Improved Sparse Least-Squares Support Vector Machines

Gavin C. Cawley [a,*,1] and Nicola L. C. Talbot

[a] *School of Information Systems, University of East Anglia, Norwich, United Kingdom, NR4 7TJ*

Abstract

Suykens *et al.* [1] describe a weighted least-squares formulation of the support vector machine for regression problems and presents a simple algorithm for sparse approximation of the typically fully dense kernel expansions obtained using this method. In this paper, we present an improved method for achieving sparsity in least-squares support vector machines, which takes into account the residuals for all training patterns, rather than only those incorporated in the sparse kernel expansion. The superiority of this algorithm is demonstrated on the motorcycle and Boston housing datasets.

Key words:

Support Vector Machines, Sparse Approximation, Kernel Ridge Regression

* Corresponding author, email : gcc@sys.uea.ac.uk

[1] This work was supported by Royal Society Research Grant RSRG-22270.

1 Least-Squares Support Vector Machines

The least-squares support vector machine (LS-SVM), given training data, $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{\ell}$, $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$, $y_i \in \mathcal{Y} \subset \mathbb{R}$, constructs a linear regression model, $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b$, in a high dimensional feature space, $\mathcal{F}(\boldsymbol{\phi} : \mathcal{X} \to \mathcal{F})$, induced by a kernel function defining the inner product $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}')$. A commonly used kernel is the Gaussian radial basis function, $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \exp\{-\sigma^{-2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2\}$. The optimal values for the weight vector, $\boldsymbol{w}$, and bias, b, are given by the minimum of an objective function,

$$W(\boldsymbol{w}, b) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \gamma \sum_{i=1}^{\ell} (y_i - \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) - b)^2, \tag{1}$$

implementing a quadratic regularisation of a sum-of-squares empirical risk. The representer theorem [2] states that the solution of this problem can be written as an expansion in terms of training patterns,

$$f(\boldsymbol{x}) = \sum_{i=1}^{\ell} \alpha_i \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b.$$

Suykens *et al.* [1] show that the optimal coefficients of this expansion, $(\boldsymbol{\alpha}, b)^T$, are given by the solution of a system of linear equations,

$$\begin{bmatrix} \boldsymbol{\Omega} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ 0 \end{bmatrix}, \tag{2}$$

where $\boldsymbol{\Omega} = (\boldsymbol{K} + \gamma^{-1}\boldsymbol{I})$, $\boldsymbol{K} = \{k_{ij} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{i,j=1}^{\ell}$, $\mathbf{1} = (1, 1, \ldots, 1)^T$, $\boldsymbol{y} = (y_i, y_2, \ldots, y_\ell)^T$ and $\boldsymbol{\alpha} = (\alpha_i, \alpha_2, \ldots, \alpha_\ell)^T$.

The kernel expansions describing least-square support vector machines are typically fully dense, i.e. $\alpha_i \neq 0, \forall\ i \in \{1, 2, \ldots, \ell\}$. Suykens *et al.* [3,1] advocate use of the following algorithm to obtain a sparse approximation: A

LS-SVM is trained on the entire dataset, yielding a vector of coefficients, $\boldsymbol{\alpha}$. A small fraction of the data (say 5%), associated with coefficients having the smallest magnitudes, are discarded and the LS-SVM retrained on the remaining data. This process is repeated until a sufficiently small kernel expansion is obtained. It is observed in [1] that model selection should be performed at each iteration to find new values for the regularisation parameter, γ and any kernel parameters, such that optimal generalisation is achieved.

2 Improved Sparsification

In this section we present two modifications to the least-squares support vector machine. The first modification is based on the observation that changing the number of training patterns alters the balance between the sum-of-square empirical risk and the quadratic regulariser in the objective function (1). This implies that model selection based on cross-validation schemes will generally lead to under-regularised models. This can easily be remedied by quadratic regularisation of a mean-squared-error empirical risk,

$$W(\boldsymbol{w}, b) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\gamma}{\ell}\sum_{i=1}^{\ell}(y_i - \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_i) - b)^2, \tag{3}$$

The optimal expansion can be found by solving the system of linear equations (2) substituting $\boldsymbol{\Omega} = (\boldsymbol{K} + \ell\gamma^{-1}\boldsymbol{I})$.

The generalisation performance of sparse approximations of least-squares support vector machines can be further improved by including the residuals of patterns not used in the kernel expansion within the objective function. The weight vector, $\boldsymbol{w}$, is then represented as a weighted sum of selected training patterns, $\boldsymbol{w} = \sum_{i\in\mathcal{S}}\beta_i\boldsymbol{\phi}(\boldsymbol{x}_i)$, where $\mathcal{S} \subset \{1, 2, \ldots \ell\}$ is the set of indices of

training patterns used to form the kernel expansion. The objective function (3) can then be written as

$$W(\boldsymbol{\beta}, b) = \frac{1}{2} \sum_{i,j \in \mathcal{S}} \beta_i \beta_j k_{ij} + \frac{\gamma}{\ell} \sum_{i=1}^{\ell} (y_i - \sum_{j \in \mathcal{S}} \beta_j k_{ij} - b)^2.$$

Setting the partial derivatives with respect to $\boldsymbol{\beta}$ and b to zero, and dividing through by $2\gamma/\ell$, yields:

$$\sum_{i \in \mathcal{S}} \beta_i \sum_{j=1}^{\ell} k_{ij} + \ell b = \sum_{j=1}^{\ell} y_j$$

and

$$\sum_{i \in \mathcal{S}} \beta_i \left(\frac{\ell}{2\gamma} k_{ir} + \sum_{j=1}^{\ell} k_{jr} k_{ji} \right) + b \sum_{i=1}^{\ell} k_{ir} = \sum_{i=1}^{\ell} y_i k_{ir}, \qquad \forall\ r \in \mathcal{S}$$

These equations can be expressed as a system of $|\mathcal{S}| + 1$ linear equations in $|\mathcal{S}| + 1$ unknowns,

$$\begin{bmatrix} \boldsymbol{\Omega} & \boldsymbol{\Phi} \\ \boldsymbol{\Phi}^T & \ell \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ b \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ \sum_{k=1}^{\ell} y_k \end{bmatrix},$$

where $\boldsymbol{\Omega} = \{\omega_{ij}\}_{i,j \in \mathcal{S}}$, $\omega_{ij} = \frac{\ell}{2\gamma} k_{ij} + \sum_{r=1}^{\ell} k_{rj} k_{ri}$, $\boldsymbol{\Phi} = (\Phi_i)_{i=1}^{|\mathcal{S}|}$, $\Phi_i = \sum_{j=1}^{\ell} k_{ij}$, and $\boldsymbol{c} = (c_i)_{i=1}^{|\mathcal{S}|}$, $c_i = \sum_{j=1}^{\ell} y_j k_{ij}$ (for notational convenience, we assume that the training data are re-ordered such that $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$).

2.1 Computational Complexity

The first modification to the least-squares support vector machine proposed in the previous section (scaling of the regularisation parameter) leaves the computational complexity unchanged as it is clearly a straightforward reparameterisation of the standard algorithm. Each iteration of Suykens' sparsification algorithm requires the solution of a system of $n + 1$ linear equations, where n is the number of training patterns retained in the kernel expansion. The

computational complexity of each iteration is therefore $\mathcal{O}(n^3)$. The complexity of reducing the kernel expansion from ℓ to n terms is then $\mathcal{O}(\ell^4)$ as we must begin by solving a system of $\ell+1$ linear equations. Each iteration of the improved sparsification procedure, again involves the solution of a system of $n+1$ linear equations, however in this case the computational complexity is dominated by the construction of the matrix $\mathbf{\Omega}$, with a complexity of $\mathcal{O}(\ell n^2)$, however the complexity of reducing the kernel expansion from ℓ to n terms remains $\mathcal{O}(\ell^4)$. The improved sparsification algorithm does not require further model selection to achieve satisfactory performance, and so it is significantly faster in practice.

3 Results

In this section, the proposed improved sparse least-squares support vector machine (implementing both of the modifications suggested in section 2) is evaluated over two well-known datasets, the motorcycle, and Boston housing benchmarks. In each case the improved training algorithm is compared with the conventional sparse least-squares support vector machine [1], including model selection based on minimisation of a 4-fold cross-validation estimate of the RMS error during each iteration of the sparsification process.

The motorcycle dataset consists of a sequence of accelerometer readings through time following a simulated motor-cycle crash during an experiment to determine the efficacy of crash-helmets (Silverman [4]). Figure 1 displays the 10-fold cross-validation error of conventional and improved sparse least-squares support vector machines for the motorcycle dataset as a function of the number of training patterns included in the kernel expansion. The RMS error for the im-

proved sparse least squares support vector machine is consistently lower than that of the conventional approach *without* recourse to further model selection during each iteration of the sparsification process.

It is interesting to note that the RMS error for the improved sparse least-squares support vector machine is almost constant until all but 9 feature vectors have been eliminated (Figure 1). This occurrs because relatively few training vectors, $\{\boldsymbol{\Phi}(\boldsymbol{x}_i)\}_{i\in\mathcal{S}}$, $\mathcal{S} \subset \{1, 2, \ldots \ell\}$, are required to form an approximate basis for the training data in the feature space $\mathcal{F}$ [5], i.e. the training data in $\mathcal{F}$ can be expressed as a linear combination of the basis vectors,

$$\boldsymbol{\Phi}(\boldsymbol{x}_i) = \sum_{j\in\mathcal{S}} \lambda_j \boldsymbol{\Phi}(\boldsymbol{x}_j), \quad \forall i \in \{1, 2, \ldots, \ell\},$$

where $\mathcal{S}$ is the set of indices corresponding to basis vectors. The representer theorem states that the weight vector $\boldsymbol{w}$ lies in the span of the training data, and so can also be written as a linear combination of these basis vectors. Provided that the set of remaining feature vectors forms a basis for the entire data set, the weight vector found by the improved sparse least-squares support vector machine coincides with that given by the fully dense least-squares support vector machine. Furthermore any basis vector that is orthogonal to the optimal weight vector can also be safely deleted without sacrificing performance. Note that some kernels, including the radial basis function kernel used in this example, are of full rank, and so a sparse basis can not generally be found [6], however in some cases, such as this, a very close approximation is possible.

The Boston housing dataset describes the relationship between the median value of owner occupied homes in the suburbs of Boston and thirteen attributes representing environmental and social factors believed to be relevant

[7]. Figure 2 displays the 10-fold cross-validation error of conventional and improved sparse least-squares support vector machines as a function of the number of training patterns included in the kernel expansion. Again the error for the improved sparse least squares support vector machine is consistently lower than that of the conventional approach. In this case a sparse basis cannot be found and so the RMS error for the improved sparse least-squares support vector machine degrades progressively as feature vectors are eliminated.

4 Conclusions

In this paper we have presented an improved training algorithm for sparse least-squares support vector machines, taking into account the residuals for all training patterns, not just those appearing in the sparse kernel expansion. The use of a scaled regularisation parameter, eliminating the dependence on the size of the training set, is also proposed. The improved method demonstrates superior generalisation over the existing pruning algorithm and also eliminates the need for further model selection and is therefore significantly faster. The method can be easily extended to the weighted least-squares support vector machine.

5 Acknowledgements

The authors would like to thank Johan Suykens for introducing them to the least-squares support vector machine and providing pre-prints of his unpublished work. The authors also thank Rob Foxall and the anonymous reviewers for their helpful comments on previous drafts of this paper.

References

[1] J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle. Weighted least squares support vector machines : robustness and sparse approximation. *Neurocomputing (*in press*)*, 2002.

[2] G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

[3] J. A. K. Suykens, L. Lukas, and J. Vandewalle. Sparse least squares support vector machine classifiers. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN-2000)*, pages 37–42, Bruges, Belgium, 26–28 April 2000.

[4] B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society, B*, 47(1):1–52, 1985.

[5] G. Baudat and F. Anouar. Kernel-based methods and function approximation. In *Proceedings, International Joint Conference on Neural Networks*, volume 3, pages 1244–1249, Washington, DC, July 2001.

[6] C. A. Micchelli. Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

[7] D. Harrison and D. L. Rubinfeld. Hedonic prices and the demand for clean air. *Journal Environmental Economics and Management*, 5:81–102, 1978.

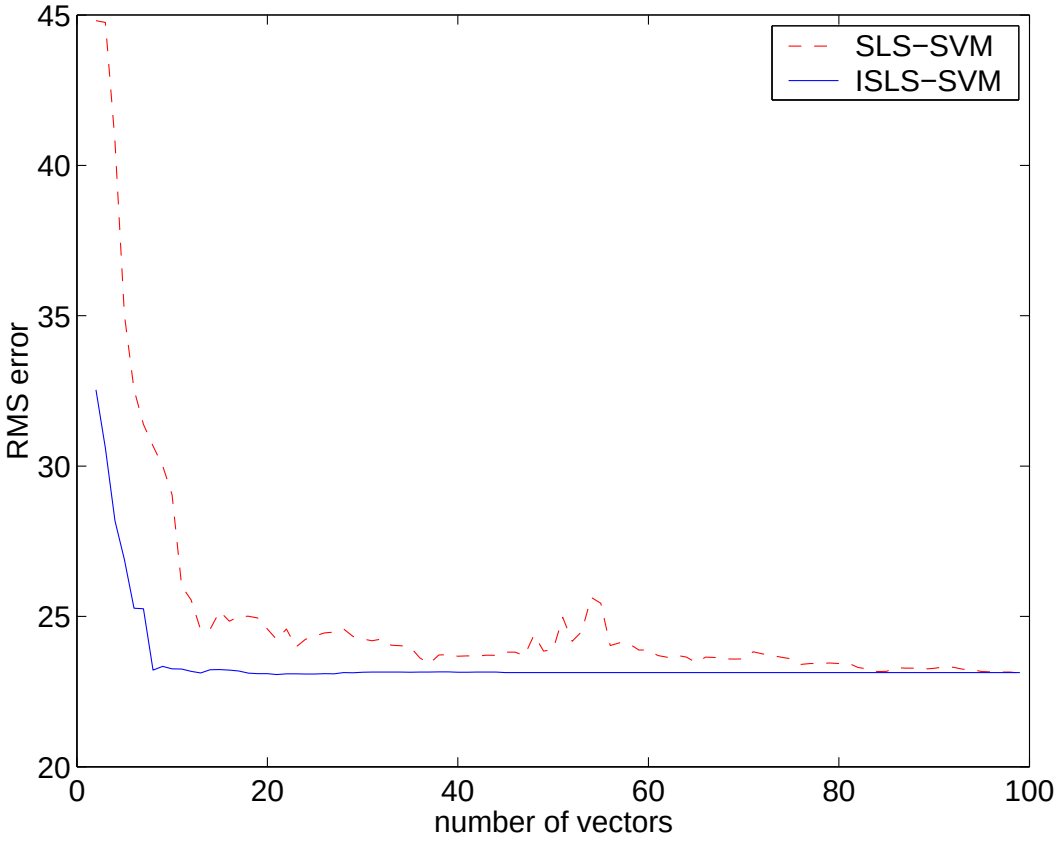

Fig. 1. Cross-validation error of conventional (SLS-SVM) and improved (ISLS-SVM) sparse least-square support vector machines, over the motorcycle dataset, as a function of the number of training patterns included in the kernel expansion.

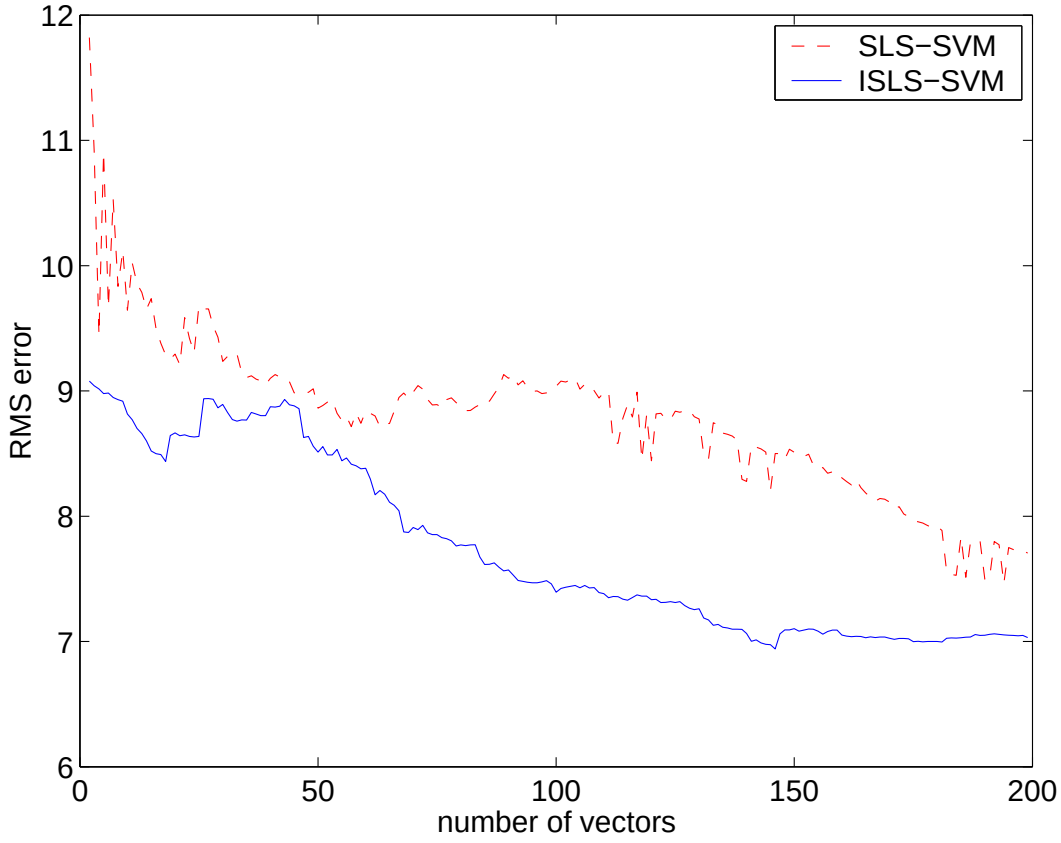

Fig. 2. Cross-validation error of conventional (SLS-SVM) and improved (ISLS-SVM) sparse least-square support vector machines, over the Boston housing dataset, as a function of the number of training patterns included in the kernel expansion.

The Authors

Dr Gavin Cawley

School of Information Systems

University of East Anglia

Norwich

Norfolk

United Kingdom

NR4 7T

Email : gcc@sys.uea.ac.uk

Telephone : (+44) 1603 593558

Fax : (+44) 1603 593345

Gavin Cawley (AMIEE, MIEEE) received a B.Eng and PhD in Electronic Systems Engineering from the University of Essex in 1990 and 1996 respectively. He is currently a lecturer in the School of Information System at the University of East Anglia. His research interests include machine learning and signal processing.

Dr Nicola Talbot

Nicola Talbot (CMath MIMA) received her B.Sc. in mathematics and Ph.D. in Electronic Systems Engineering from the University of Essex in 1991 and 1996, respectively. She formerly worked at the Institute of Food Research, funded by the Ministry of Agriculture Fisheries and Food. Her research interests include optimisation and Bayesian belief networks.