

Statistical Criteria for Early Stopping of Support Vector Machines

Tatyana V. Bandos^{a,1}, Gustavo Camps-Valls^a, and
Emilio Soria-Olivas^a

^a*Grup de Processament Digital de Senyals, Universitat de València, Spain.*

Abstract

This paper proposes the use of statistical criteria for early stopping Support Vector Machines, both for regression and classification problems. The method basically stops the minimization of the primal functional when moments of the error signal (up to fourth order) become stationary, rather than according to a tolerance threshold of primal convergence itself. This simple strategy induces lower computational efforts and no significant differences are observed in terms of performance and sparsity.

Key words: Support Vector Machines; Skewness; Kurtosis; Early Stopping

1 Introduction

Support Vector Machines (SVMs) are efficient methods for non-linear classification and regression [1], and has been used in many key application areas during the last decade [2]. The solution of an SVM reduces to solving a quadratic programming (QP) problem. This is, however, very time consuming when high number of training samples are available. Two different approaches can be followed to allow fast convergence with small memory requirements on large-scale problems: (i) *chunking* or *decomposition* methods, and (ii) *iterative* methods. We consider standard iterative minimization procedures of the primal functional, e.g. the Sequential Minimal Optimization [3] or the Iterative Re-Weighted Least Squares [4] algorithms.

All these methods include a *tolerance* threshold on the primal error to stop the iterative procedure. This parameter is set between 10^{-3} and 10^{-6} default values. Since

¹ Address: Dept. Enginyeria Electrònica. Univesitat de València. C/ Dr. Moliner, 50. Burjassot (València). Spain. Tel.: +34 96 3160197; Fax: +34 96 3160466. E-mail: Tatyana.Bandos@uv.es.

typically, at the close of optimization, primal functional attenuates slowly with iterations, these low threshold values can make the algorithms spend too much time iterating, independently of the problem at hand. For this purpose, we propose to stop the iterative process when the distribution of errors ceases changing according to moments of the error, rather than to a fixed (*ad hoc*) threshold. Early-stopping has been extensively studied, and many improvements have been proposed such as taking into account not only the training error but also model's complexity, expected or estimated noise variance, and size of the sample [5]. Our method, however, only relies on the statistics over training errors distribution at this stage, as SVMs are intrinsically regularized and have revealed excellent robustness capabilities for various signal-to-noise ratios. This, in addition, will yield a simple but powerful method, as early-stopping will control the trade-off between the empirical risk and regularization of the model (complexity) through selecting the minimum necessary number of support vectors to attain a (slightly) unchanged training error distribution.

The rest of the paper is outlined as follows. Section 2 presents the proposed methodology to allow early stopping in the iterative procedure. The results are presented in Section 3. Some concluding remarks and a proposal for further work are provided in Section 4.

2 Early stopping criteria

Let μ be the mean of the learning error at a given iteration k , $\mu_k = \frac{1}{n} \sum_{i=1}^n e_i$, with variance $\sigma_k^2 = \frac{1}{n-1} \sum_{i=1}^n (e_i - \mu)^2$, where n is the sample size. The skewness is given by $sk = \frac{1}{(n-1)\sigma^3} \sum_{i=1}^n (e_i - \mu)^3$, and the kurtosis is given by $ks = \frac{1}{(n-1)\sigma^4} \sum_{i=1}^n (e_i - \mu)^4$. The kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution associated with $ks = 3$ and zero skewness. If learning errors e_i are *i.i.d.* and normally distributed, then in the limit $n \rightarrow \infty$ their coefficient of skewness/kurtosis follows a Gaussian distribution, respectively, with mean $0/3$ and variance $\frac{6}{n}/\frac{24}{n}$ [6]. In this context, we can standardize skewness and kurtosis variance to unity as $Z_{sk} = \frac{1}{(n-1)\sigma^3} \sqrt{\frac{n}{6}} \sum_{i=1}^n (e_i - \mu)^3$ and $Z_{ks} = \frac{1}{(n-1)\sigma^4} \sqrt{\frac{n}{24}} \sum_{i=1}^n (e_i - \mu)^4$, which allow us testing normality of the residuals at a given confidence interval. So that, if $|Z_{sk}| \leq 2.57$ then the distribution of skewness is normal at the 1% level [6].

In this work, we propose to stop the iterative process before it becomes saturated when the learning error satisfies conditions imposed on the discrete version of the first or second derivatives: $\Delta Z = Z(r+1) - Z(r) \leq \Gamma$ or $Z^{(2)} = Z(r+1) - 2Z(r) + Z(r-1) \leq \Gamma$, where $Z(r)$ is the value of the corresponding error moment on the r^{th} iteration, and Γ is a fixed free parameter. To avoid dependence on the variance, we cease iterations by imposing conditions on the *normalized versions*, $\Delta Z/Z$ or $\Delta Z^{(2)}/Z$. Hence, if the change of the error moments is high (not the error itself), then the iterative process is continued; otherwise the SVM training procedure is stopped, whether the error

distribution is normal or not.

3 Experimental Results

In this section, we show the experimental results both in regression (sinc function approximation buried in noise) and pattern recognition (UCI databases) problems. We analyze the distribution of errors, computational cost, robustness to noise, and performance of different moment-based early-stopping strategies, and compare them with the standard tolerance-based iterative procedure.

3.1 Regression toy example

Here we examine the function approximation problem on the typical *sinc* function corrupted by Gaussian noise with zero mean and variance P_n^2 [7,8]. As a training dataset, we used 49 points drawn uniformly from the interval $[-10, 10]$, and 30 realizations were performed to avoid skewed results. Figure 1 shows the convergence of the primal error with and without early stopping for the approximation of the *sinc* function as well as the evolution of the error distribution in the training set. We stop iterations if the discrete version of the first (variance independent) derivative of error moment $\Delta Z/Z$ exceeds an upper bound of a joint condition on third and fourth central moments. At the end of convergence, the mean of learning error, μ , is near zero, and the standard deviation, σ , approaches the noise level value, P_n , as shown in Fig. 1. Moreover, for any P_n the observed distributions of Z_{sk} and Z_{ks} are normal at the 16.2% and 3.7% confidence levels, respectively.

In the left bottom panel of Fig. 1, we illustrate the evolution of the number of support vectors (# SVs) as a function of the number of iterations during the training. It is worth noting here that the standard training stops when the rate of the functional falls below the tolerance threshold with iterations, but the number of support vectors achieves the steady state values earlier. The solution remains almost the same up to the last iteration due to the fact that the rest of iterations serve only to fine-tune the positions of SVs, whereas their number and the learning error remain constant. Simple inspection of the error distribution shows that, after some oscillations during the initial iterations, the higher moments of error change monotonically on the final stage of the convergence. We propose to set the thresholds just on this stage, when the objective function decays slowly, to ensure gradual change of the error moments.

We carried out a number of experiments considering several joint conditions as well as separate conditions on different moments. It turned out that the performance (in terms of the SVM accuracy and complexity) under the joint condition (shown above) was close to the performance under the separate conditions on each error moment, as

far as the iterations stopped at nearly the same step. Although joined in Fig. 1, each criterion alone on third (forth) moment results in the prediction on the test stage (shown in Figure 2) because of selected threshold of 1% (10%) during the training stage. Figure 2 shows the results of all stopping strategies averaged over 30 runs in the *sinc* data. From Fig. 2(a), the mean value of the speedup factor changes from 4 up to 8 as noise level varies. Constrains on all error moments provide similar results: a higher noise level causes a higher number of SVs and prediction risk, see Fig. 2. The good results observed during training are confirmed in the testing stage (Fig. 2(d)), where accuracy and number of support vectors are very close for all strategies.

3.2 Classification on UCI datasets

The classification test for the proposed method considers five databases from the UCI *Machine Learning* repository [9]. The selected data sets (*ionosphere*, *liver-disorders*, *pima-diabets*, *wdbc*, *sonar*) have different number of patterns and attributes, which allows an overall comparison. Table 1 gives information on each dataset characteristics and shows the main results.

In order to develop SVM with RBF kernel, only the penalization parameter C and the kernel parameter σ_{ker} must be tuned. A total of 400 runs were performed for each problem and the optimum free parameters were selected through the 8-fold cross-validation method. Models were trained using both early-stopping and standard tolerance-based criteria. To compare the performance of the SVM with and without early-stopping, we took some measures of quality of performance: the cross-validation estimation of the success rate (SR[%]), the relative number of support vectors (SV[%]), the number of resulting iterations (#Iter), and the CPU time (in secs.).

Table 1 summarizes the best results obtained for each problem. It is worth to note that the computational effort is significantly lower using our proposal, sometimes, at the expense of slightly higher number of support vectors due to the early stopping. However, this does not constitute a serious drawback since the early-stopping procedure results in almost the same accuracy. For instance, if the error variance criterion is applied to the *pima-diabets* dataset (768 samples), the maximal gain in speedup is about three. Also, the early-stopping iterative procedure outperforms the standard algorithm in CPU time (performing almost 3 times faster at only 0.1% increase of # SV) on the 60-dimensional *sonar* dataset.

The most critical point when introducing early stopping in SVM procedures is, as already pointed, the possible lack of sparsity. In this sense, it is worth noting that there exists a trade-off between the number of SVs and the number of iterations in all UCI problems. Remarkably, one can obtain similar solutions (with and without early stopping) in terms of sparsity more likely by using third/forth moment-based criteria

than the first/second ones. This has a certain effect on the speedup factor, which is a bit lower using our suggestion. However, the best overall compromise between the sparsity and the number of iterations is provided by using skewness/kurtosis rather than mean/variance, e.g. see *liver-disorders* in Table 1. These results for real classification problems compare favorably in the cross-validation success rate with other studies using this dataset [10]. The interested reader can obtain more details and experimental results in [11].

4 Conclusions

This paper has presented the use of statistical criteria for early stopping in SVM both for regression and classification. The proposed method yields similar accuracy to that for the common (tolerance-based) algorithms with significant computational savings. Moreover, early-stopping criteria SVM can be implemented easily to any gradient descent techniques such as the backpropagation in a multilayer perceptron or in the framework of sparse Bayesian learning, where sparsity is intrinsically obtained. Further work will consider developing methods for the automatic searching of threshold parameters and studying the relationship to model regularization.

References

- [1] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.
- [2] G. Camps-Valls, J. L. Rojo-Álvarez, M. Martínez-Ramón, *Kernel Methods in Bioengineering, Signal and Image Processing*, Idea Group, Hershey, PA, USA, 2006.
- [3] J. Platt, Fast training of support vector machines using sequential minimal optimization, in: B. Schölkopf, C. J. C. Burges, A. J. Smola (Eds.), *Advances in Kernel Methods — Support Vector Learning*, MIT Press, Cambridge, MA, 1999, pp. 185–208.
- [4] F. Pérez-Cruz, A. Navia-Vázquez, J. L. Rojo-Álvarez, A. Artés-Rodríguez, A new training algorithm for Support Vector Machines, in: *Proceedings of the Fifth Baiona Workshop on Emerging Technologies in Telecommunications*, Baiona, Spain, 1999, pp. 116–120.
- [5] Z. Cataltepe, Y. S. Abu-Mostafa, M. Magdon-Ismail, No free lunch for early stopping, *Neural Computation* 11 (4) (1999) 995–1009.
- [6] M. G. Kendall, A. Stuart, *The Advanced Theory of Statistics*, Mcgraw Hill, 1969.
- [7] A. Chalimourda, B. Schölkopf, A. J. Smola, Experimentally optimal ν in support vector regression for different noise models and parameter settings, *Neural Networks* 17 (2004) 127–141.

- [8] V. Cherkassky, M. Yunqian, Practical selection of svm parameters and noise estimation, *Neural Networks* 17 (2004) 113–126.
- [9] C. L. Blake, C. J. Merz, UCI repository of machine learning databases, <http://www.ics.uci.edu/~mllearn/MLRepository.html>. Last modification: June, 1st 2004. Last access: June, 1st 2004. (1998).
- [10] G. Camps-Valls, J. D. Martín-Guerrero, J. L. Rojo-Álvarez, E. Soria-Olivas, Fuzzy sigmoid kernel for support vector classifiers, *Neurocomputing* 17 (2004) 113–126.
- [11] T. Bandos, G. Camps-Valls, E. Soria-Olivas, Support Vector Machines Early Stopping, Tech. Rep. TR-DIE-GPDS-02/17/2005, Dept. Enginyeria Electrònica, Universitat de València, Spain, available at <http://www.uv.es/~gcamps> (Feb 2005).

List of Figures

- 1 Error convergence in the *sinc* data for early-stopping criteria on normalized kurtosis ($\Delta Z_{ks}/Z_{ks} \leq 1\%$) and skewness ($\Delta Z_{sk}/Z_{sk} \leq 10\%$) (solid), and tolerance-based (dash-dotted) strategy. Different noise levels are illustrated: $P_n = 0.01, 0.06, 0.11, 0.16, 0.21, 0.26$. Learning parameters: $\{C = 1.58, \sigma_{ker} = 3, \varepsilon = 0.12\}$. 8
- 2 Average results (over 30 runs) in the *sinc* data. (a) CPU time (in the logarithmic scale), (b) number of support vectors, (c) number of iterations (in the logarithmic scale), and (d) test prediction risk versus standard deviation of noise, P_n . Four different stopping strategies are considered: normalized kurtosis and skewness with $\Delta Z_{ks}/Z_{ks} \leq 1\%$ and $\Delta Z_{sk}/Z_{sk} \leq 10\%$; normalized standard deviation and mean with $\Delta\sigma/\sigma \leq 1\%$ and $\Delta\mu/\mu \leq 1\%$. Learning parameters: $\{C = 1.58, \sigma_{ker} = 3, \varepsilon = 0.12\}$. 9

List of Tables

- 1 Results on the selected UCI databases. From *left* to *right* columns: database name, number of patterns(#P), number of attributes (#N), learning parameters: $\{C, \sigma_{ker}\}$, Success Rate (SR[%]), relative number of SVs (SV[%]), number of iterations (#Iter.), CPU time [s] for the standard tolerance-based SVM and the SVM stopped by the criteria on the error kurtosis, skewness, std (σ), mean (μ) (from top to bottom rows, respectively). The best scores for each performance indicator are highlighted in bold face font. 10

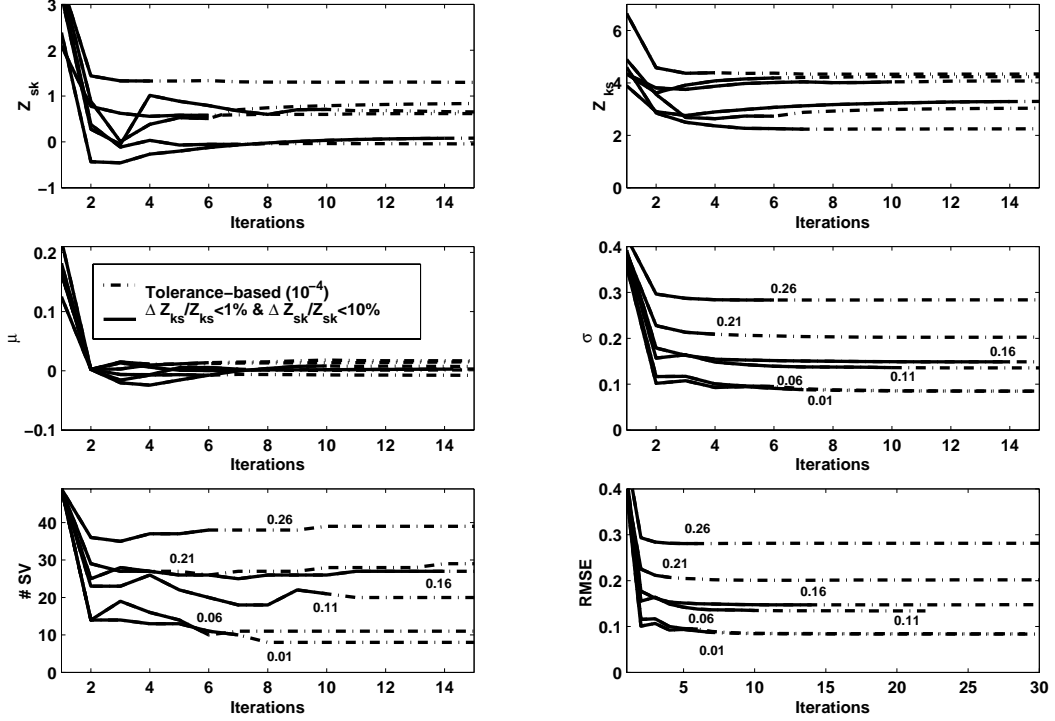


Fig. 1. Error convergence in the *sinc* data for early-stopping criteria on normalized kurtosis ($\Delta Z_{ks}/Z_{ks} \leq 1\%$) and skewness ($\Delta Z_{sk}/Z_{sk} \leq 10\%$) (solid), and tolerance-based (dash-dotted) strategy. Different noise levels are illustrated: $P_n = 0.01, 0.06, 0.11, 0.16, 0.21, 0.26$. Learning parameters: $\{C = 1.58, \sigma_{ker} = 3, \varepsilon = 0.12\}$.

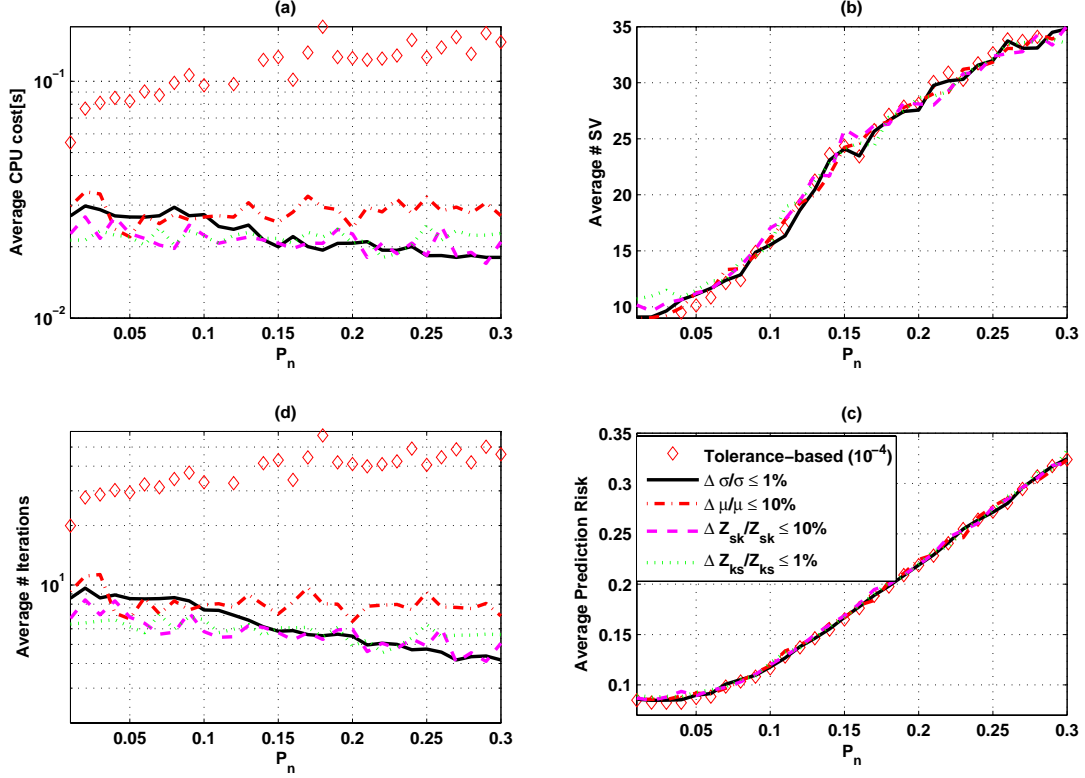


Fig. 2. Average results (over 30 runs) in the *sinc* data. (a) CPU time (in the logarithmic scale), (b) number of support vectors, (c) number of iterations (in the logarithmic scale), and (d) test prediction risk versus standard deviation of noise, P_n . Four different stopping strategies are considered: normalized kurtosis and skewness with $\Delta Z_{ks}/Z_{ks} \leq 1\%$ and $\Delta Z_{sk}/Z_{sk} \leq 10\%$; normalized standard deviation and mean with $\Delta\sigma/\sigma \leq 1\%$ and $\Delta\mu/\mu \leq 10\%$. Learning parameters: $\{C = 1.58, \sigma_{ker} = 3, \varepsilon = 0.12\}$.

Table 1

Results on the selected UCI databases. From *left* to *right* columns: database name, number of patterns (#P), number of attributes (#N), learning parameters: $\{C, \sigma_{ker}\}$, Success Rate (SR[%]), relative number of SVs (SV[%]), number of iterations (#Iter.), CPU time [s] for the standard tolerance-based SVM and the SVM stopped by the criteria on the error kurtosis, skewness, std (σ), mean (μ) (from top to bottom rows, respectively). The best scores for each performance indicator are highlighted in bold face font.

Name	#P	#N	C	σ_{ker}	SR[%]	SV[%]	#Iter.	CPU[s.]
ionosphere	351	34	5.5	2.5				
– Tolerance-based SVM					95.36	26.30	28.25	8.87
– $Z_{ks}^{(2)} / Z_{ks} \leq 1\%$					95.36	28.12	9.87	4.83
– $Z_{sk}^{(2)} / Z_{sk} \leq 0.1\%$					95.36	26.54	14.00	5.72
– $\sigma^{(2)} / \sigma \leq 1\%$					95.36	28.12	9.62	4.68
– $\mu^{(2)} / \mu \leq 0.1\%$					95.36	26.54	14.87	6.03
liver-disorders	345	6	15	65				
– Tolerance-based SVM					71.31	67.19	28.62	4.47
– $Z_{ks}^{(2)} / Z_{ks} \leq 1\%$					69.89	71.55	8.87	2.16
– $Z_{sk}^{(2)} / Z_{sk} \leq 0.1\%$					71.87	67.17	16.12	3.11
– $\sigma^{(2)} / \sigma \leq 1\%$					70.17	78.28	6.12	1.67
– $\mu^{(2)} / \mu \leq 0.1\%$					71.59	67.23	16.50	3.17
pima-diabetes	768	8	1100	200				
– Tolerance-based SVM					75.39	48.92	26.00	6.65
– $Z_{ks}^{(2)} / Z_{ks} \leq 1\%$					76.04	51.25	11.00	4.23
– $Z_{sk}^{(2)} / Z_{sk} \leq 0.1\%$					75.78	52.03	10.25	4.09
– $\sigma^{(2)} / \sigma \leq 1\%$					76.30	54.87	8.62	3.73
– $\mu^{(2)} / \mu \leq 0.1\%$					75.78	51.58	16.25	5.06
wdbc	569	30	30000	1400				
– Tolerance-based SVM					96.18	11.02	30.50	8.11
– $Z_{ks}^{(2)} / Z_{ks} \leq 1\%$					96.01	12.07	17.50	5.50
– $Z_{sk}^{(2)} / Z_{sk} \leq 0.1\%$					95.83	14.96	13.37	5.02
– $\sigma^{(2)} / \sigma \leq 1\%$					96.01	11.57	19.00	5.89
– $\mu^{(2)} / \mu \leq 0.1\%$					96.18	10.99	27.50	6.97
sonar	208	60	16.7	7.9				
– Tolerance-based SVM					87.50	56.46	25.87	7.52
– $Z_{ks}^{(2)} / Z_{ks} \leq 1\%$					87.50	56.52	7.75	2.74
– $Z_{sk}^{(2)} / Z_{sk} \leq 0.1\%$					87.50	56.52	8.00	2.92
– $\sigma^{(2)} / \sigma \leq 1\%$					87.50	56.59	8.25	3.03
– $\mu^{(2)} / \mu \leq 0.1\%$					87.50	56.46	14.12	4.28