Bias–Variance Analysis and Ensembles of SVM

Giorgio Valentini[1,2] and Thomas G. Dietterich[3]

[1] DISI - Dipartimento di Informatica e Scienze dell'Informazione,
Università di Genova, 16146 Genova, Italy,
[2] INFM, Istituto Nazionale per la Fisica della Materia, 16146 Genova, Italy
valenti@disi.unige.it
[3] Department of Computer Science,
Oregon State University, Corvallis, OR 97331, USA
tgd@cs.orst.edu

Abstract. Accuracy, diversity, and learning characteristics of base learners critically influence the effectiveness of ensemble methods. Bias–variance decomposition of the error can be used as a tool to gain insights into the behavior of learning algorithms, in order to properly design ensemble methods well-tuned to the properties of a specific base learner. In this work we analyse bias–variance decomposition of the error in Support Vector Machines (SVM), characterizing it with respect to the kernel and its parameters. We show that the bias–variance decomposition offers a rationale to develop ensemble methods using SVMs as base learners, and we outline two directions for developing SVM ensembles, exploiting the SVM bias characteristics and the bias-variance dependence on the kernel parameters.

1 Introduction

The effectiveness of ensemble methods depends on the geometry and topology of the data [8], on the relationship between diversity and accuracy of the base learners [3, 12], on their stability [1], and on their general geometrical properties [2]. The analysis of the features and properties of the base learners used in ensemble methods is crucial in order to design ensemble methods well-tuned to the characteristics of a specific base learner.

SVM are "strong" dichotomic classifiers, well-founded on the Vapnik's Statistical Learning Theory [16], and our aim is to study if and how we can exploit their specific features in the context of ensemble methods. We analysed the learning properties of SVMs using the bias–variance decomposition of the error as a tool to understand the relationships between kernels, kernel parameters, and learning processes in SVM. Historically, the bias–variance insight was borrowed from the field of regression, using squared–loss as the loss function [7]. For classification problems, where the 0/1 loss is the main criterion, several authors proposed bias–variance decompositions related to 0/1 loss [11, 1, 10, 14, 6]. All these decompositions suffer of significant shortcomings: in particular they lose the relationship to the original squared loss decomposition, forcing in most cases bias and variance to be purely additive.

We consider classification problems and the 0/1 loss function in the Domingos' unified framework of bias–variance decomposition of the error [4, 5]. In this approach bias and variance are defined for an arbitrary loss function, showing that the resulting decomposition specializes to the standard one for squared loss, but it holds also for the 0/1 loss [4]. Using this theoretical framework, we tried to answer two main questions:

- Can we characterize bias and variance in SVMs with respect to the kernel and its parameters?
- Can the bias–variance decomposition offer guidance for developing ensemble methods using SVMs as base learners?

In order to answer these two questions, we planned and performed an extensive series of experiments on synthetic and real data sets to evaluate bias variance–decomposition of the error with different kernels and different kernel parameters.

The paper is organized as follows. In Sect. 2, we summarize the main results of Domingos' unified bias–variance decomposition of error. Sect. 3 outlines the main characteristics of the data sets employed in our experiments and the main experimental tasks performed. Then we present the principal results of our experimentation about bias–variance decomposition of the error in SVM. Sect. 4 exploits the knowledge achieved by the bias–variance decomposition of the error to formulate hypotheses about the effectiveness of SVMs as base learners in ensembles of learning machines, and two directions for developing new ensemble models of SVM are proposed. An outline of future developments of this work concludes the paper.

2 Bias–Variance Decomposition of the Error

Consider a (potentially infinite) population U of labeled training data points, where each point is a pair (x_j, t_j), $t_j \in \{-1, 1\}$, $x_j \in \mathbb{R}^d$, $d \in \mathbb{N}$. Let $P(x,t)$ be the joint distribution of the data points in U. Let D be a set of m points drawn identically and independently from U according to P. We think of D as being the training sample that we are given for training a classifier. We can view D as a random variable, and we will let $E_D[\cdot]$ indicate the expected value with respect to the distribution D.

Let $\mathcal{L}$ be a learning algorithm, and define $f_D = \mathcal{L}(D)$ as the classifier produced by $\mathcal{L}$ applied to training set D. The model produces a prediction $f_D(x) = y$. Let $L(t,y)$ be the 0/1 loss function, that is $L(t,y) = 0$ if $y = t$, and $L(t,y) = 1$ otherwise.

Suppose we consider a fixed point $x \in \mathbb{R}^d$. This point may appear in many labeled training points in the population. We can view the corresponding labels as being distributed according to the conditional distribution $P(t|x)$. (Recall that it is always possible to factor the joint distribution as $P(x,t) = P(x)P(t|x)$.) Let $E_t[\cdot]$ indicate the expectation with respect to t drawn according to $P(t|x)$.

Suppose we consider a *fixed* predicted class y for a given x. This prediction will have an expected loss of $E_t[L(t,y)]$. In general, however, the prediction y is not fixed. Instead, it is computed from a model f_D which is in turn computed from a training sample D. Hence, the expected loss EL of learning algorithm $\mathcal{L}$ at point x can be written by considering both the randomness due to the choice of the training set D and the randomness in t due to the choice of a particular test point (x, t):

$$EL(\mathcal{L}, x) = E_D[E_t[L(t, f_D(x))]],$$

where $f_D = \mathcal{L}(D)$ is the classifier learned by $\mathcal{L}$ on training data D. The purpose of the bias-variance analysis is to decompose this expected loss into terms that separate the bias and the variance.

To derive this decomposition, we must define two things: the optimal prediction and the main prediction. The *optimal prediction* y_* for point x minimizes $E_t[L(t, y)]: \ y_*(x) = \arg\min_y E_t[L(t, y)]$. For 0/1 loss, the optimal prediction is equal to the label t that is observed more often in the universe U of data points. The *optimal model* $\hat{f}(x) = y_*, \ \forall x$ makes the optimal prediction at each point x. Define the noise, $N(x) = E_t[L(t, y_*)]$ to be the remaining loss that cannot be eliminated, even by the optimal prediction.

The *main prediction* y_m at point x is defined as $y_m = \arg\min_{y'} E_D[L(f_D(x), y')]$. This is a value that would give the lowest expected loss if it were the "true label" of x. In other words, it is the label for x that the learning algorithm "wishes" were correct. Foor 0/1 loss, the main prediction is the class predicted most often by the learning algorithm $\mathcal{L}$ when applied to training sets D.

Given these definitions, the *bias* $B(x)$ (of learning algorithm $\mathcal{L}$ on training sets of size m) is the loss of the main prediction relative to the optimal prediction: $B(x) = L(y_*, y_m)$. For 0/1 loss, the bias is always 0 or 1. We will say that $\mathcal{L}$ is *biased at point* x, if $B(x) = 1$.

The *variance* $V(x)$ is the average loss of the predictions relative to the main prediction: $V(x) = E_D[L(y_m, f_D(x))]$. It captures the extent to which the various predictions $f_D(x)$ vary depending on D.

Domingos shows that

$$EL(\mathcal{L}, x) = c_1 N(x) + B(x) + c_2 V(x),$$

where c_1 is $2P_D(f_D(x) = y_*) - 1$ and c_2 is $+1$ if $B(x) = 0$ and -1 if $B(x) = 1$. This can be simplified if we consider the noise free case and if we define the unbiased variance, $V_u(x)$, be the variance when $B(x) = 0$ and the biased variance, $V_b(x)$, to be the variance when $B(x) = 1$. Under these conditions,

$$EL(\mathcal{L}, x) = B(x) + V_u(x) - V_b(x).$$

If we further define $V_n(x) = V_u(x) - V_b(x)$ to be the *net variance*, then we can write

$$EL(\mathcal{L}, x) = B(x) + V_n(x).$$

To gain an understanding of this decomposition, let p_{corr} be the probability that the main prediction is correct: $p_{corr}(x) = P(y_m = t|x)$. If $p_{corr}(x) > 0.5$, then the main prediction is equal to the optimal prediction, and the (unbiased) variance is $V_u(x) = 1 - p_{corr}(x)$. If $p_{corr}(x) \leq 0.5$, then the main prediction is wrong. However, with probability $p_{corr}(x)$, the actual prediction is right (and unequal to the main prediction). The (biased) variance is therefore $V_b(x) = p_{corr}(x)$.

An interesting aspect of Domingos' decomposition is that variance hurts on unbiased points x, but it helps on biased points. Nonetheless, to obtain low overall expected loss, we want the bias to be small, and hence, we see to reduce both the bias and the unbiased variance.

This decomposition for a single point x can be generalized to the entire population by defining $E_x[\cdot]$ to be the expectation with respect to $P(x)$. Then define the *average bias* $E_x[B(x)]$, the *average unbiased variance* $E_x[V_u(x)]$, and the *average biased variance* $E_x[V_b(x)]$. In the noise-free case, the expected loss over the entire population is

$$E_x[EL(\mathcal{L}, x)] = E_x[B(x)] + E_x[V_u(x)] - E_x[V_b(x)].$$

3 Bias–Variance Analysis in SVM

The bias–variance decomposition of error represents a powerful tool to analyze learning processes in learning machines. Our aim consists in characterizing bias–variance decomposition of the error with respect to SVM kernels and their parameters.

3.1 Experimental Tasks

In the experiments, we employed 7 different two-class data sets, both synthetic and "real". Most of them are from the UCI repository [13]. We selected two synthetic data sets (*P2* and *Waveform*) and 5 "real" data sets (*Grey-Landsat*, *Letter*, reduced to the two-class problem of discriminating between the letters B and R, *Letter* with added 20% noise, *Spam*, and *Musk*). For synthetic data sets, we generated small training sets of 100 examples and reasonably large test sets (10000 examples) to perform a reliable evaluation of bias and variance. We generated 400 different training sets for *P2* and 200 training sets for *Waveform*. For real data sets, we first divided the data into training $\mathcal{T}$ and test $\mathcal{S}$ sets. We then drew 200 data sets from $\mathcal{T}$, each consisting of 100 examples drawn uniformly with replacement. The data were normalized in such a way that for each attribute the mean was 0 and the standard deviation 1.

To evaluate the bias and variance of SVMs, we conducted experiments with different kernels and different kernel parameters. In particular we considered 3 different SVM kernels: *RBF* (Gaussian Radial Basis Functions), *polynomial*, and *dot–product*. In all cases we evaluated bias and variance while varying the parameters of the kernel and the C parameter that controls the trade–off between training error and the margin. In particular we analysed:

1. The relationships between average error, bias, net–variance, unbiased and biased variance and the parameter σ (*RBF-SVM*) or the *degree* (*polynomial SVM*) of the kernel.

2. The relationships between average error, bias, net–variance, unbiased and biased variance and the parameter C (the regularization factor) of the kernel.

Each SVM model (i.e., choice of kernel, setting of kernel parameters, and C) required the training of 200 different SVMs, one for each synthesized or bootstrapped data set, for a total of about one half million training runs. For each SVM model, we computed the main prediction, bias, net-variance, biased and unbiased variance on each example of the test set, and the corresponding average quantities on the overall test set.

In all our experiments we used the *NEURObjects* [15][4] and *SVM-light* [9] programs, and we developed a C++ program `analyze_BV`[5] to perform the bias–variance decomposition of the error.

3.2 Characterization of Bias–Variance Decomposition of the Error

Space limitations preclude full description of the results. Here we present only the main observations about the experiments with *RBF-SVM*. Fig. 1 depicts the average loss, bias net–variance, unbiased and biased variance varying the values of σ and the regularization parameter C in *RBF-SVM* on the *GreyLandsat* data set. We note that σ is the most important parameter: although for very low values of C the SVM cannot learn, independently of the values of σ, (Fig. 1 a), the error, the bias, and the net–variance depend mostly on the σ parameter. In particular for low values of σ, the bias is very high and net-variance is 0, as biased and unbiased variance are equal. Then the bias suddenly drops (Fig. 1b), which lowers the average loss (Fig. 1a), and then stabilizes for higher values σ.

Fig. 2 shows the bias–variance decomposition on different data sets, varying σ, and for a fixed value of C. The plots show that average loss, bias, and variance depend significantly on σ for all the considered data sets, confirming the existence of an "high biased region" for low values of σ. In this region, biased and unbiased variance are about equal (net–variance $V = V_u - V_b$ is low). Then unbiased variance increases while biased variance decreases (Fig. 2 a,b,c and d), and finally both stabilize for relatively high values of σ. Interestingly, the average loss and the bias do not increase for high values of σ, especially if C is high.

Error, bias, net–variance, unbiased and biased variance show a common trend in the 7 data sets we used in the experimentation. Some differences, of course, arise in the different data sets, but we can distinguish three different regions in the error analysis of *RBF-SVM*, with respect to increasing values of σ (Fig. 3):

1. **High bias region**. For low values of σ, error is high: it depends on high bias. Net–variance is about 0 as biased and unbiased variance are equivalent. In this region there are no remarkable fluctuations of bias and variance: both remain constant, with high values of bias and comparable values of unbiased and biased variance, leading to net–variance values near to 0. In some cases biased and unbiased variance are about equal, but different from 0, in other cases they are equal, but near to 0.

[4] Download web site: `http://www.disi.unige.it/person/ValentiniG/NEURObjects`.
[5] The source code is available at `http://ftp.disi.unige.it/person/ValentiniG/BV`.

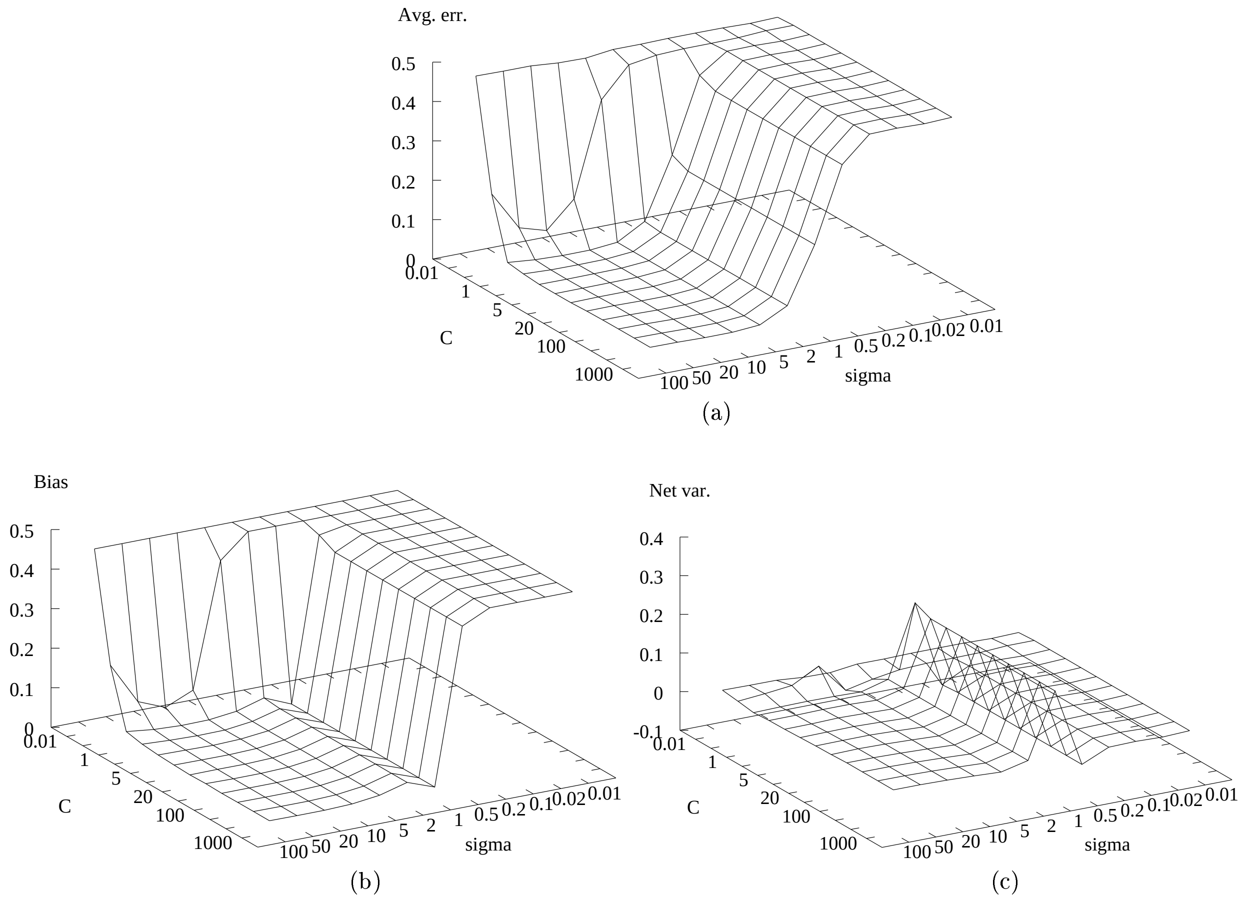

Fig. 1. Grey-Landsat data set. Error (a) and its decomposition in bias (b) and net variance (c) in RBF-SVM, varying both C and σ.

2. **Switching region**. Suddenly, for a critical value of σ, the bias decreases rapidly. This critical value depends also on C: for very low values of C, we have no learning, then for higher values the bias drops. Higher values of C cause the critical value of σ to decrease (Fig. 1 (b) and 2). In this region the increase in net–variance is less than the decrease in bias: so the average error decreases.
3. **Stabilized region**. This region is characterized by small or no variations in bias and net–variance. For high values of σ both bias and net–variance stabilize and the average error is constant.

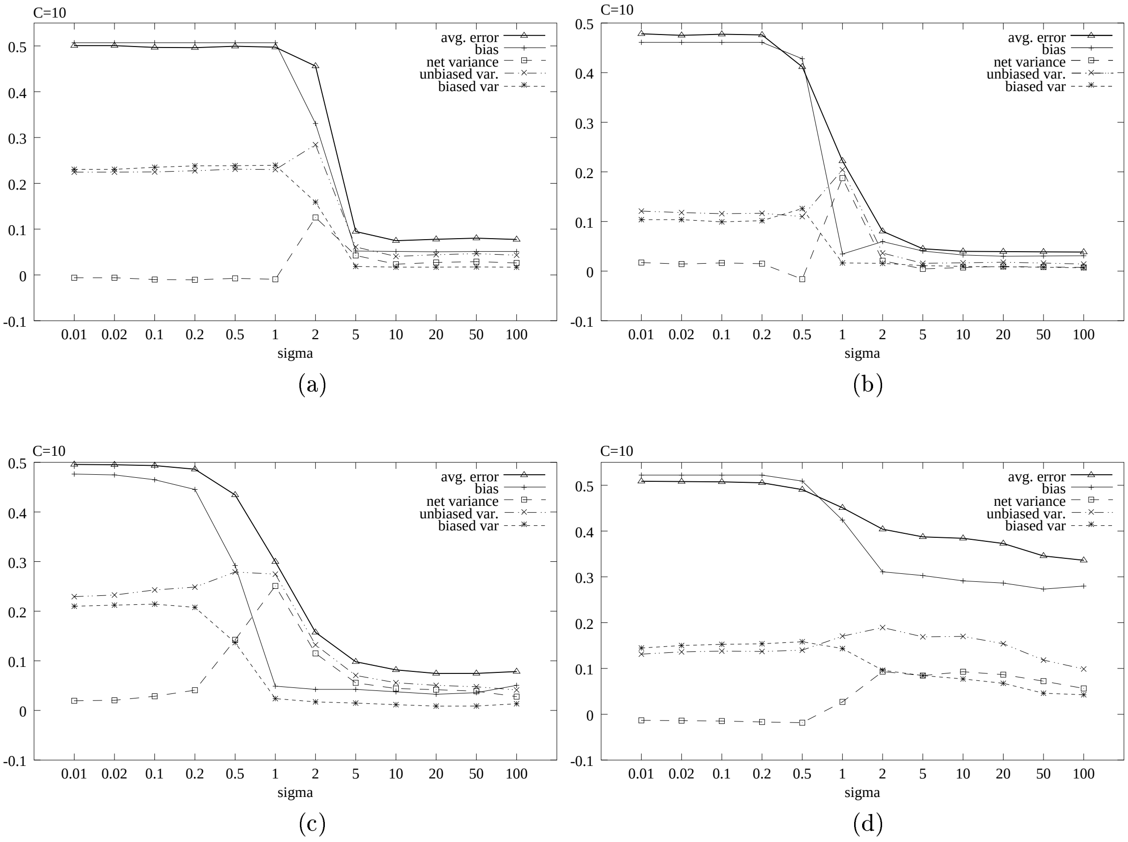

Fig. 2. Bias-variance decomposition of error in bias, net variance, unbiased and biased variance in SVM RBF, varying σ and for $C = 10$: (a) Waveform, (b) Grey-Landsat, (c) Letter-Two, (d) Letter-Two with added noise.

In the first region, bias rules SVM behavior: in most cases the bias is constant and close to 0.5, showing that we have a sort of random guessing, without effective learning. It appears that the area of influence of each support vector is too small, and the learning machine overfits the data. This is confirmed by the fact that in this region the training error is about 0 and almost all the training points are support vectors.

In the switching region, the SVM starts to learn, adapting itself to the data characteristics. Bias rapidly goes down (at the expenses of a growing net–variance), but for higher values of σ (in the second part of the switching region), sometimes net–variance also goes down, working to lower the error.

In the third region for low values of C, the error increases with σ (data not shown), as a result of the bias increment; on the whole RBF-SVMs are sensitive

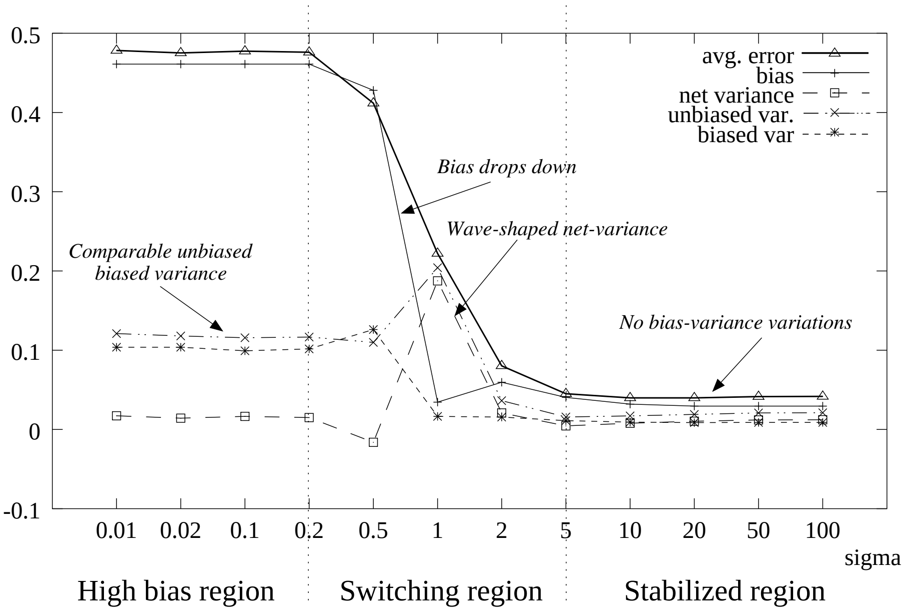

Fig. 3. The 3 regions of error in RBF-SVM with respect to σ. The data refer to Grey-Landsat data set ($C = 100$).

to low values of C: if C is too low, then bias can grow quickly. High values of C lower the bias.

For polynomial and dot–product SVMs, we have also characterized the behavior of the SVM in terms of average error, bias, net–variance, unbiased and biased variance. Here we note only that the average loss curve shows in general a U shape with respect to the polynomial degree, and this shape may depend only on unbiased variance or on both the bias and variance according to the characteristics of the data set.

4 Ensembles of SVMs based on Bias–variance Analysis

In addition to providing insight into the behavior of SVMs, the analysis of the bias–variance decomposition of the error can identify the situations in which ensemble methods might improve SVM performance. We present two possible ways of applying bias–variance analysis to develop SVM-based ensemble methods.

4.1 Bagged Ensemble of Selected Low-Biased SVM

We can schematically consider the following observations:

– We know that bagging lowers net–variance (in particular unbiased variance) but not bias [1].

- SVMs are strong, low-biased learners, but this property depends on the proper selection of the kernel and its parameters.
- If we can identify low-biased base learners with a relatively high unbiased variance, bagging can lower the error.
- Bias–variance analysis can identify SVMs with low bias.

Hence a basic high–level algorithm for a general *Bagged ensemble of selected low-bias SVMs* is the following: (a) select the SVM model with the lowest bias using bias–variance decomposition of the error; (b) perform bagging using as base learner the SVM with the estimated lowest bias. This approach combines the low bias properties of SVMs with the low unbiased variance properties of bagging and should produce ensembles with low overall error. Depending on the type of kernel and parameters considered, and on the way the bias is estimated for the different SVM models, different implementations can be given. Of course, we cannot expect high error reduction if the bias–variance analysis shows that the base learner has a high bias and a low unbiased variance.

4.2 Heterogeneous Ensembles of SVM

The analysis of bias–variance decomposition of error in SVM shows that the minimum of the overall error, bias, net–variance, unbiased and biased variance occur often in different SVM models. These different behaviors of different SVM models could be in principle exploited to produce diversity in ensembles of SVMs. Although the diversity of base learner itself does not assure the error of the ensemble will be reduced [12], the combination of accuracy and diversity in most cases does [3]. As a consequence, we could select different SVM models as base learners by evaluating their accuracy and diversity through the bias-variance decomposition of the error.

Our data show that the "optimal region" (low average loss region) is quite large in RBF-SVMs. This means that C and σ do not need to be tuned extremely carefully. From this point of view, we can avoid time-consuming model selection by combining RBF-SVMs trained with different σ values all chosen from within the "optimal region." For instance, if we know that the error curve looks like the one depicted in Fig. 3, we could try to fit a sigmoid-like curve using only few values to estimate where the stabilized region is located. Then we could train an heterogeneous ensemble of SVMs with different σ parameters (located in the low bias region) and average them according to their estimated accuracy. Bagging and boosting methods can also be combined with this approach to further improve diversity and accuracy of the base learners.

5 Conclusion and Future Works

We performed an analysis of the bias and variance of SVMs, considering gaussian, polynomial, and dot–product kernels. We discovered regular patterns in the behavior of the bias and variance, and we related those patterns to the parameters and kernel functions of the SVM.

The information supplied by bias-variance analysis suggests two promising approaches for designing ensembles of SVMs. One approach is to employ low-bias SVMs as base learners in a bagged ensemble. The other approach is to apply bias-variance analysis to construct a heterogeneous, diverse set of low-bias classifiers.

We plan to pursue both of these ensemble methods. We also plan to relate the bias and variance of SVMs to the geometrical and topological structure of the data [8].

Acknowledgments

This work has been partially funded by INFM and University of Genova.

References

[1] L. Breiman. Bias, variance and arcing classifiers. Technical Report TR 460, Statistics Department, University of California, Berkeley, CA, 1996.

[2] S. Cohen and N. Intrator. Automatic Model Selection in a Hybrid Pecep-tron/Radial Network. In In J. Kittler and F. Roli (eds.) *MCS 2001, Cambridge, UK*, pages 349–358, 2001.

[3] T.G. Dietterich. Ensemble methods in machine learning. In J. Kittler and F. Roli (eds.), *MCS 2000, Cagliari, Italy*, pages 1–15, 2000.

[4] P. Domingos. A Unified Bias-Variance Decomposition for Zero-One and Squared Loss. In *Proc. of the* 17^{th} *National Conference on Artificial Intelligence*, pages 564–569, Austin, TX, 2000.

[5] P. Domingos. A Unified Bias-Variance Decomposition and its Applications. In *Proc. of the* 17^{th} *ICML*, pages 231–238, Stanford, CA, 2000.

[6] J.H. Friedman. On bias, variance, 0/1 loss and the curse of dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77, 1997.

[7] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias-variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[8] T.K. Ho. Data Complexity Analysis for Classifiers Combination. In J. Kittler and F. Roli (eds.), *MCS 2001, Cambridge, UK*, pages 53–67, 2001.

[9] T. Joachims. Making large scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. Cambridge, MA, 1999.

[10] R. Kohavi and D.H. Wolpert. Bias plus variance decomposition for zero-one loss functions. In *Proc. of the* 13^{th} *ICML*, pages 275–283, Bari, Italy, 1996.

[11] E. Kong and T.G. Dietterich. Error - correcting output coding correct bias and variance. In *Proc. of the* 12^{th} *ICML*, pages 313–321, San Francisco, CA, 1995.

[12] L.I. Kuncheva, F. Roli, G.L. Marcialis, and C.A. Shipp. Complexity od Data Subsets Generated by the Random Subspace Method: An Experimental Investigation. In J. Kittler and F. Roli (eds.), *MCS 2001, Cambridge, UK*, pages 349–358, 2001.

[13] C.J. Merz and P.M. Murphy. UCI repository of machine learning databases, 1998. http://www.ics.uci.edu/mlearn/MLRepository.html.

[14] R. Tibshirani. Bias, variance and prediction error for classification rules. Technical report, University of Toronto, Canada, 1996.

[15] G. Valentini and F. Masulli. NEURObjects: an object-oriented library for neural network development. *Neurocomputing*. (in press).

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