Controlling the Sensitivity of Support Vector Machines

K. Veropoulos, C. Campbell, N. Cristianini
Department of Engineering Mathematics,
Bristol University, Bristol BS8 1TR,
United Kingdom

Abstract

For many applications it is important to accurately distinguish false negative results from false positives. This is particularly important for medical diagnosis where the correct balance between sensitivity and specificity plays an important role in evaluating the performance of a classifier. In this paper we discuss two schemes for adjusting the sensitivity and specificity of Support Vector Machines and the description of their performance using receiver operating characteristic (ROC) curves. We then illustrate their use on real-life medical diagnostic tasks.

1 Introduction.

Since their introduction by Vapnik and coworkers [Vapnik, 1995; Cortes and Vapnik, 1995], Support Vector Machines (SVMs) have been successfully applied to a number of real world problems such as handwritten character and digit recognition[Schölkopf, 1997; Cortes, 1995; LeCun *et al.*, 1995; Vapnik, 1995], face detection [Osuna *et al.*, 1997] and speaker identification [Schmidt, 1996] . They exhibit a remarkable resistance to overfitting, a feature explained by the fact that they directly implement the principle of Structural Risk Minimization [Vapnik, 1995]. For noise-free classification tasks they work by mapping the training points into a high-dimensional *feature space* where a separating hyperplane $(\mathbf{w}, b)$ is found which maximises the *margin* or distance from the closest data points. Hyperplanes can be represented in feature space (a Hilbert space) by means of kernel functions (dot products between mapped pairs of input points x_i):

$$K(\mathbf{x}', \mathbf{x}) = \sum_i \phi_i(\mathbf{x}')\phi_i(\mathbf{x})$$

Gaussian kernels are an example:

$$K(\mathbf{x}', \mathbf{x}) = e^{-||\mathbf{x}-\mathbf{x}'||^2/2\sigma^2}$$

For input points $\mathbf{x}_i$ mapping to targets y_i $(i = 1, \ldots, p)$, the decision function is formulated in terms of these kernels:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{p} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b\right)$$

where b is the bias and the coefficients α_i are found by maximising the Lagrangian:

$$L = \sum_{i=1}^{p} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{p} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \qquad (1)$$

subject to constraints:

$$\alpha_i \geq 0 \qquad \sum_{i=1}^{p} \alpha_i y_i = 0 \qquad (2)$$

Only those points which lie closest to the hyperplane have $\alpha_i > 0$ (the *support vectors*).

In the presence of noise, two techniques can be used to allow for, and control, a trade off between training accuracy and simplicity of the hypothesis (equivalent to a small norm of $\mathbf{w}$). One consists of setting an upper bound on the size of the α_i, so the influence of outliers is reduced. The other consists in adding a diagonal element to the matrix K, effectively mapping the data into a "separation space" where a separating hyperplane is found. Both these *soft margin* techniques were introduced by (Cortes and Vapnik 1995) and result from the following optimization problems:

$$\begin{array}{ll} \textit{minimize} & C\sum \xi_i^k + \langle \mathbf{w}, \mathbf{w} \rangle \\ \textit{subject to} & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, \text{ with } \xi_i \geq 0 \end{array} \qquad (3)$$

with $k = 1$ and $k = 2$. A theoretical analysis of both algorithms has recently been provided by Shawe-Taylor and Cristianini [Shawe-Taylor and Cristianini, 1999], based on the concept of "margin distributions".

For many decision support systems it is important to distinguish the two types of errors that can arise: a false alarm is usually not as expensive as a missed correct alarm. For example, for the detection of tumours from MRI scans it is important to avoid false negative results, but a small number of false positive results may be tolerated if the scans are subsequently re-screened by medical personnel. Similarly, for the condition monitoring of machinery, occasional false alarms may be less expensive than missing a correct alarm signalling imminant machine failure.

In this paper we will present two new techniques for controlling this trade-off between false positives and false negatives with Support Vector Machines. In section 3 we will compare these methods on real life medical datasets.

2 Sensitivity versus specificity

The performance of a binary classifier is usually quantified by its *accuracy* during the test phase, *i.e.* the fraction of misclassified points on the test set. However, as we have just remarked, the significance of the two types of misclassifications may well be different. Consequently, the performance of such systems are best described in terms of their *sensitivity* and *specificity* quantifying their performance for false positive and false negatives. Systems can then be compared by using a *ROC* (Receiver Operating Characteristic) analysis. These techniques are based on the consideration that a test point always falls into one of the following 4 categories: False Positive (FP) if the systems labels it as a positive while it is a negative; False Negative (FN) if the system labels it as a negative while it is a positive; True Positive (TP) and True Negative (TN) if the system correctly predicts the label. In the following we will use TP, TN, FP, FN to denote the number of true positives, true negatives, false positives and false negatives, respectively. Note that with this notation the number of positive points in the test set can be written as $TP+FN$, the number of negative points as $TN+FP$, and the test set size as $TP+FP+TN+FN$.

A confusion matrix can be used to summarize the performance of a learning machine:

		Expected	
		P	N
Machine	P	TP	FP
	N	FN	TN

and thus a perfect predictor would have a diagonal confusion matrix.

We then define the *sensitivity* of a learning machine as the ratio between the number of true positive predictions TP and the number of positive instances in the test set:

$$\textit{sensitivity} = \frac{TP}{TP+FN}$$

the *specificity* as the ratio between the number of true negative predictions TN and the number of negative instances in the test set:

$$\textit{specificity} = \frac{TN}{TN+FP}$$

the *accuracy* is the ratio between the number of correctly identified examples and the test set size:

$$\textit{accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

For medical diagnosis sensitivity gives the percentage of correctly classified diseased individuals and the specificity the percentage of correctly classified individuals without the disease.

ROC analysis is a classical method in Signal Detection Theory [Swets and Pickett, 1982], and is used also in statistics, medical diagnosis [Centor, 1991] and more recently in Machine Learning as an alternative method for comparing learning systems [Provost *et al.*, 1998]. *ROC* space denotes a coordinate system used for visualizing the performance of a classifier, where the true positive rate is plotted on the y-axis, and the false positive rate on the x-axis. In this way classifers are compared not by a number, but by a point in a plane. For classifiers obtained by thresholding a real valued function or depending on a real parameter, this produces a curve, called a *ROC* curve, describing the trade-off between sensitivity and specificity. Two systems can therefore be compared with the better one being the highest and leftmost one.

The objective of this paper is to outline methods for controlling the balance between the off-diagonal terms in the confusion matrix. In the following we outline and compare two schemes for doing this. The basic idea is to introduce different loss functions for positively and negatively labelled points, which translates into a bias for larger multipliers α_i for the class where the cost of misclassification is heavier. In turn, this induces a decision boundary which is much more distant from the 'critical' class than from the other. In [Shawe-Taylor, 1998] it is shown that the distance of a test point from the boundary is related to its probability of misclassification (test points further away from the hyperplane are less likely to be misclassified). This observation motivated a related technique to the one proposed in this paper. Studying the case of very imbalanced datasets (where points of one class are much more numerous than points of the other class), the authors of [Karakoulas and Shawe-Taylor, 1999] proposed an algorithm where the labels are changed in such a way as to obtain a larger margin on the side of the smaller class.

2.1. We can readily generalise the soft margin approach (3):

$$\mathbf{w}\cdot\mathbf{x}_i + b \geq 1 - \xi_i$$

where:

$$\xi_i \geq 0$$

so that the primal formulation of the Lagrangian has two loss functions for the two types of errors:

$$\begin{aligned} L_p \quad = \quad & \frac{||\mathbf{w}||^2}{2} + C^+ \sum_{\{i|y_i=+1\}}^{p} \xi_i + C^- \sum_{\{i|y_i=-1\}}^{p} \xi_i \\ & - \sum_{i=1}^{p} \alpha_i \left[y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 + \xi_i \right] - \sum_{i=1}^{p} \mu_i \xi_i \end{aligned}$$

where $\alpha_i \geq 0$ and $\mu_i \geq 0$. It is then straightforward to show that the dual formulation gives the same Lagrangian as in (1) but with the α_i constrained as follows:

$$C^+ \geq \alpha_i \geq 0$$

if $y_i = +1$ and:

$$C^- \geq \alpha_i \geq 0$$

if $y_i = -1$.

2.2. Instead of using an L_1 norm for the losses we can also use the square of the L_2 norm instead [Cortes and Vapnik, 1995]. Thus:

$$\begin{aligned} L_p &= \frac{||\mathbf{w}||^2}{2} + C^+ \sum_{\{i|y_i=+1\}}^{p} \xi_i^2 + C^- \sum_{\{i|y_i=-1\}}^{p} \xi_i^2 \\ &\quad - \sum_{i=1}^{p} \alpha_i \left[y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 + \xi_i \right] - \sum_{i=1}^{p} \mu_i \xi_i \end{aligned}$$

Using the derivatives $\partial L_p / \partial w = 0$, $\partial L_p / \partial b = 0$, and $\partial L_p / \partial \xi_i = 0$, and the Kuhn-Tucker conditions:

$$\alpha_i \left[y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 + \xi_i \right] = 0$$

and:

$$\mu_i \xi_i = 0$$

we get the dual formulation:

$$\begin{aligned} L_D &= \sum_{i=1}^{p} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{p} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ &\quad - \frac{1}{4C^+} \sum_{\{i|y_i=+1\}} \alpha_i^2 - \frac{1}{4C^-} \sum_{\{i|y_i=-1\}} \alpha_i^2 \end{aligned}$$

Let:

$$\epsilon^+ = \frac{1}{4C^+} \qquad \epsilon^- = \frac{1}{4C^-}$$

then the balance between sensitivity and specificity can be controlled using the following scheme, namely the *diagonal components* of the kernel matrices are supplemented by fixed positive contributions:

$$K(\mathbf{x}_i, \mathbf{x}_i) \leftarrow K(\mathbf{x}_i, \mathbf{x}_i) + \epsilon^+$$

for $y_i = +1$ and:

$$K(\mathbf{x}_i, \mathbf{x}_i) \leftarrow K(\mathbf{x}_i, \mathbf{x}_i) + \epsilon^-$$

for $y_i = -1$.

The addition of diagonal elements to the covariance function is a widely used in Bayesian regression theory, where the amount of noise present in the data dictates the size of the regularising diagonal elements added to the covariance function. The use of diagonal terms applies to those cases in which the amount of noise is not the same throughout the data. By analogy, the above method amounts to assuming a noise which is dependent on the class. In a sense one class requiries more regularization than the other. From the perspective of large margin classifiers we have simply enforced an asymmetric margin to minimize the risk of misclassifying the elements of one of the two classes. Algorithmically, we have achieved this by changing the covariance function, effectively choosing a special kernel which maps the data into a space where the standard hard margin algorithm can be used.

3 Numerical Experiments

Since our motivation for studying this problem comes from medical decision support we will compare the performance of these two methods on 4 medical datasets:

1. Heart Disease Diagnosis. In Figures 1 and 2 we illustrate the performance of these two methods on data from the UCI (Cleveland) heart disease dataset [Blake *et al.*, 1998]. This consists of a binary classification task with 13 input attributes and 270 examples. With test sets of 27 examples (10-fold cross-validation) we show the performance for $C^- = \infty$ and varying C^+ in Figure 1 and similarly with $\epsilon^- = 0$ and varying ϵ^+in Figure 2. For the ROC curves in Figure 3 the former method performs better on the average though the difference between the two methods is not statistically significant.

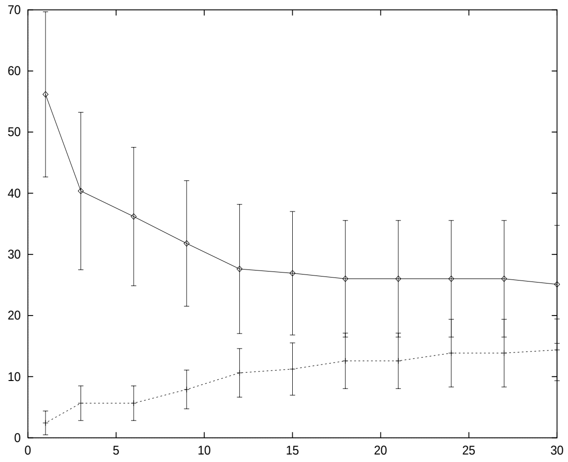

Figure 1: Heart data: the generalisation error (y-axis) versus C^+ (x-axis) for a SVM trained using Gaussian kernels with $\sigma = 4.9$ ($C^- = \infty$). As C^+ decreases the number of false positives decreases (dashed curve) but at the expense of an increase in the number of false negatives (solid curve).

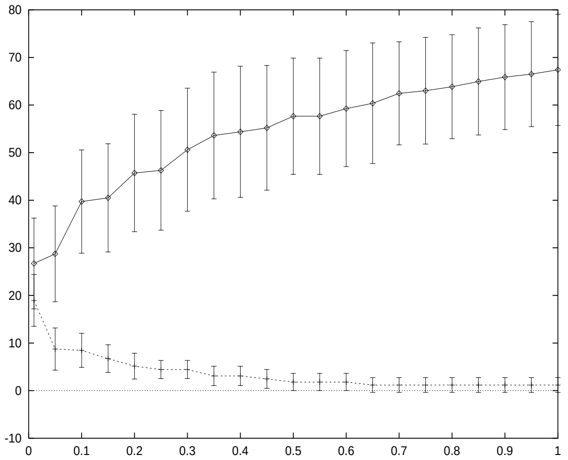

Figure 2: Heart data: the generalisation error (y-axis) versus ϵ^+ (x-axis) for a SVM trained using Gaussian kernels with $\sigma = 4.9$ ($\epsilon^- = 0.0$). As ϵ^+ increases the number of false positives decreases (dashed curve) but at the expense of an increase in the number of false negatives (solid curve).

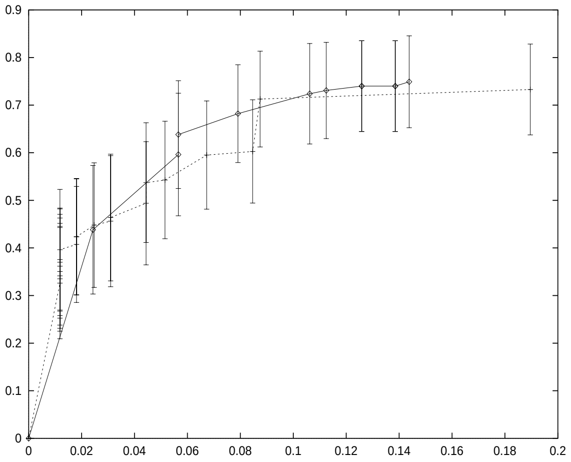

Figure 3: Heart data: true positives (y-axis) vs false positives (x-axis) ($\sigma = 4.9$). The ROC curve generated using method 2.1 (varying C^+ - solid curve) is very similar to the ROC curve generated by method 2.2 (varying ϵ^+ - dashed curve).

2. Diabetes Dataset. This dataset also came from the UCI repository [Blake *et al.*, 1998] and consists of a binary classification task with 8 input attributes. We used 768 examples and approximate 10-fold cross-validation (76 test examples). From the ROC curves (Figure 4) we see that the first method with box constraints on α_i is better on the average but not at a statistically significant level.

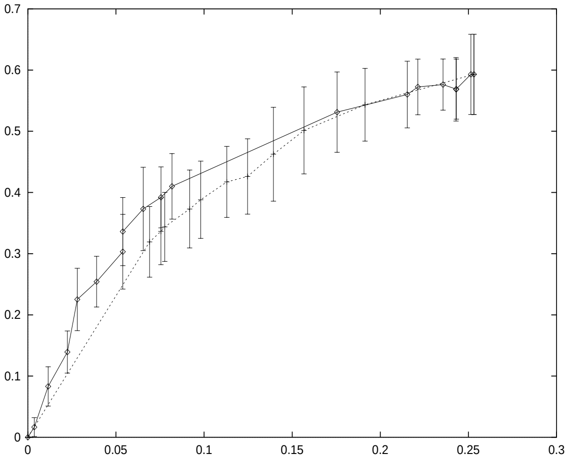

Figure 4: Diabetes data: true positives (y-axis) vs false positives (x-axis) ($\sigma = 1.1$). The ROC curve generated using method 2.1 (varying C^+ - solid curve) and 2.2 (varying ϵ^+ - dashed curve).

3. Liver Disorders Dataset. The BUPA liver disorders dataset has 6 inputs and we used approximate 10-fold cross validation (34 test examples). There is little to distinguish the two methods from the ROC curves (Figure 5) though prohibitive training times for the second method meant only part of the latter curve could be plotted though this feature would depend on the training algorithm used [Friess *et al.*, 1998].

4. TB dataset. This dataset derives from one of our own projects [Veropoulos *et al.*, 1998; 1999]. The task involves classification of image objects (TB bacilli or non-bacilli) on images captured using a microscope. It is intended as part of a system being developed for semi-automated diagnosis to increase the volume of samples investigated and improve accuracy (current screening programmes may miss up to 30-50% of active cases [WHO, 1998]). A total of 1140 examples were used (each with 14 input attributes) together with 10-fold cross validation. Again (Figure 6) there is little to distinguish the two methods though we were not able to complete the ROC curve for the second method for the same reason as occured in experiment **3**.

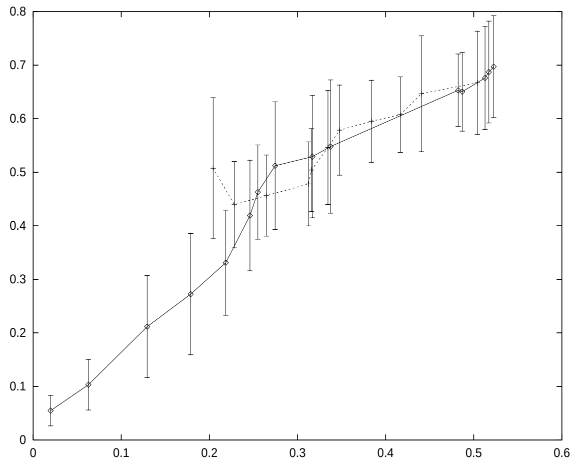

Figure 5: Liver data: true positives (y-axis) vs false positives (x-axis) ($\sigma = 0.7$). The ROC curve generated using method 2.1 (varying C^+ - solid curve) and 2.2 (varying ϵ^+ - dashed curve)

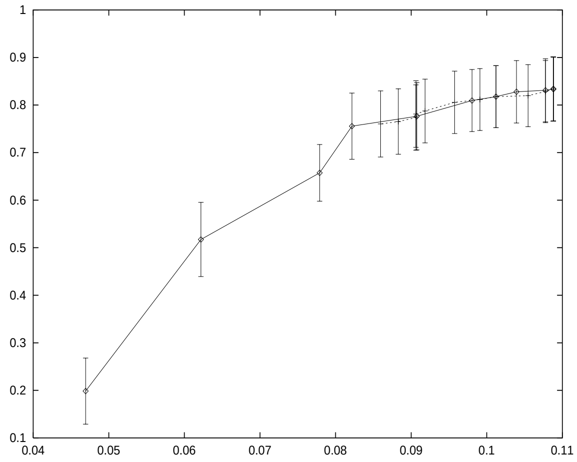

Figure 6: TB data: true positives (y-axis) vs false positives (x-axis) ($\sigma = 0.3$). The ROC curve generated using method 2.1 (varying C^+ - solid curve) and 2.2 (varying ϵ^+ - dashed curve)

4 Conclusion.

In this paper we have outlined 2 schemes for controlling the balance between false positives and false negatives. In our numerical experiments the method based on using box constraints (C^+, C^-) is better on the average for 2 datasets (heart and diabetes). However, the difference between the two methods is not statistically significant for any of the 4 datasets considered. For the algorithm used [Friess *et al.*, 1998] the method based on additions to the kernel diagonal proved to be prohibitively slow for large values of (ϵ^+, ϵ^-). However, the extent of this problem is likely to be related to the type of QP routine used.

For many data mining applications the problem of imbalanced data sets or asymmetric loss functions is common (with sensitivity and specificity more important than overall performance). The techniques proposed here are sufficiently simple to be promptly implemented while adding little further computational load to the algorithm. As demonstrated in the above case studies, they have the power to effectively control the sensitivity of the learning machine.

References

[Blake *et al.*, 1998] C. Blake, E. Keogh, and C.J. Merz. *UCI Repository of Machine Learning Databases.* Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[Centor, 1991] R.M. Centor. Signal detectability: The use of roc curves and their analyses. *Medical Decision Making*, 11:102–6, 1991.

[Cortes and Vapnik, 1995] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[Cortes, 1995] C. Cortes. *Prediction of Generalisation Ability in Learning Machines.* PhD Thesis, Department of Computer Science, University of Rochester, 1995.

[Friess *et al.*, 1998] T. Friess, N. Cristianini, and C. Campbell. The kernel-adatron: a fast and simple learning procedure for support vector machines. In J. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 188–196, 1998.

[Karakoulas and Shawe-Taylor, 1999] G. Karakoulas and J. Shawe-Taylor. Optimizing classifiers for imbalanced training sets. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11.* MIT Press, 1999.

[LeCun *et al.*, 1995] Y. LeCun, L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J.S. Denker, H. Drucker, I. Guyon, U.A. Muller, E. Sackinger, P. Simard, and V. Vapnik. Comparison of learning algorithms for handwritten digit recognition. In F. Fogelman and P. Gallinari, editors, *International Conference on Artificial Neural Networks*, pages 53–60, 1995.

[Osuna *et al.*, 1997] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *Proc. Computer Vision and Pattern Recognition '97*, pages 130–136, 1997.

[Provost *et al.*, 1998] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comapring induction algorithms. In J. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning (ICML98)*, 1998.

[Schmidt, 1996] M. Schmidt. Identifying speakers with support vector machines. In *Proceedings of Interface '96, Sydney*, 1996.

[Schölkopf, 1997] B. Schölkopf. *Support Vector learning.* R. Oldenbourg Verlag, Munich, 1997.

[Shawe-Taylor and Cristianini, 1999] J. Shawe-Taylor and N. Cristianini. Further results on the margin distribution. In *Proceedings of the 12th Conference on Computational Learning Theory*, 1999.

[Shawe-Taylor, 1998] J. Shawe-Taylor. *Algorithmica*, 22:157–172, 1998.

[Swets and Pickett, 1982] J.A. Swets and R.M. Pickett. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* Academic Press, New York, 1982.

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

[Veropoulos *et al.*, 1998] K. Veropoulos, C. Campbell, G. Learmonth, B. Knight, and J. Simpson. The automated identification of tubercle bacilli using image processing and recognition techniques. In L. Niklasson, M. Boden, and T. Ziemke, editors, *Perspectives in Neural Computing, ICANN98*, volume 2, pages 797–802. ICANN98, Skovde, Sweden, (Springer), 1998.

[Veropoulos *et al.*, 1999] K. Veropoulos, G. Learmonth, C. Campbell, B. Knight, and J. Simpson. The automated identification of tubercle bacilli in sputum: A preliminary investigation. *Analytical and Quantitative Cytology and Histology*, to appear, 1999.

[WHO, 1998] WHO. Fact sheet no 104: Tuberculosis, 1998.