Asymmetrical Margin Approach to Surveillance of Nosocomial Infections Using Support Vector Classification

Gilles Cohen
Medical Informatics Division
University Hospital of Geneva
1211 Geneva, Switzerland
E-mail: gilles.cohen@dim.hcuge.ch

Mélanie Hilario
Artificial Intelligence Laboratory
University of Geneva
1211 Geneva, Switzerland
E-mail: Melanie.Hilario@unige.ch

Stéphane Hugonnet and **Hugo Sax**
Department of Internal Medicine
University Hospital of Geneva
1211 Geneva, Switzerland
E-mail: {Stephane.Hugonnet|Hugo.Sax}@hcuge.ch

Abstract

Nosocomial or hospital-acquired infections (NIs) have become a major concern not only in health care institutions but also among the general public. Since 1994 the Geneva University Hospital has been undertaking yearly prevalence studies in order to monitor and detect NIs. This paper describes a retrospective analysis of the results of one such study. Our goal is to identify patients with one or more NIs on the basis of clinical and other data collected during the survey. In this classification task, the main difficulty resides in the significant imbalance between positive or infected (11%) and negative (89%) cases. To cope with class imbalance, we investigate a support vector algorithm in which asymmetrical margins are tuned to improve recognition of rare positive cases. Experiments have shown this approach to be effective for the NI detection problem: we obtained a sensitivity rate of 92%, significantly better than the highest sensitivity (87%) obtained via novel resampling strategies in a previous study.

1 INTRODUCTION

The problem of nosocomial[1] or hospital-acquired infections has become a major public concern in the wake of several dramatic cases that have been widely disseminated by the mass media. Surveillance is the key element in the prevention and control of infections, regardless of origin: it provides data to assess the magnitude of the problem, detect outbreaks, identify risk factors, target control measures on high-risk patients or wards, or evaluate prevention programs. Ultimately, the goal of surveillance is to decrease infection risk and consequently improve patients' safety. Hospital-wide prospective surveillance is considered the gold standard. However, because it requires resources that are beyond what any hospital can afford, this strategy is recommended only in selected wards, such as intensive care units. As an alternative and more feasible approach, prevalence surveys are considered a valid surveillance strategy and are increasingly performed. Their major limitations are their retrospective nature, the dependency on readily available data, a prevalence bias, the inability to detect outbreak (depending on the frequency the surveys are performed), and the limited capacity to identify risk factors. However, they provide sufficiently good data to measure the magnitude of the problem, evaluate a prevention program, and help allocate resources. They give a snapshot of clinically active NIs during a given index day and provide information about the frequency and characteristics of these infections. The efficacy of infection control policies can be easily measured by repeated prevalence surveys [French *et al.*, 1983].

[1] Nosocomial (from the Greek word *nosokomeion* for hospital) are those which were not present or incubating at the time of admission but have been acquired during hospitalization. Usually, infections occuring more than 48 hours after admission are considered nosocomial.

2 DATA COLLECTION AND PREPARATION

The University Hospital of Geneva (HUG) has been performing yearly prevalence studies since 1994 [Harbarth *et*

al., 1999]. These surveys are undertaken every year at the same period and last approximately three weeks. All patients hospitalized at time of the survey for at least 48 hours are assessed for the presence of an active nosocomial infection. Data are extracted from medical records, kardex, X-ray and microbiology reports, and interviews with nurses and physicians in charge of the patient, if necessary. All nosocomial infections active during the 6 days preceding the day of survey are recorded and identified according to modified Centres for Disease Control (CDC) criteria. Collected variables include administrative information, demographic characteristics, admission diagnosis, comorbidities and severity of illness scores, type of admission, exposure to various risk factors for infection (surgery, intensive care unit stay, invasive devices, antibiotics, antacids, immunosuppressive treatments), clinical and paraclinical information, and data related to infection, when present.

This type of hospital-wide prevalence survey has been favoured over prospective surveillance, as it is less time-consuming. However, it still requires considerable resources, as about 800 hours are needed for data collection only. Consequently, we cannot afford performing this surveillance more than once a year. The aim of this pilot study is to apply data mining techniques to data collected in the 2002 prevalence study in order to detect nosocomial infected patients on the basis of the factors described above.

The dataset consisted of 688 patient records and 83 variables. With the help of hospital experts on nosocomial infections, we filtered out spurious records as well as irrelevant and redundant variables, reducing the data to 683 cases and 49 variables. In addition, several variables had missing values, due mainly to erroneous or missing measurements. These values were assumed to be missing at random, as domain experts did not detect any clear correlation between the fact that they were missing and the data (whether values of the incomplete variables themselves or of others). We replaced these missing values with the class-conditional mean for continuous variables and the class-conditional mode for nominal ones. These preprocessing operations are often necessary in such retrospective analyses where data collection has not been engineered specifically for data mining purposes.

3 THE IMBALANCED DATA PROBLEM

The major difficulty inherent in the data (as in many medical diagnostic applications) is the highly skewed class distribution. Out of 683 patients, only 75 (11% of the total) were infected and 608 were not. The problem of imbalanced datasets is particularly crucial in applications where the goal is to maximize recognition of the minority class [2]. The issue of class imbalance, which has been actively investigated and remains largely open, is handled in a number of ways [Japkowicz, 2002], including, oversampling the minority class, building cost-sensitive classifiers [Domingos, 1999] that assign higher cost to misclassifications of the minority class, stratified sampling on the training instances to balance the class distribution [Kubat and Matwin, 1997] and rule-based methods that attempt to learn high confidence rules for the minority class [Ali *et al.*, 1997]. In this paper we investigate another way of biasing the inductive process to boost sensitivity (i.e., capacity to recognize positives) based on asymmetrical soft margin support vector machines. Experiments conducted to assess this approach are described in Section 5 and results are discussed in Section 6.

[2] For convenience we identify positive cases with the minority and negative cases the majority class.

4 CLASSIFICATION

4.1 Support vector classification

Support vector machines [Vapnik, 1998; Cortes and Vapnik, 1995] (SVM) are learning machines based on the *Structural Risk Minimization principle* (SRM) from statistical learning theory. SRM principle seeks to minimize an upper bound of the generalization error rather than minimizing the training error (Empirical Risk Minimization (ERM)). This approach results in better generalization than conventionnal techniques generally based on the ERM principle.

Consider a labelled training set $\{\mathrm{x}_i, y_i\}, \quad i = 1, \ldots, n, \quad y_i \in \{-1, +1\}, \quad \mathrm{x}_i \in \mathcal{R}^d$. For a separable classification task, there exists a separating hyperplane, defined by $\mathrm{w}.\mathrm{x} + b$, with w the weight vector and b the bias, which maximises the $margin$ or distance between the hyperplane and the closest data points belonging to the different classes. This optimum separating hyperplane is given by the solution to the problem :

$$\begin{aligned} &minimize && \frac{1}{2}||\mathrm{w}||^2 \\ &subject\ to && y_i(\mathrm{w}.\mathrm{x}_i + b) \geq 1 \quad \forall_i \end{aligned} \tag{1}$$

where $\frac{b}{||\mathrm{w}||}$ is the distance between origin and hyperplane. This is a quadratic programming problem (QP), solved by Karush-Kuhn-Tucker theorem. Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ be the n non negative Lagrange multipliers associated with the constraints, the solution to the problem is equivalent to determining the solution of the *Wolfe dual* [Fletcher, 1987] problem :

$$\begin{aligned} &maximize && \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j (\mathrm{x}_i.\mathrm{x}_j) y_i y_j. \\ &subject\ to && \sum_i \alpha_i y_i = 0 \quad \alpha_i \geq 0 \end{aligned} \tag{2}$$

The solution for w is

$$\mathrm{w} = \sum_i \alpha_i y_i \mathrm{x}_i. \tag{3}$$

There is a Lagrange multiplier α_i for each training point and only those training examples that lie close to the decision boundary have nonzero α_i. These vectors are called the *support vectors*. The classifier decision function $f(x)$ is :

$$f(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i (\mathrm{x}.\mathrm{x}_i) + b\right). \tag{4}$$

4.2 Soft margin

While the above method is fine for separable data points, very often noisy data or sampling problems will lead to no linear separation in the feature space. Very often, the data points will be almost linearly separable in the sense that only a few of the members of the data points cause it to be non linearly separable. Such data points can be accommodated into the theory with the introduction of slack variables that allow particular vectors to be misclassified. The hyperplane margin is then relaxed by penalising the training points misclassified by the system. Formally the optimal hyperplane is defined to be the hyperplane which maximizes the margin and minimizes some functional $\theta(\xi) = \sum_{i=1}^{n} \xi_i^{\sigma}$, where σ is some small positive constant. Usually the value $\sigma = 1$ is used since it is a QP and the corresponding dual does not involve ξ and therefore offers a simple optimization problem. The constraint in (1) now assumes the form

$$y_i(\mathrm{w}.\mathrm{x}_i + b) \geq 1 - \xi_i,\ \forall_i\ \xi_i \geq 0\ . \tag{5}$$

If we select $\sigma = 1$ the optimization problem becomes

$$\begin{aligned} &minimize && \frac{1}{2}||\mathrm{w}||^2 + C\sum_{i=1}^{n}\xi_i \\ &subject\ to && y_i(\mathrm{w}.\mathrm{x}_i + b) \geq 1 - \xi_i, \quad \forall_i\ \xi_i \geq 0. \end{aligned} \tag{6}$$

where ξ_i introduces a positive slack variable that measures the degree of violation of the constraint. The penalty C is a regularisation parameter that controls the trade-off between maximizing the margin and minimizing the training error. This is called the soft margin approach.
Again, instead of solving directly optimization problem (6) we consider the corresponding dual problem

$$\begin{aligned} &maximize && \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j(\mathrm{x}_i.\mathrm{x}_j)y_iy_j. \\ &subject\ to && \sum_i \alpha_i y_i = 0 \quad 0 \leq \alpha_i \leq C \end{aligned} \tag{7}$$

The entire construction can be extended rather naturally to include nonlinear decision boundaries. Each data point x in input space is mapped into a vector $\mathrm{z} = \phi(\mathrm{x})$ in a higher dimensional feature space. We can then substitute the dot product $(\phi(\mathrm{x}).\phi(\mathrm{x}_i))$ in feature space with a non linear function $K(\mathrm{x}, \mathrm{x}_i)$, also called a *kernel*. Conditions for a function to be a kernel are expressed in a theorem by Mercer [Burges, 1998; Christianini and J.S., 2000]. The final classifier $f(x)$ is then expressed in term of $K(\mathrm{x}, \mathrm{x}_i)$

$$f(x) = sign\left(\sum_{i=1}^{n}\alpha_i y_i K(\mathrm{x}, \mathrm{x}_i) + b\right)\ . \tag{8}$$

4.3 Asymmetrical soft margin

The above formulation of the SVM is inappropriate in two common situations : in case of unbalanced distributions, or whenever misclassifications must be penalized more heavily for one class than for the other. In order to adapt the SVM algorithm to these cases [Karakoulas and Shawe-Taylor, 1999; Veropoulos *et al.*, 1999] the basic idea is to introduce different error weights C^+ and C^- for the positive and the negative class, which results in a bias for larger multipliers α_i of the critical class. This induces a decision boundary which is more distant from the smaller class than from the other. This transforms (6) into the following optimization problem:

$$\begin{aligned} &minimize && \frac{1}{2}||\mathrm{w}||^2 + C^-\sum_{i:y_i=-1}^{n}\xi_i^- + C^+\sum_{i:y_i=+1}^{n}\xi_i^+ \\ &s.t && (\mathrm{w}.\mathrm{x}_i + b) \geq 1 - \xi_i^+, \\ & && (\mathrm{w}.\mathrm{x}_i + b) \leq -1 + \xi_i^- \end{aligned} \tag{9}$$

5 EXPERIMENTAL SETUP

5.1 Performance metrics

In classification tasks, performance of a classifier is commonly quantified by is predictive accuracy, i.e. the fraction of misclassified data points on the test set. However the significance of positives and negatives misclassification may well be different or the class distribution may be imbalanced making this metric close to meaningless. To see this, consider a dataset consisting of 5% positive and 95% negatives. The simple rule of assigning a case to the majority class would result in an impressive 95% accuracy whereas the classifier would have failed to recognize a single positive case—an inacceptable situation in medical diagnosis. The reason for this is that the contribution of a class to the overall accuracy rate is a function of its cardinality, with the effect that rare positives have an almost insignificant impact on the performance measure.

To discuss alternative performance criteria we adopt the standard definitions used in binary classification. TP and TN stand for the number of true positives and true negatives respectively, i.e., positive/negative cases recognized as such by the classifier. FP and FN represent respectively the number of misclassified positive and negative cases. In two-class problems, the accuracy rate on the positives, called sensitivity, is defined as :

$$sensitivity : TP/(TP + FN), \tag{10}$$

whereas the accuracy rate on the negative class, also known as specificity, is :

$$specificity : TN/(TN + FP). \tag{11}$$

Classification accuracy is simply :

$$accuracy : (TP + TN)/N, \tag{12}$$

where $N = TP + TN + FP + FN$ is the total number of cases.

5.2 ROC curves

In medical diagnosis [Centor, 1991] , biometrics and recently machine learning [Provost *et al.*, 1998], the usual way of assessing a classification method is the receiver operating characteristic (ROC) curve. A ROC curve plots sensitivity versus $1 - specificity$ for different thresholds of the classifier output. Based on the ROC curve, one can

decide how many false positives (respectively false negatives) one is willing to tolerate and tune the classifier threshold to best suit a certain application. A random assignment of classes to data would result in a ROC Curve in form of a diagonal line from (0,0) to (1,1).

5.3 Evaluation strategy

The experimental goal was to measure the performance of an SVM asymmetrical soft margin approach to cope with uneven datasets. To train our SVM classifiers we use a radial basis kernel of the form $K(x_i, x_j) = \exp^{\frac{-||x_i - x_j||^2}{2\sigma^2}}$. To obtain the optimal values for the hyperparameters σ, C, C^+ and C^- we experimented with different SVM classifiers using a range of values. Given the limited amount of data, 5-fold stratified cross-validation was applied to find the best classifier based on validation error. The performance of the selected SVMs was quantified based on its sensitivity, specificity and accuracy. For our experiments we fixed C^- at 1, and to determine the best C^+ parameter we learned several SVM classifiers using different values for C^+.

6 RESULTS

Table 1 summarizes performance results for symmetrical and asymmetrical SVMs on the original skewed class distribution and illustrates clearly the inadequacy of the former for this task. These are the best results from a selection of configurations used for training the classifiers.

Table 1: Performance for different SVM configurations (original class distribution: 0.11 pos, 0.89 neg)

SVM Classifier RBF Gaussian	Hyperpar. ($\sigma = 0.1$)	Accuracy	Sensitivity	Specificity
sym. margin	$C = 4$	0.893	0.026	1
sym. margin	$C = 20$	0.906	0.44	0.964
sym. margin	$C = 45$	0.896	0.506	0.944
asym. margin	$C^+ = 3$	0.876	0.586	0.912
asym. margin	$C^+ = 5$	0.828	0.76	0.837
asym. margin	$C^+ = 11$	0.816	0.88	0.809
asym. margin	$C^+ = 29$	0.744	0.92	0.722

In the first experiment based on symmetrical margins, accuracy rates hover constantly around 90% whereas even the best sensitivity remains barely higher than 50% (see Figure 1). This clearly illustrates the inadequacy of the symmetrical soft margin approach as well as the inappropriateness of accuracy as a performance criterion for the nosocomial application.

To explore the effect of asymmetrical soft margins, we trained SVMs with σ fixed at 0.1 and C^- fixed at 1 for a wide range of C^+ values. Figure 2 illustrates the effect of upper bound C^+ on the α_i of the positive (i.e. infected) class. For example, as C^+ increases, the number of false positives is increased but at the detriment of a decrease in

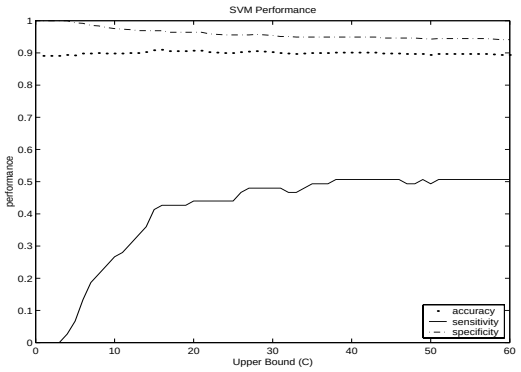

Figure 1: Generalization performance of the symmetrical-margin SVM classifier against different C values

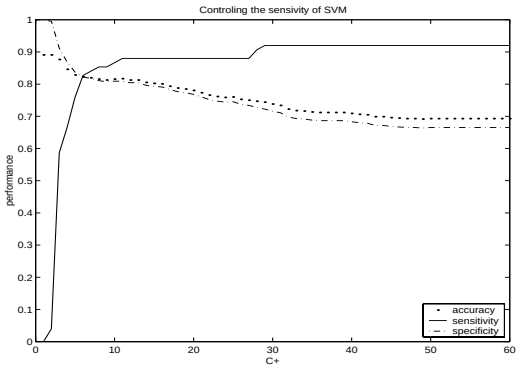

Figure 2: Generalization performance of the asymmetrical-margin SVM classifier against different C^+ values

the number of false negatives. Sensitivity increases while specificity decreases with increasing values of C^+ (at least up to 29), but as shown clearly in the figure, the gain in sensitivity far outdistances loss of specificity—a fact occluded by the concomitant decrease in accuracy.

In a previous study on the same nosocomial dataset [Cohen *et al.*, 2003], we compared the predictive performance of 5 learning algorithms (including symmetrical-margin SVMs) after applying novel methods based on synthetic example generation to correct class imbalance. One method consisted in downsizing the majority (negative) class, not by randomly choosing existing cases, but by generating N_m prototypes via K-means clustering, where N_m is the size of the minority (positive) class. These N_m prototypes then replaced all the original instances of the majority class. In an alternative approach, the minority class was oversampled by applying agglomerative hierarchical clustering to the original cases and computing the prototypes of the resulting clusters at all merge levels. These prototypes were then added to the original cases in order to expand the minority class. Finally, we combined these two methods into a hybrid over/undersampling approach. All three

Table 2: Over/undersampling via synthetic example generation (0.5 pos 0.5 neg). Bracketed figures are baseline sensitivity rates obtained prior to class balancing.

Classifier	Accu.	Sensitivity	Specif.	Method
IB1	0.84	0.56 [0.19]	0.88	KMU
NaiveBayes	0.75	0.87 [0.57]	0.74	HYB
C4.5	0.68	0.72 [0.28]	0.67	KMU
AdaBoost	0.75	0.84 [0.45]	0.74	KMU
SVM	0.75	0.83 [0.43]	0.74	KMU

methods were shown to lead to significantly higher sensitivity than random oversampling and subsampling. Table 2 shows the best performance measures obtained in these previous experiments, together with the method used (KMU for K-means based undersampling and HYB for the hybrid method). Between brackets, the baseline sensitivity rates (those attained by the classifiers on the original class distribution) give an idea of the inherent difficulty of the NI detection problem.

A comparison of Tables 1 and 2 shows that the asymmetrical margin approach leads to better sensitivity than all our previously proposed methods, provided that the appropriate hyperparameters are used. The best sensitivity rate in these previous experiments was 0.87, attained by Naive Bayes coupled with hybrid over/undersampling via prototype generation. SVMs using asymmetrical margins and a C+ parameter of 29 perform remarkably better with a sensitivity rate of 0.92.

In order to visualize and assess the behaviour of the SVM classifiers throughout a whole range of the output threshold values, the ROC curve shown in Figure 3 has been produced. This allows experts to easily choose the model best suited to their purpose. The model corresponding to the circled point on the ROC curve (Figure 3) has been retained by our experts for our NI classification application. It corresponds to the highest sensitivity 92% reached for a specificity of 72.2% which has been judged completely acceptable.

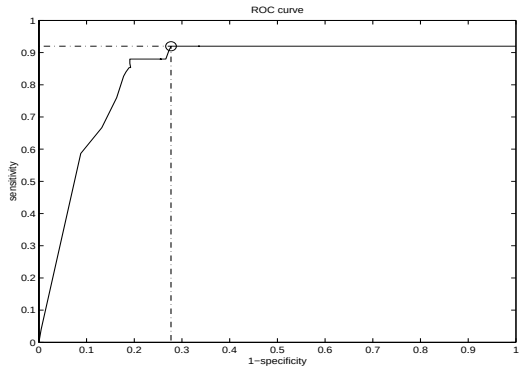

Figure 3: ROC Curve for SVM classifiers varying error weight values for the positive class C^+

7 CONCLUSION AND FUTURE WORK

We analysed the results of a prevalence study of nosocomial infections in order to predict infection risk on the basis of patient records. The major hurdle, typical in medical diagnosis, is the problem of rare positives. To address this problem we investigated the feasibility of an algorithm proposed by [Karakoulas and Shawe-Taylor, 1999; Veropoulos *et al.*, 1999] where class dependent regularization parameters are introduced in such a way as to obtain a larger margin on the side of the smaller class (asymmetrical soft margin). The results obtained are basically encouraging: whereas the sensitivity range of symmetrical soft margin SVMs was $[2.6 - 50.6]\%$, it increased to $[58.6 - 92]\%$ with asymmetrical soft margin SVMs. The maximal sensitivity rate of 92% represents a significant improvement over the best sensitivity of 87% attained previously by the same authors using class balancing with synthetic examples.

On the research agenda for the immediate future, we intend to prospectively validate the classification model obtained by performing in parallel a standard prevalence survey and then to improve it in order to classify site-specific infections.

Acknowledgements

The authors thank Profs. A. Geissbuhler and D. Pittet (University of Geneva Hospitals) for having given them the opportunity to undertake this study.

References

[Ali *et al.*, 1997] K. Ali, S. Manganaris, and R. Srikant. Partial classification using association rules. In *Proc. 3rd International Conference on Knowledge Discovery in Databases and Data Mining*, 1997.

[Burges, 1998] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, pages 121–167, 1998.

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

[Christianini and J.S., 2000] N. Christianini and Taylor J.S. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[Cohen *et al.*, 2003] G. Cohen, M. Hilario, S. Hugonnet, and H. Sax. Data imbalance in surveillance of nosocomial infections. In *International Symposium on Medical Data Analysis*, 2003.

[Cortes and Vapnik, 1995] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[Domingos, 1999] P. Domingos. A general method for making classifiers cost-sensitive. In *Proc. 5th International Conference on Knowledge Discovery and Data Mining*, pages 155–164, 1999.

[Fletcher, 1987] R. Fletcher. *Practical Methods of Optimization*. John Wiley and Sons, 1987.

[French *et al.*, 1983] G. G. French, A. F. Cheng, S. L. Wong, and S. Donnan. Repeated prevalence surveys for monitoring effectiveness of hospital infection control. *Lancet*, 2:1021–23, 1983.

[Harbarth *et al.*, 1999] S. Harbarth, Ch. Ruef, P. Francioli, A. Widmer, D. Pittet, and Swiss-Noso Network. Nosocomial infections in swiss university hospitals: a multicentre survey and review of the published experience. *Schweiz Med Wochenschr*, 129:1521–28, 1999.

[Japkowicz, 2002] N. Japkowicz. The class imbalance problem: A systematic study. *Intelligent Data Analysis Journal*, 6(5), 2002.

[Karakoulas and Shawe-Taylor, 1999] G. Karakoulas and J. Shawe-Taylor. Optimizing classifiers for imbalanced training sets. In *Advances in Neural Information Processing Systems (NIPS-99)*. MIT Press, 1999.

[Kubat and Matwin, 1997] M. Kubat and S. Matwin. Addressing the curse of imbalanced data sets: One-sided sampling. In *Procsòf the Fourteenth International Conference on Machine Learning*, pages 179–186, 1997.

[Provost *et al.*, 1998] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proc. Fifteenth International Conference on Machine Learning (ICML98)*, 1998.

[Vapnik, 1998] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[Veropoulos *et al.*, 1999] K. Veropoulos, N. Cristianini, and C. Campbell. Controlling the sensitivity of support vector machines. In *Proc. International Joint Conference on Artificial Intelligence*, 1999.