Tuning L1-SVM Hyperparameters with Modified Radius Margin Bounds and Simulated Annealing

Javier Acevedo, Saturnino Maldonado, Philip Siegmann, Sergio Lafuente, and Pedro Gil

University of Alcala, Teoría de la señal, Alcala de Henares, Spain
javier.acevedo@uah.es
http://www2.uah.es/teose

Abstract. In the design of support vector machines an important step is to select the optimal hyperparameters. One of the most used estimators of the performance is the Radius-Margin bound. Some modifications of this bound have been made to adapt it to soft margin problems, giving a convex optimization problem for the L2 soft margin formulation. However, it is still interesting to consider the L1 case due to the reduction in the support vector number. There have been some proposals to adapt the Radius-Margin bound to the L1 case, but the use of gradient descent to test them is not possible in some of them because these bounds are not differentiable. In this work we propose to use simulated annealing as a method to find the optimal hyperparameters when the bounds are not differentiable, have multiple local minima or the kernel is not differentiable with respect to its hyperparameters.

1 Introduction

Support vector machines (SVM) [1] have been applied with a satisfactory level of success to many different binary classification problems. In order to achieve a good performance and generalization some parameters have to be optimized, such us the regularization constant or the width if gaussians kernels are used.

The adjustment of these hyperparameters is usually done minimizing an error estimator. Usually, the error is estimated by testing the adjusted SVM with external data not used for training. The problem with this approach is the bias of the estimator and a lack of generalization. There are some estimators based only on the training set, like the bootstrap, the cross-validation or the leave-one-out (LOO). Although all of them provide a statistical estimation of the error, the computational requirements needed make them to be not commonly used. In the last years there was an intensive research to find bounds to the LOO estimator. In [2] and [3] it was exposed the $\alpha\xi$-estimator and the span bound respectively, based on a deep knowledge of the SVM. These methods are inexpensive from a computational point of view, but the bias they present is very high in most of the problems and lead us to not select the optimal hyperparameters of the SVM.

Another bound that is more related to the Statistical Learning Theory (SLT) is the Radius-Margin bound [1]. However, this bound is formulated for the

F. Sandoval et al. (Eds.): IWANN 2007, LNCS 4507, pp. 284–291, 2007.

hard-margin formulation of the SVM and can not be directly applied to the L1-norm soft-margin formulation. In [4] it was demonstrated that for the L2-norm soft margin formulation, the Radius-Margin bound can be modified and the optimization problem is still convex. In [5] it was reinforced this idea by comparison with other methods, but also introduced a modified Radius-Margin bound to be applied for the L1-norm soft margin formulation. Motivated by the fact that with this last formulation the number of support vectors is more reduced, in [6] new bounds were proposed for the L1-soft margin case.

The final target of these methods is to obtain an automatic method to tune the hyperparameters. As it has been mentioned, in the L2 case the problem is convex and also differentiable, so a gradient descent approach is a good strategy, but in the L1 case, the proposed bounds are not differentiable with respect to the hyperparameters or have multiple local minima. Moreover, gradient descent methods, even in the L2 case, are limited to kernels that can be differentiable with respect to their hyperparameters, as the gaussian case. This limitation makes that other kernels like the inhomogeneous polynomial one [7] can not be adjusted.

Simulated Annealing (SA) [8] is a useful method to optimize functions based on the Statistical Local Search (SLS) principle. The proposal of this work is to use a variant of this method to obtain the hyperparameters, based on the L1-SVM Radius-Margin bounds proposed in the literature. In fact, with the SA method, there is no need in the optimization function to be differentiable, so the second part of the work studies the hyperparameter selection of inhomogeneous polynomial kernel in the L1-SVM and L2-SVM cases with the proposed bounds.

2 The Radius-Margin Bound

Given a binary classification problem with l training vectors $\mathbf{x_i} \in \mathbb{R}^{\mathbf{m}}$ and labels $y_i \in \{-1, 1\}$, the dual L1-Norm soft margin SVM formulation can be written as follows:

$$\begin{array}{ll} \min\limits_{\mathbf{w},\boldsymbol{\xi}} & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\mathbf{e}^T\boldsymbol{\xi} \\ \text{subjet to} & \\ & y_i\left(\langle \mathbf{w}, \phi\left(\mathbf{x_i}\right)\rangle + b\right) \geq 1 - \xi_i \qquad i = 1, 2, ..., l \ . \end{array} \tag{1}$$

Where $\mathbf{e}$ is the vector of all ones , $\mathbf{w}$ is the normal vector to the discriminating hyperplane, $\phi\left(\mathbf{x}_i\right)$ is a map to a higher dimension of the training vectors, C is a regularization parameter and $\boldsymbol{\xi}$ is a vector of slack variables allowing some of the training patterns to be into the margin region or misclassified. In order to find the solution of this optimization problem it is easier to solve the dual problem:

$$\begin{array}{l} \min W\left(\boldsymbol{\alpha}\ \right) = \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \mathbf{e^T}\boldsymbol{\alpha} \\ \text{subject to:} \\ \quad \alpha_i \geq 0 \\ \quad \mathbf{y}^T\boldsymbol{\alpha} = 0 \ . \end{array} \tag{2}$$

Where $\mathbf{Q}$ is an $l \times l$ matrix and $Q_{i,j} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, being $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x_i}), \phi(\mathbf{x_j}) \rangle$ a kernel function that satisfies Mercer's conditions. In this work, we have considered two kernels:

$$\begin{array}{l} K(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}, \sigma \in \mathbb{R} \\ K(x,y) = (\langle x, y \rangle + c)^d, c \in \mathbb{R}, d \in \mathbb{N} \ . \end{array} \tag{3}$$

The first one is the gaussian kernel and it has been widely used in the SVM field. Most of the works published related to the hyperparameter selection are focused on this kernel. The σ parameter has to be set a priori and thus, it is the kernel hyperparameter to be tuned. The second one is the inhomogeneous polynomial kernel, where the exponent is a natural number. This kernel has its counterpart in the homogeneous one, where the degree can be a real number, but the inhomogeneous one is interesting for many real applications where the kernel operations have to be as simpler as possible. The hyperparameters to be estimated in this case are the bias constant c and the degree d. As well as tuning this kernel hyperparameters the C regualarization constant has to be adjusted in all the cases.

The L2-Norm SVM soft margin formulation change the upper part of (1) by

$$\min_{\mathbf{w},\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\boldsymbol{\xi}^T\boldsymbol{\xi} \ . \tag{4}$$

and the dual becomes

$$\begin{array}{l} \min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T \left(\mathbf{Q} + \frac{\mathbf{I}}{C}\right)\boldsymbol{\alpha} \\ \text{subject to:} \\ \quad \alpha_i \geq 0 \\ \quad \mathbf{y}^T\boldsymbol{\alpha} = 0 \ . \end{array} \tag{5}$$

The Radius-Margin Bound describes how for the separable case, the LOO error is bounded by:

$$\text{LOO} \leq \frac{1}{4l} D^2 \|\mathbf{w}\|^2 \ . \tag{6}$$

Where $D = 2R$ is the diameter of the smallest sphere containing all the training points in the transformed space, where the inputs have been mapped. The radius of this sphere is calculated by means of a dual optimization problem as described in [9]. Although this bound is not valid for the soft margin formulations, it is simple and is related to the maximum margin classifiers principle. In the case of the L2-SVM a change in the variables R and $\|\mathbf{w}\|^2$ can be done to take into account the slack variables and the modified $\widetilde{R}^2 \|\widetilde{\mathbf{w}}\|^2$ [4] has the property of being a convex problem to be solved. However, for the L1 case it is not possible to make this change in the variables, but it is still interesting to use the L1-SVM because it provides less support vectors. Due to this reason some modifications in the original Radius-Margin bound have been proposed in [5] and [6]:

$$\begin{array}{l} D^2\mathbf{e}^T\boldsymbol{\alpha} + \mathbf{e}^T\boldsymbol{\xi} \\ \left(R^2 + \frac{1}{C}\right)\mathbf{e}^T\boldsymbol{\alpha} \\ \left(R^2 + \frac{1}{C}\right)\left(\|\mathbf{w}\|^2 + 2C\mathbf{e}^T\boldsymbol{\xi}\right) \ . \end{array} \tag{7}$$

Only the third of them is differentiable with respect to C, that make this bound the only one that can be used when tuning hyperparameters by gradient descent but there can be multiple local minima. On the other hand, as the first and the second are not differentiable with respect to C, in the published works they have been evaluated by means of a grid sampling search making the use of these bounds unpractical. In the following, we refer to these bounds as L1BOUND1, L1BOUND2 and L1BOUND3 respectively.

On the other hand, although a gradient descent can be applied in the L2-SVM formulation and L1BOUND3 case, it requires the kernel to be differentiable with respect to the kernel hyperparameters. While this is true in the gaussian case, in the inhomogeneous polynomial kernel this condition fails.

3 Simulated Annealing

Simulated annealing (SA) is a random-search technique which exploits an analogy between the way in which a metal cools and freezes into a minimum energy crystalline structure (the annealing process) and the search for a minimum in a more general system. SA has been applied to many optimization problems and it has been proved that by carefully controlling the rate of cooling of the temperature, SA can find the global optimum. However, this requires infinite time. In this work we have focused on a variant of the original algorithm called Adapted Simulated Annealing (ASA) [10], that applied to our optimization problem can be summarized in the following steps:

1. Randomly select a value of the C regularization constant and for the kernel hyperparameters θ. Make the selection from a uniform distribution taking the values from given ranks.
2. Calculate the objective function f to be minimized. In the L1-SVM select one of the three bounds proposed in (7) whereas in the L2 case compute $\widetilde{R}^2 \left\|\widetilde{\mathbf{w}}\right\|^2$.
3. Assign to the initial temperature T_0 and to the initial energy E_0 the value f_0 obtained in the previous step. Assign to the final temperature T_f a value $T_f = T_0\epsilon$ where ϵ is a fixed constant. In our examples we have considered $\epsilon = 10^{-4}$. Set f^{global} to be f^0.
4. Select a candidate solution $\left[C^N, \theta^N\right]$ based on increments or decrements of the previous solution. Each increment/decrement of the variables y_i of this candidate solution is found by a neighborhood function:

$$\begin{aligned} &y_j = sgn(u_j - 0.5)T\left[\left(1+\tfrac{1}{T}\right)^{|2u_j-1|} - 1\right]\left(\text{Range}_j\right) \qquad j = 1, 2, ..D \\ &u_j \in [0,1] \ \text{ uniform randomized variable } . \end{aligned} \tag{8}$$

   Compute the objective function f^N with the candidate solution.
5. if $f^N < f^{\text{global}}$, then $\left[C^{\text{global}}, \theta^{\text{global}}\right] \leftarrow \left[C^N, \theta^N\right]$ and $f^{\text{global}} \leftarrow f^N$. Else accept the candidate solution with a probability

$$p = e^{\frac{-|f^{global} - f^N|}{T}} \ . \tag{9}$$

6. Update the temperature following a cooling schedule

$$T(k) = T_0 e^{\frac{-hk}{D}} \quad . \tag{10}$$

   Where h is a constant calculated with the final temperature and D is the number of variables to be adjusted.
7. If the maximal epochs allowed has been reached, return $\left[C^{\text{global}}, \theta^{\text{global}}\right]$ and f^{global}.

The advantages of the proposed method are that there are no multiple parameter to adjust and the proposed cooling scheme search for all the solution space at the beginning and refines the solution at the final epochs.

4 Results and Discussion

The proposed application of the SA optimization algorithm was evaluated on different databases, all of them available from the U.C.I. repository and detailed in Table(1).

Table 1. Description of the datasets used

	DATASET			
	BANANA	IMAGE	TREE	SPLICE
Number of Features	2	18	18	60
Number of Training Samples	400	1300	700	1000
Number of Test Samples	4900	1010	11692	2175

4.1 Gaussian Kernel and L1 Soft Margin-SVM

In this first section we have focused in the gaussian kernel, because most part of the published works have used only this kernel. The first task to do was to compare the behavior of the proposed algorithm against the gradient descent when the bound is differentiable, as it is the case of the L1BOUND3. To make this comparison both methods were tested with the same number of function evaluations, starting from the same initial solution chosen randomly. Following the published literature, a logarithmic transformation is applied to the hyperparameters C and σ. In Fig.(1) it is shown the results obtained for the Image dataset, where the contour lines have been calculated by intensive grid search with a cluster of computers. It can be appreciated how the gradient descent solutions when fall into the convergence region are closer than in the SA method, but there are more trials that have failed. It is important to note that we are looking for a region and not for a exact minimization point. The experiment was repeated with different datasets and different number of function evaluations and it can be said that the SA behavior is as good as the gradient descent approach.

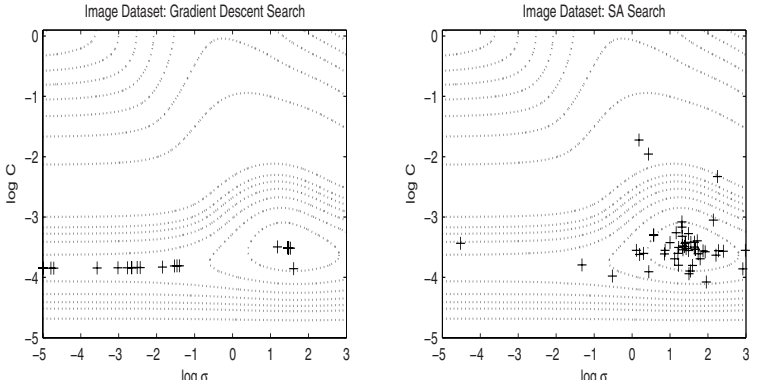

Fig. 1. Comparison of Gradient Descent Search and Simulated Annealing with L1BOUND3 in the Image dataset

The advantage of using SA is that we can also use bounds not differentiable to tune the hyperparameters. One of the major issues in the optimization search is the number of function evaluations that are needed to achieve the desired minimum. In Fig.(2) it is shown the number of function evaluations needed per each bound with different datasets. It has been considered that SA has reached the minimum if the proposed solution found by the algorithm varies in less than a 1% of the near optimum solution calculated with intensive search. These curves provide the probability of reaching the minimum with a number of function evaluations. To make each of the curves a process evaluated the SA 1000 times.

Table 2. Comparison of the different bounds with respect to test samples

Dataset	L1BOUND1			L1BOUND2			L1BOUND3		
	log C, log σ	N^o SV	Test Error	log C, log σ	N^o SV	Test Error	log C, log σ	N^o SV	Test Error
BANANA	(-2.8,-1.2)	324	0.1361	(-0.8,-1.4)	163	0.1102	(-0.6,-1.8)	162	0.1129
IMAGE	(-0.8,0)	705	0.0466	(6.8,-1.2)	294	0.0228	(-3.5,1.5)	935	0.1337
TREE	(-2.16,0.2)	409	0.1388	(-1.6,0.8)	337	0.1360	(-3.1,1)	392	0.1498
SPLICE	(1.6,2.2)	975	0.1995	(1.9,3.1)	837	0.1002	(1.8,2.7)	944	0.1205

Once it has been obtained the number of function evaluations needed to have success with a high probability, a comparison between the three bounds has been made. The procedure was to train different L1-SVM with the hyperparameters found by each of the bounds. The test samples are then classified with the trained SVM. Results of this task are shown in Table (2), where it can be appreciated that L1BOUND2 and L1BOUND3 give similar results, but there is a reduction in the number of support vectors used when using L1BOUND2. The worst behavior in all cases is the one provided by L1BOUND1, because the hyperparameters provided have generated SVM with more support vectors and give less accuracy.

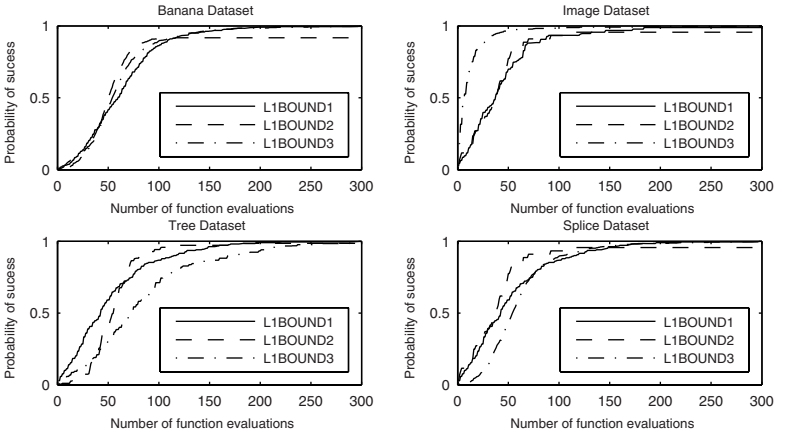

Fig. 2. Number of function evaluations against probability of success in different datasets

4.2 Inhomogeneous Polynomial Kernel

Another advantage of SA is the use of not differentiable kernels with respect to the kernel hyperparameters. In this section we have chosen an example of one of this kernels like the inhomogeneous polynomial one. Again, SLS methods seems to be a good solution to solve these kind of problems. The experiment done was to compare the same datasets used in the previous section but in this case, the L1BOUND1 has been replaced by the L2 bound. It is interesting how the number of support vectors is very much higher in the L2 case, with no better accuracy.

Table 3. Results obtained for the inhomogeneous polynomial kernel

Dataset	L1BOUND2			L1BOUND3			L2 BOUND		
	log C, c, d	N^o SV	Test Error	log C, c, d	N^o SV	Test Error	log C, c, d	N^o SV	Test Error
BANANA	(4.60, 2.699, 4)	110	0.1178	(3.16, 2.6534, 4)	109	0.1176	(3.912 2.1, 4)	179	0.1271
IMAGE	(3.83 0.177,1)	564	0.1416	(3.86 2.5366, 1)	565	0.1416	(3.88 0.8817,1)	973	0.1723
TREE	(4.37 1.832,4)	159	0.1670	(3.86 0.733, 5)	206	0.1124	(4.25 1.765,2)	243	0.1132
SPLICE	(4.69 1.7086,4)	484	0.1159	(4.1746 2.12, 1)	0.1536	0.4257	(3.88 0.8817,1)	976	0.1713

5 Conclusion

In this work we have presented an adapted algorithm based on simulated annealing to find the optimal hyperparameters. This approach is needed in some

of the bounds for the L1 soft margin formulation or when we are working with no differentiable kernels. Results show that it is possible to optimize these hyperparameters with the proposed method.

Acknowledgement

This work was supported by Comunidad of Madrid projects CAM-UAH 2005/031 and CCG06-UAH/TIC0695.

References

1. Vapnik, N.V.: The Natureof Statistical Learning Theory, 1ed: 1998. Springer, Berlin (2000)
2. Joachims, T.: Estimating the generalization performance of a SVM efficiently. In: Langley, P. (ed.) Proc. of ICML-00, pp. 431–438. Morgan Kaufmann, San Francisco, US (2000)
3. Vapnik, V., Chapelle, O.: Bounds on error expectation for support vector machines. Neural Computation 12(9), 2013–2036 (2000)
4. Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S.: Choosing multiple parameters for support vector machines. Machine Learning 46(1), 131–159 (2002)
5. Duan, K., Sathiya, S., Poo, A.: Evaluation of simple performance measures for tuning the svm hyperparameters. Neurocomputing 51, 41–59 (2003)
6. Chung, K.M., Kao, W.C., Sun, C.L., Wang, L.L., Lin, C.J.: Radius margin bounds for support vector machines with the rbf kernel. Neural Computation (15) , pp. 2643–2681, (2003)
7. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)
8. Kirkpatrick, S., Gellat, C., Vecchi, M.: Optimization by simulated annealing. Science (220) , pp. 671–680 (1983)
9. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, New York, USA (2004)
10. Ingber, A.L.: Adaptive simulated annealing (asa): Lessons learned. Control and Cybernetics 25(1), 33–54 (1996)