Model Selection for Support Vector Machines via Adaptive Step-Size Tabu Search

Gavin C. Cawley*[1]

*School of Information Systems, University of East Anglia, Norwich, Norfolk, U.K. NR4 7TJ. E-mail: gcc@sys.uea.ac.uk.

Abstract

The generalisation properties of a support vector classification network are typically governed by a regularisation parameter, C, and a small number of parameters specifying the kernel function. The process by which the optimal values of these parameters are obtained is known as model selection. This paper describes an automated model selection procedure based on minimisation of an upper bound on the leave-one-out cross-validation error, via a simple tabu search strategy with adaptive step size adjustment.

1 Introduction

Support vector machines have demonstrated impressive performance in a wide range of notable real world classification problems. The major parameters of the support vector classification network are given by the solution of a linearly constrained quadratic optimisation problem, for which efficient algorithms are available (e.g. Platt [1]). However the optimal choice of kernel function and the values of a small number of hyper-parameters, consisting of the kernel parameters and the regularisation parameter C, must also be determined. This task, known as model selection, is most often performed by training a number of classifiers with different permutations from a range of kernel functions and hyper-parameters, and retaining the configuration resulting in optimal performance on an independent set of validation patterns. In this paper we present a simple and efficient tabu search method, with a robust adaptive step size adjustment heuristic, that can be used to find the value of these hyper-parameters, via minimisation of a recent upper bound on the leave-one-out cross-validation error rate. Model selection is then fully automated and allows all of the available data to be used during training.

[1]This work was supported by the European Commission, grant number IST-99-11764, as part of its Framework V IST programme.

The remainder of this paper is structured as follows: Section 2 briefly describes the support vector classifier and introduces the notation used. Section 3 describes a suitable model selection criteria based on a recent upper bound on the leave-one-out cross-validation error. Section 4 details a model selection procedure, based on Tabu search for the minimiser of this criteria. Initial results obtained on a small, but real-world classification task are presented in section 5, and the work summarised in section 6.

2 Support Vector Classification

The support vector machine [2, 3], given labelled training data

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}, \quad \boldsymbol{x}_i \in \boldsymbol{X} \subset \mathbb{R}^d, \quad y_i \in \{-1, +1\},$$

generates a maximal margin linear decision rule of the form $h(\boldsymbol{x}) = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x} - b)$. The weight vector, $\boldsymbol{w}$, is given by the solution of the primal optimisation problem: minimise

$$V(\boldsymbol{w}, \boldsymbol{\xi}) = \boldsymbol{w} \cdot \boldsymbol{w} + C \sum_{i=1}^{\ell} \xi_i \tag{1}$$

subject to

$$y_i[\boldsymbol{w} \cdot \boldsymbol{x}_i - b] \geq 1 - \xi_i, \qquad i = 1, 2, \ldots, \ell, \tag{2}$$

and

$$\xi_i \geq 0, \qquad i = 1, 2, \ldots, \ell. \tag{3}$$

The slack parameters, ξ_i, allow training patterns to be misclassified in the case of linearly non-separable problems. The parameter C sets the penalty applied to margin-errors, and therefore can be viewed as a regularisation parameter, controlling the trade-off between the width of the margin and training set error. A non-linear decision rule can be constructed using a maximal margin linear classifier in a high dimensional feature space, $\Phi(\boldsymbol{x})$, defined by a positive definite kernel function, $k(\boldsymbol{x}, \boldsymbol{x}')$, specifying an inner product in the feature space,

$$\Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}').$$

A common kernel is the Gaussian radial basis function (RBF),

$$k(\boldsymbol{x}, \boldsymbol{x}') = e^{-\gamma ||\boldsymbol{x}-\boldsymbol{x}'||^2}.$$

The function implemented by a support vector machine is given by

$$f(\boldsymbol{x}) = \left\{ \sum_{i=1}^{\ell} \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) \right\} - b. \tag{4}$$

The optimal coefficients, $\boldsymbol{\alpha}$, of this expansion are given by the solution of the Wolfe dual of the primal optimisation problem: maximise

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{5}$$

in the non-negative quadrant,

$$0 \leq \alpha_i \leq C, \qquad i = 1, 2, \ldots, \ell, \tag{6}$$

subject to the constraint,

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0. \tag{7}$$

The Karush-Kuhn-Tucker (KKT) conditions can be stated as follows:

$$\alpha_i = 0 \quad \Rightarrow \quad y_i f(\boldsymbol{x}_i) \geq 1, \tag{8}$$

$$0 < \alpha_i < C \quad \Rightarrow \quad y_i f(\boldsymbol{x}_i) = 1, \tag{9}$$

$$\alpha_i = C \quad \Rightarrow \quad y_i f(\boldsymbol{x}_i) \leq 1. \tag{10}$$

These conditions are satisfied for the set of feasible Lagrange multipliers, $\boldsymbol{\alpha}^0 = \{\alpha_1^0, \alpha_2^0, \ldots, \alpha_\ell^0\}$, maximising the objective function given by equation 5. The bias parameter, b, is selected to ensure that the second KKT condition is satisfied for all input patterns corresponding to non-bound Lagrange multipliers. Note that in general only a limited number of Lagrange multipliers, $\boldsymbol{\alpha}$, will have non-zero values; the corresponding input patterns are known as support vectors. Let $\mathcal{I}$ be the set of indices corresponding to non-bound Lagrange multipliers,

$$\mathcal{I} = \{i \quad : \quad 0 < \alpha_i^0 < C\},$$

and similarly $\mathcal{J}$ be the set of indices corresponding to Lagrange multipliers at the upper bound C,

$$\mathcal{J} = \{i \quad : \quad \alpha_i^0 = C\}.$$

Equation 4 can then be written as an expansion over support vectors,

$$f(\boldsymbol{x}) = \left\{ \sum_{i \in \{\mathcal{I}, \mathcal{J}\}} \alpha_i^0 y_i k(\boldsymbol{x}_i, \boldsymbol{x}) \right\} - b. \tag{11}$$

For a full exposition of the support vector method, see the excellent books by Vapnik [4] or Cristianini and Shawe-Taylor [5].

3 Model Selection Criteria

Model selection is performed for most classifiers on the basis of validation set error. An alternative approach is possible in the case of support vector classifiers as theoretical bounds on generalisation performance are available. In this paper we adopt the latter approach. Joachims [6] demonstrates that the leave-one-out cross-validation error of a stable soft-margin support vector classifier is bounded by,

$$\mathrm{Err}_{\xi\alpha}^{\ell} = \frac{d}{\ell}, \; d = |\{i \; : \; (\rho \alpha_i^0 R_\Delta^2 + \xi_i) \geq 1\}|, \tag{12}$$

where ρ equals 2, and R_Δ^2 is an upper bound on $k(\boldsymbol{x}, \boldsymbol{x}) - k(\boldsymbol{x}, \boldsymbol{x}')$, $\forall \boldsymbol{x}, \boldsymbol{x}'$. The inequality $\rho \alpha_i^0 R_\Delta^2 + \xi_i \geq 1$ holds for any training pattern corresponding to an error in the leave-one-out procedure, equation 12 therefore provides an upper bound on the leave-one-out error that can be efficiently computed from the solution of the primal and dual optimisation problems given by equations 1-3 and 5-7 respectively. Unlike other bounds on the generalisation ability of support vector classifiers, this bound is directly applicable to soft-margin support vector machines incorporating a bias parameter.

The upper bound on the leave-one-out cross-validation error (12) is discrete, and therefore is less than ideal for model selection based on most heuristic search methods, as a small change in the hyperparameters in general will not produce a change in the value of the bound. Instead we minimise the following continuous model selection criteria:

$$E = \sum_{i \in \mathcal{I}, J} e_i,$$

where

$$e_i = \begin{cases} \rho \alpha_i^0 R_\Delta^2 + \xi_i - 1, & \rho \alpha_i^0 R_\Delta^2 + \xi_i \geq 1 \\ 0, & \rho \alpha_i^0 R_\Delta^2 + \xi_i < 1 \end{cases}.$$

This criteria also penalises patterns that may correspond to leave-one-out errors, but the penalty is linear in the deviation from the boundary between definite correctness and possible error in the leave-one-out cross-validation procedure.

4 Model Selection via Tabu Search

The tabu search procedure (e.g. Glover and Laguna [7]) used to minimise the cost function described in the previous section is based on a simple iterative local search heuristic. At each step

the cost function is evaluated at the set of points given by positive and negative perturbations of each parameter around the current solution. The point minimising the cost function then forms the starting point for the subsequent iteration. Tabu search heuristics disallow moves likely to recover recently encountered configurations. In this case, the simplest tabu produces good results; the direction of a legal step for the current iteration must not oppose the most recently accepted step. For kernels with more than one parameter a more substantial tabu is likely to further reduce the computational expense of model selection.

4.1 Adaptive Step Size Adjustment

A separate step size parameter is associated with each hyper-parameter optimised during the model selection process. Each step size parameter is adjusted adaptively at the end of every epoch according to a procedure based on that used in the RPROP algorithm [8]. Each time the value of the cost function is reduced the step size is multiplied by a factor greater than unity (in this case 1.1), if the cost increases the step size is multiplied by a factor less than unity (in this case 0.1). The step size is only updated if the corresponding hyper-parameter was modified during that epoch. It is assumed that the kernel parameters, like the regularisation parameter C, are strictly positive, and so the step size parameters are moderated to ensure that a hyper-parameter cannot be reduced below a fixed fraction (in this case 0.5) of its current value during the subsequent epoch.

5 Results

This section presents initial results obtained using the model selection procedure outlined in the previous section on a small, but real-world pattern recognition task. The well known Iris data set (Fisher [9]) consists of 150 records describing the lengths and widths of the sepals and petals of three varieties of Iris (Setosa, Versicolour and Virginica). In this work we aim to find the optimal classifier separating examples of Versicolour from Setosa and Virginica varieties, using only the petal length and width attributes. Figure 1 shows the decision boundary formed by the support vector classifier used as the initial estimate in the model selection procedure (Gaussian radial basis kernel, $C = 100, \gamma = 0.5$). This classifier achieves a training set error of 4% and a $\xi\alpha$ bound on the leave-one-out error of 8.7%.

Model selection procedures based on local and tabu search heuristics, both using adaptive step size selection, were then performed. Table 1 summarises the number of times the cost function is evaluated, the number of iterations performed and the total elapsed time for each search method. Note that tabu search, as might be expected, out-performs local search in every respect. The tabu heuristic helps to prevent evaluation of the cost function for configurations unlikely to lead to a reduction in cost.

Search	Epochs	Evals	Time
Local	36	144	18.9913 sec
Tabu	23	70	8.7215 sec

Table 1. Summary of performance statistics for model selection methods based on local and tabu search heuristics.

Figure 2 shows the true leave-one-out cross validation error and $\xi\alpha$-bound during model selection using tabu search. Note it appears that the bound is not always tight, and that a decrease in the cost function does not always result in a decrease in the bound on the leave-one-out error. It should be noted that the Iris data set is fairly small, with only a small number of patterns close to the decision boundary. Only a small number of patterns are then likely to correspond to leave-one-out errors for any sensible classifier. The resulting highly quantised nature of the leave-one-out cross-validation error may in part explain the variation in the quality of the $\xi\alpha$ bound.

Figure 3 shows the decision boundary formed by the support vector classifier resulting from the tabu search model selection procedure ($C = 0.04462, \gamma = 0.3345$). Note that this classifier is far more heavily regularised, having

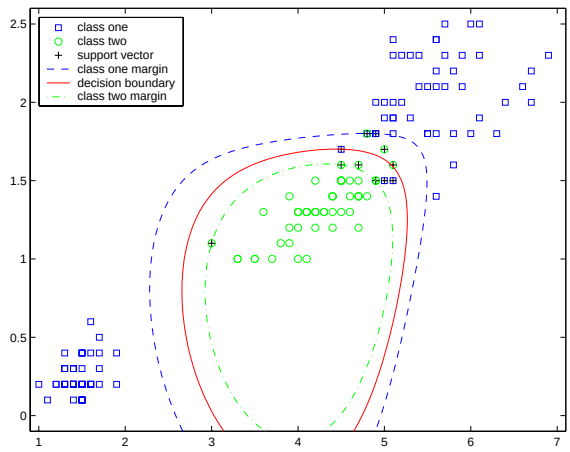

Fig. 1. Decision surface for the support vector classifier used as the starting point for the model selection process.

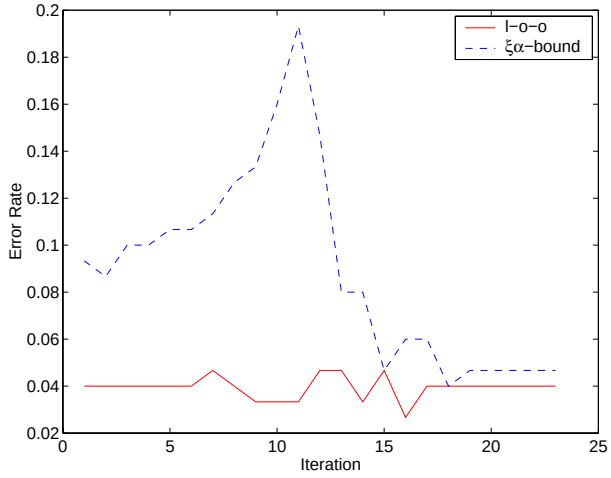

Fig. 2. True leave-one-out cross-validation error and $\xi\alpha$-bound during model selection via tabu search.

a broader margin, but also having a much larger number of (bound) support vectors. This classifier achieves a training set error and $\xi\alpha$ bound on the leave-one-out error of 4%. Although the true leave-one-out error has not decreased, minimising the $\xi\alpha$-bound has produced a subjectively better solution as Occam's Razor tells us that it is sensible to prefer heavily regularised decision rules.

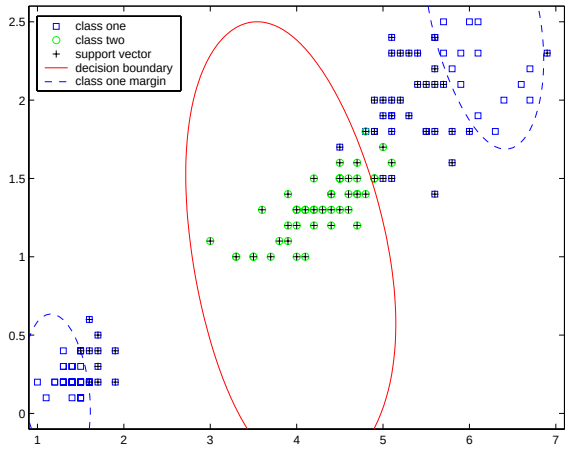

Fig. 3. Decision surface for the support vector classifier resulting from the tabu search model selection process.

6 Summary

This paper describes a practical automatic model selection procedure for support vector classifiers based on tabu search with adaptive step size selection. This heuristic seems to behave efficiently and robustly for a range of initial conditions and search parameters. Further work is needed to further refine the cost function and to evaluate performance on large-scale benchmark pattern recognition tasks.

Acknowledgements

The author would like to thank Rob Foxall for his helpful comments on previous drafts of this manuscript.

References

[1] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Tech. Rep. MSR-TR-98-14, Microsoft Research, 1998.

[2] B. Boser, I. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on computational learning theory*, (Pittsburgh), pp. 144–152, ACM, 1992.

[3] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 1–25, 1995.

[4] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[5] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines (and other kernel-based learning methods)*. Cambridge, U.K.: Cambridge University Press, 2000.

[6] T. Joachims, "Estimating the generalization performance of a SVM efficiently," Tech. Rep. LS-8 number 25, Univerität Dortmund, Fachbereich Informatik, 1999.

[7] F. Glover and M. Laguna, "Tabu search," in *Modern Heuristic Techniques for Combinatorial Problems* (C. R. Reeves, ed.), Advanced Topics in Computer Science, ch. 3, pp. 70–150, Mc Graw Hill, 1995.

[8] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *Proceedings of the IEEE International Conference on Neural Networks* (H. Ruspini, ed.), (San Francisco, CA), pp. 586–591, 1993.

[9] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annual Eugenics*, vol. 7, no. II, pp. 179–188, 1936.