The LCCP for Optimizing Kernel Parameters for SVM

Sabri Boughorbel[1], Jean Philippe Tarel[2], and Nozha Boujemaa[1]

[1] IMEDIA Group, INRIA Rocquencourt, 78153 Le Chesnay, France
[2] DESE, LCPC, 58 Bd Lefebvre, 75015 Paris, France

Abstract. Tuning hyper-parameters is a necessary step to improve learning algorithm performances. For Support Vector Machine classifiers, adjusting kernel parameters increases drastically the recognition accuracy. Basically, cross-validation is performed by sweeping exhaustively the parameter space. The complexity of such grid search is exponential with respect to the number of optimized parameters. Recently, a gradient descent approach has been introduced in [1] which reduces drastically the search steps of the optimal parameters. In this paper, we define the LCCP (Log Convex Concave Procedure) optimization scheme derived from the CCCP (Convex ConCave Procedure) for optimizing kernel parameters by minimizing the radius-margin bound. To apply the LCCP, we prove, for a particular choice of kernel, that the radius is log convex and the margin is log concave. The LCCP is more efficient than gradient descent technique since it insures that the radius margin bound decreases monotonically and converges to a local minimum without searching the size step. Experimentations with standard data sets are provided and discussed.

1 Introduction

Support Vector Machine (SVM) [2] is one of the most successful algorithms of machine learning. SVM is flexible since various kernels can be plugged for different data representations. Besides RBF and Polynomial kernels only few other kernels have been used. An interesting and important issue for kernel design consists of assigning, for instance, different scales for each feature component. This is refereed as adaptive metrics [3]. On the other hand, the classical method for tuning the learning algorithm parameters is to select parameters that minimize an estimation or a bound on the generalization error such as cross validation or the radius margin [2]. The latter has been shown to be a simple and predictive enough "estimator" of the generalization error. In this paper, we define the LCCP for optimizing kernel parameters by minimizing the radius margin bound. The LCCP is the direct application of the CCCP [4] to our optimization case.

2 The Log Convex Concave Procedure (LCCP)

The convex concave procedure (CCCP) has been recently introduced [4] for optimizing a function that can be written as a sum of convex and concave functions.

The advantage of the CCCP compared with gradient descent techniques is that it insures the monotonic decrease of the objective function without searching the size step. In the following, we summarize the main results of the CCCP optimization framework.

Theorem 1. *[4]*

- *Let $E(\boldsymbol{\theta})$ be an objective function with bounded Hessian $\frac{\partial^2 E(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}$. Thus, we can always decompose it into the sum of convex and concave functions.*
- *We consider the minimization problem of a function $E(\boldsymbol{\theta})$ of form $E(\boldsymbol{\theta}) = E_{vex}(\boldsymbol{\theta}) + E_{cave}(\boldsymbol{\theta})$ where E_{vex} is convex and E_{cave} is concave. Then the discrete iterative CCCP algorithm: $\boldsymbol{\theta}_p \rightarrow \boldsymbol{\theta}_{p+1}$ given by $\boldsymbol{\nabla} E_{vex}(\boldsymbol{\theta}_{p+1}) = -\boldsymbol{\nabla} E_{cave}(\boldsymbol{\theta}_p)$ decreases monotonically the objective function $E(\boldsymbol{\theta})$ and hence converges to a minimum or a saddle point of $E(\boldsymbol{\theta})$.*
- *The update rule for $\boldsymbol{\theta}_{p+1}$ can be formulated as a minimization of a convex function $\boldsymbol{\theta}_{p+1} = \arg\min_{\boldsymbol{\theta}} E_{p+1}(\boldsymbol{\theta})$ where the convex function $E_{p+1}(\boldsymbol{\theta})$ is defined by*
$$E_{p+1}(\boldsymbol{\theta}) = E_{vex}(\boldsymbol{\theta}) + \boldsymbol{\theta}^\top \boldsymbol{\nabla} E_{cave}(\boldsymbol{\theta_p}).$$

We define the LCCP by applying the CCCP to the case of the minimization of a positive function $J(\boldsymbol{\theta})$ that can be written as a product of log convex and log concave functions $J(\boldsymbol{\theta}) = J_{lvex}(\boldsymbol{\theta}) J_{lcave}(\boldsymbol{\theta})$ where $J_{lvex}(\boldsymbol{\theta}) > 0$ is log convex and $J_{lcave}(\boldsymbol{\theta}) > 0$ is log concave. In $\log(J(\boldsymbol{\theta})) = \log(J_{lvex}(\boldsymbol{\theta})) + \log(J_{lcave}(\boldsymbol{\theta}))$, we set $E(\boldsymbol{\theta}) = \log(J(\boldsymbol{\theta}))$, $E_{vex}(\boldsymbol{\theta}) = \log(J_{lvex}(\boldsymbol{\theta}))$ and $E_{cave}(\boldsymbol{\theta}) = \log(J_{lcave}(\boldsymbol{\theta}))$. Hence, we obtain $E(\boldsymbol{\theta}) = E_{vex}(\boldsymbol{\theta}) + E_{cave}(\boldsymbol{\theta})$ where $E_{vex}(\boldsymbol{\theta})$ is convex and $E_{cave}(\boldsymbol{\theta})$ is concave. Moreover, the minima location of $E(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$ are the same since the log function is strictly increasing.

3 Parameters selection procedure

The optimization of SVM parameters can be performed by minimizing an estimator of the generalization error. The simplest strategy consists in performing an exhaustive search over all possible parameters. When the number of parameters is high, such a technique becomes intractable. In [1], gradient descent framework is introduced for kernel parameter's optimization. Powerful results on the differentiation of various error estimators and generalization bounds are provided. Based of this work, we apply the LCCP framework for optimizing multiple kernel parameters by minimizing the radius margin bound [2]. Indeed, for good choice of kernels, the optimizing problem can be expressed under the condition of LCCP, in particular for the multi-parameters L_1-distance kernel.

3.1 Distance kernel

In [1], tests with multiple parameters for polynomial and RBF kernels have been successfully carried without over-fitting. From the L_1-distance kernel:

$$K_{L_1}(\boldsymbol{x}, \boldsymbol{x}') = -\sum_{k=1}^{n} |x^k - x'^k|, \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are in $\mathbb{R}^n$ with components x^k and x'^k, we propose its following multiple parameters extension:

$$K_{L_1,\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{x}') = -\sum_{k=1}^{n} \frac{|x^k - x'^k|}{\theta^k}, \tag{2}$$

where $\boldsymbol{\theta}$ is in $\mathbb{R}^{+^n}$ with components θ^k. This kernel is conditionally positive definite, see [5]. We prove that it is possible to use the LCCP for minimizing radius-margin bound $R^2\|\boldsymbol{w}\|^2$, with respect to $\boldsymbol{\theta}$. To do so, we prove the log convexity of the radius R^2 and the log concavity of $\|\boldsymbol{w}\|^2$. Another proof may be used for another kernel. More precisely, for R^2, we will prove that it can be written as a sum of log convex functions. For $\|\boldsymbol{w}\|^2$, it is sufficient to prove that it is concave since the concavity implies the log concavity.

3.2 The log convexity of R^2

First, we recall from [6] a useful result on convex functions that we need in the proof of the log convexity of the radius R^2.

Lemma 1. *If for each $\boldsymbol{y} \in \mathcal{A}$, $f(\boldsymbol{x},\boldsymbol{y})$ is convex in $\boldsymbol{x}$, then the function g, defined as $g(\boldsymbol{x}) = \max_{\boldsymbol{y}\in\mathcal{A}} f(\boldsymbol{x},\boldsymbol{y})$ is convex in $\boldsymbol{x}$.*

This result can be easily extended to the case of log convex functions. The radius R^2 can be written for the kernel (2) as the following:

$$R^2(\boldsymbol{\theta}) = \max_{\boldsymbol{\beta}\in\mathcal{B}} J_{R^2}(\boldsymbol{\beta},\boldsymbol{\theta}), \tag{3}$$

where $\mathcal{B} = \{\beta_i \geq 0, \sum_{i=1}^{\ell} \beta_i = 1\}$ and J_{R^2} is the following function:

$$J_{R^2}(\boldsymbol{\beta},\boldsymbol{\theta}) = -\sum_{i=1}^{\ell} \beta_i \sum_{k=1}^{n} F_{a_{ii}^k}(\boldsymbol{\theta}) + \sum_{i,j=1}^{\ell} \beta_i\beta_j \sum_{k=1}^{n} F_{a_{ij}^k}(\boldsymbol{\theta}), \tag{4}$$

with $F_{a_{ij}^k}(\boldsymbol{\theta}) = f_{a_{ij}^k}(\theta^k) = \frac{a_{ij}^k}{\theta^k}$ and $a_{ij}^k = |x_i^k - x_j^k|$. Since $a_{ii}^k = 0$, the first sum in J_{R^2} is zero. Next, we prove that F is log convex. To do so, it is necessary and sufficient [6] to prove that $\boldsymbol{\nabla}^2 F(\boldsymbol{\theta})F(\boldsymbol{\theta}) - \boldsymbol{\nabla} F(\boldsymbol{\theta})\boldsymbol{\nabla} F(\boldsymbol{\theta})^\top$ is a positive definite matrix. By computing the gradient $\boldsymbol{\nabla} F$ and the Hessian $\boldsymbol{\nabla}^2 F$, it turns out that the obtained matrix is diagonal. Thus the necessary and sufficient condition for the log convexity becomes $f''_{a_{ij}^k}(\theta^k) f_{a_{ij}^k}(\theta^k) - f'^2_{a_{ij}^k}(\theta^k) \geq 0$. We have:

$$f_a(t) = \frac{a}{t},\ f'_a(t) = -\frac{a}{t^2},\ f''_a(t) = \frac{2a}{t^3},\ f''_a(t)f_a(t) - {f'_a}^2(t) = \frac{a^2}{t^4} \geq 0.$$

So J_{R^2} is log convex with respect to $\boldsymbol{\theta}$, as a sum of log convex functions [6]. Lemma 1 implies that R^2 is log convex.

3.3 Log concavity of $\|\boldsymbol{w}\|^2$

A similar result to Lemma 1, for the concave case, can be derived [6]:

Lemma 2. *Assume that $\mathcal{A}$ is a convex set, if $f(\boldsymbol{x}, \boldsymbol{y})$ is concave in $(\boldsymbol{x}, \boldsymbol{y})$, then the function g, defined by $g(\boldsymbol{x}) = \max_{\boldsymbol{y} \in \mathcal{A}} f(\boldsymbol{x}, \boldsymbol{y})$ is concave in $\boldsymbol{x}$.*

We also need two extra lemmas, which are proved with details in [5]:

Lemma 3. *We define the function f by*

$$f(\boldsymbol{a}, t) = \frac{1}{t} \boldsymbol{a}^\top \boldsymbol{K}\, \boldsymbol{a},\ \boldsymbol{a} \in \mathbb{R}^\ell,\ t \in \mathbb{R}_+,\ \boldsymbol{K} \in \mathbb{R}^{\ell \times \ell}.$$

If $\boldsymbol{K}$ is a positive definite matrix then f is convex in $(\boldsymbol{a}, t)$.

Lemma 4. *We define the function g for $\boldsymbol{t} \in \mathbb{R}^n, \boldsymbol{a} \in \mathbb{R}^\ell$*

$$g(\boldsymbol{a}, \boldsymbol{t}) = \sum_{k=1}^{n} f_k(\boldsymbol{a}, t^k).$$

If each f_k is convex in $(\boldsymbol{a}, t^k)$, then g is convex in $(\boldsymbol{a}, \boldsymbol{t})$.

The expression of $\|\boldsymbol{w}\|^2$ is the following:

$$\|\boldsymbol{w}\|^2(\boldsymbol{\theta}) = \max_{\boldsymbol{\alpha} \in \Lambda} J_{\|\boldsymbol{w}\|^2}(\boldsymbol{\alpha}, \boldsymbol{\theta}),$$

where $\Lambda = \{\alpha_i \geq 0, \sum_{i=1}^{\ell} \alpha_i y_i = 0\}$ and

$$J_{\|\boldsymbol{w}\|^2}(\boldsymbol{\alpha}, \boldsymbol{\theta}) = 2 \sum_{i=1}^{\ell} \alpha_i - \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

It is obvious that Λ is a convex set. The first term in $J_{\|\boldsymbol{w}\|^2}$ is linear with respect to $\boldsymbol{\alpha}$, thus it does not affect the convex or concave nature of $J_{\|\boldsymbol{w}\|^2}$. We thus only focus on:

$$J'_{\|\boldsymbol{w}\|^2}(\boldsymbol{\alpha}, \boldsymbol{\theta}) = -\sum_{k=1}^{n} \frac{1}{\theta^k} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

where $K_k(\boldsymbol{x}_i, \boldsymbol{x}_j) = -|x_i^k - x_j^k|$ is conditionally positive definite. We introduce the kernel $\tilde{K}_k$ defined by $\tilde{K}_k(\boldsymbol{x}, \boldsymbol{x}') = K_k(\boldsymbol{x}, \boldsymbol{x}') - K_k(\boldsymbol{x}, \boldsymbol{x}_0) - K_k(\boldsymbol{x}', \boldsymbol{x}_0) + K_k(\boldsymbol{x}_0, \boldsymbol{x}_0)$ where $\boldsymbol{x}_0$ is chosen arbitrary. It is known that $\tilde{K}_k$ is positive definite and that it can be substituted to K_k in the dual SVM problem, see [5]. Similarly, we can substitute K_k by $\tilde{K}_k$ in $J'_{\|\boldsymbol{w}\|^2}$ according to the constraint $\sum_{i=1}^{\ell} \alpha_i y_i = 0$ and rewrite it as $J'_{\|\boldsymbol{w}\|^2}(\boldsymbol{\alpha}, \boldsymbol{\theta}) = -\sum_{k=1}^{n} \frac{1}{\theta^k} \boldsymbol{\alpha_y}^\top \tilde{\boldsymbol{K}}_k\ \boldsymbol{\alpha_y} = -\sum_{k=1}^{n} f_k(\boldsymbol{\alpha}, \theta^k)$, where $\tilde{\boldsymbol{K}}_k$ is the Gram matrix of $\tilde{K}_k$, $\boldsymbol{\alpha_y}$ denotes the vector $[\alpha_1 y_1 \ldots \alpha_\ell y_\ell]^\top$, and $f_k(\boldsymbol{\alpha}, \theta^k) = \frac{1}{\theta^k} \boldsymbol{\alpha}^\top \tilde{\boldsymbol{K}}_{\boldsymbol{y},k}\ \boldsymbol{\alpha}$ with $\left[\tilde{\boldsymbol{K}}_{\boldsymbol{y},k}\right]^{ij} = y_i y_j \left[\tilde{\boldsymbol{K}}_k\right]^{ij}$. We have that $\tilde{\boldsymbol{K}}_{\boldsymbol{y},p}$

is positive definite. Therefore, lemma 3 implies that f_k is convex in $(\boldsymbol{\alpha}, \theta^k)$ and lemma 4 implies that the sum over f_k is convex in $(\boldsymbol{\alpha}, \boldsymbol{\theta})$. Therefore, we have the concavity of $J'_{\|\boldsymbol{w}\|^2}$ in $(\boldsymbol{\alpha}, \boldsymbol{\theta})$ and lemma 2 implies the concavity of $\|\boldsymbol{w}\|^2$ with respect to $\boldsymbol{\theta}$. The log concavity is always obtained when the concavity of positive functions is insured [6]. The conditions of LCCP are all fulfilled. We can thus apply it for the optimization of the L_1-distance kernel parameters.

4 Experiments

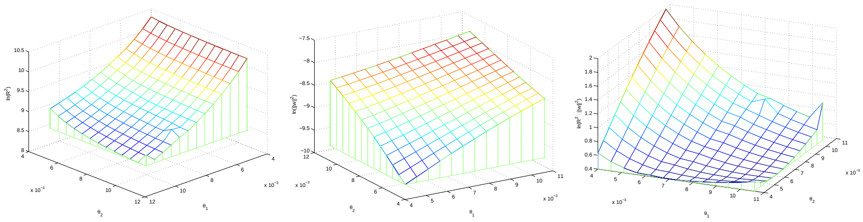

Fig. 1. The left, middle and right figures plot respectively $\log(R^2)$, $\log(\|\boldsymbol{w}\|^2)$ and $\log(R^2\|\boldsymbol{w}\|^2)$ with respect to the L_1-distance kernel parameters (θ^1, θ^2) on banana dataset which is a set of 2D points [7].

Fig. 1 shows the variation of $\log(R^2)$, $\log(\|\boldsymbol{w}\|^2)$ and $\log(R^2\|\boldsymbol{w}\|^2)$ with respect to θ_1 and θ_2 for the kernel (2). It illustrates the log convexity of R^2 and the log concavity of $\|\boldsymbol{w}^2\|$.

	Thyroid	Titanic	Heart	Breast-cancer
K_{L_1} (1)	**5.77%**	22.68%	20.65%	28.97%
K_{L_2} (5)	11.21%	22.56%	18.23%	29.77%
$K_{L_1,\boldsymbol{\theta}}$ (2)	6.20%	**22.08%**	**17.34%**	**27.12%**
n	5	3	13	9

Table 1. Test error's comparison of the single parameter L_1 distance kernel (1), L_2 distance kernel (5) and L_1 distance kernel with multiple parameters. n denotes the number of parameters for multi-parameter's kernel. LCCP is used for optimizing the radius margin bound.

In order to evaluate the performance of the LCCP for optimizing multiple parameters, we performed experiments on datasets obtained from [7]. We compare the L_1-distance kernel without parameters (1), the L_2-distance kernel:

$$K_{L_2}(\boldsymbol{x}, \boldsymbol{x}') = -\sum_{k=1}^{n} (x^k - x'^k)^2, \tag{5}$$

and the L_1-distance kernel with multiple parameters (2). Initial starting point is set to 1 for all θ_i, as in [1]. The stopping criterion is $|log(E_{p+1}(\theta_{p+1})) - log(E_p(\theta_p))| < \epsilon$. The data sets contain 100 realizations of training and test examples. For each realization, we optimize the kernel parameters on the training sample using the LCCP. The obtained parameters are used to estimate the generalization error on the test sample by a 5-fold cross-validation. Tab. 1 summarizes average test errors for different data sets. The L_2-distance kernel is equivalent to the linear kernel when used within SVM. We observe that the L_1-distance kernel performs better or similarly than L_2-distance kernel except on heart dataset. Tab. 1 shows that the use of multiple parameters in L_1-distance kernel allows us most of the time to decrease the test error, despite the weightening of each dataset. This shows clearly the interest of the introduction of multiple parameters in kernels.

5 Conclusion

In this paper, we propose an original way for optimizing of kernel multiple parameters by minimizing the radius margin bound, we named LCCP. The LCCP is derived directly from CCCP optimizing framework. The LCCP approach is more efficient than the gradient descent technique since it converges to a local minimum without searching the size step. We prove that, for the multi-parameters L_1-distance kernel, the radius margin fulfills the conditions for application of the LCCP. Comparison on standard data set leads to improved recognition performance compared to single parameter L_1-distance kernel. The multi-parameters L_1-distance kernel is only one example of kernel which fulfills the conditions for application of the LCCP, but other exists. The formal definition of the set of all kernels that fulfills these conditions is the subject of our future researches.

References

1. O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, “Choosing multiple parameters for support vector machines,” *Machine Learning*, vol. 46, no. 1-3, pp. 131–159, 2002.
2. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, 2nd edition, New York, 1999.
3. K. Tsuda, “Optimal hyperplane classifier with adaptive norm,” Technical report tr-99-9, ETL, 1999.
4. Yuille A.L. and Rangarajan A., “The concave-convex procedure,” *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.
5. S. Boughorbel, *Kernels for Image Classification with SVM*, Ph.D. thesis, submited to University of Paris Sud, Orsay, 2005.
6. S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.
7. G. Ratsch, T. Onoda, and K. R. Muller, “Soft margins for adaboost,” *Machine Learning*, vol. 42, no. 3, pp. 287–320, 2001.