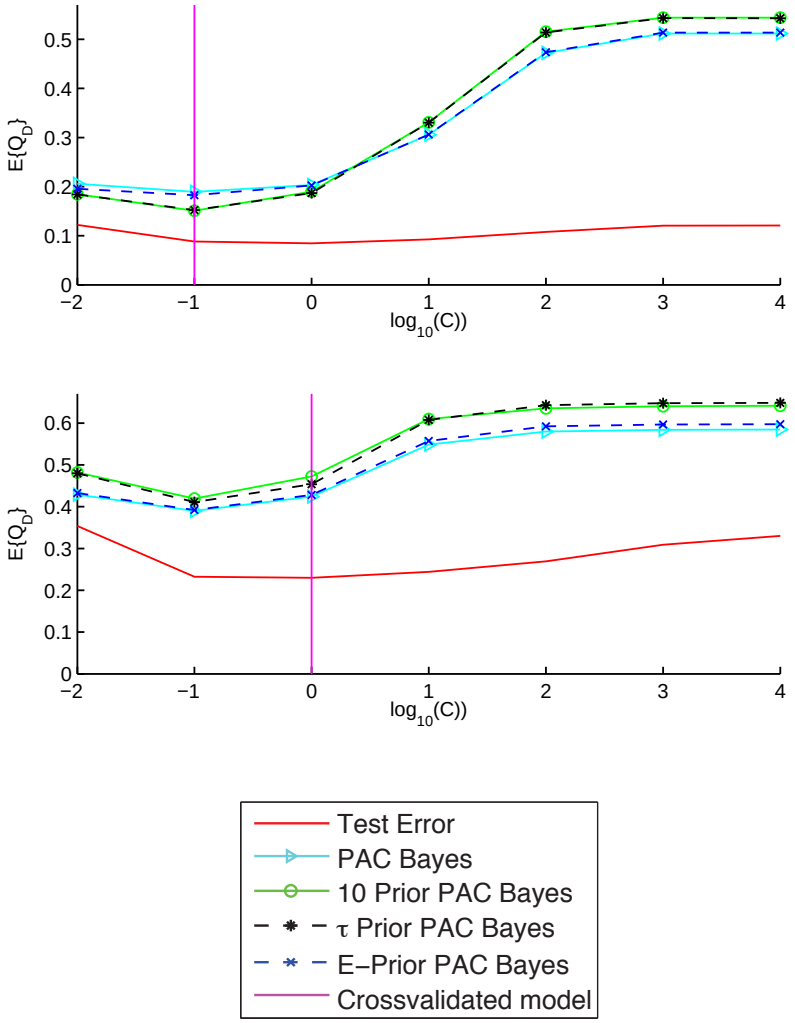

Figure 4: Values of bounds and observed test error rate as a function of C for data sets `wav` (top plot) and `pim` (bottom).

says that high values of C are likely to result in overfit, since the SVM is keener in reducing the training set error. However, our experiments seem to show that the bounds are overreacting to that behavior.

6. Conclusions

In this paper we have presented some strategies to tighten the already tight PAC-Bayes bound for binary classifiers by learning an informative prior distribution on classifiers. We have studied the SVM case, considering multivariate Gaussian priors and using some training data to infer their

mean and/or covariance matrix. The first strategy, named prior PAC Bayes bound, considers an identity covariance matrix. Then, an SVM learn on a separated subset of training samples serves as a direction along which to place the mean of the prior. This prior can be further refined in the τ-prior PAC Bayes bound case, where this direction is also used to stretch the covariance matrix. The second strategy, named expectation prior PAC-Bayes bound also considers identity covariances, but expresses the direction to place the prior as an statistic of the training data distribution and uses all the training samples to estimate such statistic. The expectation prior can also be refined stretching the covariance along the direction of the mean, yielding the τ-expectation prior PAC-Bayes bound.

The experimental work shows that these prior PAC-Bayes bounds achieve estimations of the expected true error of SVMs significantly tighter than those obtained with the original PAC-Bayes bound. It is remarkable that the prior PAC Bayes bounds improve the tightness of the PAC-Bayes bound even when the size of the training set experiences reductions of up to an 80% of its size.

The structure of the prior PAC-Bayes bound: learn a prior classifier using some data and then consider the SVM to be a posterior classifier inspired the design of new algorithms to train SVM-like classifiers. The prior SVM proposes a set of prior parts (fixed scalings along a prior direction learnt with separate data) and then fits a posterior part to each prior. The overall prior SVM classifier is the prior-posterior couple that yields a lower value of the bound. The η-prior SVM learns the scaling of the prior part and the posterior in the same quadratic program, thus significantly reducing the computational burden of the training. The analysis of these classifiers under the prior PAC-Bayes framework shows that the achieved bounds are dramatically tighter than those obtained for the original SVM under the same framework. Moreover, if the bound drives the selection of the hyperparameters of the classifiers, the observed empirical test error rate is similar to that observed in the SVM when the hyperparameters are tuned via ten fold cross validation.

Moreover, the prior SVM enables the use of different values of the regularisation constant C for both prior and posterior parts, which further tightens the bounds. The prior SVM classifiers with hyperparameters selected by minimising the τ-prior PAC Bayes bound achieve classification accuracies comparable to those obtained by an SVM with its parameters fixed by ten fold cross validation; with the great advantage that the theoretical bound on the expected true error provided by the τ-prior PAC Bayes bound is tightly close to the empirically observed.

All in all, the final message from this work is that the use informative priors can significantly improve the analysis and design of classifiers within the PAC-Bayes framework. We find the study of ways of extracting relevant prior domain knowledge from the available data and incorporating such knowledge in the form of the prior distribution to be a really promising line of research.

Acknowledgments

This work was partially supported by the IST Programme of the European Community under the PASCAL2 Network of Excellence IST-2007-216886. E. Parrado-Hernandez acknowledges support from Spain CICYT grant TIN2011-24533. Shiliang Sun is supported in part by the National Natural Science Foundation of China under Project 61075005, and the Fundamental Research Funds for the Central Universities. This publication only reflects the authors' views.

Appendix A.

The first step is to construct a Lagrangian functional to be optimized by the introduction of the constraints with multipliers α_i and ν_i, $i = 1,\ldots,m-r$,

$$L_P = \frac{1}{2}\|\mathbf{w} - \eta\mathbf{w}_r\|^2 + C\sum_{i=1}^{m-r}\xi_i - \sum_{i=1}^{m-r}\alpha_i\left(y_i\mathbf{w}^T\phi(\mathbf{x}_i) - 1 + \xi_i\right) - \sum_{i=1}^{m-r}\nu_i\xi_i\,, \quad \nu_i,\alpha_i \geq 0. \tag{19}$$

Taking the gradient of (19) with respect to $\mathbf{w}$ and derivatives with respect to ξ_i we obtain the optimality conditions:

$$\mathbf{w} - \eta\mathbf{w}_r = \sum_{j=1}^{m-r}\alpha_j y_j\phi(\mathbf{x}_j), \tag{20}$$

$$C - \alpha_i - \nu_i = 0 \Rightarrow 0 \leq \alpha_i \leq C \qquad i = 1,\ldots,m-r. \tag{21}$$

Plugging Equation (20) in functional (19) and applying the optimality condition (21) we arrive at the dual problem

$$\max_{\alpha_i}\frac{1}{2}\left\|\sum_{j=1}^{m-r}\alpha_j y_j\phi(\mathbf{x}_j)\right\|^2 - \sum_{i=1}^{m-r}\alpha_i\left(y_i\left(\eta\mathbf{w}_r^T + \sum_{j=1}^{m-r}\alpha_j y_j\phi^T(\mathbf{x}_j)\right)\phi(\mathbf{x}_i) - 1\right)$$

subject to

$$0 \leq \alpha_i \leq C \qquad i = 1,\ldots,m-r.$$

Now we can replace the prior $\mathbf{w}_r$ by its corresponding combination of mapped input vectors, $\mathbf{w}_r = \sum_{k=m-r+1}^{m} y_k\tilde{\alpha}_k\phi(\mathbf{x}_k)$ (with $\tilde{\alpha}_k$ being the scaled version of the Lagrange multipliers that yield a unit vector $\mathbf{w}_r$), and substitute kernel functions ($\kappa(\cdot,\cdot)$) for the inner products to arrive at

$$\max_{\alpha_i}\sum_{i=1}^{m-r}\alpha_i - \sum_{i=1}^{m-r}\eta\sum_{k=m-r+1}^{m}\alpha_i y_i\tilde{\alpha}_k y_k\kappa(\mathbf{x}_i,\mathbf{x}_k) - \frac{1}{2}\sum_{i,j=1}^{m-r}\alpha_i\alpha_j y_i y_j\kappa(\mathbf{x}_i,\mathbf{x}_j)$$

subject to

$$0 \leq \alpha_i \leq C \qquad i = 1,\ldots,m-r.$$

Grouping terms we have

$$\max_{\alpha_i}\sum_{i=1}^{m-r}\alpha_i\left(1 - y_i\eta\sum_{k=m-r+1}^{m}\tilde{\alpha}_k y_k\kappa(\mathbf{x}_i,\mathbf{x}_k)\right) - \frac{1}{2}\sum_{i,j=1}^{m-r}\alpha_i\alpha_j y_i y_j\kappa(\mathbf{x}_i,\mathbf{x}_j) \tag{22}$$

subject to

$$0 \leq \alpha_i \leq C \qquad i = 1,\ldots,m-r.$$

Now we can introduce the following matrix identifications to further compact Equation (22)

$$\mathbf{Y}_{(m-r),(m-r)} = \mathrm{diag}(\{y_i\}_{i=1}^{m-r}),$$

$$\mathbf{K}_{(m-r),(m-r)} = (\mathbf{K}_{(m-r),(m-r)})_{ij} = \kappa(\mathbf{x}_i,\mathbf{x}_j) \qquad i,j = 1,\ldots,m-r,$$

$$\mathbf{H}_{(m-r),(m-r)} = \mathbf{Y}_{(m-r),(m-r)}\mathbf{K}_{(m-r),(m-r)}\mathbf{Y}_{(m-r),(m-r)}, \tag{23}$$

$$\mathbf{v} = (\mathbf{v})_i = \left(1 - y_i \eta \sum_{k=m-r+1}^{m} \tilde{\alpha}_k y_k \kappa(\mathbf{x}_i, \mathbf{x}_k)\right) \qquad i = 1, \ldots, m-r, \tag{24}$$

$$\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{m-r}]^T. \tag{25}$$

Plugging (23), (24) and (25) in (22), we arrive at its final form that can be solved by off-the-shelf quadratic programming methods:

$$\max_{\alpha} \mathbf{v}^T \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{H}_{(m-r),(m-r)} \boldsymbol{\alpha} \tag{26}$$

with box constraints

$$0 \leq \alpha_i \leq C \qquad i = 1, \ldots, m-r.$$

Once (26) is solved, the overall prior SVM classifier $\mathbf{w}$ can be retrieved from (20):

$$\mathbf{w} = \sum_{i=1}^{m-r} \alpha_i y_i \phi(\mathbf{x}_i) + \eta \sum_{k=m-r+1}^{m} \tilde{\alpha}_k y_k \phi(\mathbf{x}_k). \tag{27}$$

References

A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 9–16. MIT Press, Cambridge, MA, 2007.

P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

C. L. Blake and C. J. Merz. *UCI Repository of Machine Learning Databases*. Department of Information and Computer Sciences, University of California, Irvine, http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.

B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Conference on Computational Learning Theory*, COLT '92, pages 144–152, 1992.

O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.

P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 353–360, 2009.

J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6(Mar):273–306, 2005.

J. Langford and J. Shawe-Taylor. PAC-Bayes and margins. In *Advances in Neural Information Processing Systems*, volume 14, Cambridge MA, 2002. MIT Press.

D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*, COLT '99, pages 164–170, 1999.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge (MA), 2002.

M. Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

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

J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44:1926–1940, 1998.

V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.