Least Squares Support Vector Machine Classifiers: a Large Scale Algorithm *

J.A.K. Suykens[1], **L. Lukas**[1], **P. Van Dooren**[2], **B. De Moor**[1], **J. Vandewalle**[1]
[1] K.U. Leuven, Dept. of Electrical Engineering ESAT-SISTA,
Kardinaal Mercierlaan 94, B-3001 Leuven, Belgium
Email: johan.suykens@esat.kuleuven.ac.be
[2] Universite Catholique de Louvain
Dept. of Mathematical Engineering, Batiment Euler
4, avenue Georges Lemaitre, B-1348 Louvain la Neuve, Belgium

Abstract

Support vector machines (SVM's) have been introduced in literature as a method for pattern recognition and function estimation, within the framework of statistical learning theory and structural risk minimization. A least squares version (LS-SVM) has been recently reported which expresses the training in terms of solving a set of linear equations instead of quadratic programming as for the standard SVM case. In this paper we present an iterative training algorithm for LS-SVM's which is based on a conjugate gradient method. This enables solving large scale classification problems which is illustrated on a multi two-spiral benchmark problem.
Keywords. Support vector machines, classification, neural networks, RBF kernels, conjugate gradient method.

1 Introduction

Support vector machines have been introduced in [16] for solving pattern recognition and nonlinear function estimation problems. In this method one maps the data into a higher dimensional input space in which one constructs an optimal separating hyperplane. As kernel functions one can use polynomials, splines, radial basis function networks and multilayer perceptrons. For the mapping into the higher dimensional input space and kernels one makes use of Mercer's condition. While classical neural network techniques suffer from the existence of many local minima [1, 3, 8, 19], SVM solutions are obtained from quadratic programming problems possessing a global solution. Kernel functions and parameters can be chosen such that a bound on the VC dimension is minimized [3, 16, 17, 18, 15]. Being based on the structural risk minimization principle and capacity concept with pure combinatorial definitions, the quality and complexity of the SVM solution does not depend directly on the dimensionality of the input space [16, 17, 18]. Links between SVM's, regularization theory and sparse approximations have been shown in [12, 6].

In the support vector method of function estimation one typically employs Vapnik's epsilon insensitive loss function or Huber's loss function. In [14] a least squares version of SVM's for classification has been proposed, which is related to the LS version for function estimation reported in [10]. In this LS-SVM version one finds the solution by solving a linear system instead of quadratic programming. This is due to the use of equality instead of inequality constraints in the problem formulation. In [2, 5, 13] such linear systems have been called augmented systems or Karush-Kuhn-Tucker (KKT) systems and their numerical stability has been investigated. In this paper we present an iterative solution to LS-SVM's based on the conjugate gradient method [7]. This method enables solving large scale classification problems. As an example we show the excellent performance on a multi two-spiral benchmark problem, which is known to be a difficult test case for neural network classifiers [9].

This paper is organized as follows. In Section 2 we discuss LS-SVM classifiers. In Section 3 we present an iterative method for training large scale LS-SVM's. In Section 4 an illustrative example is given on a multi two-spiral benchmark problem.

*This research work was carried out at the ESAT laboratory and the Interdisciplinary Center of Neural Networks ICNN of the Katholieke Universiteit Leuven, in the framework of the FWO project G.0262.97 *Learning and Optimization: an Interdisciplinary Approach*, the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture (IUAP P4-02 & IUAP P4-24), the Concerted Action Project MIPS (*Modelbased Information Processing Systems*) of the Flemish Community. Johan Suykens is a postdoctoral researcher with the National Fund for Scientific Research FWO - Flanders. Bart De Moor is a senior research associate with the FWO - Flanders.

2 Least Squares Support Vector Machines

Given a training set of N data points $\{y_k, x_k\}_{k=1}^N$, where $x_k \in \mathbb{R}^n$ is the k-th input pattern and $y_k \in \mathbb{R}$ is the k-th output pattern, the support vector method approach aims at constructing a classifier of the form:

$$y(x) = \text{sign}[\sum_{k=1}^N \alpha_k \, y_k \, \Psi(x, x_k) + b] \qquad (1)$$

where α_k are support values and b is a real constant. For $\Psi(\cdot,\cdot)$ one typically has the following choices: $\Psi(x, x_k) = x_k^T x$ (linear SVM); $\Psi(x, x_k) = (x_k^T x + 1)^d$ (polynomial SVM of degree d); $\Psi(x, x_k) = \exp\{-\|x - x_k\|_2^2/\sigma^2\}$ (RBF SVM); $\Psi(x, x_k) = \tanh[\kappa \, x_k^T x + \theta]$ (MLP SVM), where σ, κ and θ are constants.

For the case of two classes, one assumes

$$\begin{cases} w^T \varphi(x_k) + b \geq +1 & , \quad \text{if } y_k = +1 \\ w^T \varphi(x_k) + b \leq -1 & , \quad \text{if } y_k = -1 \end{cases} \qquad (2)$$

which is equivalent to

$$y_k[w^T \varphi(x_k) + b] \geq 1, \qquad k = 1, ..., N \qquad (3)$$

where $\varphi(\cdot)$ is a nonlinear function which maps the input space into a higher dimensional space. LS-SVM classifiers as introduced in [14] are obtained as solution to the following optimization problem:

$$\min_{w,b,e} \mathcal{J}_{LS}(w, b, e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{k=1}^N e_k^2 \qquad (4)$$

subject to the equality constraints

$$y_k \, [w^T \varphi(x_k) + b] = 1 - e_k, \; k = 1, ..., N. \qquad (5)$$

One defines the Lagrangian

$$\mathcal{L}(w, b, e; \alpha) = \mathcal{J}_{LS} - \sum_{k=1}^N \alpha_k \{y_k[w^T \varphi(x_k) + b] - 1 + e_k\} \qquad (6)$$

where α_k are Lagrange multipliers, which can be either positive or negative due to the equality constraints as follows from the Kuhn-Tucker conditions [4].

The conditions for optimality

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \quad \rightarrow \quad w = \sum_{k=1}^N \alpha_k y_k \varphi(x_k) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \quad \rightarrow \quad \sum_{k=1}^N \alpha_k y_k = 0 \\ \frac{\partial \mathcal{L}}{\partial e_k} = 0 \quad \rightarrow \quad \alpha_k = \gamma e_k \\ \frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \quad \rightarrow \quad y_k[w^T \varphi(x_k) + b] - 1 + e_k = 0 \end{cases} \qquad (7)$$

for $k = 1, ..., N$ can be written as the linear system [4]

$$\left[\begin{array}{ccc|c} I & 0 & 0 & -Z^T \\ 0 & 0 & 0 & -Y^T \\ 0 & 0 & \gamma I & -I \\ \hline Z & Y & I & 0 \end{array}\right] \left[\begin{array}{c} w \\ b \\ e \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ \hline \vec{1} \end{array}\right] \qquad (8)$$

where $Z = [\varphi(x_1)^T y_1; ...; \varphi(x_N)^T y_N]$, $Y = [y_1; ...; y_N]$, $\vec{1} = [1; ...; 1]$, $e = [e_1; ...; e_N]$, $\alpha = [\alpha_1; ...; \alpha_N]$. Elimination of w and e gives

$$\left[\begin{array}{c|c} 0 & Y^T \\ \hline Y & ZZ^T + \gamma^{-1} I \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \vec{1} \end{array}\right]. \qquad (9)$$

Mercer's condition is applied to the matrix $\Omega = ZZ^T$ with

$$\begin{aligned} \Omega_{kl} &= y_k y_l \, \varphi(x_k)^T \varphi(x_l) \\ &= y_k y_l \, \Psi(x_k, x_l). \end{aligned} \qquad (10)$$

The parameters of the kernels, such as σ for the RBF kernel, can be optimally chosen by optimizing an upper bound on the VC dimension, which involves solving a quadratic programming problem [3, 16, 17, 18]. The support values α_k are proportional to the errors at the data points in the LS-SVM case, while in the standard SVM case many support values are typically equal to zero. Hence one could rather speak of a support value spectrum in the LS-SVM case.

3 A Large Scale Algorithm for LS-SVM's

The matrix in (9) is of dimension $(N+1) \times (N+1)$. For large value values of N this matrix cannot be stored, such that an iterative solution method for solving (9) is needed. A Hestenes-Stiefel conjugate gradient algorithm for solving $\mathcal{A}x = \mathcal{B}$ with $\mathcal{A} \in \mathbb{R}^{n \times n}$ symmetric positive definite and $\mathcal{B} \in \mathbb{R}^n$ is given by (see [7] p.523):

Conjugate Gradient Method

$$\begin{array}{l} i = 0; x_0 = 0; r_0 = \mathcal{B}; \\ \text{while } r_i \neq 0 \\ \quad i = i + 1 \\ \quad \text{if } i = 1 \\ \quad\quad p_1 = r_0 \\ \quad \text{else} \\ \quad\quad \beta_i = r_{i-1}^T r_{i-1} / r_{i-2}^T r_{i-2} \\ \quad\quad p_i = r_{i-1} + \beta_i p_{i-1} \\ \quad \text{end} \\ \quad \lambda_i = r_{i-1}^T r_{i-1} / p_i^T \mathcal{A} p_i \\ \quad x_i = x_{i-1} + \lambda_i p_i \\ \quad r_i = r_{i-1} - \lambda_i \mathcal{A} p_i \\ \text{end} \\ x = x_i \end{array} \qquad (11)$$

A convergence property of this method is that if $\mathcal{A} = I + \mathcal{C}$ is symmetric positive definite and $\mathrm{rank}(\mathcal{C}) = r$ then the algorithm converges in at most $r+1$ steps [7].

The problem (9) is of the form

$$\begin{bmatrix} 0 & Y^T \\ Y & H \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} \tag{12}$$

with $H = \Omega + \gamma^{-1} I$, $\xi_1 = b$, $\xi_2 = \alpha$, $d_1 = 0$, $d_2 = \vec{1}$. The matrix in (12) on the other hand is not positive definite. Hence in this form it cannot be solved by (11). However, (9)(12) is equivalent to solving

$$\begin{bmatrix} s & 0 \\ 0 & H \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 + H^{-1} Y \xi_1 \end{bmatrix} = \begin{bmatrix} -d_1 + Y^T H^{-1} d_2 \\ d_2 \end{bmatrix} \tag{13}$$

with $s = Y^T H^{-1} Y > 0$ ($H = H^T > 0$).

Finally, (12) can be solved then as follows

LS-SVM - Large Scale Algorithm

1. Solve η, ν from $H\eta = Y$ and $H\nu = d_2$ using (11).
2. Compute $s = Y^T \eta$.
3. Find solution
   $b = \xi_1 = \eta^T d_2 / s$
   $\alpha = \xi_2 = \nu - \eta \xi_1$.

Note that in (11), $\mathcal{A}$ isn't stored in the case of this large scale algorithm. The computational complexity is $\mathcal{O}(Nr^2)$.

4 Example: a Multi Two-Spiral Problem

The two-spiral problem [9] is a well-known benchmark problem for testing the quality of neural network classifiers. In [14] the excellent training and generalization performance of LS-SVM's with RBF kernel on this problem has been shown. In order to illustrate the large scale LS-SVM version we define here a more complicated multi two-spiral classification problem as shown in Figure 1. Given are 1000 training data where the training data of the two classes are indicted by '*' and 'o'. A RBF kernel was used with $\sigma = 1$ and $\gamma = 10$. The resulting classifier (1) with support values α_k and bias term b obtained from the large scale algorithm is shown on Figure 1. Taking 1000 support values one has no misclassification on the training set, together with excellent generalization as is clear from the decision boundary between the black and white regions on Figure 1. The support value spectrum is shown on Figure 2, which are the obtained support values sorted from largest to smallest. Figure 3 shows the performance of the classifier on the training data in terms of the number of misclassified data as a function of the amount of most significant support values.

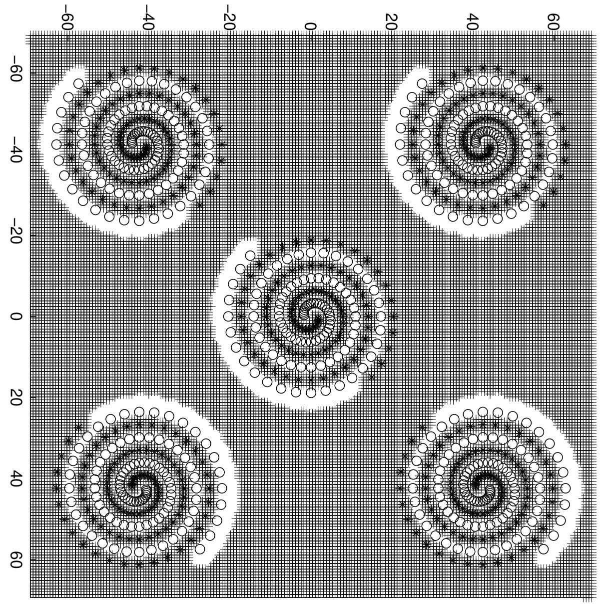

Figure 1: *Multi two-spiral classification problem with 1000 training data (data points of the two classes are indicated by '*' and 'o'). The excellent generalization performance of the LS-SVM with RBF kernel is clear by visual inspection from the black and white regions which determine the decision boundary between the two classes.*

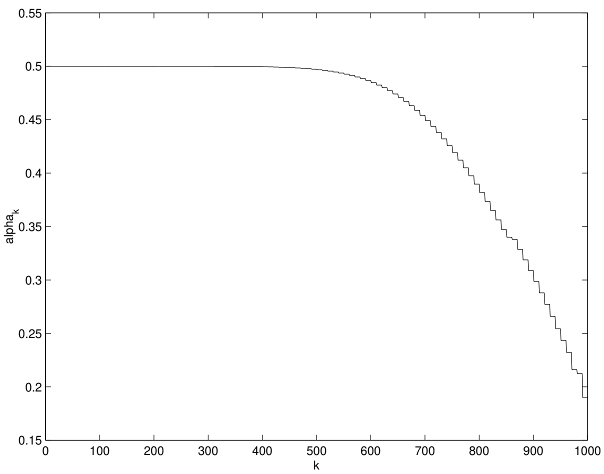

Figure 2: *Support Value Spectrum related to the Figure 1, with support values α_k sorted from largest to smallest values for the given training data set.*

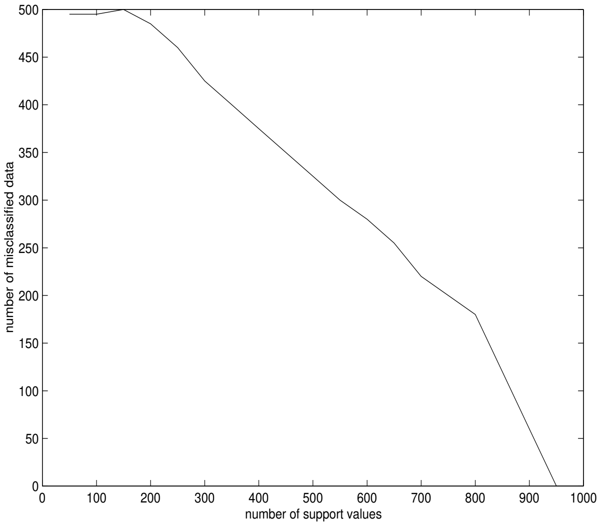

Figure 3: *Performance on the training set as a function of the number of the sorted support values, related to previous Figures.*

5 Conclusions

In this paper we proposed a large scale algorithm for least squares support vector machines based on a conjugate gradient method. The application to a difficult multi two-spiral classification problem shows that excellent generalization performance can be obtained using the LS-SVM approach in the separable case. This approach involves solving a linear system instead of quadratic programming for the standard SVM case and allows to apply Mercer's condition with use of several type of kernels functions. The performance of the classifier turns out to be quite robust as well with respect to tuning parameters of the algorithm.

References

[1] Bishop C.M., *Neural networks for pattern recognition*, Oxford University Press, 1995.

[2] Björk A., Paige C.C., "Solution of augmented linear systems using orthogonal factorizations," *BIT* 34, 1-24, 1994.

[3] Cherkassky V., Mulier F., *Learning from data: concepts, theory and methods*, John Wiley and Sons, 1998.

[4] Fletcher R., *Practical methods of optimization*, Chichester and New York: John Wiley and Sons, 1987.

[5] Fletcher R., Johnson T., "On the stability of null-space methods for KKT systems," *SIAM J. Matrix Anal. Appl.*, Vol.18, No.4, 938-958, 1997.

[6] Girosi F., "An equivalence between sparse approximation and support vector machines," *Neural Computation*, 10(6), 1455-1480, 1998.

[7] Golub G.H., Van Loan C.F., *Matrix Computations*, Baltimore MD: Johns Hopkins University Press, 1989.

[8] Haykin S., *Neural Networks: a Comprehensive Foundation*, Macmillan College Publishing Company: Englewood Cliffs, 1994.

[9] Ridella S., Rovetta S., Zunino R., "Circular backpropagation networks for classification," *IEEE Transactions on Neural Networks*, Vol.8, No.1, pp.84-97, 1997.

[10] Saunders C., Gammerman A., Vovk V., "Ridge Regression Learning Algorithm in Dual Variables," *Proc. of the 15th Int. Conf. on Machine Learning ICML-98*, Madison-Wisconsin, 1998.

[11] Schölkopf B., Sung K.-K., Burges C., Girosi F., Niyogi P., Poggio T., Vapnik V., "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *IEEE Transactions on Signal Processing,* Vol.45, No.11, pp.2758-2765, 1997.

[12] Smola A., Schölkopf B., Müller K.-R., "The connection between regularization operators and support vector kernels," *Neural Networks*, 11, 637-649, 1998.

[13] Sun J.-G., "Structured backwards errors for KKT systems," *Linear Algebra and its Applications*, 288, 75-88, 1999.

[14] Suykens J.A.K., Vandewalle J., "Least squares support vector machine classifiers," *Neural Processing Letters*, 1999, to appear.

[15] Suykens J.A.K., Vandewalle J., "Training multilayer perceptron classifiers based on a modified support vector method", *IEEE Transactions on Neural Networks*, 1999, to appear.

[16] Vapnik V., "The nature of statistical learning theory," Springer-Verlag, New-York, 1995.

[17] Vapnik V., "Statistical learning theory," John Wiley, New-York, 1998.

[18] Vapnik V., "The support vector method of function estimation," In *Nonlinear Modeling: advanced black-box techniques*, Suykens J.A.K., Vandewalle J. (Eds.), Kluwer Academic Publishers, Boston, pp.55-85, 1998.

[19] Zurada J.M., *Introduction to Artificial Neural Systems*, West Publishing Company, 1992.