

Data Dependent Priors in PAC-Bayes Bounds

John Shawe-Taylor¹, Emilio Parrado-Hernández², and Amiran Ambroladze

¹ Dept. of Computer Science & CSML, University College London
London, WC1E 6BT, UK, *jst@cs.ucl.ac.uk*

² Dept. of Signal Processing and Communications, University Carlos III of
Madrid
Leganés, 28911, Spain, *emipar@tsc.uc3m.es*

Abstract. One of the central aims of Statistical Learning Theory is the bounding of the test set performance of classifiers trained with i.i.d. data. For Support Vector Machines the tightest technique for assessing this so-called generalisation error is known as the PAC-Bayes theorem. The bound holds independently of the choice of prior, but better priors lead to sharper bounds. The priors leading to the tightest bounds to date are spherical Gaussian distributions whose means are determined from a separate subset of data. This paper gives another turn of the screw by introducing a further data dependence on the shape of the prior: the separate data set determines a direction along which the covariance matrix of the prior is stretched in order to sharpen the bound. In addition, we present a classification algorithm that aims at minimizing the bound as a design criterion and whose generalisation can be easily analysed in terms of the new bound.

The experimental work includes a set of classification tasks preceded by a bound-driven model selection. These experiments illustrate how the new bound acting on the new classifier can be much tighter than the original PAC-Bayes Bound applied to an SVM, and lead to more accurate classifiers.

Keywords: PAC Bayes Bound, Support Vector Machines, Generalization prediction, Model Selection

1 Introduction

Support vector machines (SVM) [3] 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 countered by maximising the margin of the classifier on the training examples. For this reason there has been considerable interest in bounds on the generalisation 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 generalisation, especially in what concerns worse-case performance. In this sense, the widely used cross-validation generalization measures say 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 distribution of the error of these sets of samples. 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 [6], while later bounds have considered Rademacher complexity. The tightest bounds for practical applications appear to be the PAC-Bayes bound [5] and in particular the one given in [1], with a data dependent prior.

Another issue affected by the ability to predict the generalisation 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 more reliable in most experiments, despite the fact that it is statistically poorly justified and relatively expensive.

The PAC-Bayes Bounds (overviewed in Section 2) use a Gaussian prior with zero mean and identity covariance matrix. The Prior PAC-Bayes Bound [1] tightens the prediction of the generalisation error of an SVM by using a separate subset of the training data to learn the mean of the Gaussian prior. The key to the new bound introduced in this work to come up with even more informative priors by using the separate data to also learn a stretching of the covariance matrix. Then section 4 presents a classification algorithm, named η -Prior SVM, that introduces a regularization term that tries to optimise a PAC-Bayes Bound.

The new bounds and algorithms are evaluated in some classification tasks after parameters selection in Section 5. The experiments illustrate the capabilities of the Prior PAC-Bayes Bound to (i) select an acceptable model (hyperparameter estimation) for an SVM and (ii) to provide tighter predictions of the generalisation of the resulting classifier.

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

2 PAC-Bayes bound for SVM

This section is devoted to a brief review of the PAC-Bayes Bound of [4] and the Prior PAC-Bayes Bound of [1]. Let us consider a distribution \mathcal{D} of patterns \mathbf{x} lying in a certain input space \mathcal{X} , with their corresponding output labels y , $y \in \{-1, 1\}$. In addition, let us also consider a distribution Q over the classifiers c . For every classifier c , one can define the *True error*, as the probability of misclassifying a pair pattern-label (\mathbf{x}, y) selected at random from \mathcal{D} , $c_{\mathcal{D}} \equiv \Pr_{(\mathbf{x}, y) \sim \mathcal{D}}(c(\mathbf{x}) \neq y)$. In addition, the *Empirical error* \hat{c}_S of a classifier c on a sample S of size m is defined as 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 a function equal to 1 if the argument is true and equal to 0 if the argument is false.

Now we can define two error measures on the distribution of classifiers: the 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 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 1. (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 is the Kullback-Leibler divergence, $KL(p||q) = q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p}$ and $KL(Q(c)||P(c)) = \mathbb{E}_{c \sim Q} \ln \frac{Q(c)}{P(c)}$.

The proof of the theorem can be found in [4].

This bound can be particularised for the case of linear classifiers in the following way. The m training patterns define a linear classifier that can be represented by the following equation¹:

$$c(\mathbf{x}) = \text{sign}(\mathbf{w}^T \phi(\mathbf{x})) \tag{1}$$

where $\phi(\mathbf{x})$ is a nonlinear projection to a certain feature space² where the linear classification actually takes place, and \mathbf{w} is a vector from that feature space that determines the separating hyperplane.

For any vector \mathbf{w} we can define a stochastic classifier in the following way: we choose the distribution $Q = Q(\mathbf{w}, \mu)$ to be a spherical Gaussian with identity covariance matrix centred on the direction given by \mathbf{w} at a distance μ from the origin. Moreover, we can choose the prior $P(c)$ to be a spherical Gaussian with identity covariance matrix centred on the origin. Then, for classifiers of the form in equation (1) performance can be bounded by

Corollary 1. (PAC-Bayes Bound for margin classifiers [4]) *For all distributions \mathcal{D} , for all $\delta \in (0, 1]$, we have*

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

¹ We are considering here unbiased classifiers, i.e., with $b = 0$.

² This projection is induced by a kernel $\kappa(\cdot)$ satisfying $\kappa(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$

It can be shown (see [4]) that

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

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

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

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

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

Note that the SVM is a thresholded linear classifier expressed as (1) computed by means of the kernel trick [3]. The generalisation error of such a classifier can be bounded by at most twice the true (stochastic) error $Q_{\mathcal{D}}(\mathbf{w}, \mu)$ in Corollary 1, (see [5]);

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

for all μ .

These bounds were further refined in [1] by the introduction of data dependent priors.

Theorem 2. (Multiple Prior PAC-Bayes Bound) *Let $\{P_j(c)\}_{j=1}^J$ be a set of possible priors that can be selected with positive weights $\{\pi_j\}_{j=1}^J$ so that $\sum_{j=1}^J \pi_j = 1$. Then, for all $\delta \in (0, 1]$,*

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

In the standard application of the bound the prior is chosen to be a spherical Gaussian centred at the origin. We now consider learning a different prior based on training an SVM on a subset R 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 these to be the last r examples $\{\mathbf{x}_k, y_k\}_{k=m-r+1}^m$ in the description below. With these left-out r examples we can determine an SVM classifier, \mathbf{w}_r and form a set of potential priors $P_j(\mathbf{w} | \mathbf{w}_r)$ by centering spherical Gaussian distributions along \mathbf{w}_r , at distances $\{\eta_j\}$ from the origin, where $\{\eta_j\}_{j=1}^J$ are positive real numbers. In such a case, we obtain

Corollary 2. (Multiple Prior based PAC-Bayes Bound for margin classifiers) *Let us consider a set $\{P_j(\mathbf{w}|\mathbf{w}_r, \eta_j)\}_{j=1}^J$ of prior distributions of classifiers consisting in spherical Gaussian distributions with identity covariance matrix centered on $\eta_j \tilde{\mathbf{w}}_r$, where $\{\eta_j\}_{j=1}^J$ are real numbers. Then, for all distributions \mathcal{D} , for all $\delta \in (0, 1]$, we have*

$$Pr_{S \sim \mathcal{D}^m} \left(\begin{array}{l} \forall \mathbf{w}, \mu : KL(\hat{Q}_{S \setminus R}(\mathbf{w}, \mu) \| Q_{\mathcal{D}}(\mathbf{w}, \mu)) \leq \\ \min_j \frac{\frac{\|\eta_j \tilde{\mathbf{w}}_r - \mu \tilde{\mathbf{w}}\|^2}{2} + \ln(\frac{m-r+1}{\delta}) + \ln J}{m-r} \end{array} \right) \geq 1 - \delta \quad (5)$$

3 Stretched Prior PAC-Bayes Bound

The first contribution of this paper consists in a new data dependent PAC-Bayes Bound where not only the mean, but the covariance matrix of the Gaussian prior can be shaped by the data distribution. 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 $\tilde{\mathbf{w}}_r$.

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

Theorem 3. (τ -Prior PAC-Bayes Bound for linear classifiers) *Let us consider a prior $P(\mathbf{w}|\mathbf{w}_r, \tau, \eta)$ distribution of classifiers consisting of a Gaussian distribution centred on $\eta \tilde{\mathbf{w}}_r$, with identity covariance matrix in all directions except $\tilde{\mathbf{w}}_r$ in which the variance is τ^2 . Then, for all distributions \mathcal{D} , for all $\delta \in (0, 1]$, for all posteriors (\mathbf{w}, μ) we have that with probability greater or equal than $1 - \delta$*

$$KL(\hat{Q}_{S \setminus R}(\mathbf{w}, \mu) \| Q_{\mathcal{D}}(\mathbf{w}, \mu)) \leq \frac{0.5(\ln(\tau^2) + \tau^{-2} - 1 + P_{\tilde{\mathbf{w}}_r}^{\parallel}(\mu \mathbf{w} - \eta \tilde{\mathbf{w}}_r)^2 / \tau^2 + P_{\tilde{\mathbf{w}}_r}^{\perp}(\mu \mathbf{w})^2) + \ln(\frac{m-r+1}{\delta})}{m-r} \quad (6)$$

Proof. The application of the PAC-Bayes theorem follows that of [4] except that we must recompute the KL divergence. Note that the quantity

$$\hat{Q}_{S \setminus R}(\mathbf{w}, \mu) = \mathbb{E}_{m-r}[\tilde{F}(\mu \gamma(\mathbf{x}, y))] \quad (7)$$

remains unchanged as the posterior distribution is still a spherical Gaussian centred at \mathbf{w} . Using the expression for the KL divergence between two Gaussians

$$KL(\mathcal{N}(\mu_0, \Sigma_0) \| \mathcal{N}(\mu_1, \Sigma_1)) = \frac{1}{2} \left(\ln \left(\frac{\det \Sigma_1}{\det \Sigma_0} \right) + \text{tr}(\Sigma_1^{-1} \Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1} (\mu_1 - \mu_0) - N \right), \quad (8)$$

we obtain

$$\text{KL}(Q(\mathbf{w}, \mu) \| P(\mathbf{w} | \mathbf{w}_r, \tau, \eta)) = \frac{1}{2} \left(\ln(\tau^2) + \left(\frac{1}{\tau^2} - 1 \right) + \frac{P_{\mathbf{w}_r}^{\parallel} (\mu \mathbf{w} - \eta \tilde{\mathbf{w}}_r)^2}{\tau^2} + P_{\mathbf{w}_r}^{\perp} (\mu \mathbf{w})^2 \right)$$

and the result follows.

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 with the Prior PAC-Bayes Bound required a range of scalings of $\tilde{\mathbf{w}}_r$ from 1 to 100. For this we can choose $\eta = 50$ and $\tau = 50$, giving an increase in the bound over the factor $P_{\mathbf{w}_r}^{\perp} (\mu \mathbf{w})^2$ directly optimised in the algorithm of

$$\frac{0.5(\ln(\tau^2) + \tau^{-2} - 1) + P_{\mathbf{w}_r}^{\parallel} (\mu \mathbf{w} - \eta \tilde{\mathbf{w}}_r)^2 / \tau^2}{m - r} \leq \frac{\ln(\tau) + 0.5\tau^{-2}}{m - r} \approx \frac{3.912}{m - r}. \quad (9)$$

4 η -Prior SVM

The good performance of the Prior PAC-Bayes bound as a means to select good hyperparameters for SVMs reported in [1] motivates the exploration of new classifiers that incorporate the optimisation of the bound as a design criterion.

The Prior PAC-Bayes Bound defined the prior distribution as mixture of Gaussians along the direction given by \mathbf{w}_r and searched for the component in the mixture that yielded the tightest bound. The τ -Prior PAC-Bayes Bound presented above replaces the covering of the prior direction with a mixture of Gaussians by a stretching of the prior along \mathbf{w}_r . This motivates the introduction of the following classifier, termed η -Prior SVM. The η -Prior SVM is a combination of a prior classifier, \mathbf{w}_r and a posterior one, \mathbf{v} : $\mathbf{w} = \mathbf{v} + \eta \tilde{\mathbf{w}}_r$. The prior \mathbf{w}_r is determined as in the Prior PAC-Bayes framework, running an SVM on a subset of r training patterns. The posterior part \mathbf{v} and the scaling of the prior η come out of the following optimisation problem:

$$\min_{\mathbf{v}, \eta, \xi_i} \left[\frac{1}{2} \|\mathbf{v}\|^2 + C \sum_{i=1}^{m-r} \xi_i \right] \quad (10)$$

subject to

$$y_i (\mathbf{v} + \eta \tilde{\mathbf{w}}_r)^T \phi(\mathbf{x}_i) \geq 1 - \xi_i \quad i = 1, \dots, m - r \quad (11)$$

$$\xi_i \geq 0 \quad i = 1, \dots, m - r \quad (12)$$

Since the tightness of the bound depend on the KL divergence between the prior and posterior distribution, the proposed minimisation of the norm of \mathbf{v}

brings the posterior classifier \mathbf{w} close to the prior \mathbf{w}_r . Besides, the constraints push towards a reduced stochastic training error on the samples used to learn the posterior. Therefore, these two factors pursue an optimisation of the PAC-Bayes bound.

The statistical analysis of the η -Prior SVM can be performed in two ways. On the one hand, 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. On the other hand, the analysis can be performed based on the τ -Prior PAC-Bayes Bound.

5 Experimental Work

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 a classification preceded by model selection task using some UCI [2] datasets: `handwritten digits`, `waveform`, `pima`, `ringnorm` and `spam filtering`.

For every dataset, we prepare 10 different training/test set partitions where 80% of the samples form the training set and the remaining 20% form the test set. With each of the partitions we learn 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 maximisation of the margin and the minimisation 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 5×5 grid of parameters where $C \in \{1, 10, 100, 1000, 10000\}$ and $\sigma \in \{\frac{1}{4}\sqrt{d}, \frac{1}{2}\sqrt{d}, \sqrt{d}, 2\sqrt{d}, 4\sqrt{d}\}$, where d is the input space dimension. With respect to the parameters needed by the Prior PAC-Bayes bounds, the experiments reported in [1] suggest that $J = 10$ priors and leaving half of the training set to learn the prior direction lead to reasonable results.

The results presented in the sequel correspond to the combination of a model selection method plus a classification algorithm. Model selection methods refer to the fitness function (usually a bound) used in the grid search for the optimal pair of hyperparameters. The studied combinations are:

- Using the regular SVM as classifier:
 - PAC-SVM** Regular SVM with model selection driven by the PAC-Bayes bound of [4].
 - Prior-PAC** Regular SVM and Multiple Prior PAC-Bayes Bound.
 - 2FCV** Regular SVM with the model selection made through two fold cross-validation. This setting involves a computational burden similar to the bound based ones, where half of the training data are used to learn the prior and the other half to learn the posterior).
- Using the η -Prior SVM as classifier we have the following two configurations:

Prior-PAC- η -PSVM η -Prior SVM and Multiple Prior PAC-Bayes Bound considering η comes from a multiple prior setting of $J = 50$ priors $\eta_j \mathbf{w}_r / \|\mathbf{w}_r\|$ with the η_j equally spaced between $\eta_1 = 1$ and $\eta_{50} = 100$. This setting minimises the penalty term in the Prior PAC-Bayes Bound as we are not actually using these priors to learn the posterior.

τ -PriorPAC η -PSVM η -Prior SVM and the new τ -Prior PAC-Bayes Bound.

The displayed values of training set bounds (PAC-Bayes and (Multiple) Prior PAC-Bayes) are obtained according to the following setup. For each one of the 10 partitions we train an instance of the corresponding classifier for each position of the grid of hyperparameters and compute the bound. For that partition we select the classifier with the minimum value of the bound found through the whole grid and compute its test error rate. Then we display the sample average and standard deviation of the 10 values of the bound and of the test error. Note that proceeding this way we select a (possibly) different pair of parameters for every partition. That is the reason why we name this task as model selection plus classification.

Moreover, the reported values of the PAC-Bayes and the Multiple 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 assuming a Gaussian distribution of variance equal to one.

Table 1 displays values of the bounds for the studied configurations of bound plus classifier. Notice that most of the configurations involving the new bounds achieve a significant cut in the value of the PAC-Bayes Bound, which indicates that learning the prior distribution helps to improve the PAC-Bayes bound. The tightest values of the bound correspond to the η -Prior SVM, it was expected since this algorithm aims at optimising the bound as well as at reducing the classification error. However, for the explored datasets and range of hyperparameters, the tightness of the bound presented in this paper acting on the η -Prior SVM is very close to the Multiple Prior PAC Bayes Bound one.

Table 1 also displays the test error rates averaged over the 10 partitions plus the sample standard deviation. The results illustrate that although cross-validation seems to be the safest model selection option, the PAC-Bayes bounds are catching up. It is remarkable how the η -Prior SVM achieves a fairly good trade-off between good model selection (reduced test error) and tightness of the generalisation error prediction (low bound).

6 Concluding remarks

We have presented a new bound (the τ Prior PAC-Bayes Bound) on the performance of SVMs based on the estimation of a Gaussian prior with an

		Classifier				
		SVM			η Prior SVM	
Problem		2FCV	PAC	PrPAC	PrPAC	τ -PrPAC
digits	Bound	–	0.175	0.107	0.050	0.047
		–	± 0.001	± 0.004	± 0.006	± 0.006
	CE	0.007	0.007	0.014	0.010	0.009
		± 0.003	± 0.002	± 0.003	± 0.005	± 0.004
waveform	Bound	–	0.203	0.185	0.178	0.176
		–	± 0.001	± 0.005	± 0.005	± 0.005
	CE	0.090	0.084	0.088	0.087	0.086
		± 0.008	± 0.007	± 0.007	± 0.006	± 0.006
pima	Bound	–	0.424	0.420	0.428	0.416
		–	± 0.003	± 0.015	± 0.018	± 0.020
	CE	0.244	0.229	0.229	0.233	0.233
		± 0.025	± 0.027	± 0.026	± 0.027	± 0.028
ringnorm	Bound	–	0.203	0.110	0.053	0.050
		–	± 0.000	± 0.004	± 0.004	± 0.004
	CE	0.016	0.018	0.018	0.016	0.016
		± 0.003	± 0.003	± 0.003	± 0.003	± 0.003
spam	Bound	–	0.254	0.198	0.186	0.178
		–	± 0.001	± 0.006	± 0.008	± 0.008
	CE	0.066	0.067	0.077	0.070	0.072
		± 0.006	± 0.007	± 0.011	± 0.009	± 0.010

Table 1. Values of the bounds and Test Classification Error Rates (CE) for various settings.

stretched covariance matrix of the distribution of classifiers given a particular dataset, and the use of this prior in the PAC-Bayes generalisation bound.

The new bound has motivated the development of a classification algorithm (η -Prior SVM), that automatically determines the position of the mean of the prior as part of the optimisation. Empirical results show that the statistical analysis of this new algorithms yields tighter values of the Multiple Prior PAC-Bayes Bound, even when compared to this bound applied to a regular SVM. Moreover, if we use the bounds to guide the model selection (we select the values of C and σ that yield a minimum value of the bound), the new algorithms combined with the bound arrive at better models in terms of classification error that the original SVM and PAC-Bayes Bound. In fact, the performance of the Multiple Prior PAC-Bayes Bound model selection plus η -Prior SVM is comparable to that of a regular SVM using cross-validation for model selection.

The work presented in this paper is being continued in two main directions. On the one hand, we are studying new structures for the prior along the lines of the bound of equation (6), based on multivariate Gaussian

distributions with more sophisticated covariance matrices that help tighten the bounds. On the other hand, we are envisaging the implementation of a meta-classifier consisting in a series of prior and posterior classifiers within a dynamic programming framework. This second line of research aims at the application of these ideas in incremental learning scenarios.

Acknowledgments

This work was partially supported by the IST Programme of the European Community under the PASCAL2 Network of Excellence IST-2007-216886. E. P-H. acknowledges support from Spain CICYT grant TEC2008-02473/TEC. This publication only reflects the authors' views.

References

1. Amiran Ambroladze, Emilio Parrado-Hernandez, and John 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.
2. C L Blake and C J Merz. *UCI Repository of machine learning databases*. University of California, Irvine, Dept. of Information and Computer Sciences, [<http://www.ics.uci.edu/~mlearn/MLRepository.html>], 1998.
3. Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learning Theory*, pages 144–152, 1992.
4. J Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6(Mar):273–306, 2005.
5. J Langford and J Shawe-Taylor. PAC-Bayes & Margins. In *Advances in Neural Information Processing Systems*, volume 14, Cambridge MA, 2002. MIT Press.
6. J Shawe-Taylor, P L Bartlett, R C Williamson, and M Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Information Theory*, 44(5):1926 – 1940, 1998.