One-Class Support Vector Machines with a Conformal Kernel. A Case Study in Handling Class Imbalance

Gilles Cohen[1,2], Mélanie Hilario[2], and Christian Pellegrini[2]

[1] Medical Informatics Service, University Hospital of Geneva,
1211 Geneva, Switzerland
Gilles.Cohen@sim.hcuge.ch
[2] Artificial Intelligence Laboratory, University of Geneva,
1211 Geneva, Switzerland
Melanie.Hilario@cui.unige.ch

Abstract. Class imbalance is a widespread problem in many classification tasks such as medical diagnosis and text categorization. To overcome this problem, we investigate one-class SVMs which can be trained to differentiate two classes on the basis of examples from a single class. We propose an improvement of one-class SVMs via a conformal kernel transformation as described in the context of binary SVM classifiers by [2, 3]. We tested this improved one-class SVM on a health care problem that involves discriminating 11% nosocomially infected patients from 89% non infected patients. The results obtained are encouraging: compared with three other SVM-based approaches to coping with class imbalance, one-class SVMs achieved the highest sensitivity recorded so far on the nosocomial infection dataset. However, the price to pay is a concomitant decrease specificity, and it is for domain experts to decide the proportion of false positive cases they are willing to accept in order to ensure treatment of all infected patients.

1 The Imbalanced Data Problem

Data imbalance is a crucial problem in applications where the goal is to maximize recognition of the minority class, as is typically the case in medical diagnosis. The issue of class imbalance has been actively investigated and remains widely open; it is handled in a number of ways [14], including: oversampling the minority class, building cost-sensitive classifiers [10] that assign higher cost to misclassifications of the minority class, stratified sampling on the training instances to balance the class distribution [15] and rule-based methods that attempt to learn high confidence rules for the minority class [1]. In this paper we investigate another way of biasing the inductive process to boost sensitivity (i.e., capacity to recognize positives). This approach, based on one-class support vector machines (SVMs) with a conformal kernel, is described in Section 2 and its application to nosocomial infection detection is discussed in Section 3. Experiments conducted to assess this approach as well as results are described in Section 4.

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 850–858, 2004.
© Springer-Verlag Berlin Heidelberg 2004

2 Using Conformal Kernels in One-Class SVMs

2.1 One-Class Classification

While the majority of classification problems consist in discriminating between two or more classes, some other problems are best formulated as *one-class* or *novelty detection* problems. In a probabilistic sense, one-class classification is equivalent to deciding whether an unknown case has been produced by the distribution underlying the training set of normal cases. In one-class classification the classifier is trained exclusively on cases from the majority class and never sees those from the minority class. It must estimate the boundary that separates two classes and minimize misclassification based only on data lying on one side of it.

The one-class approach is particularly attractive in situations where cases from one class are expensive or difficult to obtain for model construction (i.e. imbalanced datasets). The most straigthforward method for detecting novel or abnormal cases is to estimate the density of the training data and to set a threshold on this density [4, 17]. However, it is much simpler to model the support of a data distribution, i.e., to create a binary-valued function which is positive in those regions of input space containing most of the data and negative elsewhere; the following section describes this approach.

2.2 One-Class Support Vector Machines

Support vector machines [18, 8] are learning machines based on the *Structural Risk Minimization principle* (SRM) from statistical learning theory. They were originally introduced to solve the two-class pattern recognition problem. An adaptation of the SVM methodology in order to handle classification problems using data from only one class has been proposed by [16]. This adapted method, termed one-class SVM, identifies "abnormal" cases amongst the known cases and assumes them to belong to the complement of the "normal" cases. Schölkopf et al. formulate the one-class SVM approach as follows:
Consider a training set $\mathcal{X} = \{\mathbf{x}_i\},\ i = 1, \ldots, n,\ \mathbf{x}_i \in \mathbb{R}^d$, and suppose its instances are distributed according to some unknown underlying probability distribution P. We want to know if a test example $\mathbf{x}$ is distributed according to P or not. This can be done by determining a region R of the input space X such that the probability that a test point drawn from P lies outside of R is bounded by some a priori specified value $\nu \in (0, 1)$. This problem is solved by estimating a decision function f which is positive on R and negative elsewhere.

$$f(\mathbf{x}) > 0 \text{ if } \mathbf{x} \in R \text{ and } f(\mathbf{x}) < 0 \text{ if } \mathbf{x} \in R \tag{1}$$

A non linear function $\Phi : \mathcal{X} \rightarrow \mathcal{F}$ maps vector $\mathbf{x}$ from the input vector space $\mathcal{X}$ endowed with an inner product to a Hilbert space $\mathcal{F}$ termed feature space. In this new space, the training vectors follow an underlying distribution P', and the problem is to determine a region R' of $\mathcal{F}$ that captures most of this probability mass distribution. In other words the region R' corresponds to the

part of the feature space where most of the data vectors lie. To separate as many as possible of the mapped vectors from the origin in feature space $\mathcal{F}$ we construct a hyperplane $H(\mathbf{w}, \rho)$ in a feature space $\mathcal{F}$ defined by

$$H(\mathbf{w}, \rho) : \langle \mathbf{x}, \Phi(\mathbf{x}) \rangle - \rho \tag{2}$$

with $\mathbf{w}$ the weight vector and ρ the offset, as illustrated in Fig 1.

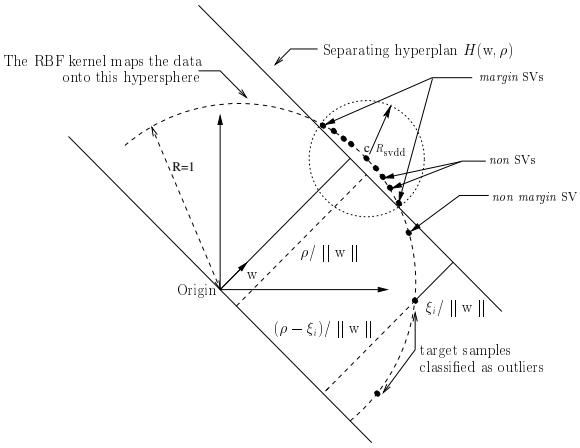

Fig. 1. Schematic 2D overview of a one-class SVM classifier with an RBF kernel. In the feature space, the vectors are located on a hypersphere. The hyperplane $H(\mathbf{w}, \rho)$ separates the training vectors from the rest of the surface of the hypersphere.

The maximum margin from the origin is found by solving the following quadratic optimization problem.

$$Minimize \quad \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle - \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i \tag{3}$$

$$subject\ to \quad \langle \mathbf{w}, \Phi(\mathbf{w}) \rangle \geq \rho - \xi_i, \quad \forall_i\ \xi_i \geq 0 \tag{4}$$

where ξ_i are so-called slack variables that penalize the objective function but allow some of the points to be on the wrong side of the hyperplane, i.e. located between the origin and $H(\mathbf{w}, \rho)$ as depicted in Fig.1. $\nu \in (0, 1)$ is a parameter that controls the trade off between maximizing the distance from the origin and containing most of the data in the region created by the hyperplane. It is proved in [16] that ν is an upper bound on the fraction of outliers i.e. training errors, and also a lower bound on the fraction of support vectors.

Let $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ be n non negative Lagrange multipliers associated with the constraints, the solution to the problem is equivalent to the solution of the Wolfe dual [11] problem:

$$Maximize \quad \frac{1}{2}\alpha_i\alpha_j\langle\Phi(\mathbf{x}_i),\Phi(\mathbf{x}_i)\rangle \tag{5}$$

$$subject\ to \quad 0 \leq \alpha_i \leq \frac{1}{\nu n}, \quad \sum_{i=1}^{n}\alpha_i = 1 \tag{6}$$

the solution for $\mathbf{w}$ is $\sum_{i=1}^{n}\Phi(\mathbf{x}_i)$ where $0 \leq \alpha_i \leq \frac{1}{\nu n}$ and the corresponding decision function is :

$$f(\mathbf{x}_j) = sgn\left(\sum_{i=1}^{n}\alpha_i\langle\Phi(\mathbf{x}_i),\Phi(\mathbf{x}_j)\rangle - \rho\right) \tag{7}$$

All training data vectors $\mathbf{x}_i$ for which $f(\mathbf{x}_i) \leq 0$ are called support vectors (SVs); these are the only vectors for which $\alpha_i \neq 0$. SVs are divided in two sets : the *margin* SVs, for which $f(\mathbf{x}_i) = 0$, and the *non-margin* SVs, for which $f(\mathbf{x}_i) < 0$. Notice that in (5) only inner products between data are considered; for certain particular maps $\mathcal{F}$, there is no need to actually compute $\Phi(\mathbf{x}_i)$ and $\Phi(\mathbf{x}_j)$; the inner product can be derived directly from $\mathbf{x}_i$ and $\mathbf{x}_j$ by means of the so-called "kernel trick". A kernel K is a symmetric function that fulfills Mercer's [18, 9] conditions. The main property of functions satisfying these conditions is that they implicitly define a mapping from $\mathcal{X}$ to a Hilbert space $\mathcal{F}$ such that

$$K(\mathbf{x}_i,\mathbf{x}_j) = \langle\Phi(\mathbf{x}_i),\Phi(\mathbf{x}_j)\rangle \tag{8}$$

and thus can be used in algorithms using inner products. Accordingly, the hyperplane (2) in feature space $\mathcal{F}$ becomes a non linear function in the input space $\mathcal{X}$.

$$f(\mathbf{x}) = sgn(\sum_{i=1}^{n}\alpha_i K(\mathbf{x}_i,\mathbf{x}) - \rho) \tag{9}$$

There are many admissible choices for the kernel function $K(\mathbf{x}_i,\mathbf{x}_j)$. The most widely used in one-class SVMs is the Gaussian Radial Basis Function RBF kernel:

$$K(\mathbf{x}_i,\mathbf{x}_j) = \exp^{-\|\mathbf{x}_i-\mathbf{x}_j\|^2/2\sigma^2} \tag{10}$$

where σ is a parameter which controls the width of the kernel function around $\mathbf{x}_i$. Since $\langle\Phi(\mathbf{x}_i),\Phi(\mathbf{x}_i)\rangle = K(\mathbf{x}_i,\mathbf{x}_i) = \exp^0 = 1$ with an RBF kernel, the training data in $\mathcal{F}$ lie on a region on the surface of a hypersphere centered at the origin of $\mathcal{X}$ with radius 1 as depicted in Fig. 1. Finally one has the decision function of Eq. (9) with $\rho = \sum_{i=1}^{n}\alpha_i K(\mathbf{x}_i,\mathbf{x}_j)$ for any $\mathbf{x}_j$ such that α_i satisfies $0 < \alpha_j < \frac{1}{\nu n}$ which defines the contour of the region R in input space by cutting the hypersurface defined by the weighted addition of SVM kernels at a given altitude ρ.

2.3 Accuracy Improvement

The accuracy of the one-class classifier can be improved by enhancing the resolution in the support vector region boundaries. One way to reach this goal is

via a *conformal transformation*[1] of the kernel. This approach has been described in the context of binary SVMs classifier by [2, 3], but the basic principle is also applicable to the one-class SVM. From a geometrical point of view the mapped data lie on a surface S in $\mathcal{F}$ with the same dimensionality as the input space $\mathcal{X}$ [6]. In the case of an RBF kernel function, the associated surface S in $\mathcal{F}$ can be considered as a Riemannian manifold [5] and a Riemannian metric thereby induced and expressed in the closed form in terms of the kernel [6, 2, 3]. A Riemannian metric, also called tensor, is a function which computes the intrinsic distance measured along the surface S itself between any two points lying on it. Its components can be viewed as multiplication factors which must be placed in front of the differential displacements dx_i in $\mathcal{X}$ to compute the distance ds of an element dz in $\mathcal{F}$ in a generalized Pythagorean theorem,

$$ds^2 = \sum_{i,j} g_{ij} dx_i dx_j \tag{11}$$

where g_{ij} is the induced metric, and the surface S is parametrized by the x_i. Let $\mathbf{x}$ be a point in $\mathcal{X}$ and $\mathbf{z}$ it corresponding mapping by Φ in $\mathcal{F}$. Letting $d\mathbf{x}$ represent a small but finite displacement, we have

$$\begin{aligned} ds^2 &= \|d\mathbf{z}\|^2 = \|\Phi(\mathbf{x} + d\mathbf{x}) - \Phi(\mathbf{x})\|^2 \\ &= K(\mathbf{x} + d\mathbf{x}, \mathbf{x} + d\mathbf{x}) - 2K(\mathbf{x}, \mathbf{x} + d\mathbf{x}) + K(\mathbf{x}, \mathbf{x}) \\ &= \sum_{i,j} \left(\frac{\partial^2 K(\mathbf{x}, \mathbf{y})}{\partial x_i \partial y_j} \right)_{\mathbf{y}=\mathbf{x}} dx_i dx_j \end{aligned}$$

From Eq. 11 we see that the Riemannian metric induced on S can be defined as

$$g_{ij} = \left(\frac{\partial^2 K(\mathbf{x}, \mathbf{z})}{\partial x_i \partial y_j} \right)_{\mathbf{y}=\mathbf{x}} \tag{12}$$

Note how a local area around $\mathbf{x}$ in $\mathcal{X}$ is magnified in $\mathcal{F}$ under the mapping $\Phi(\mathbf{x})$. The principle of conformal mapping is to increase the metric $g_{ij}(\mathbf{x})$ around the boundary and to reduce it everywhere else. To do this the non linear mapping Φ is modified in such a way that $g_{ij}(\mathbf{x})$ is enlarged around the boundary. This can be done by introducing a conformal transformation of the kernel [2, 3],

$$\tilde{K}(\mathbf{x}, \mathbf{y}) = c(\mathbf{x}) c(\mathbf{y}) K(\mathbf{x}, \mathbf{y}) \tag{13}$$

where $c(\mathbf{x})$ is a defined positive function. The modified kernel satisfies the Mercer positivity condition [9]. From Eq. 12, we obtain the new Riemannian metric $\tilde{g}_{ij}$

$$\tilde{g}_{ij} = c(\mathbf{x})^2 g_{ij}(\mathbf{x}) + c_i(\mathbf{x}) c_j(\mathbf{x}) + 2 c_i(\mathbf{x}) c(\mathbf{x}) K_i(\mathbf{x}, \mathbf{x}) \tag{14}$$

where $K_i(\mathbf{x}, \mathbf{x}) = \partial K(\mathbf{x}, \mathbf{y}) / \partial x_i |_{\mathbf{x}=\mathbf{y}}$ and $c_i(\mathbf{x}) = \partial c(\mathbf{x}) / \partial x_i$. For the Gaussian RBF kernel the last term is zero.

[1] A transformation that preserves the magnitude and orientation of the angle between any two curves intersecting at a given point is *conformal* at that point. A transformation is called conformal in a domain D if it is conformal at every point in D.

To expand the spatial resolution in the margin of a support vector $c(\mathbf{x})$ should be chosen in a way such that the metric $\tilde{g}_{ij}(\mathbf{x})$ has greater values around the decision boundary. However, in practice, we do not know where the boundary is, so an initial estimate is done by prior training of one-class SVMs.
A possible conformal transformal transformation $c(\mathbf{x})$ is

$$c(\mathbf{x}) = \sum_{i \in \hat{SV}} \hat{\alpha}_i e^{\frac{-\|\mathbf{x}-\mathbf{x}_i\|^2}{2\tau^2}} \tag{15}$$

where $\hat{SV}$ is the set of margin support vectors, $\hat{\alpha}_i$ is a positive number representing the contribution of the ith support vector, $\mathbf{x}_i$ is the ith support vector and τ is a free parameter; $\hat{\cdot}$ refers to a one-class SVM previously trained on the same dataset.

3 Application to Nosocomial Infection Detection

We tested the performance of one-class SVMs with a conformal kernel on a medical problem, the detection of nosomial infections. A nosocomial infection (from the Greek word *nosokomeion* for hospital) is an infection that develops during hospitalization whereas it was not present nor incubating at the time of the admission. Usually, a disease is considered a nosocomial infection if it develops 48 hours after admission.

The University Hospital of Geneva (HUG) has been performing yearly prevalence studies to detect and monitor nosocomial infections since 1994 [13]. Their methodology is as follows: the investigators visit every ward of the HUG over a period of approximately three weeks. All patients hospitalized for 48 hours or more at the time of the study are included. Medical records, kardex, X-ray and microbiology reports are reviewed, and additional information is eventually obtained by interviewing nurses or physicians in charge. Collected variables include demographic characteristics, admission date, admission diagnosis, comorbidities, McCabe score, type of admission, provenance, hospitalization ward, functional status, previous surgery, previous intensive care unit (ICU) stay, exposure to antibiotics, antacid and immunosuppressive drugs and invasive devices, laboratory values, temperature, date and site of infection, fulfilled criteria for infection.

The resulting 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. 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. This application was thus an excellent testbed for assessing the efficacy of one-class SVMs with a conformal kernel in the presence of data imbalance.

4 Experimentation

4.1 Evaluation Strategy

The experimental goal was to assess the impact of a conformal kernel on the the ability of one-class SVMs to cope with imbalanced datasets. To train one-class SVM classifiers we used an RBF kernel (Eq. (10)) and experimented with different values for the parameters ν and σ. Generalization error was estimated using 5-fold cross-validation (10-fold cross-validation would have resulted in an extremely small number of infected test cases per fold). The complete dataset was randomly partitioned into five subsets. On each iteration, one subset (comprising 20% of the data samples) was held out as a test set and the remaining four (80% of the data) were concatenated into a training set. The training sets consisted only of non infected patients whereas the test sets contained both infected and non infected patients according to the original class distribution. Error rates estimated on the test sets were then averaged over the five iterations. The following strategy was followed for conformal mapping:

1. Train one-class SVM with the primary RBF kernel K to get the SVs. Then change the kernel K according to Eq. 13,15.
2. Train one-class SVM with the modified kernel $\tilde{K}$.
3. Apply the above two steps (1. and 2.) until the best performances is attained.

For Eq. 15 we took $\tau = \sigma/\sqrt{n}$ which is the optimal value reported in [3].

4.2 Results

Table 1 summarizes performance results for one-class SVMs. It shows the best results obtained by training classifiers using different parameter configurations on non infected cases only. The last three columns show results based on three performance metrics. Accuracy is the percentage of correctly classified cases, sensitivity is the number of true positives over all positive cases, while specificity is the number of true negatives over all negative cases. Clearly, for both RBF and conformal kernels, highest sensitivity is attained when σ is very small: the system

Table 1. Performance of one-class SVMs for different parameter settings using (1) an RBF Gaussian kernel and (2) a conformal kernel.

One-class SVM	ν	σ	Accuracy %	Sensitivity %	Specificity %
RBF kernel	0.05	10^{-4}	74.6	92.6	43.73
	0.05	0.10	75.49	80.60	65.60
	0.2	10^{-4}	75.69	79.28	68.27
	0.2	0.10	74.36	74.67	72.27
Conformal kernel	0.05	10^{-4}	75.6	93.4	43.15
	0.05	0.1	76.65	82.35	64.1
	0.2	10^{-4}	77.3	81.1	69.7
	0.2	0.06	76.25	79.1	69.2

Table 2. Best performance of four SVM-based approaches to class imbalance.

SVM Classifier	Accuracy	Sensitivity	Specificity
Binary with symm. margin	89.6%	50.6%	94.4%
Binary with asymm. margin	74.4%	92%	72.2%
One-class with RBF kernel	74.6%	92.6%	43.73%
One-class with conformal kernel	75.6%	93.4%	43.15%

puts a Gaussian of narrow width around each data point and hence most of the infected test cases are correctly recognized as abnormal. The price is that many non infected cases are equally labelled abnormal, thus yielding low specificity. Larger values of σ in the RBF kernel are required to achieve tight approximations for the region R (non infected patients). Therefore the kernel parameter σ is crucial in determining the balance between normality and abnormality as there is no explicit penalty for false positive in one-class classification, contrary to the two class formulation [7]. Since the goal of this study is to identify infected cases, the solution retained is that which achieves maximal sensitivity.

In a previous study on the same nosocomial dataset [7], we investigated a support vector algorithm in which asymmetrical margins are tuned to improve recognition of rare positive cases. Table 2 compares the best performance measures obtained in previous and the latest experiments. Classical binary SVMs with a symmetrical margin attain a baseline sensitivity 50.6%; with the use of asymmetrical margins, sensitivity jumps to 92%. This is further improved by one-class SVMs with an RBF kernel (92.6%) and with a conformal kernel (93.4%). Note however that this progress in sensitivity comes at the cost of a corresponding decrease in specificity.

5 Conclusion and Future Work

We proposed one-class SVMs with a conformal kernel as a novel way of handling class imbalance in classification tasks. We showed that this approach achieves higher sensitivity than all SVM models previously applied to this problem. However, the price paid in terms of loss in specificity is quite exhorbitant, and domain experts must decide if the high recognition rate is worth the cost of treating false positive cases. From this point of view, asymmetrical-margin SVMs might prove preferable in that they maintain a more reasonable sensitivity-specificity tradeoff. In the near future, we intend to prospectively validate the classification model obtained by performing in parallel a standard prevalence survey. Overall we feel that one-class SVMs with a conformal kernel are a promising approach to the detection of nosocomial infections and can become a reliable component of an infection control system.

Acknowledgements

The authors thank members of the University of Geneva Hospitals – Profs. A. Geissbuhler and D. Pittet for their support and Drs. H. Sax and S. Hugonnet for the nosocomial infection data.

References

1. 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.
2. S. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.
3. S. Amari and S. Wu. An information-geometrical method for improving the performance of support vector machine classifiers. In *ICANN99*, pages 85–90, 1999.
4. C. Bishop. Novelty detection and neural network validation. *IEEE Proceedings on Vision, Image and Signal Processing*, 141(4):217–222, 1994.
5. W.M. Boothby. *An introduction to differential manifolds and Riemannian geometry.* Academic Press, Orlando, 1986.
6. C. Burges. Geometry and invariance in kernel based methods. In MIT Press, editor, *Adv. in kernel methods: Support vector learning*, 1999.
7. G. Cohen, M. Hilario, H. Sax, and S. Hugonnet. Asymmetrical margin approach to surveillance of nosocomial infections using support vector classification. In *Intelligent Data Analysis in Medicine and Pharmacology*, 2003.
8. C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20(3):273–297, September 1995.
9. N. Cristianini and Taylor J.S. *An Introduction to Support Vector Machines.* Cambridge University Press, 2000.
10. 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.
11. R. Fletcher. *Practical Methods of Optimization.* John Wiley and Sons, 1987.
12. 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.
13. S. Harbarth, Ch. Ruef, P. Francioli, A. Widmer, D. Pittet, and Swiss-Noso Network. Nosocomial infections in Swiss university hospitals: a multi-centre survey and review of the published experience. *Schweiz Med Wochenschr*, 129:1521–28, 1999.
14. N. Japkowicz. The class imbalance problem: A systematic study. *Intelligent Data Analysis Journal*, 6(5), 2002.
15. M. Kubat and S. Matwin. Addressing the curse of imbalanced data sets: One-sided sampling. In *Procsóf the Fourteenth International Conference on Machine Learning*, pages 179–186, 1997.
16. B. Scholkopf, R.C. Williamson, A.J. Smola, J. Shawe-Taylor, and J.Platt. Estimating the support of a high-dimensional distribution. In *Neural Computation*, volume 13, pages 1443–1471. MIT Press, 1999.
17. L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Proceedings of the 4th IEE International Conference on Artificial Neural Networks (ICANN'95)*, pages 442 – 447, 1995.
18. V. Vapnik. *Statistical Learning Theory.* Wiley, 1998.