Multiclass Least Squares Support Vector Machines

J.A.K. Suykens and J. Vandewalle

Katholieke Universiteit Leuven, Dept. of Electr. Eng., ESAT-SISTA
Kardinaal Mercierlaan 94, B-3001 Leuven (Heverlee), Belgium
Email: johan.suykens@esat.kuleuven.ac.be

Abstract

In this paper we present an extension of least squares support vector machines (LS-SVM's) to the multiclass case. While standard SVM solutions involve solving quadratic or linear programming problems, the least squares version of SVM's corresponds to solving a set of linear equations, due to equality instead of inequality constraints in the problem formulation. In LS-SVM's Mercer condition is still applicable. Hence several type of kernels such as polynomial, RBF's and MLP's can be used. The multiclass case that we discuss here is related to classical neural net approaches for classification where multi classes are encoded by considering multiple outputs for the network. Efficient methods for solving large scale LS-SVM's are available.

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 high 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, 2, 7], 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 [2, 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, 13, 5].

Many cost functions have been studied in the context of standard SVM methodology [18, 13]. 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 (LS) version of SVM's for classification has been proposed, which is related to the LS version for function estimation reported in [9]. 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 this paper we discuss the present an extension of LS-SVM's to the multiclass case. It is related to classical neural net approaches for classification where multi classes are encoded by considering multiple outputs for the network. The large scale method for LS-SVM's of [15] is also applicable to this multiclass case.

This paper is organized as follows. In Section 2 we review LS-SVM classifiers for two classes. In Section 3 we present multiclass LS-SVM's. In Section 4 we give an illustrative example.

2 Least Squares SVM's

In this Section we consider first the case of two classes. Given a training set of N data points $\{y_k, x_k\}_{k=1}^{N}$, where $x_k \in \mathbb{R}^n$ denotes the k-th input pattern and $y_k \in \mathbb{R}$ the k-th output pattern, the support vector method approach aims at constructing a classifier of the form:

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

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

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

which is equivalent to

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

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:

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

subject to the equality constraints

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

One defines the Lagrangian

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

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 [3].

The conditions for optimality

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

for $k=1,...,N$ can be written after elimination of w and e as the linear system [3, 4]:

$$\left[\begin{array}{c|c} 0 & Y^T \\ \hline Y & ZZ^T+\gamma^{-1}I \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right]=\left[\begin{array}{c} 0 \\ \hline \vec{1} \end{array}\right] \tag{8}$$

where $Z=[\varphi(x_1)^T y_1;...;\varphi(x_N)^T y_N]$, $Y=[y_1;...;y_N]$, $\vec{1}=[1;...;1]$, $e=[e_1;...;e_N]$, $\alpha=[\alpha_1;...;\alpha_N]$. Mercer's condition is applied to the matrix $\Omega=ZZ^T$ with

$$\begin{aligned} \Omega_{kl} &= y_k y_l\,\varphi(x_k)^T\varphi(x_l) \\ &= y_k y_l\,\Psi(x_k,x_l). \end{aligned} \tag{9}$$

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 [2, 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.

3 Multiclass LS-SVM Classifiers

For the multiclass case we consider given training data $\{y_k^{(i)},x_k\}_{k=1,i=1}^{k=N,i=m}$. We make use now of additional outputs in order to encode multi classes, where $y_k^{(i)}$ denotes the output of the ith output unit for pattern k. This is one possible approach which is very similar to classical neural network methodology. The m outputs can in principle encode 2^m different classes (we don't discuss the issue of optimal coding in this paper).

The derivation of the multiclass LS-SVM is based upon the formulation

$$\min_{w_i,b_i,e_{k,i}} \mathcal{J}_{LS}^{(m)}(w_i,b_i,e_{k,i})=\frac{1}{2}\sum_{i=1}^{m}w_i^T w_i+\gamma\,\frac{1}{2}\sum_{k=1}^{N}\sum_{i=1}^{m}e_{k,i}^2 \tag{10}$$

subject to the equality constraints

$$\begin{cases} y_k^{(1)}\,[w_1^T\varphi_1(x_k)+b_1]=1-e_{k,1},\; k=1,...,N \\ y_k^{(2)}\,[w_2^T\varphi_2(x_k)+b_2]=1-e_{k,2},\; k=1,...,N \\ \ldots \\ y_k^{(m)}\,[w_m^T\varphi_m(x_k)+b_m]=1-e_{k,m},\; k=1,...,N. \end{cases} \tag{11}$$

One defines the Lagrangian

$$\begin{aligned} &\mathcal{L}^{(m)}(w_i,b_i,e_{k,i};\alpha_{k,i})=\mathcal{J}_{LS}^{(m)}- \\ &\textstyle\sum_{k,i}\alpha_{k,i}\{y_k^{(i)}[w_i^T\varphi_i(x_k)+b_i]-1+e_{k,i}\} \end{aligned} \tag{12}$$

which gives as conditions for optimality:

$$\begin{cases} \frac{\partial\mathcal{L}}{\partial w_i}=0 & \rightarrow \quad w_i=\sum_{k=1}^{N}\alpha_{k,i}y_k^{(i)}\varphi_i(x_k) \\ \frac{\partial\mathcal{L}}{\partial b_i}=0 & \rightarrow \quad \sum_{k=1}^{N}\alpha_{k,i}y_k^{(i)}=0 \\ \frac{\partial\mathcal{L}}{\partial e_{k,i}}=0 & \rightarrow \quad \alpha_{k,i}=\gamma e_{k,i} \\ \frac{\partial\mathcal{L}}{\partial \alpha_{k,i}}=0 & \rightarrow \quad y_k^{(i)}[w_i^T\varphi_i(x_k)+b_i]=1-e_{k,i} \end{cases} \tag{13}$$

for $k=1,...,N$ and $i=1,...,m$. Elimination of w_i and $e_{k,i}$ gives the linear system:

$$\left[\begin{array}{c|c} 0 & Y_M^T \\ \hline Y_M & \Omega_M \end{array}\right]\left[\begin{array}{c} b_M \\ \hline \alpha_M \end{array}\right]=\left[\begin{array}{c} 0 \\ \hline \vec{1} \end{array}\right] \tag{14}$$

with given matrices

$$Y_M = \text{blockdiag}\{ \begin{bmatrix} y_1^{(1)} \\ \vdots \\ y_N^{(1)} \end{bmatrix}, ..., \begin{bmatrix} y_1^{(m)} \\ \vdots \\ y_N^{(m)} \end{bmatrix} \}$$

$$\Omega_M = \text{blockdiag}\{\Omega^{(1)}, ..., \Omega^{(m)}\}$$

$$\Omega_{kl}^{(i)} = y_k^{(i)} y_l^{(i)} \Psi_i(x_k, x_l) + \gamma^{-1} I$$

and solution vector

$$b_M = [b_1; ...; b_m]$$

$$\alpha_M = [\alpha_{1,1}; ...; \alpha_{N,1}; ....; \alpha_{1,m}; ...; \alpha_{N,m}].$$

Mercer condition is again applied here where one has for the RBF kernel case:

$$\Psi_i(x_k, x_l) = \varphi_i^T(x_k)\varphi_i(x_l) = \exp\{-\|x_k - x_l\|_2^2/\sigma_i^2\} \tag{15}$$

for $i = 1, ..., m$. The large scale algorithm proposed in [15] for the two-class case, which is based on a conjugate gradient method, is also applicable to (14) up to some minor modifications which exploit the block structure of the matrices Ω_M, Y_M. Storage of the matrix Ω_M is avoided in this algorithm and (14) is reformulated such that two linear subsystems involving positive definite matrices can be solved. The latter is needed in order to apply iterative methods for linear systems [6].

4 Example

In [14, 15] some examples are given on two-spiral and multi spiral problems which are known to be hard for neural classifiers. The LS-SVM's show excellent generalization performance on these data sets.

Here we present a small illustrative example on a simple spiral problem for which four classes have been defined. Figure 1 shows $N = 60$ training data points (indicated by 'o') with 4 classes that each contain 15 data points (outputs equal to $[+1; +1], [+1; -1], [-1; +1], [-1; -1]$). The four classes have been encoded by taking $m = 2$. A RBF kernel has been taken with $\sigma_1^2 = \sigma_2^2 = 0.1$ and $\gamma = 1$. Figure 1 shows the resulting LS-SVM classifier by solving (14). Figure 2 shows the support values $\alpha_{k,1}$ and $\alpha_{k,2}$ (sorted) which illustrates that sparsity is lost compared to standard SVM's due to the least squares formulation with equality constraints. The generalization performance of the LS-SVM classifier is robust with respect to the choice of σ, γ.

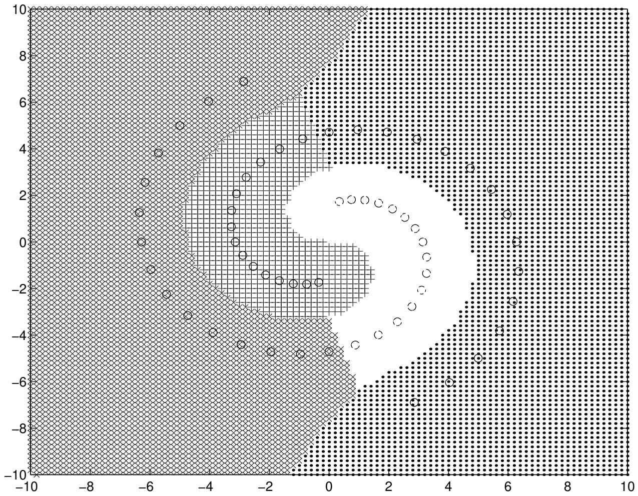

Figure 1: *Illustrative example of a LS-SVM with RBF kernel on a four class classification problem. The classifier is obtained by solving a set of linear equations. It shows a good generalization performance with respect to the choice of* σ, γ.

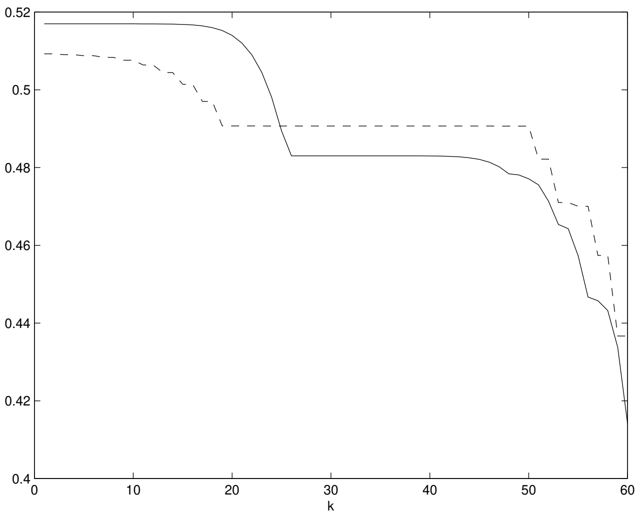

Figure 2: *Support values (sorted) related to Figure 1 for the first output (solid line) and the second output (dashed line). Sparsity is lost for LS-SVM's compared to standard SVM's.*

5 Conclusions

We presented an extension of least squares support vector machines to the multiclass case. This has been done with encoding the classes by defining additional outputs. As for the binary case the resulting multiclass LS-SVM classifier is obtained by solving a set of linear equations. This is due to equality constraints in the problem formulation. On the other hand sparsity is lost in the least squares case which corresponds to a form of ridge regression. Mercer's condition is still applicable such that several possible kernel functions can be employed as in standard SVM methodology. Iterative methods for solving large scale problems are available and simulation results suggest that LS-SVM's are robust with respect to the choice of the smoothing parameter and the choice of the regularization parameter. For many real life applications it may offer a fast method for obtaining classifiers with good generalization performance.

Acknowledgements
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) and 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.

References

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

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

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

[4] 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.

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

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

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

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

[9] 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.

[10] 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.

[11] Schölkopf B., *Support Vector Learning*, PhD Thesis, published by: R. Oldenbourg Verlag, Munich, 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] Smola A., *Learning with Kernels*, PhD Thesis, published by: GMD, Birlinghoven, 1999.

[14] Suykens J.A.K., Vandewalle J., "Least squares support vector machine classifiers," *Neural Processing Letters*, Vol.9, No.3, 1999.

[15] Suykens J.A.K., Lukas L., Van Dooren P., De Moor B., Vandewalle J., "Least squares support vector machine classifiers: a large scale algorithm," *ECCTD'99 European Conf. on Circuit Theory and Design*, August 1999.

[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.