APPLIED STOCHASTIC MODELS IN BUSINESS AND INDUSTRY
Appl. Stochastic Models Bus. Ind., 2005; **21**:199–214
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asmb.534

A comparative study of multi-class support vector machines in the unifying framework of large margin classifiers

Yann Guermeur[1,*,†], André Elisseeff[2,‡] and Dominique Zelus[3,§]

[1] *LORIA-CNRS, Campus Scientifique, BP 239 Vandœuvre-lès-Nancy Cedex 54506, France*
[2] *MPI for Biological Cybernetics, Spemannstraße 38, Tübingen 72076, Deutschland*
[3] *Wiener Lab, CIBIO, Av Presidente Juan D. Perón 2991, Rosario 2000, Argentina*

SUMMARY

Vapnik's statistical learning theory has mainly been developed for two types of problems: pattern recognition (computation of dichotomies) and regression (estimation of real-valued functions). Only in recent years has multi-class discriminant analysis been studied independently. Extending several standard results, among which a famous theorem by Bartlett, we have derived distribution-free uniform strong laws of large numbers devoted to multi-class large margin discriminant models. The capacity measure appearing in the confidence interval, a covering number, has been bounded from above in terms of a new generalized VC dimension. In this paper, the aforementioned theorems are applied to the architecture shared by all the multi-class SVMs proposed so far, which provides us with a simple theoretical framework to study them, compare their performance and design new machines. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: multi-class support vector machines (M-SVMs); generalization error bounds; large margin classifiers; extended VC dimensions

1. INTRODUCTION

One of the pioneering contributions to the study of the generalization capabilities of infinite sets of functions is the work of Vapnik and Chervonenkis [1] relating the consistency of the empirical risk minimization (ERM) inductive principle to the finiteness of a simple combinatorial quantity called the Vapnik–Chervonenkis dimension. Since then, the consistency of the ERM principle has been analysed in various contexts [2, 3]. Concomitantly, many studies have been devoted to deriving bounds on the expected risk (computing sample complexities), or implementing the structural risk minimization (SRM) inductive principle [4]. Among the main contributions are [5, 6]. However, the case of multi-class discrimination has seldom been studied independently

*Correspondence to: Yann Guermeur, LORIA-CNRS, Campus Scientifique, BP 239 Vandœuvre-lès-Nacy Cedex 54506, France.
[†] E-mail: yann.guermeur@loria.fr
[‡] E-mail: andre.elisseeff@tuebingen.mpg.de
[§] E-mail: dzelus@wiener-lab.com.ar

[7, 8], although we pointed out in Reference [9] the fact that it is inappropriate to tackle it in the straightforward manner, by plugging a generalized VC dimension as capacity measure in a standard uniform convergence bound. In Reference [10], we have extended to the multi-class case a lemma by Bartlett [6] expressing the sample complexity of pattern classification models in terms of a margin-based covering number. In Reference [11], a generalized Sauer's lemma has been established to bound such covering numbers in terms of a new variant of the VC dimension, the M-fat-shattering dimension. In this article, the corresponding guaranteed risk is computed for the multi-class support vector machines (M-SVMs). By M-SVMs, we mean the machines the architecture of which is a multivariate affine model, with as many hyperplanes as there are categories, i.e. those which perform the discrimination task in one step. This excludes all the standard decomposition schemes based on biclass SVMs. The reason for this restrictive choice is that the capacity of these latter architectures can be studied thanks to straightforward extensions of classical results [12]. Using our unifying framework makes it possible to justify *a posteriori* the principle of the training algorithms, which can thus be simply cast in the framework of the SRM inductive principle, compare performance and pave the way for the specification of new machines. The organization of the paper is as follows. Section 2 briefly summarizes our uniform convergence result. In Section 3, we explain the way the covering numbers of interest can be bounded in terms of the corresponding M-fat-shattering dimension. In Section 4, this bound is applied to the architecture shared by all the M-SVMs described in literature. At last, Section 5 deals with alternative possibilities to compute sample complexities.

2. GUARANTEED RISK FOR MULTI-CLASS DISCRIMINANT MODELS

We consider the case of a Q-category pattern recognition problem, where $Q \geqslant 3$. Let $\mathscr{X}$ be the space of description and $\mathscr{C}$ the set of categories. We make the assumption, standard in statistical learning theory, that there is a joint probability distribution P, fixed but unknown, on $\mathscr{X} \times \mathscr{C}$. Our goal is to find, in a given set $\mathscr{H}$ of functions $h = [h_k]$ from $\mathscr{X}$ into $\mathbb{R}^Q$, a function with the lowest 'error rate'. The 'error rate' of a function h is the error rate or *risk* of the corresponding discrimination function, obtained by assigning each pattern x to the category C_k in $\mathscr{C}$ satisfying: $h_k(x) = \max_l h_l(x)$. This discriminant function, hereafter denoted by f, must thus be as close as possible to Bayes' decision rule. In the common case where the outputs of the function selected are estimates of the class posterior probabilities, which happens for instance when $\mathscr{H}$ is the set of functions computed by a multi-layer perceptron and the training criterion has been adequately chosen (see, for instance, Reference [13]), applying this decision function simply amounts to implementing Bayes' estimated decision rule. Hereafter, $C(x_i)$ will denote indifferently the category of pattern x_i, or the index of this category. $\mathscr{Y} = \{y\}$ will be the set of canonical codings of the categories in $\{-1, 1\}^Q$ vectors. The uniform convergence result we established is based on the following definitions.

Definition 1 (Expected risk)

The *expected risk* of a function f from $\mathscr{X}$ into $\mathscr{C}$ is the probability that $f(x) \neq C(x)$ for a labelled example $(x, C(x))$ chosen randomly according to P, i.e.:

$$R(f) = \int_{\mathscr{X} \times \mathscr{C}} \mathbf{I}_{\{f(x) \neq C\}} \, \mathrm{d}P(x, C) \tag{1}$$

where the indicator function $\mathbf{I}_{\{f(x)\neq C\}}$ is defined as follows: $\mathbf{I}_{\{f(x)\neq C\}} = 1$ if $f(x)\neq C$, $\mathbf{I}_{\{f(x)\neq C\}} = 0$ otherwise.

Definition 2 (Empirical risk)

Let $s_m = \{(x_i, C(x_i))\} \in (\mathscr{X} \times \mathscr{C})^m$. The *empirical risk* of f on s_m is defined as

$$R_{s_m}(f) = \frac{1}{m}|\{(x_i, C(x_i)) \in s_m / f(x_i) \neq C(x_i)\}| \tag{2}$$

As stated above, the expected risk (resp. empirical risk) of a function h from $\mathscr{X}$ to $\mathbb{R}^Q$ is the expected risk (resp. empirical risk) of the corresponding discriminant function f.

Definition 3 (ε-cover or ε-net)

Let (E, ρ) be a pseudo-metric space, and $B(v, r)$ the closed ball of centre v and radius r in E. Let H be a subset of E. An ε-cover of H is a subset $\bar{H}$ of E such that

$$H \subset \bigcup_{v\in\bar{H}} B(v, \varepsilon)$$

Definition 4 (Covering numbers)

Let (E, ρ) be a pseudo-metric space. If $H \subset E$ has an ε-cover of finite cardinality, then its *covering number* $\mathscr{N}(\varepsilon, H, \rho)$ is the smallest cardinality of its ε-covers. If there is no such finite cover, then the covering number is defined to be ∞.

Definition 5

Let $\mathscr{H}$ be a set of functions from $\mathscr{X}$ into $\mathbb{R}^Q$. For a set s of points in $\mathscr{X}$, define the pseudo-metric $d_{l_\infty, l_\infty(s)}$ on $\mathscr{H}$ as

$$\forall (h, \bar{h}) \in \mathscr{H}^2, d_{l_\infty, l_\infty(s)}(h, \bar{h}) = \max_{x\in s} \max_{k\in\{1,\ldots,Q\}} |h_k(x) - \bar{h}_k(x)|$$

Definition 6 (Delta operator)

Let $h = [h_k]$ be a function from $\mathscr{H}$. Define $\Delta h = [\Delta h_k]$, $(1 \leqslant k \leqslant Q)$, as the function from $\mathscr{X}$ into $\mathbb{R}^Q$ given by

$$\forall k \in \{1, \ldots, Q\}, \quad \Delta h_k(x) = \frac{1}{2}\left\{h_k(x) - \max_{l\neq k} h_l(x)\right\}$$

Extending a definition from Bartlett [6], we introduced in Reference [10] the following definition:

Definition 7 (Empirical margin risk)

The empirical risk with margin $\gamma \in (0, 1]$ of h on a set s_m of size m is

$$R^{\gamma}_{s_m}(h) = \frac{1}{m}|\{(x_i, C(x_i)) \in s_m / \Delta h_{C(x_i)}(x_i) < \gamma\}|$$

For $\gamma \in (0,1]$, let $\pi_\gamma : \mathbb{R} \to [-\gamma, \gamma]$ be the piecewise-linear squashing function defined as

$$\pi_\gamma(x) = \begin{cases} \gamma \operatorname{sign}(x) & \text{if } |x| \geqslant \gamma \\ x & \text{otherwise} \end{cases}$$

$\forall h \in \mathscr{H},\ \Delta h^\gamma = [\Delta h_k^\gamma] = [\pi_\gamma \circ \Delta h_k],\ (1 \leqslant k \leqslant Q).\ \Delta\mathscr{H}^\gamma = \{\Delta h^\gamma / h \in \mathscr{H}\}.$ Let $\mathscr{N}_{\infty,\infty}(\varepsilon, \Delta\mathscr{H}^\gamma, m) = \max_{s_m \in \mathscr{X}^m} \mathscr{N}(\varepsilon, \Delta\mathscr{H}^\gamma, d_{l_\infty, l_\infty(s_m)})$. With these hypotheses and definitions at hand, extending Lemma 4 and Corollary 9 from Reference [6], as well as the basic lemma of Theorem 4.1 in Reference [14], we established in Reference [11] (see also Reference [10]) the following theorem:

Theorem 1

Let s_m be an m-sample of examples drawn independently from P. With probability at least $1 - \delta$, for every value of γ in $(0,1]$, the risk $R(h)$ of a function h computed by a numerical Q-class discriminant model $\mathscr{H}$ is bounded above by

$$R(h) \leqslant R_{s_m}^\gamma(h) + \sqrt{\frac{2}{m}\left(\ln(2\mathscr{N}_{\infty,\infty}(\gamma/4, \Delta\mathscr{H}^\gamma, 2m)) + \ln\left(\frac{2}{\gamma^\delta}\right)\right)} + \frac{1}{m} \tag{3}$$

3. COVERING NUMBERS AND EXTENDED FAT-SHATTERING/ GRAPH DIMENSION

In this section, the covering numbers of interest are bounded using the strategy advocated in Reference [6]. The bound springs from the extension of several lemmas in Reference [3] to the case of vector-valued functions. It involves an original capacity measure, the M-fat-shattering dimension, which is concomitantly an extension of the fat-shattering dimension to the multivariate case and a scale-sensitive variant of the graph dimension.

3.1. Definitions

To formulate the bound, and pave the way for the next section, we must first introduce some definitions.

Definition 8 (Growth function [1])

Let $\mathscr{F}$ be a set of indicator (binary-valued) functions of a set $\mathscr{X}$. Let $\Pi_\mathscr{F}$ be the function which maps any set s of points in $\mathscr{X}$ to the number of dichotomies $\Pi_\mathscr{F}(s)$ computed on it by the functions in $\mathscr{F}$. Then, the *growth function* of $\mathscr{F}, G_\mathscr{F}$, is the function from the non-negative integers to the non-negative integers given by

$$G_\mathscr{F}(m) = \max_{s_m \in \mathscr{X}^m} \Pi_\mathscr{F}(s_m)$$

Definition 9 (Vapnik–Chervonenkis (VC) dimension [1])

Let $\mathscr{F}$ be a set of indicator functions on a set $\mathscr{X}$. A subset s_m of $\mathscr{X}^m$ is said to be *shattered* by $\mathscr{F}$ if $\Pi_{\mathscr{F}}(s_m) = 2^m$, i.e. if each dichotomy on s_m is computed by a function of $\mathscr{F}$. The *VC dimension* of $\mathscr{F}$, VC-dim($\mathscr{F}$), is the largest value of m such that $G_{\mathscr{F}}(m) = 2^m$, if this value is finite, of infinity otherwise. If the VC dimension is finite, it is thus the size of the largest set of points shattered by $\mathscr{F}$.

Pollard's pseudo-dimension extends the notion of VC dimension to the case of sets of real-valued functions.

Definition 10 (Pollard's pseudo-dimension [15, 16])

Let $\mathscr{H}$ be a set of real-valued functions on a set $\mathscr{X}$. A subset $s_m = \{x_i\}$, $(1 \leqslant i \leqslant m)$ of $\mathscr{X}$ is said to be *P-shattered* by $\mathscr{H}$ if there is a vector $v_b = [b_i] \in \mathbb{R}^m$ such that, for each binary vector $v_y = [y_i] \in \{-1, 1\}^m$, there is a function $h_y \in \mathscr{H}$ satisfying

$$\forall i \in \{1, \ldots, m\} \begin{cases} h_y(x_i) - b_i \geqslant 0 & \text{if } y_i = 1 \\ h_y(x_i) - b_i < 0 & \text{if } y_i = -1 \end{cases}$$

The *P-dimension* of $\mathscr{H}$, P-dim($\mathscr{H}$), is the maximal cardinality of a subset of $\mathscr{X}$ P-shattered by $\mathscr{H}$, if it is finite, or infinity otherwise.

The fat-shattering dimension of Kearns and Schapire is a scale-sensitive version of the pseudo-dimension.

Definition 11 (Fat-shattering dimension [17, 18])

Let $\mathscr{H}$ be a set of real-valued functions on a set $\mathscr{X}$. For $\gamma > 0$, a subset $s_m = \{x_i\}$, $(1 \leqslant i \leqslant m)$ of $\mathscr{X}$ is said to be γ-*shattered* by $\mathscr{H}$ if there is a vector $v_b = [b_i] \in \mathbb{R}^m$ such that, for each binary vector $v_y = [y_i] \in \{-1, 1\}^m$, there is a function $h_y \in \mathscr{H}$ satisfying

$$(h_y(x_i) - b_i) y_i \geqslant \gamma \quad (1 \leqslant i \leqslant m)$$

The vector v_b is then said to *witness* the γ-shattering of s_m by $\mathscr{H}$. The *fat-shattering dimension* of the set $\mathscr{H}$, $\text{fat}_{\mathscr{H}}$, is a function from the positive real numbers to the integers which maps a value γ to the size of the largest set γ-shattered by functions of $\mathscr{H}$, if this size is finite, or to infinity otherwise.

We propose to extend this definition to the case of vector-valued functions in the following manner.

Definition 12 (M-fat-shattering dimension)

Let $\mathscr{H}$ be a set of functions on a set $\mathscr{X}$ taking their values in $\mathbb{R}^Q$. For $\gamma > 0$, a subset $s_m = \{x_i\}$, $(1 \leqslant i \leqslant m)$ of $\mathscr{X}$ is said to be M-γ-*shattered* by $\mathscr{H}$ if there is a vector $v_b = [b_i] \in \mathbb{R}^m$ and a vector $v_c = [c_i] \in \{1, \ldots, Q\}^m$ such that, for each binary vector $v_y = [y_i] \in \{-1, 1\}^m$, there is a function $h_y = [h_{yk}]$, $(1 \leqslant k \leqslant Q) \in \mathscr{H}$ satisfying

$$(h_{yc_i}(x_i) - b_i) y_i \geqslant \gamma, \quad (1 \leqslant i \leqslant m)$$

The couple (v_b, v_c) is then said to *witness* the M-γ-shattering of s_m by $\mathscr{H}$. The *M-fat-shattering dimension* of the set $\mathscr{H}$, $M\text{-fat}_{\mathscr{H}}$, is a function from the positive real numbers to the integers which maps a value γ to the size of the largest set M-γ-shattered by functions of $\mathscr{H}$, if this size is finite, or to infinity otherwise.

In References [3, 19], the authors define the V_γ dimension, denoted V_γ-dim, as a simplified variant of the fat-shattering dimension in which all the components of vector v_b are required to be equal. Taking our inspiration from this example, we introduce the uniform M-fat-shattering dimension, which will prove useful in the subsequent computations.

Definition 13 (Uniform M-fat-shattering dimension)

Let $\mathscr{H}$ be a set of functions on a set $\mathscr{X}$ taking their values in $\mathbb{R}^Q$. For $\gamma > 0$, the *uniform M-fat-shattering dimension* of $\mathscr{H}$, $UM\text{-fat}_{\mathscr{H}}$, is simply $M\text{-fat}_{\mathscr{H}}$ in the case where the components of vector v_b are allowed to take only Q different values, one for each category. In other words, if two components of the vector v_c are equal, then the corresponding components of the vector v_b are also equal.

Bounding the M-fat-shattering dimension in terms of the uniform M-fat-shattering dimension and conversely is easily performed thanks to the pigeonhole principle (see, for instance, Reference [11, Lemma 6]).

As stated in the introduction of the section, the M-fat-shattering dimension can be seen alternatively as a straightforward scale-sensitive extension of the graph dimension, introduced independently in References [20, 21] (see also Reference [7]).

Definition 14 (Graph dimension)

Let $\mathscr{F}$ be a set of functions on a set $\mathscr{X}$ taking their values in a countable set. For any $f \in \mathscr{F}$, the *graph* $\mathscr{G}$ of f is $\mathscr{G}(f) = \{(x, f(x))/x \in \mathscr{X}\}$ and the *graph space* of $\mathscr{F}$ is $\mathscr{G}(\mathscr{F}) = \{\mathscr{G}(f)/f \in \mathscr{F}\}$. Then the *graph dimension* of $\mathscr{F}$, $G\text{-dim}(\mathscr{F})$, is defined to be the VC dimension of the space $\mathscr{G}(\mathscr{F})$.

In the context of our study, the most natural way to handle this dimension consists in making use of the general scheme developed in Reference [2], which rests on the notion of Ψ-dimension.

Definition 15 (Ψ-shattering)

Let $\mathscr{F}$ be a set of functions on a set $\mathscr{X}$ taking their values in the finite set $\{1, \ldots, Q\}$. Let Ψ be a family of mappings ψ from $\{1, \ldots, Q\}$ into $\{-1, 1, *\}$, where $*$ is thought of as a null element. A subset $s_m = \{x_i\}$, $(1 \leqslant i \leqslant m)$ of $\mathscr{X}$ is said to be Ψ-*shattered* by $\mathscr{F}$ if there is a mapping $\psi^m = \{\psi^{(1)}, \ldots, \psi^{(m)}\}$ in Ψ^m such that for each vector v_y of $\{-1, 1\}^m$, there is a function f_y in $\mathscr{F}$ satisfying $[\psi^{(1)} \circ f_y(x_1), \ldots, \psi^{(i)} \circ f_y(x_i), \ldots, \psi^{(m)} \circ f_y(x_m)]^{\mathrm{T}} = v_y$.

Definition 16 (Ψ-dimension)

Let $\mathscr{F}$ and Ψ be defined as above. The Ψ-*dimension* of $\mathscr{F}$, denoted by $\Psi\text{-dim}(\mathscr{F})$, is the maximal cardinality of a subset of $\mathscr{X}$ Ψ-shattered by $\mathscr{F}$, if it is finite, or infinity otherwise.

When the functions in $\mathscr{F}$ have a finite range, the graph dimension appears as a particular case of Ψ-dimension, as can be seen with the following alternative definition.

Definition 17 (Graph dimension)

Let $\mathscr{F}$ be a defined as above. The graph dimension of $\mathscr{F}$ is the Ψ-dimension of $\mathscr{F}$ in the specific case where Ψ is the set of Q mappings ψ_k, $(1 \leqslant k \leqslant Q)$, such that ψ_k takes the value 1 if its argument is equal to k, and the value -1 otherwise. Reformulated in the context of multi-class discriminant analysis, the functions ψ_k are the indicator functions of the categories.

To understand the way the M-fat-shattering dimension can be seen as a scale-sensitive extension of the graph dimension, suffice it to notice two things, which are expressed here, for the sake of simplicity, and without loss of generality, in the restricted framework of multi-class discriminant analysis. First, the functions f involved in the definition of the graph dimension can be seen as the discriminant functions associated with the multivariate functions h, by application of the 'max rule' defined in the beginning of Section 2. Second, the choice of the vector v_c plays in the case of the M-fat-shattering dimension the role played by the choice of the set of mappings ψ^m in the case of the graph dimension. To sum up, the M-fat-shattering dimension is related to the fat-shattering dimension through the parameters γ and v_b, which deal with the margin, whereas it is related to the graph dimension through the vector v_c, which aims at focusing, for each of the points considered, on the behaviour of one specific component of the vector-valued function. From a computational point of view, the following theorem can be used to reformulate the problem of bounding the M-fat-shattering dimension of a set of vector-valued functions in terms of the fat-shattering dimensions of the sets of real-valued functions corresponding to the components considered separately.

Theorem 2

Let $\mathscr{H}$ be a set of vector-valued functions $h = [h_k]$, $(1 \leqslant k \leqslant Q)$, from a set $\mathscr{X}$ into $\mathbb{R}^Q$. Let $\mathscr{H}_k$, $(1 \leqslant k \leqslant Q)$, be the sets of real-valued functions h_k corresponding to the different components of the functions h. Then, for $\gamma > 0$, the following bound holds true:

$$M\text{-fat}_{\mathscr{H}}(\gamma) \leqslant \sum_{k=1}^{Q} \text{fat}_{\mathscr{H}_k}(\gamma) \tag{4}$$

Proof

Let $s_m = \{x_1, \ldots, x_i, \ldots, x_m\}$ be a subset of $\mathscr{X}$ of cardinality $m = M\text{-fat}_{\mathscr{H}}(\gamma)$ M-γ-shattered by $\mathscr{H}$. Let the couple (v_b, v_c) witness this M-γ-shattering. Let m_k be the number of components of v_c equal to k and let $s(k) = \{x_{\sigma(1)}, \ldots, x_{\sigma(i)}, \ldots, x_{\sigma(m_k)}\}$ be the corresponding set of examples in s_m. According to the definition of the M-fat-shattering dimension, for each binary vector $v_y = [y_i] \in \{-1, 1\}^m$, there is a function $h_y = [h_{yk}]$, $(1 \leqslant k \leqslant Q) \in \mathscr{H}$ satisfying

$$(h_{yc_i}(x_i) - b_i)y_i \geqslant \gamma, \quad (1 \leqslant i \leqslant m)$$

and thus, more specifically

$$(h_{yk}(x_{\sigma(i)}) - b_{\sigma(i)})y_{\sigma(i)} \geqslant \gamma, \quad (1 \leqslant i \leqslant m_k)$$

Since by construction all the real-valued functions h_{yk} belong to $\mathscr{H}_k$, it springs from Definition 11 that $\mathscr{H}_k$ γ-shatters $s(k)$ and, by way of consequence, $m_k \leqslant \mathrm{fat}_{\mathscr{H}_k}(\gamma)$. Summing over the index k concludes the proof. □

In the case of discriminant analysis, the use of the max rule implies that the quantity of interest is not the M-fat-shattering dimension of $\mathscr{H}$, or $\mathscr{H}^\gamma$, but the (uniform) M-fat-shattering dimension of $\Delta\mathscr{H}^\gamma$.

The following generalization of Sauer's lemma was proved in Reference [11].

Theorem 3

Let $\mathscr{H}$ be a set of functions from $\mathscr{X}$ into $\mathbb{R}^Q$. For every value of γ in (0,1] and every value of m satisfying $M\text{-fat}_{\Delta\mathscr{H}^\gamma}(\gamma/8) < 2m$, the following bound is true:

$$\mathscr{N}_{\infty,\infty}(\gamma/2, \Delta\mathscr{H}^\gamma, 2m) \leqslant 2(2mQ9^Q)^{d \log_2(18emQ/d)} \tag{5}$$

where $d = M\text{-fat}_{\Delta\mathscr{H}^\gamma}(\gamma/8)$.

From this theorem, it springs that the problem of bounding the covering numbers of interest (appearing in the confidence interval of (3)), can actually be reduced to the problem of bounding the M-fat-shattering dimension of $\Delta\mathscr{H}^\gamma$. Note that we have used the hypothesis that twice the size of the sample available was superior to the extended VC dimension considered, in the sole aim to highlight the fact that if this hypothesis is not satisfied, then different (simpler) results can be derived, giving birth to tighter bounds.

4. BOUNDS ON ERROR EXPECTATION FOR M-SVMS

4.1. Pattern recognition SVMs

Support vector machines (SVMs) are learning systems which have been introduced by Vapnik and co-workers [22, 23] as a non-linear extension of the maximal margin hyperplane [4]. Originally, they were designed to perform pattern recognition (compute dichotomies). In this context, the principle on which they are based is very simple. First, the examples are mapped into a high-dimensional Hilbert space called the *feature space* thanks to a non-linear transform, usually denoted by Φ. Second, the maximal margin hyperplane is computed in that space, to separate the two categories.

4.2. Architecture and training of the M-SVMs

The problem of performing multi-class discriminant analysis with SVMs was initially tackled through decomposition schemes [14, 24, 25]. Only later came the multi-class SVMs obtained by combining a multivariate affine model with the non-linear mapping Φ into the feature space. Formally, the family $\mathscr{H}$ of functions $h = [h_k]$ computed by these machines is defined by

$$\forall k \in \{1, \ldots, Q\}, \quad h_k(x) = \langle w_k, \Phi(x) \rangle + b_k \tag{6}$$

As usual, the mapping Φ does not appear explicitly in the computations. Thanks to the 'kernel trick', it is replaced with the *reproducing kernel function* K, which computes the l_2 dot product in

the feature space, i.e.

$$\forall (x^{(1)}, x^{(2)}) \in \mathcal{X}^2, \quad K(x^{(1)}, x^{(2)}) = \langle \Phi(x^{(1)}), \Phi(x^{(2)}) \rangle \tag{7}$$

Hence, the 'linear part' of each component of the model is a function of x belonging to a reproducing kernel Hilbert space (RKHS) (see, for instance, References [26–28]). The kernel satisfies Mercer's conditions [29].

In its primal formulation, training thus consists in finding the values of the vectors w_k and the reals b_k. This amounts to solving a quadratic programming (QP) problem. The different M-SVMs published differ in the nature of this problem. For the sake of completeness, we detail them below. The first multi-class SVM published was proposed independently and under different forms by several teams (see, for instance, References [14, 30–32]).

Problem 1 (*M-SVM*1):

$$\begin{aligned} &\min_{h \in \mathcal{H}} \quad \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik} \right\} \\ &\text{s.t.} \quad \begin{cases} \langle w_{C(x_i)} - w_k, \Phi(x_i) \rangle + b_{C(x_i)} - b_k \geqslant 1 - \xi_{ik} & (1 \leqslant i \leqslant m),\ (1 \leqslant k \neq C(x_i) \leqslant Q) \\ \xi_{ik} \geqslant 0 & (1 \leqslant i \leqslant m),\ (1 \leqslant k \neq C(x_i) \leqslant Q) \end{cases} \end{aligned}$$

A variant of this machine can be found in Reference [33]. The model described in Reference [34] (see also Reference [35]), makes an original use of the empirical margin risk in the objective function.

Problem 2 (*M-SVM*2):

$$\begin{aligned} &\min_{h \in \mathcal{H}} \quad \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \xi_i \right\} \\ &\text{s.t.} \quad \langle w_{C(x_i)} - w_k, \Phi(x_i) \rangle + b_{C(x_i)} - b_k + \delta_{C(x_i),k} \geqslant 1 - \xi_i, \quad (1 \leqslant i \leqslant m),\ (1 \leqslant k \leqslant Q) \end{aligned}$$

The bound on the generalization error provided is directly borrowed from a tree-based decomposition approach called DAGSVM [12]. In References [36, 37], the machine is devised to asymptotically implement the Bayes rule.

Problem 3 (*M-SVM*3):

$$\begin{aligned} &\min_{h \in \mathcal{H}} \quad \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik} \right\} \\ &\text{s.t.} \quad \begin{cases} \langle w_k, \Phi(x_i) \rangle + b_k \leqslant -1/(Q-1) + \xi_{ik} & (1 \leqslant i \leqslant m),\ (1 \leqslant k \neq C(x_i) \leqslant Q) \\ \xi_{ik} \geqslant 0 & (1 \leqslant i \leqslant m),\ (1 \leqslant k \neq C(x_i) \leqslant Q) \\ \sum_{k=1}^{Q} w_k = 0, \ \sum_{k=1}^{Q} b_k = 0 \end{cases} \end{aligned}$$

At last, we evaluated in References [38, 39], as ensemble method, an M-SVM the specification of which resulted form our early works on uniform convergence bounds [10].

Problem 4 (*M-SVM*4):

$$\min_{h\in\mathscr{H}} \quad \left\{\frac{1}{2}t^2 + C\sum_{i=1}^{m}\sum_{k=1}^{Q}\xi_{ik}\right\}$$

$$\text{s.t.} \quad \begin{cases} \|w_k - w_l\|^2 \leqslant t^2, \ (1 \leqslant k < l \leqslant Q) \\ \text{Constraints of Problem 1} \\ \sum_{k=1}^{Q} w_k = 0 \end{cases}$$

For both theoretical and technical reasons, linked for instance to the use of the kernel trick, the QP problems above are solved in their Wolfe dual form [40]. The corresponding representer theorems all involve the same functional expression of the optimal solution. The parameters to be optimized are the coefficients $\beta_{i,k}$ appearing in the following expansions:

$$\forall k \in \{1,\ldots,Q\}, \quad h_k(k) = \sum_{i=1}^{m} \beta_{i,k} K(x_i, x) + b_k \tag{8}$$

where the x_i, $(1 \leqslant i \leqslant m)$ are the covariates of the points in the training set.

4.3. Motivations and hypotheses

In view of the summary made in the preceding subsection, the bound on the error expectation of M-SVMs, the subject of the whole section, can be obtained by bounding the confidence interval of Theorem 1 in the particular case of kernel machines taking their values in $\mathbb{R}^Q$. Owing to Theorem 3, this can boil down to studying the behaviour of the corresponding M-fat-shattering dimension as a function of the constraints on the parameters and the nature of the kernel. In performing this task, our goal is primarily to study in a common framework the existing training algorithms and make it possible to specify new ones, as an implementation of the (*data-dependent*) SRM inductive principle [5]. With this aim in mind, we do not attempt to establish the tightest possible bound, or even to present a single master theorem. We simply sketch a straightforward pathway highlighting the dependence of the capacity measure on the penalty terms appearing in the objective functions of the different training algorithms.

We make no specific hypothesis regarding the set $\mathscr{X}$ of covariates. On the contrary, the feature space $E_{\Phi(\mathscr{X})}$ is assumed to be a Hilbert space endowed with the l_2 dot product. This standard hypothesis is a prerequisite to compute linear boundaries. $E_{\Phi(\mathscr{X})}$ can be infinite dimensional, so that no restriction is induced on the nature of the kernel used, which can for instance be Gaussian. Furthermore, $\Phi(\mathscr{X})$ is supposed to be bounded in $E_{\Phi(\mathscr{X})}$, which will be needed to bound the M-fat-shattering dimension.

4.4. Uniform M-fat-shattering dimension of M-SVMs

The M-fat-shattering dimension has been defined as a straightforward extension of the fat-shattering dimension to the multivariate case. As a consequence, its use is relevant for any kind of model taking its values in $\mathbb{R}^Q$, not only discriminant ones. However, in the specific case of multi-class supervised learning, the quantity of interest is primarily the difference between the scores associated with the different labels. More precisely, the difference between the output corresponding to the category of the example and the second highest output must be as large as possible. Thus, the degree of freedom provided by the vector v_b appearing in the definition of the

fat-shattering dimension and its variants does not seem useful but to cope with pathological situations, such as a sampling bias [37]. We do not address this type of problems here. Consequently, in what follows, the study deals with the uniform M-fat-shattering dimension of the M-SVMs, or more precisely the V_γ dimension of the corresponding sets of functions $\Delta\mathscr{H}_k$, computed under the additional constraint $v_b = 0$. This is appropriate indeed, since the adaptation of Theorem 3 to this specific situation rises no difficulty. For lack of place, details are omitted here. For the sake of simplicity, we also consider a multivariate linear model instead of the affine architecture described in Section 4.2. The connection between the capacities of these two architectures is characterized in Reference [39], Theorem 5.

The following theorem provides us with the desired bound:

Theorem 4

Let $\mathscr{H}$ be the set of functions h computed by the linear variant of the *M-SVM* architecture described in Section 4.2 ($\forall k \in \{1, \ldots, Q\}$, $b_k = 0$). Suppose that $\Phi(\mathscr{X})$ is included in the ball of radius $\Lambda_{\Phi(\mathscr{X})}$ in $E_{\Phi(\mathscr{X})}$ and that the vectors w_k in (6) satisfy the constraint: $\max_{1 \leqslant k < l \leqslant Q} \|w_k - w_l\|_2 \leqslant \Lambda_w$. Then, for all couple (γ, ε) such that $0 < \varepsilon < \gamma$,

$$UM\text{-fat}_{\Delta\mathscr{H}^\gamma}(\varepsilon) \leqslant Qm \tag{9}$$

where m is the largest integer satisfying

$$\frac{Qm}{(Q-1)\sqrt{m} + \frac{Q-2}{2}m} \leqslant \frac{\max_{1 \leqslant k < l \leqslant Q} \|w_k - w_l\| \Lambda_{\Phi(\mathscr{X})}}{\varepsilon} \tag{10}$$

Proof

To get rid of the γ parameter, one can make use of the following result, the proof of which is trivial:

$$\forall (\gamma, \varepsilon)/0 < \varepsilon < \gamma, \quad \text{UM-fat}_{\Delta\mathscr{H}^\gamma}(\varepsilon) \leqslant \text{UM-fat}_{\Delta\mathscr{H}}(\varepsilon) \tag{11}$$

With this intermediate result at hand, the rest of the proof is inspired form the proof of Theorem 4.6 in Reference [41] (see also Reference [19, Theorem 2]). More precisely, we make use of the following lemma:

Lemma 1 (Lemma 4.3 in Reference [41])

Let $s_m = \{x_i\}$, $(1 \leqslant i \leqslant m)$, be a set of vectors included in the ball of radius $\Lambda_{\mathscr{X}}$ in a Hilbert space $E_{\mathscr{X}}$. Then s_m can be split into two subsets s_m^+ and s_m^- such that

$$\left\| \sum_{x_i \in s_m^+} x_i - \sum_{x_j \in s_m^-} x_j \right\| \leqslant \sqrt{m} \Lambda_{\mathscr{X}} \tag{12}$$

With the notations of Theorem 2, we define $\Delta\mathscr{H}_k$ as the set of functions Δh_k such that $h = [h_l]$, $(1 \leqslant l \leqslant Q)$ belongs to $\mathscr{H}$. From Theorem 2, we obtain

$$\text{UM-fat}_{\Delta\mathscr{H}}(\varepsilon) \leqslant \sum_{k=1}^{Q} V_\varepsilon\text{-dim}(\Delta\mathscr{H}_k) \tag{13}$$

($\text{fat}_{\Delta\mathscr{H}_k}(\varepsilon)$ is identically equal to V_ε-dim$(\Delta\mathscr{H}_k)$ when $v_b = 0$).

Let $s(k) = \{x_1, \ldots, x_i, \ldots, x_{m_k}\}$ be a subset of $\mathscr{X}$ of cardinality $m_k = V_\varepsilon$-dim$(\Delta\mathscr{H}_k)$, V_ε-shattered by $\Delta\mathscr{H}_k$ (with the bias equal to 0). According to Lemma 1, $s(k)$ can be split into two subsets $s(k)^+$ and $s(k)^-$ such that

$$\left\| \sum_{x_i \in s(k)^+} \Phi(x_i) - \sum_{x_j \in s(k)^-} \Phi(x_j) \right\| \leqslant \sqrt{m_k}\Lambda_{\Phi(\mathscr{X})} \tag{14}$$

Let $m_k^+ = |s(k)^+|$ and $m_k^- = |s(k)^-|$. Without loss of generality, we make the additional hypothesis that $m_k^+ \geqslant m_k^-$. Since $s(k)$ is V_ε-shattered by $\Delta\mathscr{H}_k$, there exists a function h in $\mathscr{H}$ such that $\Delta h_k(x_i) \geqslant \varepsilon$ if x_i belongs to $s(k)^+$ and $\Delta h_k(x_i) \leqslant -\varepsilon$ otherwise. Let this function be defined by the vectors w_l, $(1 \leqslant l \leqslant Q)$. By definition, we obtain

$$\forall x_i \in s(k)^+, \quad \forall l \in \{1, \ldots, Q\} \backslash \{k\}, \quad \langle w_k - w_l, \Phi(x_i) \rangle \geqslant 2\varepsilon \tag{15}$$

According to the pigeonhole principle, there is at least one index of category, say l, such that there is a subset $s(k,l)$ of $s(k)^-$ of cardinality m_{kl} at least equal to $\lceil m_k^- / (Q-1) \rceil$ and satisfying

$$\forall x_i \in s(k,l), \quad \langle w_k - w_l, \Phi(x_i) \rangle \leqslant -2\varepsilon \tag{16}$$

Combining (15) and (16) gives

$$\left\langle w_k - w_l, \sum_{x_i \in s(k)^+} \Phi(x_i) - \sum_{x_j \in s(k,l)} \Phi(x_j) \right\rangle \geqslant 2(m_k^+ + m_{kl})\varepsilon \tag{17}$$

By the Cauchy–Schwarz inequality, (17) implies

$$2(m_k^+ + m_{kl})\varepsilon \leqslant \|w_k - w_l\| \left\| \sum_{x_i \in s(k)^+} \Phi(x_i) - \sum_{x_j \in s(k,l)} \Phi(x_j) \right\| \tag{18}$$

The right-hand side of (18) can be bounded thanks to the triangular inequality

$$\begin{aligned} \left\| \sum_{x_i \in s(k)^+} \Phi(x_i) - \sum_{x_j \in s(k,l)} \Phi(x_j) \right\| &= \left\| \sum_{x_i \in s(k)^+} \Phi(x_i) - \sum_{x_j \in s(k)^-} \Phi(x_j) + \sum_{x_j \in s(k)^-} \Phi(x_j) - \sum_{x_q \in s(k,l)} \Phi(x_q) \right\| \\ &\leqslant \left\| \sum_{x_i \in s(k)^+} \Phi(x_i) - \sum_{x_j \in s(k)^-} \Phi(x_j) \right\| + \left\| \sum_{x_i \in s(k)^-} \Phi(x_i) - \sum_{x_j \in s(k,l)} \Phi(x_j) \right\| \end{aligned}$$

From (14), it springs that the first term of this last expression is bounded from above by $\sqrt{m_k}\Lambda_{\Phi(\mathscr{X})}$. Furthermore, keeping in mind that $s(k,l)$ is a subset of $s(k)^-$ of cardinality m_{kl}, the second term is trivially upperbounded by $(m_k^- - m_{kl})\Lambda_{\Phi(\mathscr{X})}$. Substituting the resulting upper bound of $\| \sum_{x_i \in s(k)^+} \Phi(x_i) - \sum_{x_j \in s(k,l)} \Phi(x_j) \|$ in (18) finally gives

$$2(m_k^+ + m_{kl})\varepsilon \leqslant (\sqrt{m_k} + m_k^- - m_{kl}) \|w_k - w_l\| \Lambda_{\Phi(\mathscr{X})} \tag{19}$$

Since $m_k = m_k^+ + m_k^-$, $m_k^+ \geqslant m_k^-$ and $m_{kl} \geqslant \lceil m_k^- / (Q-1) \rceil$, (19) still implies

$$\frac{Q m_k}{(Q-1)\sqrt{m_k} + ((Q-2)/2)\, m_k} \leqslant \frac{\|w_k - w_l\| \Lambda_{\Phi(\mathscr{X})}}{\varepsilon} \tag{20}$$

Since the left-hand side of (20) is an increasing function of m_k, we have exhibited an upper bound of V_ε-dim$(\Delta\mathscr{H}_k)$ in terms of $\max_{l \neq k} \|w_k - w_l\|$, $\Lambda_{\Phi(\mathscr{X})}$ and ε, which is non-trivial for large enough values of ε. Note further that the bound on the fat-shattering dimension of hyperplanes established in Reference [41] appears as a special case of this inequality, in the degenerate case where $Q = 2$. Given the bound on V_ε-dim$(\Delta\mathscr{H}_k)$, the bound on UM-fat$_{\Delta\mathscr{H}'}(\varepsilon)$ then directly results from (13) and (11). □

4.5. *Discussion*

Theorem 4 highlights the fact that the functional $\max_{k<l} \|w_k - w_l\|^2$ (or alternatively $\sum_{k<l} \|w_k - w_l\|^2$) plays for M-SVMs a role similar to the one played by $\|w\|^2$ for the standard binary SVMs. This is satisfactory indeed, since both functions are convex. Their use as control term in the objective function of the training procedure, as was done in References [32, 39], is thus once more justified. In References [14, 30, 34, 36], the functional selected to perform the capacity control is slightly different, since it is $\sum_{k=1}^{Q} \|w_k\|^2$, whereas in Reference [31], the authors used instead $\sum_{k<l}^{Q} \|w_k - w_l\|^2 + \sum_{k=1}^{Q} \|w_k\|^2$. Can the theorems derived here justify these choices as well? This is the case indeed. For instance, it was proved in Reference [39] (see also Reference [33]) that the machines introduced in References [14, 30–32], in spite of their different formulations, are utterly equivalent, since they all generate the same optimal solution, provided the value of their soft margin parameter C is selected appropriately. Furthermore, variants of Theorem 4 can easily be derived, to fit more precisely a given training algorithm (penalty term). We have thus endowed all the M-SVMs published so far with a well founded theoretical justification, which makes it possible to compare their performance on a sound basis.

5. ALTERNATIVE APPROACHES

In Sections 3 and 4, the guaranteed risk of interest has been studied according to a standard strategy, which can be summarized as follows. First, express the confidence interval in terms of a capacity measure (Theorem 1). Second, relate this capacity measure to an extended notion of VC dimension, by means of a generalized Sauer's lemma (Theorem 3). Third, characterize the behaviour of this VC dimension as a function of the constraints on the model parameters (Theorem 4). Recently, Williamson and co-workers have introduced an alternative approach in Reference [42] (see also References [43–45]). It is based on functional analysis results on the compactness of operators (see, for instance, Reference [46]). The covering numbers are determined via the entropy numbers of a linear operator. The main advantage of this strategy rests on the fact that it makes no use of a combinatorial dimension, and is thus more 'direct'. With fewer partial bounds, the confidence interval should *a priori* be tighter. We already built on this work in Reference [10], to pave the way for a theoretical study of M-SVMs. A comparison with the results of this paper is currently underway. A more diverging possibility consists in deriving bounds based on data-dependent capacity measures such as the empirical VC entropy. In this field, the most promising studies are probably those dealing with concentration

inequalities, and especially References [47, 48]. More generally, the study of model selection based on penalized empirical loss minimization, as presented for instance in Reference [49], should also prove particularly fruitful.

6. CONCLUSIONS AND FUTURE WORK

This paper has described a pathway to bound the covering numbers of sets of vector-valued functions used to perform multi-class discriminant analysis. The resulting bound, involving an extended notion of fat-shattering dimension, has been applied to the architecture shared by the different multi-class SVMs developed so far. This has enabled us to cast them into a unified theoretical framework and highlight the part played by their penalty term. From there, one could compare these machines, both theoretically and empirically, or put forward new arguments to justify *a posteriori* the choice of the structure on which they are based, i.e. the choice of their objective functions. Our results could also be used to design new machines.

Major benefits should result from deriving more direct bounds on the confidence interval of M-SVMs. Indeed, reducing the number of steps should produce less conservative guaranteed risks, telling us more about the precise behaviour of these machines. In that respect, significant improvements should be expected from extending to the multi-class case the approaches listed in the preceding section. These extensions, and the subsequent comparisons, are the subject of an ongoing work.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewer for his comments.

REFERENCES

1. Vapnik VN, Chervonenkis AYa. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications* 1971; **16**:264–280.
2. Ben-David S, Cesa-Bianchi N, Haussler D, Long PM. Characterizations of learnability for classes of $\{0,\ldots,n\}$-valued functions. *Journal of Computer and System Sciences* 1995; **50**:74–86.
3. Alon N, Ben-David S, Cesa-Bianchi N, Haussler D. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM* 1997; **44**:615–631.
4. Vapnik VN. *Estimation of Dependences Based on Empirical Data*. Springer: New York, 1982.
5. Shawe-Taylor J, Bartlett PL, Williamson RC, Anthony M. Structural risk minimization over data-dependent hierarchies. *Technical Report* NC-TR-96-053, NeuroCOLT, 1996.
6. Bartlett PL. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory* 1998; **44**(2):525–536.
7. Shawe-Taylor J, Anthony M. Sample sizes for multiple-output threshold networks. *Network: Computation in Neural Systems* 1991; **2**:107–117.
8. Anthony M. Probabilistic analysis of learning in artificial neural networks: the PAC model and its variants. *Neural Computing Surveys* 1997; **1**:1–47.
9. Guermeur Y, Elisseeff A, Paugam-Moisy H. Estimating the sample complexity of a multi-class discriminant model. In *ICANN'99*. IEE: London, 1999; 310–315.
10. Elisseeff A, Guermeur Y, Paugam-Moisy H. Margin error and generalization capabilities of multi-class discriminant models. *Technical Report NC-TR-99-051-R,* NeuroCOLT2, 1999 (revised in 2001).
11. Guermeur Y. A simple unifying theory of multi-class support vector machines. *Technical Report RR-4669,* INRIA, 2002.
12. Platt JC, Cristianini N, Shawe-Taylor J. Large margin DAGs for multiclass classification. In *NIPS'12*, 2000; 547–553.

13. Richard MD, Lippmann RP. Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation* 1991; **3**:461–483.
14. Vapnik VN. *Statistical Learning Theory*. Wiley: New York, 1998.
15. Pollard D. Empirical processes: theory and applications. In *NFS-CBMS Regional Conference Series in Probability and Statistics*, vol. 2. Institute of Math. Stat. and Am. Stat. Assoc., 1990.
16. Haussler D. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation* 1992; **100**:78–150.
17. Kearns MJ, Schapire RE. Efficient distribution-free learning of probabilistic concepts. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*, vol. 1. IEEE Computer Society Press: Silver Spring, MD, 1990; 382–391.
18. Kearns MJ, Schapire RE. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences* 1994; **48**(3):464–497.
19. Gurvits L. A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. *Theoretical Computer Science* 2001; **261**(1):81–90.
20. Dudley RM. Universal Donsker classes and metric entropy. *The Annals of Probability* 1987; **15**(4):1306–1326.
21. Natarajan BK. On learning sets and functions. *Machine Learning* 1989; **4**:67–97.
22. Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. In *COLT'92*, 1992; 144–152.
23. Cortes C, Vapnik VN. Support-vector networks. *Machine Learning* 1995; **20**:273–297.
24. Schölkopf B, Burges C, Vapnik V. Extracting support data for a given task. In *ICKDDM'95*, 1995; 252–257.
25. Mayoraz E, Alpaydin E. Support vector machines for multi-class classification. *Technical Report 98-06*, IDIAP, 1998.
26. Saitoh S. *Theory of Reproducing Kernels and its Applications*. Longman Scientific & Technical: Harlow, England, 1988.
27. Wahba G. Spline models for observational data. In *SIAM*, vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, 1990.
28. Wahba G. Support vector machines, reproducing Kernel Hilbert spaces, and randomized GACV. In *Advances in Kernel Methods*, *Support Vector Learning*, Schölkopf B, Burges CJC, Smola AJ (eds). MIT Press: Cambridge, MA, 1999; 69–88.
29. Aizerman M, Braverman E, Rozonoer L. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control* 1964; **25**:821–837.
30. Weston J, Watkins C. Multi-class support vector machines. *Technical Report CSD-TR-98-04*, Royal Holloway, Department of Computer Science, University of London, 1998.
31. Bredensteiner EJ, Bennett KP. Multicategory classification by support vector machines. *Computational Optimization and Applications* 1999; **12**(1/3):53–79.
32. Guermeur Y, Elisseeff A, Paugam-Moisy H. A new multi-class SVM based on a uniform convergence result. In *IJCNN'00*, vol. IV, 2000; 183–188.
33. Hsu C-W, Lin C-J. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks* 2002; **13**:415–425.
34. Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research* 2001; **2**:265–292.
35. Crammer K, Singer Y. On the learnability and design of output codes for multiclass problems. In *Proceedings of the Thirteen Annual Conference on Computational Learning Theory* (*COLT*), 2000; 35–46.
36. Lee Y, Lin Y, Wahba G. Multicategory support vector machines. *Technical Report 1043*, Department of Statistics, University of Wisconsin, Madison, 2001.
37. Lee Y. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Technical Report 1063*, Department of Statistics, University of Wisconsin, Madison, 2002.
38. Guermeur Y, Zelus D. Combining protein secondary structure prediction models with ensemble methods of optimal complexity. In *JOBIM'01*, 2001; 97–104.
39. Guermeur Y. Combining discriminant models with new multi-class SVMs. *Pattern Analysis and Applications* 2002; **5**(2):168–179.
40. Fletcher R. *Practical Methods of Optimization*. Wiley: New York, 1987.
41. Bartlett PL, Shawe-Taylor J. Generalization performance of support vector machines and other pattern classifiers. In *Advances in Kernel Methods*, *Support Vector Learning*, Schölkopf B, Burges CJC, Smola A (eds). MIT Press: Cambridge, MA, 1999; 43–54.
42. Williamson RC, Smola AJ, Schölkopf B. Generalization performance of regularization networks and support vector machines *via* entropy numbers of compact operators. *IEEE Transactions on Information Theory* 2001; **47**(6): 2516–2532.
43. Smola AJ. Learning with kernels. *Ph.D. Thesis*, Technische Universität Berlin, 1998.
44. Williamson RC, Smola AJ, Schölkopf B. Entropy numbers of linear function classes. In *COLT'00*, 2000; 309–319.
45. Schölkopf B, Smola AJ. *Learning with Kernels*, *Support Vector Machines*, *Regularization*, *Optimization and Beyond*. MIT Press: Cambridge, MA, 2002.

46. Carl B, Stephani I. *Entropy*, *Compactness*, *and the Approximation of Operators*. Cambridge University Press: Cambridge, UK, 1990.
47. Boucheron S, Lugosi G, Massart P. A sharp concentration inequality with applications. *Technical Report NC2-TR-1999-057*, NeuroCOLT2, 1999.
48. Bousquet O. Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms. *Ph.D. Thesis*, Ecole Polytechnique, 2002.
49. Bartlett PL, Boucheron S, Lugosi G. Model selection and error estimation. *Machine Learning* 2002; **48**:85–113.

