Sparse Least Squares Support Vector Machine Classifiers

J.A.K. Suykens, L. Lukas and J. Vandewalle

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

Abstract. In least squares support vector machine (LS-SVM) classifiers the original SVM formulation of Vapnik is modified by considering equality constraints within a form of ridge regression instead of inequality constraints. As a result the solution follows from solving a set of linear equations instead of a quadratic programming problem. However, a drawback is that sparseness is lost in the LS-SVM case due to the choice of 2-norms. In this paper we propose a method for imposing sparseness to the LS-SVM solution. This is done by pruning the support value spectrum which is revealing the relative importance of the training data points and is immediately available as solution to the linear systems.

Keywords. Support vector machines, classification, ridge regression, dual problem, sparse approximation, pruning.

1. Introduction

Support vector machines (SVM's) have been successfully applied in classification and function estimation problems [2, 8] after its introduction by Vapnik within the context of statistical learning theory and structural risk minimization [13]. The SVM classifier typically follows from the solution to a quadratic programming (QP) problem. Several types of kernels can be used within SVM's such as linear, polynomial, splines, radial basis functions (RBF) and one hidden layer multilayer perceptrons (MLP). The kernel based SVM representation is motivated by the Mercer condition. Normally, many of the support values which are the solution to the QP problem will be equal to zero. The non-zero values are related to support vector data and are contributing to the construction of the classifier.

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.

A modified version of SVM classifiers in a least squares sense has been proposed in [10]. In this case the solution is given by a linear system instead of a QP problem. Taking into account the fact that the computational complexity strongly increases with the number of training data least squares support vector machines (LS-SVM's) can be efficiently estimated using iterative methods [4, 11]. A straightforward extension of LS-SVM's to the multiclass problem has been made in [12]. Related work on ridge regression type SVM's is [7] (but without considering a bias term, which has serious implications concerning algorithms) [2].

A drawback of LS-SVM's on the other hand is that sparseness is lost due to the form of ridge regression. This is important in the context of an equivalence between sparse approximation and support vector machines [3]. Now, in this paper we demonstrate how sparseness can be imposed by pruning the support value spectrum. The sorted support values are indeed available as solution to the linear system. The support values reveal the relative importance of each of the training data points. In the case of RBF kernels a small support value indicates that this point can be omitted from the training set which is equivalent then to removing the hidden unit which corresponds to this data point. While pruning of classical neural networks involves the computation of an inverse Hessian matrix [1, 5, 6], the LS-SVM pruning can be done immediately based upon the support value spectrum. The pruning method could be potentially improved based upon the insights of [9].

This paper is organized as follows. In Section 2 we discuss LS-SVM's. In Section 3 we present the pruning method in order to impose sparseness. In Section 4 an illustrative example is given.

2. Least Squares SVM Classifiers

Given a training set $\{x_k, y_k\}_{k=1}^N$ with input patterns $x_k \in \mathbb{R}^n$ and output values $y_k \in \{-1, +1\}$ indicating the class, SVM formulations [13] start from the assumption that

$$\left\{ \begin{array}{lcl} w^T\varphi(x_k) + b & \geq & +1\,, \text{ if } y_k = +1 \\ w^T\varphi(x_k) + b & \leq & -1\,, \text{ if } y_k = -1 \end{array} \right. \tag{1}$$

which is equivalent to $y_k[w^T\varphi(x_k) + b] \geq 1 \, (k = 1, ..., N)$. Here the nonlinear mapping $\varphi(\cdot)$ maps the input data into a so-called higher dimensional feature space. In LS-SVM's [10] an equality constraint based formulation is made within the context of ridge regression [4] as follows

$$\min_{w,e} \mathcal{J}(w, e) = \frac{1}{2}w^T w + \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 \;\; \text{s.t} \;\; y_k[w^T\varphi(x_k) + b] = 1 - e_k\,,\; k = 1, N \tag{2}$$

with Lagrangian

$$\mathcal{L}(w, b, e; \alpha) = \mathcal{J}(w, e) - \sum_{k=1}^{N} \alpha_k \{y_k[w^T\varphi(x_k) + b] - 1 + e_k\} \tag{3}$$

and Lagrange multipliers (support values) α_k. The conditions for optimality $\partial\mathcal{L}/\partial w = 0$, $\partial\mathcal{L}/\partial b = 0$, $\partial\mathcal{L}/\partial e_k = 0$, $\partial\mathcal{L}/\partial \alpha_k = 0$ give $w = \sum_{k=1}^{N} \alpha_k y_k \varphi(x_k)$, $\sum_{k=1}^{N} \alpha_k y_k = 0$, $\alpha_k = \gamma e_k$, $y_k[w^T \varphi(x_k) + b] = 1 - e_k (k = 1, N)$, respectively. By eliminating e, w one obtains the KKT system

$$\left[\begin{array}{c|c} 0 & Y^T \\ \hline Y & \Omega + \gamma^{-1} I \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \vec{1} \end{array}\right] \tag{4}$$

where $Y = [y_1; ...; y_N]$, $\vec{1} = [1; ...; 1]$, $\alpha = [\alpha_1; ...; \alpha_N]$ and

$$\begin{aligned} \Omega_{kl} &= y_k y_l \, \varphi(x_k)^T \varphi(x_l), \qquad k, l = 1, ..., N \\ &= y_k y_l \, \Psi(x_k, x_l) \end{aligned} \tag{5}$$

after application of the Mercer condition. This finally results into the following LS-SVM classifier

$$y(x) = \mathrm{sign}[\sum_{k=1}^{N} \alpha_k y_k \, \Psi(x, x_k) + b] \tag{6}$$

where α, b are the solution to (4). For the choice of the kernel function $\Psi(\cdot, \cdot)$ one has several possibilities including the RBF kernel $\Psi(x, x_k) = \exp\{-\|x - x_k\|_2^2/\sigma^2\}$. Note that σ, γ are to be considered as additional tuning parameters for the LS-SVM which do not follow as a solution to the linear system.

3. Imposing Sparseness by Pruning

A drawback of the LS-SVM classifier in comparison with the original SVM formulation is that sparseness is lost in the LS case. This immediately follows from the choice of the 2-norm and is also revealed by the fact that the support values are proportional to the errors at the data points, namely $\alpha_k = \gamma e_k$. However, by plotting the spectrum of the sorted $|\alpha_k|$ values one can evaluate which data are most significant for contribution to the LS-SVM classifier. Sparseness is imposed then by gradually omitting the least important data from the training set and re-estimating the LS-SVM (Fig.1):

1. Train LS-SVM based on N points.
2. Remove a small amount of points (e.g. 5% of the set) with smallest values in the sorted $|\alpha_k|$ spectrum.
3. Re-train the LS-SVM based on the reduced training set.
4. Go to 2, unless the user-defined performance index degrades.

This procedure corresponds to pruning of the LS-SVM. An advantage in comparison with pruning for classical neural network techniques [1, 6, 5] is that now the pruning does not involve a computation of a Hessian matrix but is immediately done based upon the physical meaning of the solution vector α. In this paper we do not discuss the important issue of selecting the value of γ and σ [13, ?, 14, 8] for RBF kernels in the context of this pruning procedure.

4. Example

We give an illustrative example for the LS-SVM pruning procedure. Training data ($N = 500$) are generated from Gaussian distributions (250 points for each class) with centra $\mu_1 = [-0.5; -0.5]$, $\mu_2 = [0.5; 0.5]$ (small overlap) (Fig.2) and centra $\mu_1 = [-0.3; -0.3]$, $\mu_2 = [0.3; 0.3]$ (larger overlap) (Fig.4) with covariance matrices $\Sigma_1 = \Sigma_2 = 0.25I$ in both cases. Fig.3 is similar to the case of Fig.2 but with a modification of 10 misclassifications in the data. Assuming equal prior probabilities for the two classes the optimal decision boundary in the sense of Bayes rule is given by a straight line [1] shown on the Figs. In all simulations we employ a RBF kernel with $\sigma = 3$; $\gamma = 10$ in Fig.2; $\gamma = 1$ in Fig.3-4 (more emphasis on regularization term $\|w\|$). The sorted spectrum is gradually pruned by leaving out 5% of the training data which are least significant according to the SV spectrum. The number of hidden units can be reduced from 500 to at least 100 without loss of performance in Fig.2-3. Snapshots of the shifted SV spectrum are shown for 250 and 100 SV's (only support vectors are shown on Figs). One observes that SV's are both near and far from the decision line which is different from standard SVM classifiers. For the case of a larger overlap (Fig.4) between the distributions, less SV's can be pruned.

5. Conclusions

We proposed a pruning method for achieving sparse least squares SVM classifiers. Pruning is done based upon the support value spectrum. Examples for RBF kernels illustrate how a significant amount of hidden units, i.e. support vectors can be reduced without loss of performance in the case of small and large overlap of the underlying distributions and misclassified data.

References

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

[2] Cristianini N., Shawe-Taylor J., *An Introduction to Support Vector Machines and Other Kernel Based Learning Methods*, Cambridge University Press, 2000.

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

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

[5] Hassibi B., Stork D.G., "Second order derivatives for network pruning: optimal brain surgeon," In Hanson, Cowan, Giles (Eds.) *Advances in Neural Information Processing Systems*, Vol.5, pp.164-171, San Mateo, CA: Morgan Kaufmann, 1993.

[6] Le Cun Y., Denker J.S., Solla S.A., "Optimal brain damage," In Touretzky (Ed.) *Advances in Neural Information Processing Systems*, Vol.2, pp.598-605, San Mateo, CA: Morgan Kaufmann, 1990.

[7] Saunders C., Gammerman A., Vovk V., "Ridge regression learning algorithm in dual variables," In J. Shavlik (Ed.) *Machine Learning: Proceedings of the Fifteenth International Conference*. Morgan Kaufmann, 1998.

[8] Schölkopf B., Burges C., Smola A. (Eds.), *Advances in Kernel Methods - Support Vector Learning*, MIT Press, 1998.

[9] Schölkopf B., Mika S., Burges C., Knirsch P., Müller K.R., Rätsch G., Smola A., "Input Space vs. Feature Space in Kernel-Based Methods," *IEEE Transactions on Neural Networks*, Vol.5, No.10, pp.1000-1017, 1999.

[10] Suykens J.A.K., Vandewalle J., "Least squares support vector machine classifiers," *Neural Processing Letters*, Vol.9, No.3, pp.293-300, June 1999.

[11] Suykens J.A.K., Lukas L., Van Dooren P., De Moor B., Vandewalle J., "Least squares support vector machine classifiers: a large scale algorithm," *European Conference on Circuit Theory and Design, ECCTD'99*, pp.839-842, Stresa Italy, August 1999.

[12] Suykens J.A.K., Vandewalle J., "Multiclass Least Squares Support Vector Machines," *Int. Joint Conference on Neural Networks IJCNN'99*, Washington DC, July 1999.

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

[14] 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.

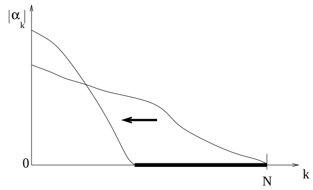

Fig.1: *Pruning of the LS-SVM spectrum.*

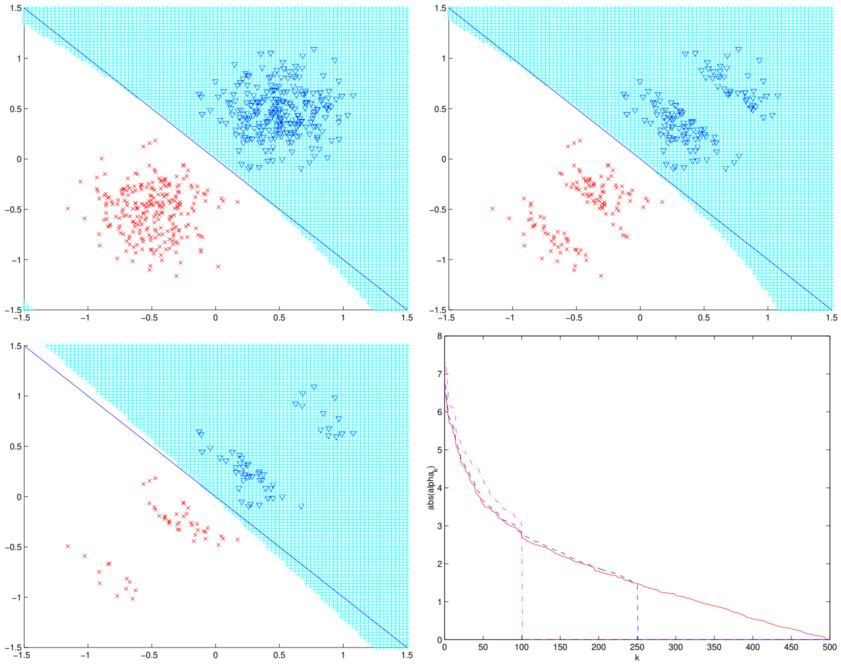

Fig.2: Pruning of LS-SVM classifier with RBF kernel: separable case. (Top-Left) $N = 500$ data; (Top-Right) 250 SV's; (Bottom-Left) 100 SV's; (Bottom-Right) Pruned SV spectrum.

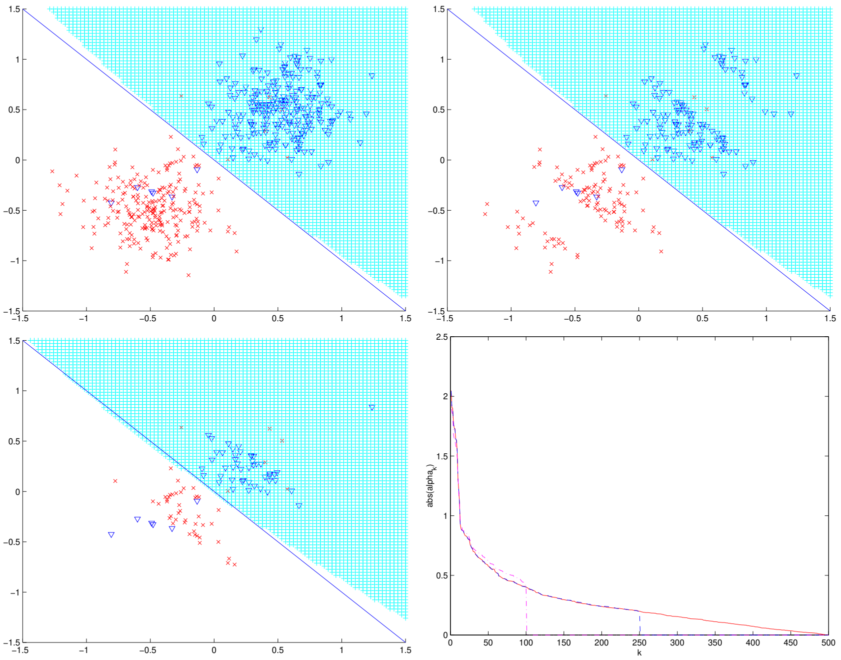

Fig.3: Pruning of LS-SVM classifier: data set with misclassifications

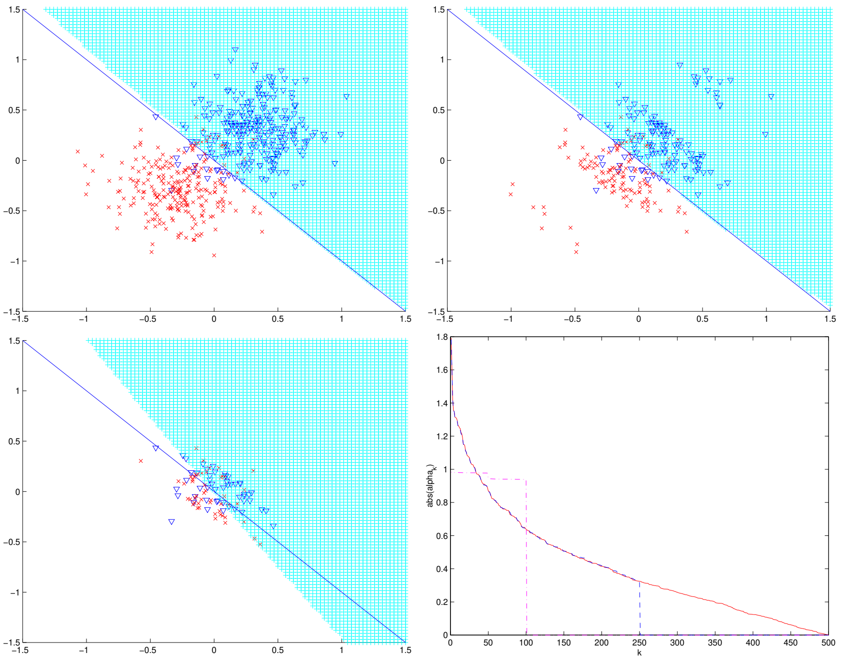

Fig.4: Pruning of LS-SVM classifier: case of a larger overlap between the underlying Gaussian distributions.