# Estimating the Costs Associated with Worthwhile Predictions of Poor Air Quality

Gavin C. Cawley*[1] Stephen R. Dorling† Robert J. Foxall* Danilo P. Mandic*

*School of Information Systems, University of East Anglia, Norwich, Norfolk, U.K. NR4 7TJ. E-mail: {gcc,rjf,mandic}@sys.uea.ac.uk †School of Environmental Sciences, University of East Anglia, Norwich, Norfolk, U.K. NR4 7TJ. E-mail: s.dorling@uea.ac.uk .

**Abstract**

In this study we investigate the effect of varying the ratio of false-positive and false-negative misclassification costs on the sensitivity and selectivity of binary predictions of exceedences of atmospheric pollutants. This allows us to determine a window of values for this ratio for which it is worthwhile making definite rather than probabilistic predictions. The support vector machine provides a suitable statistical pattern recognition method for this work.

## 1 Introduction

It is rarely the case in real world classification tasks that the penalties associated with false-negative and false-positive misclassifications are exactly equal, although this is frequently an implicit assumption of practical statistical pattern recognition algorithms. For instance in diagnosis of a medical disorder a false-positive result is likely to be cause for some concern for the patient and may incur financial costs due to the conduct of further unnecessary tests, however a false-negative result may lead to a potentially serious disorder developing undetected, with far more severe consequences. Likewise misclassifications in the prediction of episodes of poor air quality also incur asymmetric social, healthcare and financial penalties. These costs are complex, difficult to adequately quantify and vary for different end users. As the prior probability of an exceedence of a given pollutant is relatively low, for a prediction of poor air quality to be possible, either there must be little overlap in the distributions of patterns representing good and poor air quality, or the penalty associated with false-negative misclassifications must be sufficiently higher than that associated with false-positive errors. In this work, we aim to estimate the range of values the ratio of misclassification costs can take for which worthwhile predictions of poor air quality can still be made.

## 2 Support Vector Classification

The support vector machine [1, 2], 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 $h(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{x} - b)$. The weight vector, $\boldsymbol{w}$, is given by the solution of the primal optimisation problem: minimise

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

subject to

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

and

$$\xi_i \geq 0, \qquad i = 1, 2, \ldots, \ell. \qquad (3)$$

The slack parameters, $\xi_i$, 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),

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

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

The function implemented by a support vector machine is given by

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

To find the optimal coefficients, $\boldsymbol{\alpha}$, of this expansion it is sufficient to maximise the functional,

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

in the non-negative quadrant,

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

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, $\boldsymbol{\alpha}^0$, will have non-zero values; the corresponding input patterns are known as support vectors. Equation 4 can then be written as an expansion over support vectors,

$$f(\boldsymbol{x}) = \left\{ \sum_{\text{support vectors}} \alpha_i^0 y_i k(\boldsymbol{x}_i, \boldsymbol{x}) \right\} - b. \tag{11}$$

For a full exposition of the support vector method, see Vapnik [3].

## 3 Support Vector Machines and Asymmetric Misclassification Costs

In the case of binary classification, for any risk functional that is a linear combination of penalties for each observation, the imposition of asymmetric false-positive and false-negative misclassification costs is equivalent to an unequal replication of positive and negative training examples. Consider a generalised empirical risk functional,

$$R_{\text{Emp}}^* = \frac{1}{\ell} \sum_{i=1}^{\ell} C_i \theta(y_i, f(\boldsymbol{x}_i, \boldsymbol{\alpha})), \tag{12}$$

where $C_i$ is the cost associated with the error for pattern $i$. For binary pattern recognition, where $y_i, f \in \{-1, +1\}$, typically

$$\theta(y, f(\boldsymbol{x}, \boldsymbol{\alpha})) = \left\{ \begin{array}{ll} 0 & y = f(\boldsymbol{x}, \boldsymbol{\alpha}) \\ 1 & y \neq f(\boldsymbol{x}, \boldsymbol{\alpha}) \end{array} \right. .$$

To implement asymmetric misclassification costs for positive and negative examples,

$$C_i = \left\{ \begin{array}{ll} C^+ & y_i = +1 \\ C^- & y_i = -1 \end{array} \right. ,$$

where $C^+$ is the cost associated with false-negative and $C^-$ the cost associated with false-positive misclassifications. Clearly the generalised risk functional given by equation 12 is equivalent to the standard empirical risk,

$$R_{\text{Emp}} = \frac{1}{\ell'} \sum_{i=1}^{\ell'} \theta(y_i, f(\boldsymbol{x}_i, \boldsymbol{\alpha})),$$

evaluated over a second dataset consisting of $C^+$ replicates of each positive training example and $C^-$ replicates of each negative example.

For the support vector machine, the symmetry of the optimisation problem given by equations 5-7 suggests that identical training patterns can safely be assigned identical Lagrange multipliers. A notional resampling of training patterns is then implemented by the solution of a modified optimisation problem, maximise

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

in the non-negative quadrant,

$$0 \leq \alpha_i^* \leq C, \qquad i = 1, \ldots, \ell,$$

subject to the constraint,

$$\sum_{i=1}^{\ell} y_i \zeta_i \alpha_i^* = 0,$$

where $\zeta_i$ is the replication factor for pattern $i$. A change of variables, such that $\alpha_i = \zeta_i \alpha_i^*$, reveals that the solution of the modified optimisation problem is identical to that of the original problem subject to the modified box constraint,

$$0 \leq \alpha_i \leq \zeta_i C, \qquad i = 1, \ldots, \ell.$$

Unequal misclassification costs can therefore be accommodated using the modified box constraint,

$$\begin{cases} 0 \leq \alpha_i \leq C^+ C & y_i = +1 \\ 0 \leq \alpha_i \leq C^- C & y_i = -1 \end{cases}$$

(c.f. Lin *et al.* [4], Veropoulos *et al.* [5]).

## 4 Method

For the majority of air quality time series, exceedences of a given pollutant are likely to be relatively rare. As a result it may be the case that it is only worthwhile predicting exceedences if the cost of false-negative predictions outweighs that of false positives. In order to investigate the effect of asymmetric misclassification costs on prediction of exceedences, radial basis function support vector machines can be trained using a range of misclassification costs, in this case for the task of predicting $SO_2$ exceedences in Belfast. The input vector for the support vector network consists of variables representing todays mean $SO_2$ concentration, sin and cosine components representing the day of the week and the Julian day and also meteorological variables representing tomorrows' weather, namely mean temperature, sea level pressure, wind speed and wind direction. The network is trained to predict the existence of an exceedence twenty-four hours in advance. The support vector machines were trained using a freely available MATLAB toolbox [6]. The value of the overall regularisation parameter, $C$, and kernel parameter, $\gamma$, are chosen in accordance with the model selection procedure [7], which attempts to minimise an upper bound on the leave-one-out cross-validation error [8]. The $\xi\alpha$ bound on the leave-one-out error of a support vector machine is given by,

$$\mathrm{Err}^{\ell}_{\xi\alpha} = \frac{d}{\ell}, \; d = |\{i \; : \; (\rho\alpha_i^0 R_\Delta^2 + \xi_i) \geq 1\}|, \qquad (13)$$

where $\rho$ equals 2, and $R_\Delta^2$ 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 13 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 (equations 1-3 and 5-7). For support vector machines with a Gaussian radial basis kernel, $R_\Delta^2 = 1$.

The performance of classifiers will be reported in terms of three statistics,

$$\begin{aligned} \text{recall} &= 1 - \frac{e_+}{n_+}, \\ \text{precision} &= \frac{n_+ - e_+}{n_+ - e_+ + e_-}, \\ \text{accuracy} &= 1 - \frac{e_+ + e_-}{n_+ + n_-}, \end{aligned}$$

where $n_+$ and $n_-$ represent the number of positive and negative examples respectively, and $e_+$ and $e_-$ represent the number of false-negative and false-positive leave-one-out errors respectively,

$$\begin{aligned} e_+ &= |\{i : y_i = +1 \wedge (\rho\alpha_i R_\Delta^2 + \xi_i) \geq 1\}|, \\ e_- &= |\{i : y_i = -1 \wedge (\rho\alpha_i R_\Delta^2 + \xi_i) \geq 1\}|, \\ n_+ &= |\{i : y_i = +1\}|, \\ n_- &= |\{i : y_i = -1\}|. \end{aligned}$$

## 5 Results

Figure 1 shows a graph of recall against the ratio of misclassification costs, over the training set, using the $\xi\alpha$ bound on the leave-one-out cross-validation error and the true leave-one-out error. It can be seen that for symmetric misclassification costs, it is only marginally worthwhile to make any positive classifications. For a ratio of costs of just over 6:1 or more, all exceedences are reliably predicted.

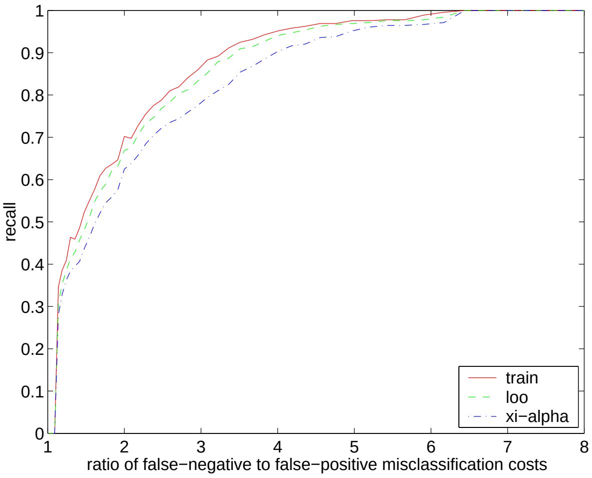

**Fig. 1.** Graph of recall against ratio of misclassification costs.

Figure 2 shows a graph of precision against the ratio of misclassification costs. Naturally the number of false positive errors increases with an increasing ratio of misclassification costs. If a ratio of more than approximately 6:1 is used, all patterns are predicted as exceedences.

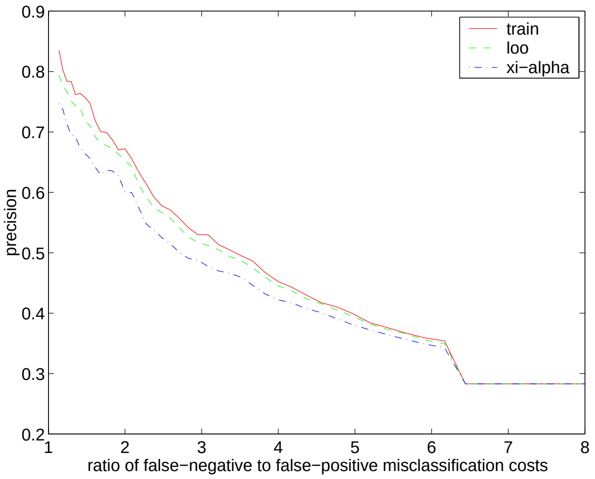

**Fig. 2.** Graph of precision against ratio of misclassification costs.

Figure 3 shows a graph of accuracy against the ratio of misclassification costs. Note the large discontinuities at either extreme of the graph. It appears that the model selection criterion used favours the simplest possible classifier, with a constant output, rather too strongly for marginally worthwhile predictions. Better classifiers for the extremes may result from a different model selection criterion.

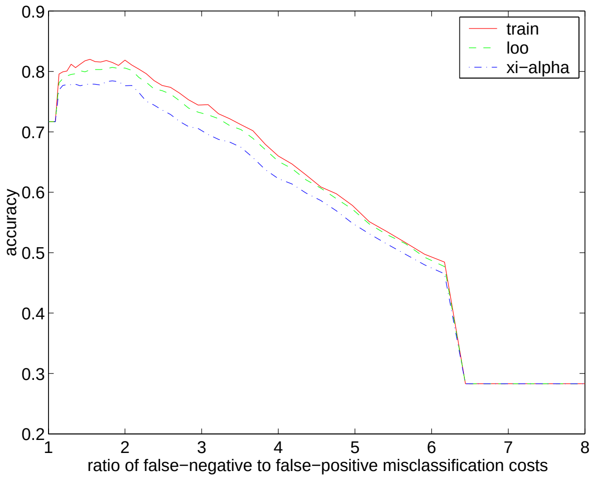

**Fig. 3.** Graph of accuracy against ratio of misclassification costs.

## 6 Summary

In this paper we have demonstrated that the performance of a classifier strongly depends on the costs associated with false-positive and false-negative misclassification errors. We have also experimentally determined the range of misclassification costs for which it is worthwhile actively making predictions of poor air quality due to $SO_2$ in Belfast.

## References

[1] 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.

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

[3] V. N. Vapnik, *Statistical Learning Theory*. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications and Control, New York: Wiley, 1998.

[4] Y. Lin, Y. Lee, and G. Wahba, “Support vector machines for classification in nonstandard situations,” Tech. Rep. 1016, Department of Statistics, University of Wisconsin, 1210 West Drayton St., Madison, WI 53706, March 2000.

[5] K. Veropoulos, C. Campbell, and N. Cristianini, “Controlling the sensitivity of support vector machines,” in *Proceedings of the International Joint Conference on Artificial Intelligence (Workshop ML3)*, (Stockholm, Sweeden), pp. 17–21, 1999.

[6] G. C. Cawley, “MATLAB SVM toolbox (v0.50) [`http://theoval.sys.uea.ac.uk/~gcc/svm/toolbox`].” University of East Anglia, School of Information Systems, Norwich, Norfolk, U.K. NR4 7TJ, 2000.

[7] G. C. Cawley, “Model selection for support vector machines via adaptive step-size tabu search,” in *Proceedings of the International Conference on Artificial Neural Networks and Genetic Algorithms (*accepted for publication*)*, (Prague), April 2001.

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