Pattern Analysis & Applications (2002)5:168–179

Combining Discriminant Models with New Multi-Class SVMs

Yann Guermeur

LORIA, Campus Scientifique, Vandœvre-lès-Nancy, France

Abstract: The idea of performing model combination, instead of model selection, has a long theoretical background in statistics. However, making use of theoretical results is ordinarily subject to the satisfaction of strong hypotheses (weak error correlation, availability of large training sets, possibility to rerun the training procedure an arbitrary number of times, etc.). In contrast, the practitioner is frequently faced with the problem of combining a given set of pre-trained classifiers, with highly correlated errors, using only a small training sample. Overfitting is then the main risk, which cannot be overcome but with a strict complexity control of the combiner selected. This suggests that SVMs should be well suited for these difficult situations. Investigating this idea, we introduce a family of multi-class SVMs and assess them as ensemble methods on a real-world problem. This task, protein secondary structure prediction, is an open problem in biocomputing for which model combination appears to be an issue of central importance. Experimental evidence highlights the gain in quality resulting from combining some of the most widely used prediction methods with our SVMs rather than with the ensemble methods traditionally used in the field. The gain increases when the outputs of the combiners are post-processed with a DP algorithm.

Keywords: Classifier fusion; Generalisation performance; Hierarchical sequence processing systems; Protein secondary structure prediction; Statistical learning theory; Support Vector Machines

1. INTRODUCTION

Since the early 1960s, and precisely the studies of Bates and Granger [1,2], model combination has proved to be an efficient alternative to model selection for a wide range of statistical inference problems. Theory in the field has made rapid strides [2–9], however, until recently, theoretical evidence had been mainly developed in the framework of regression, whereas discrimination was seldom considered independently. In the last decade, many studies have dealt with the specific problems of discrimination, such as the estimation of Bayes error [10,11], or variance reduction [12], the link between error correlation and error reduction [13] (see also Clemen and Winkler [14]), as well as the decomposition of the error into a bias and a variance term [15]. The success of methods such as *bagging* [16] and *boosting* [17] has highlighted the usefulness of implementing bootstrap algorithms to improve the performance of 'weak classifiers'. This is indeed of primary importance, since the theory of boosting meets Vapnik's theory of bounds through the fundamental notion of a *maximal margin classifier*. Classifier combination is thus currently endowed with a rich theoretical framework, which is very useful as long as the problem at hand satisfies the hypotheses on which it is grounded. Unfortunately, in many real-life situations, the practitioner is faced with the worst configuration one can think of when combining models (pre-trained experts with different types of outputs, errors highly correlated, small set of labelled data available for training, etc.). These difficulties prevent him from making the best of the potential of the theory, and his main concern is to avoid overfitting. As a consequence, in this context, the problem to be solved primarily consists in finding a combiner of adequate complexity, so that, with high probability, the training error observed could constitute an estimate precise enough of the generalisation error, and the gain in prediction accuracy, small as it should be, could be 'guaranteed'. This is precisely the type of situations for which Support Vector Machines (SVMs) have been developed. SVMs have been introduced by Vapnik and coworkers [18,19] as a direct implementation of the Structural Risk Minimisation (SRM) inductive principle [20]. The aim of the support vector method, a description of which can be found elsewhere [21–24], is to maximise the generalisation capabilities by minimising an upper bound on the *expected*

Received: 15 November 2000
Received in revised form: 26 October 2001
Accepted: 13 December 2001

risk (or generalisation error) with respect to the values of the model parameters. This bound is systematically made up of two terms. The first one is the *empirical risk* (training error), the second one, that Vapnik calls a confidence interval, is a growing function of the *capacity* of the model, capacity which can be expressed in terms of different measures. For instance, in the case of dichotomy computation, the most common one is the Vapnik–Chervonenkis (VC) dimension [25]. Simple introductions to the theory of bounds applied to neural networks can be found in Haussler [26] and Anthony [27]. With this structure of the bound in mind, it appears immediately that the SRM inductive principle can be implemented by minimising the control term for different levels of the empirical risk, in order to find a minimum of the *guaranteed risk* functional with a linesearch. Indeed, this aim is reached with the support vector methods developed for estimating indicator or real-valued functions. Unfortunately, although many multi-class discriminant models have been developed around the support vector method, none of them owns this property. Initially, multi-class discrimination was implemented with SVMs through the so-called *one-against-the-rest* or *one-per-class* approaches [28,29]. Later on came the *pairwise-coupling* decomposition scheme [30,31] and the *k*-class SVM proposed independently by Vapnik [21], Weston and Watkins [31] and Bredensteiner and Bennett [32], among others. Strictly speaking, these three approaches fail to implement the SRM inductive principle, since they are not related, at least explicitly, to a uniform convergence result, or guaranteed risk, which makes it impossible to characterise a satisfactory compromise between training performance and complexity. In this paper, building upon the uniform strong law of large numbers introduced in Elisseeff et al [33], we develop a theoretical framework which leads to the specification of a family of Multi-class SVMs (M-SVMs). They differ either in the expression of the guaranteed risk or in the specification of the structure. This enables us to provide Vapnik's *k*-class SVM with a theoretical grounding. Two of these SVMs are assessed as classifier combiners on an open real-world problem: the problem of protein secondary structure prediction. This task is of central importance in predictive structural biology. Numerous methods have been proposed to predict the secondary structure (see elsewhere [34–36] for reviews on the subject). *A priori*, implementing a combination of models appears particularly relevant in this context, since most of the prediction systems developed so far ordinarily use, in addition to the amino acid sequences (or profiles of multiple alignments), data from different knowledge sources (physicochemical properties, homology, etc.). Consequently, whenever secondary structure is to be predicted, several sets of conformational scores are available, which can be expected not to be utterly correlated. Indeed, most of the current best prediction methods already implement conformational score combinations at one stage or another. These combinations can take many forms, ranging from the simple linear opinion pool [37] to the more complex non-linear regression schemes performed by neural networks [38,39]. Symbolic methods based on empirical results have also been implemented, such as the algorithm combine [40]. However, a constant of these studies is that the choice of a particular combiner is hardly ever justified, although it appears to have a crucial effect on performance. Furthermore, the scores combined are systematically homogeneous, (i.e. they represent estimates of the same quantities), whereas the practitioner who needs to make his own prediction based on the results of several methods has most often to deal with inhomogeneous scores. Last but not least, the gain resulting from the combination is seldom significantly superior to that resulting from a simple averaging of the outputs of the base classifiers. This phenomenon is indeed acknowledged by leading experts in the domain (B. Rost and G. Pollastri, personal communications). A first attempt to overcome these limitations was described in our earlier work [41]. In this paper, we establish that noticeable benefits can spring from combining protein secondary structure models with M-SVMs. The gain in prediction accuracy over other standard ensemble methods becomes statistically significant, with confidence exceeding 0.98 when the outputs are post-processed with a simple Dynamic Programming (DP) algorithm borrowed from the field of speech processing, which suggests that our SVMs would perform best when incorporated in hierarchical prediction systems. The organisation of this paper is as follows. In Section 2, we briefly summarise our uniform convergence result, and explain how it can be of practical use to study the generalisation capabilities of multi-class discriminant models (to bound the expected risk). The corresponding theorems and formulae are then applied to the multivariate linear (or more precisely affine) regression model, which leads to the specification of the new M-SVMs. Mathematical details are reported in the Appendix. Initial experimental results, regarding the sole combination, are given in Section 3. The comparative study is developed in Section 4, where the possibility of post-processing the outputs is assessed.

2. FROM UNIFORM STRONG LAWS OF LARGE NUMBERS TO MULTI-CLASS SVMs

2.1. Framework of the Study

We consider the case of a Q-category pattern recognition problem, where $Q \geq 3$. Let $\mathcal{X}$ be the space of description (or input space) and $\mathcal{C}$ the set of categories. We make the assumption, standard in statistical learning theory, that there is a joint probability, fixed but unknown, on $\mathcal{X} \times \mathcal{C}$. Our goal is then to find, in the set $\mathcal{H} = \{h\}$ of functions implemented by a statistical model, a function which corresponds to the lowest error rate. The decision function associated with this function must thus be as close as possible to Bayes' decision rule. We make further the hypothesis that the elements of $\mathcal{H}$ are multivariate real-valued functions. Precisely, for each example x in $\mathcal{X}$ and each category C_k in $\mathcal{C}$, $(1 \leq k \leq Q)$, a function h_k of x taking its values in $\mathbb{R}$ is computed. The discriminant function associated with these regression functions is obtained by assigning each pattern x to the category C_k satisfying $h_k(x) = \max_l h_l(x)$. This frame-

work is very common indeed. In the case where the $h_k(x)$ are estimates of the class posterior probabilities, which occurs for instance when the model is a neural network and the training criterion is adequately selected [42,43], choosing this decision function simply amounts to implementing Bayes' estimated decision rule. In what follows, $C(x_i)$ will denote indifferently the category of pattern x_i, or the index of this category, while y_i will be the corresponding canonical coding, i.e. $C(x_i) = C_l \Leftrightarrow y_i = [y_{ik}] \in \{-1, 1\}^Q$, where $y_{ik} = -1^{1-\delta_{kl}}$, and δ is Kronecker's symbol.

2.2. Uniform Strong Law of Large Numbers based on Covering Numbers

In this context, we have established a uniform strong law of large numbers which is based of the following definitions.

Definition 1 (Covering numbers). *Let (E, ρ) be a pseudo-metric space, and $B(v, r)$ the closed ball of radius r and centre v in E. The* covering number $\mathcal{N}(\epsilon, H, \rho)$ *of a set $H \subset E$ is the smallest cardinality of the sets $\bar{H} \subset E$ such that*

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

The sets $\bar{H}$ satisfying this property are called ϵ-covers of H: each element in H is at a distance less than ϵ of an element in $\bar{H}$.

See Kolmogorov and Tihomirov [44] and Carl and Stephani [45] for the fundamental results regarding covering numbers.

Definition 2. *Let $\mathcal{F}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For a set s of points in $\mathcal{X}$, define the pseudo-metric $d_{l_\infty, l_1(s)}$ on $\mathcal{F}$ as*

$$\forall (f, \bar{f}) \in \mathcal{F}^2,\ d_{l_\infty, l_1(s)}(f, \bar{f}) = \max_{x \in s} \sum_{k=1}^{Q} |f_k(x) - \bar{f}_k(x)|$$

Definition 3. *For all $h \in \mathcal{H}$ and all $x \in \mathcal{X}$, let $M_1(h, x)$ be the smallest index l such that $h_l(x) = \max_k h_k(x)$ and $M_2(h, x)$ the smallest index $l \neq M_1(h, x)$ such that $h_l(x) = \max_{k \neq M_1(x)} h_k(x)$. Define $\Delta h = [\Delta h_k]$, $(1 \leq k \leq Q)$, as the function from $\mathcal{X}$ into $\mathbb{R}^Q$, satisfying*

$$\Delta h_k(x) = \begin{cases} \frac{1}{2}(h_k(x) - h_{M_2(h,x)}(x)) & \text{if } k = M_1(h, x) \\ \frac{1}{2}(h_k(x) - h_{M_1(h,x)}(x)) & \text{otherwise} \end{cases}$$

Note that this function is directly related to the notion of margin introduced by Schapire and co-workers [17] to extend to the multi-class case the uniform convergence results established for boosting algorithms. Define the threshold function *sign*: $\mathbb{R} \to \{-1, 1\}$ as

$$sign(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

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

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

$\forall h \in \mathcal{H}$, $\Delta h^\gamma = [\Delta h_k^\gamma] = [\pi_\gamma \circ \Delta h_k]$, $(1 \leq k \leq Q)$. $\Delta\mathcal{H}^\gamma = \{\pi_\gamma(\Delta h)/h \in \mathcal{H}\}$. With these definitions at hand, we denote

Definition 4.

$$\mathcal{N}_{\infty,1}(\gamma/2, \Delta\mathcal{H}^\gamma, 2N) = \max_{s_{2N} \in \mathcal{X}^{2N}} \mathcal{N}(\gamma/2, \Delta\mathcal{H}^\gamma, d_{l_\infty, l_1(s_{2N})})$$

To select an optimal function $h \in \mathcal{H}$, we make the assumption that a training set $S_N = \{(x_i, y_i)\}$, $(1 \leq i \leq N)$, made up of labelled examples, iid according to the joint distribution on $\mathcal{X} \times C$ to be inferred, is available. Extending a definition from Bartlett [46], we introduce the following definition:

Definition 5. *The empirical risk with margin $\gamma \in (0, 1]$ on a training set S_N of size N is*

$$R_{s_N}^\gamma(h) = \frac{1}{N} \left| \{x_i, C(x_i)) \in s_N / \Delta h_{C(x_i)}(x_i) < \gamma \} \right|$$

Studies on the use of margins in statistical learning theory date back from the early works of Vapnik [20]. Different illustrations of the richness of this approach can be found elsewhere [47,48,17,33]. In this context, extending Lemma 4 and Corollary 9 from Bartlett [46], as well as Theorem 4.1 from Vapnik [21], we established [33] the following theorem (Corollary 2):

Theorem 1. *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 $\mathcal{H}$ trained on a set of size N is bounded above by*

$$R(h) \leq R_{s_N}^\gamma(h) + \sqrt{\frac{1}{2N}\left(\ln(2\mathcal{N}_{\infty,1}(\gamma/2, \Delta\mathcal{H}^\gamma, 2N)) + \ln\left(\frac{2}{\gamma\delta}\right)\right)} + \frac{1}{N} \quad (1)$$

Similar theorems can be derived for different pseudo-metrics (see, for instance, Guermeur et al [49]). One of the most interesting possibilities is to consider the pseudo-metric $d_{l_\infty, l_\infty(s)}$ given by:

Definition 6. *Let $\mathcal{F}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For a set s of points in $\mathcal{X}$, define the pseudo-metric $d_{l_\infty, l_\infty(s)}$ on $\mathcal{F}$ as*

$$\forall (f, \bar{f}) \in \mathcal{F}^2,\ d_{l_\infty, l_\infty(s)}(f, \bar{f}) = \max_{x \in s} \max_{l \leq k \leq Q} |f_k(x) - \bar{f}_k(x)|$$

Note that, contrary to other well known bounds, these theorems do not rest on the hypothesis that the functions in $\mathcal{H}$ take their values in $[-1, 1]^Q$, which makes them quite general. They apply, for instance, to multi-layer perceptrons, even when they have linear output units, and consequently also to SVMs. The bounds are significantly tighter than those obtained by using as a capacity measure multi-class extensions of the VC dimension such as the *graph dimension* or *Natarajan dimension* [50]. A preliminary comparative study on this question can be found elsewhere [51]. To implement the SRM inductive principle, the main term the value of which must be determined, except for the empirical margin risk, is the covering number characterising the model capacity. Expressing this measure in terms of the model parameters can thus provide us with the objective function

of an optimisation problem corresponding to a training algorithm.

2.3. Bounds on the Covering Numbers

Several methods have been proposed to bound covering numbers [52,53,46]. In this section, we outline a strategy to derive an upper bound on the covering numbers for general multi-class models (unspecified families of functions $\mathcal{H}$ taking their values in $\mathbb{R}^Q$), using the method introduced in Williamson et al [54] and Smola [55]. A key feature of this approach is that it directly bounds the covering numbers of interest rather than making use of a combinatorial dimension such as the extensions of the VC dimension cited before or the *fat-shattering dimension* [56]. To that end, we make the additional assumption that $\mathcal{H}$ is included in a finite-dimensional Banach space $E_{\mathcal{H}}$. We assume further that $\mathcal{H}$ is bounded in $E_{\mathcal{H}}$ (with respect to the corresponding norm). This is indeed a mild hypothesis, since it is satisfied among others by SVMs, regularisation networks [57] and linear models, when prior information is assumed or is given. As a consequence, $\mathcal{H}$ is *precompact* (see Carl and Stephani [45] for a proof of this proposition), which means that the covering numbers of interest will always be finite. For the set $s_{2N} \in \mathcal{X}^{2N}$ aforementioned, let us define the following linear operator:

$$T_{s_{2N}}: E_{\mathcal{H}} \rightarrow M_{2N,Q(Q-1)/2}(\mathbb{R})$$
$$g = [g_k] \mapsto T_{s_{2N}}(g)$$

with

$$T_{s_{2N}}(g) = \begin{bmatrix} g_1(x_1) - g_2(x_1) & \cdots & g_k(x_1) - g_l(x_1) & \cdots & g_{Q-1}(x_1) - g_Q(x_1) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ g_1(x_i) - g_2(x_i) & \cdots & g_k(x_i) - g_l(x_i) & \cdots & g_{Q-1}(x_i) - g_Q(x_i) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ g_1(x_{2N}) - g_2(x_{2N}) & \cdots & g_k(x_{2N}) - g_l(x_{2N}) & \cdots & g_{Q-1}(x_{2N}) - g_Q(x_{2N}) \end{bmatrix}$$

Let $B_{\mathcal{H}}$ be a closed ball of $E_{\mathcal{H}}$ in which $\mathcal{H}$ is included. We endow $M_{2N,Q(Q-1)/2}(\mathbb{R})$ with a norm $\|.\|$ chosen in accordance with the choice of the pseudo-metric on $\mathcal{H}$. For instance, in the case where the pseudo-metric is $d_{l_\infty, l_\infty(s)}$, we set

$$\forall A \in M_{2N,Q(Q-1)/2}(\mathbb{R}), \quad A = [a_{ij}],$$
$$\|A\| = \|A\|_{l_\infty, l_\infty(2N)} = \max_{1 \le i \le 2N} \max_{j=1}^{Q(Q-1)/2} |a_{ij}|$$

After some algebra (see the Appendix), one obtains:

$$\mathcal{N}(\gamma/2, \Delta\mathcal{H}, d_{s_{2N}}) \le f_{d_{s_{2N}}}(\mathcal{N}(\gamma/2, T_{s_{2N}}(B_{\mathcal{H}}), \|.\|))$$

where $f_{d_{s_{2N}}}$ is an increasing function which depends upon the choice of the pseudo-metric, represented here by the generic notation $d_{s_{2N}}$. Since π_γ satisfies the Lipschitz condition with constant 1, one finally derives:

Theorem 2. *For all $s_{2N} \in \mathcal{X}^{2N}$ and for all $\gamma \in (0, 1]$,*

$$\mathcal{N}(\gamma/2, \Delta\mathcal{H}^\gamma, d_{s_{2N}}) \le f_{d_{s_{2N}}}(\mathcal{N}(\gamma/2, T_{s_{2N}}(B_{\mathcal{H}}), \|.\|)) \quad (2)$$

We have thus reduced the problem of bounding the covering number appearing in Eq. (1), or other similar formulae based on different pseudo-metrics, to the problem of finding an upper bound of $\mathcal{N}(\gamma/2, T_{s_{2N}}(B_{\mathcal{H}}), \|.\|)$, when s_{2N} describes the whole set $\mathcal{X}^{2N}$. This can be done readily thanks to functional analysis results. To detail the corresponding process, we must first introduce additional definitions.

Definition 7 (Entropy numbers). *Let (E, ρ) be a pseudo-metric space. The* nth entropy number $\epsilon_n(H)$ *of a set $H \subset E$ is defined as the smallest real ϵ such that there exists an ϵ-cover of H of cardinality at most n.*

Let T be a linear operator from a Banach space E into a Banach space F. Let U_E be the closed unit ball of E. The nth entropy number *of T is defined as*

$$\epsilon_n(T) = \epsilon_n(T(U_E))$$

Definition 8 (operator norm). *Let T be a linear operator acting between arbitrary real Banach spaces E and F. $\|T\|_F$ is the operator norm given by $\|T\|_F = \sup_{f \in U_E} \|T(f)\|_F$.*

The idea underlying the introduction of the entropy numbers is simple. On the one hand, there is a simple relationship between a bound on an entropy number and a bound on a covering number. On the other hand, functional analysis provides us with results, such as Maurey's theorem (see Williamson et al [58] and below), to bound the entropy numbers of linear operators. For the sake of simplicity, we illustrate the first point in the (univariate) linear case, borrowing our example from Williamson et al [54]. Similarly, the formulation of Maurey's theorem we give should be adapted to apply to the specific context at hand.

Theorem 3. *Let $F_{\Lambda w}$ be the set of linear applications f_w, from E_X, Banach space containing $\mathcal{X}$, into $\mathbb{R}$, satisfying $\|w\|_{l_2} \le \Lambda_w$, where $\|.\|_{l_2}$ is the Euclidean norm. Then*

$$\epsilon_n(T: \|.\|_{l_2} \rightarrow \|.\|_{l_\infty(s_{2N})}) \le \epsilon_0 \Rightarrow \mathcal{N}_\infty(\epsilon_0, \mathcal{F}_{\Lambda w}, 2N) \le n \quad (3)$$

Theorem 4 (Maurey). *Let $T \in \mathcal{L}(E, l_{\infty(s_m)})$, where E is an Hilbert space. Then there exists a constant $c > 0$ such that, for all $n, m \in \mathbb{N}$,*

$$\epsilon_n(T) \le c\|T\| \left((\log n + 1)^{-1} \log \left(1 + \frac{m}{\log n + 1} \right) \right)^{1/2} \quad (4)$$

In a nutshell, applying Eq. (4), or more precisely the adequate extension of this formula, it is possible to derive an upper bound on the entropy numbers of the linear operator $T_{s_{2N}}$. From this bound, an upper bound on the covering number of interest, $\mathcal{N}_{\infty,1}(\gamma/2, \Delta\mathcal{H}^\gamma, 2N)$, $\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2N)$, etc. can then be derived, by making use of Eq. (2), or the appropriate extension of Eq. (3). A last difficulty must be overcome, which springs from the fact that by hypothesis, the functions in $\mathcal{H}$ live in $B_{\mathcal{H}}$, whereas

the aforementioned results involve $U_{E_{\mathcal{H}}}$, the unit ball of $E_{\mathcal{H}}$. This is done very simply, thanks to the following proposition:

Proposition 1. *Let T be a linear operator acting between arbitrary real Banach spaces $(E, \|.\|_E)$ and $(F, \|.\|_F)$. For all balls B_Λ of radius Λ centred in a:*

$$T(B_\Lambda) = \Lambda T(U_E) + T(a) \tag{5}$$

This means that, once one has been able to characterise a ball of $E_{\mathcal{H}}$ in which the functions of $\mathcal{H}$ live, the nature of its parameters (centre and radius) raise no (theoretical) difficulty to bound the covering numbers of interest (bearing in mind that the smaller the radius is, the better the bound will be).

2.4. Architecture of the M-SVMs

The results presented in the previous sections apply to any multi-class discriminant system obtained by combining a multivariate model with Bayes' estimated decision rule. In this section, we turn to the specific case of M-SVMs. The study of the standard (bi-class) SVMs is usually done in two steps: first, the linear case (optimal hyperplane), then the non-linear one (by introduction of kernels satisfying Mercer's conditions [59]). Indeed, the specification of the training procedure does not take into account explicitly the nature of the kernel, although bounds on the generalisation error of kernel machines have been derived (see, for instance, Williamson et al [54] for a very powerful theoretical framework on the subject). In the same way as a 'linear' SVM shares the architecture of the perceptron, a multi-class linear SVM is a multivariate linear regression model (a set of hyperplanes of cardinality equal to the number of classes). We thus have $\mathcal{H} = \{h\}$, with

$$\forall x \in \mathcal{X},\ h(x) = Wx + b = \begin{bmatrix} w_1^T \\ \vdots \\ w_k^T \\ \vdots \\ w_Q^T \end{bmatrix} x + \begin{bmatrix} b_1 \\ \vdots \\ b_2 \\ \vdots \\ b_Q \end{bmatrix}$$

Given the results reported in the preceding subsections, to apply the SRM inductive principle to M-SVMs, and consequently to determine the objective function of the training procedure, we must thus bound the covering numbers of the multivariate linear (affine) model.

2.5. Covering Numbers of the Multivariate Linear Regression Model

A bound on the covering numbers of the model of interest can be deduced from the following two theorems. Note that, for the sake of simplicity, both have been expressed for a specific choice of the pseudo-metric on $\mathcal{H}$, which induces no loss of generality.

Theorem 5. *Let $\tilde{\mathcal{F}}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$ and $\mathcal{F}$ a set of functions satisfying $\forall f \in \mathcal{F}, \exists(\tilde{f}, b) \in \tilde{\mathcal{F}} \times [-B, B]^Q / f = \tilde{f} + b$. Let $\Delta\mathcal{F}$ and $\Delta\tilde{\mathcal{F}}$ be the sets of functions derived from $\mathcal{F}$ and $\tilde{\mathcal{F}}$, respectively, by applying Definition 3. Then the following bounds hold:*

$$\mathcal{N}_{\infty,1}(Q\epsilon, \mathcal{F}, 2N) \leq \left(\left\lceil \frac{2B}{\epsilon} \right\rceil + 1\right)^Q \mathcal{N}_{\infty,1}(\epsilon, \tilde{\mathcal{F}}, 2N)$$

$$\mathcal{N}_{\infty,1}(Q\epsilon, \Delta\mathcal{F}, 2N) \leq \left(\left\lceil \frac{4B}{\epsilon} \right\rceil + 1\right)^Q \mathcal{N}_{\infty,1}(\epsilon, \Delta\tilde{\mathcal{F}}, 2N) \tag{6}$$

Theorem 6. *Let $\mathcal{H}$ be the multivariate linear model from $\mathcal{X} \subset \mathbb{R}^d$ into $\mathbb{R}^Q$. $\mathcal{H}$ and $\mathbb{R}^d$ are endowed with the Euclidean norm $\|.\|_{l_2}$. If $\mathcal{X}$ is included in a ball of radius $\Lambda_{\mathcal{X}}$ about the origin, then the following bound holds:*

$$\forall h \in \mathcal{H},\ \|T_{s_{2N}}(h)\|_{l_\infty, l_1} \leq \Lambda_{\mathcal{X}} \sqrt{\frac{Q(Q-1)}{2}} \sqrt{\sum_{k<l} \|w_k - w_l\|_{l_2}^2} \tag{7}$$

To sum up, combining the results exposed in the previous subsections with Eqs (6) and (7), it was possible to express the confidence interval which constitutes, with the empirical risk, the expression of the guaranteed risk, as an increasing function of $\Sigma_{k<l} \|w_k - w_l\|^2$. Since this sum is a convex functional, this result nicely extends Vapnik's well known bound on the VC dimension of canonical hyperplanes in terms of the square of the norm of the corresponding vector (see Vapnik [21], Theorem 10.3).

2.6. Unification of the Multi-class SVMs Proposed so Far

Making use of Theorem 6, we can readily specify a multi-class SVM, the objective function of which takes the confidence interval into account through the term $\Sigma_{k<l} \|w_k - w_l\|^2$. The training procedure associated with this model consists in solving the following quadratic programming problem:

Problem 1

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \sum_{1 \leq k < l \leq Q} \|\mathrm{w}_k - \mathrm{w}_l\|^2 + \mathrm{C} \sum_{i=1}^{N} \sum_{k=1}^{Q} \xi_{ik} \right\}$$

$$s.t. \begin{cases} (w_{C(x_i)} - w_k)^T x_i + b_{C(x_i)} - b_k \geq 1 - \xi_{ik}, (1 \leq i \leq N), (1 \leq k \neq C(x_i) \leq Q) \\ \xi_{ik} \geq 0, \qquad (1 \leq i \leq N), (1 \leq k \neq C(x_i) \leq Q) \end{cases}$$

As usual, the non-negative slack variables ξ_{ik} have been introduced to take into account the fact that the data could be non-separable by the multivariate linear model. Their values characterise the empirical risk. Many algorithms are available to find an optimal solution (see, for instance, Fletcher [60], Smola [55] and Elisseeff [61]). In fact, additional specifications are required to ensure the unicity of the optimal solution, due to the following result. Let $(W^{(1)}, b^{(1)})$ be an optimal solution of Problem 1. Then the couple $(W^{(2)}, b^{(2)})$ such that $w_k^{(2)} = \mathrm{w}_k^{(1)} + \mathrm{v}$, $(1 \leq k \leq Q)$, where v is an arbitrary vector of $\mathbb{R}^d$, and $b_k^{(2)} = b_k^{(1)} + c$, $(1 \leq k \leq Q)$, where c is an arbitrary

real, is also an optimal solution of Problem 1. To ensure the unicity, we thus impose the following additional constraints:

$$\begin{cases}\Sigma_{k=1}^{Q}\ w_k = 0_d \\ \Sigma_{k=1}^{Q}\ b_k = 0\end{cases}$$

Taking into account these constraints, the SVM specified compares directly with the other multi-category SVMs developed so far. Indeed, the only difference between these SVMs lies in the expression of the objective function, as can be seen in Table 1.

In fact, all these models are utterly equivalent. A sketch of the proof can be found in the Appendix.

To sum up, starting from a uniform strong law of large numbers, we have been able to derive in the framework of statistical learning theory the specifications of the sole multi-class SVM published so far, which had been 'discovered' independently by several people. This result is interesting in its own right, since this rigorous justification was lacking, the reasons used by the aforementioned authors to support their choice being mainly the analogy with the bi-class case [21,31,32], and considerations regarding the regularisation theory [32]. In the following section, we establish that our framework can be used to specify other models.

2.7. Different Models Obtained by Changing the Metric

The specifications of the M-SVM considered in the previous subsection are based on the use of the $d_{L_\infty,l_1(s_{2N})}$ pseudo-distance and $\|.\|_{L_\infty,l_1}$ norm. We have seen that this is not a compulsory choice. Furthermore, to optimise performance, selecting a specific (pseudo)-metric should result from a study of the nature of the problem at hand. This is precisely one of the degrees of freedom which generates the family of models we have been dealing with. The primary limitation, to extend nicely Vapnik's bi-class SVM, is that the resulting training procedure must still amount to solving a convex programming problem. An exhaustive study of the different possibilities goes beyond the scope of this paper. In this section, we focus on the use of the pseudo-metric $d_{L_\infty,l_\infty(s)}$ (see Definition 6), thus specifying the other M-SVM which will be used in the experiments detailed in the following sections. For this definition of the pseudo-distance, and the corresponding norm, the following result, established by A. Elisseeff, holds:

Theorem 7. *If $\mathcal{H}$ is the multivariate linear model, $\mathcal{X} \subset \mathbb{R}^d$ and $\max_{k<l} \|w_k - w_l\|_{l_2}$ is bounded, then the expression of $\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2N)$ with respect to the norm $\|.\|_{l_{\infty,\infty}}$ satisfies*

$$\ln(\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2N)) = O\left(Qd \ln\left(\frac{1}{\gamma}\right)\right) \tag{8}$$

Building upon this formula, we can then specify the new multi-class SVM, the objective function of which takes the confidence interval into account through the convex function $\max_{k<l} \|w_k - w_l\|^2$. Its parameters are still the solution of a convex programming problem.

Problem 2

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

$$s.t. \begin{cases} \textit{constraints of Problem } 1 \\ \|w_k - w_l\|^2 \le t^2,\ (1 \le k < l \le Q) \end{cases}$$

The choice between this model and the former one can, for instance, be based on the knowledge available regarding the domain in which the data live.

3. IMPLEMENTATION OF M-SVMs TO COMBINE PROTEIN SECONDARY STRUCTURE PREDICTION METHODS

3.1. Characterisation of the Problem

To estimate the generalisation capabilities of M-SVMs, we used those specified above to combine protein secondary structure prediction methods. Prediction of protein 3D structure from the primary sequence of amino acids is a very challenging task, for which no satisfactory solution is currently available. A step forward is to predict the local conformation of the polypeptide chain, which is called the secondary structure. Protein secondary structure prediction is usually treated as a three-class discrimination task, which consists in assigning a conformational state α-helix, β-strand or aperiodic (coil), to each residue (amino acid) of a sequence. Apart from the fact that this problem is of central importance in structural biology, it presents characteristics which make it highly attractive from the point of view of pattern recognition. First, large databases of protein chains are available. It is thus possible to assess the models developed to process them in a wide spectrum, ranging from small samples to asymptotic behaviour. Secondly, many different methods have been proposed to predict the secondary structure, as was already pointed out in Section 1. Thirdly, combining these methods is not an easy task, since the risk of decreasing the training error

Table 1. Specifications of the different multi-category SVMs published so far

SVM	Objective function	Add. constraints
Vapnik [21]	$\Sigma_{k=1}^{Q} \|w_k^2\|^2$	–
Bredensteiner and Bennett [32]	$\Sigma_{k<l}^{Q} \|w_k - w_l\|^2 + \Sigma_{k=1}^{Q} \|w_k\|^2$	–
This work	$\Sigma_{k<l}^{Q} \|w_k - w_1\|^2$	$\Sigma_{k=1}^{Q}\ w_k = 0_d$

while increasing the test error has been stressed by many specialists of the field. For all these reasons, we consider this problem to be a touchstone to assess ensemble methods.

3.2. Experimental Protocol

We have implemented the SVMs associated with the $\|.\|_{l_\infty,l_1}$ norm (M-SVM1) and $\|.\|_{l_\infty,l_\infty}$ norm (M-SVM2) to combine the outputs of three of the most widely used secondary structure prediction methods: SOPMA [62], which uses multiple alignments, GOR IV [63], which is based on the formalism of the information theory, and SIMPA96 [64], a nearest-neighbour method. To assess the resulting predictions, we compared them with those of majority voting, a weighted average, optimal with respect to the least squares criterion, a Multi-Layer Perceptron (MLP) and the Multivariate Linear Regression Combiner (MLRC) introduced by Guermeur et al [11,41]. Two learning architectures involving multiple binary pattern recognition SVMs were also implemented. The first, involving three SVMs, was implementing the widely used one-against-all method (see, for instance, Vapnik [21]). The second one was the new DAGSVM of Platt and co-workers [65]. The MLR combiner requires the outputs of the experts to be class posterior probability estimates, and precisely to be non-negative and sum to unity. This is not the case with the prediction methods used here. To compare the combiners in a fair way, the outputs of the base classifiers were thus preliminary post-processed with the structure-to-structure filtering neural network described in Guermeur et al [41]. To constitute the training and test sets, we selected a release of the PDBSELECT database [66] containing 629 chains. These chains are made up of 147,518 residues. The secondary structure assignment was carried out according to DSSP [67]. To obtain unbiased estimates of the accuracy of the predictions, a variant of *stacked generalisation* [68] was applied, to train in sequence the filtering networks and the combiners. The database was divided into seven disjoint parts of roughly equal size. Based on this splitting, a two-stage cross-validation procedure was implemented. Each subset was iteratively used as the test set. The six remaining sets were then grouped by three in six different ways, to constitute as many pairs of disjoint training sets for the filtering networks and combiners. In this variant, the initial leave-one-out cross-validation procedure was thus replaced with a more computationally efficient six-fold cross-validation. This implementation of stacked generalisation, although suboptimal, has been observed not to deteriorate the generalisation performance, or more precisely the test error, which is consistent with other results (for instance, those reported in Breiman [69]). The prediction accuracy was assessed by means of four standard measures: the percentage of correctly predicted residues Q_3 for a three-state description of secondary structure (helix, extended and aperiodic); Pearson's/Matthews' correlation coefficient C [70]; the segment overlap measure Sov'94 [71,72]; and the standard deviation in the secondary structure content σ. The Sov measure plays a specific role, since it evaluates the quality of the prediction with respect to the conformational segments, which is a criterion of primary importance for the task. The figures characterising the behaviour of the individual methods, before and after filtering, have been gathered in Tables 2 and 3.

Table 2. Initial relative prediction accuracy of individual experts on a set of 629 non-homologous globular proteins from the PDBSELECT database

	GOR IV	SOPMA	SIMPA
Q_3	64.1	68.4	69.2
C_α	0.47	0.55	0.56
C_β	0.39	0.48	0.49
C_c	0.44	0.49	0.49
Sov	0.66	0.72	0.71
Sov_α	0.63	0.72	0.74
Sov_β	0.67	0.73	0.67
Sov_c	0.68	0.72	0.72
σ_α	13.9	10.8	10.8
σ_β	11.5	10.3	11.2
σ_c	9.4	9.9	11.6

Table 3. Relative prediction accuracy of individual experts on a set of 629 non-homologous globular proteins from the PDBSELECT database. Initial scores have been post-processed as was done in Guermeur et al [41]

	GOR IV	SOPMA	SIMPA
Q_3	66.5	69.7	69.4
C_α	0.51	0.58	0.57
C_β	0.43	0.49	0.49
C_c	0.46	0.50	0.49
Sov	0.68	0.71	0.70
Sov_α	0.67	0.73	0.72
Sov_β	0.64	0.68	0.66
Sov_c	0.70	0.72	0.71
σ_α	12.5	10.7	10.6
σ_β	11.6	11.1	10.7
σ_c	10.1	10.6	11.1

3.3. Raw Results of the Combinations

Table 4 summarises the relative performance of the different combiners. Figures given here correspond to SVMs and M-SVMs with radial basis kernels ($\sigma = 0.1$) and $C = 10$. These values of the parameters were selected, since they appeared to be 'satisfactory' for all the models. However, systematic experiments should be conducted in order to assess more precisely the influence of the parameterisation. Furthermore, additional experiments performed with two different polynomial kernels seem to suggest that the choice of a particular kernel could have significant incidence on the prediction accuracy (data not shown). The training procedure of the M-SVMs consisted in a slight modification of the algorithm described in Elisseeff [61]. The comparison of the predictive success of native methods and combinations illustrates the usefulness of the best combiners, which succeed in significantly increasing the recognition rate, even though the spectrum of quality among the classifiers

Table 4. Relative prediction accuracy of combiners on a set of 629 non-homologous globular proteins from the PDBSELECT database

	Vote	Average	MLP	MLRC	$SVM_{\alpha+\beta+c}$	DAGSVM	M-SVM1	M-SVM2
Q_3	70.2	70.9	71.2	71.3	71.4	71.4	71.7	71.6
C_α	0.59	0.60	0.60	0.60	0.60	0.60	0.61	0.60
C_β	0.49	0.50	0.52	0.52	0.52	0.52	0.52	0.53
C_c	0.51	0.50	0.52	0.52	0.52	0.52	0.52	0.52
Sov	0.72	0.71	0.72	0.72	0.72	0.72	0.73	0.72
Sov_α	0.73	0.72	0.73	0.74	0.73	0.74	0.74	0.74
Sov_β	0.69	0.70	0.70	0.68	0.69	0.69	0.68	0.68
Sov_c	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.74
σ_α	10.5	10.0	10.1	10.3	10.6	10.7	10.6	10.6
σ_β	10.3	10.2	10.1	10.9	10.9	10.9	10.8	10.8
σ_c	10.1	10.3	10.5	11.4	11.3	11.3	11.2	11.1

is wide. The M-SVMs obtain the best results, although the difference with the MLR combiner is too low to be statistically significant (the corresponding confidence is slightly over 0.9).

4. POST-PROCESSING OF THE CONFORMATIONAL SCORES

Promising as they may seem, these results are not sufficient to determine to what extent the conformational scores computed by the M-SVMs can be useful for the biologist. A property one usually expects from these scores is the possibility to use them in higher-level treatments, or simply to provide a measure of reliability of the predictions, as was done elsewhere [37,73,39]. To evaluate the quality of the combiners with respect to these criteria, we post-processed their outputs with a dynamic programming algorithm inspired by Ramesh and Wilpon [74], and first assessed for protein secondary prediction in earlier work [11]. To that end, the outputs of the SVMs and M-SVMs had to be preliminary standardised. This could be performed simply for each architecture, except the DAGSVM, which was thus discarded. The outputs of the weighted average, the MLR combiner and the MLP were already class posterior probability estimates. The underlying Inhomogeneous Hidden Markov Model (IHMM) is depicted in Fig. 1.

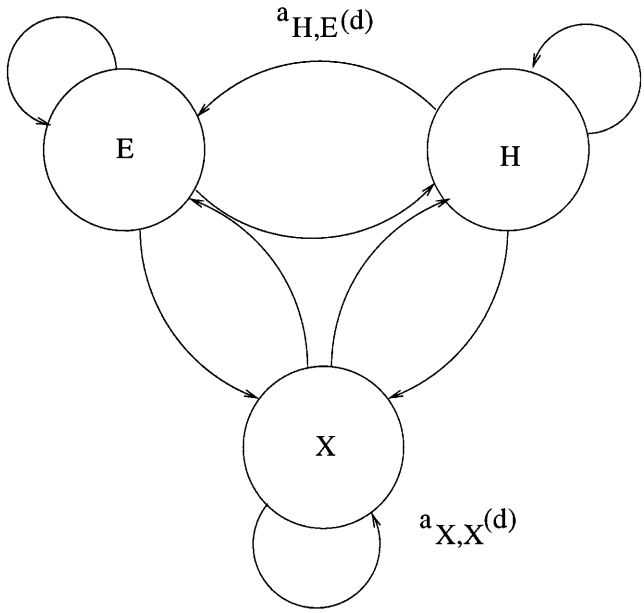

Fig. 1. Topology of the IHMM used to post-process the outputs of the combiners.

It has three states, one for each conformational state. The observations are the residues of the primary structure. The specificity of the algorithm lies in the modelling of state durations. Instead of the standard stationary (first order) state transition probabilities, the terms $a_{ij}(d)$ are used, where the extra parameter d represents the duration spent in the current (conformational) state i. These probabilities are estimated by the corresponding frequencies observed on the training set, whereas the observation probability density functions are derived from the outputs of the combiners by means of Bayes' theorem.

As can be seen in Table 5, the main advantage of such a post-processing is to narrow the gap between the length distributions of observed and predicted structural segments. However, its use also induces an improvement of the other standard measures of prediction accuracy. Indeed, the overall increase in recognition rate compared to the best results obtained so far on the same database, with the same experimental procedure [41], is now statistically significant with confidence exceeding 0.99. Once more, the M-SVMs obtain the best results among the combiners, the difference in performance between M-SVM1 and MLRC being statistically significant with confidence exceeding 0.98. This means that the values of their outputs carry some valuable information, and can for instance be used to estimate the class posterior probabilities. This property could prove to be very useful. One could, for instance, think about incorporating M-SVMs in hierarchical models far more complex than the one described here, such as the hybrid systems used in speech processing, user modelling or handwriting recognition.

Table 5. Quality of the predictions when the outputs of the combiners have been post-processed by an inhomogeneous DP algorithm

	Average	MLP	MLRC	$SVM_{\alpha+\beta+c}$	M-SVM1	M-SVM2
Q_3	71.1	71.5	71.5	72.0	72.3	72.2
C_α	0.60	0.61	0.61	0.61	0.62	0.60
C_β	0.50	0.52	0.52	0.51	0.53	0.51
C_c	0.52	0.52	0.52	0.52	0.53	0.52
Sov	0.72	0.74	0.74	0.74	0.74	0.73
Sov_α	0.72	0.74	0.74	0.74	0.75	0.74
Sov_β	0.68	0.70	0.70	0.70	0.70	0.69
Sov_c	0.72	0.73	0.75	0.75	0.75	0.75
σ_α	10.6	11.8	10.8	10.9	10.7	10.7
σ_β	10.4	10.6	11.1	11.0	10.9	10.9
σ_c	10.6	10.8	11.6	11.8	11.8	11.7

5. CONCLUSION AND FUTURE WORK

We have introduced a new family of multi-class SVMs. Unlike the previous work in the field [21,31,32], this study grounds the specification of the models directly on a uniform strong law of large numbers, with the consequence that the training procedure corresponds to an explicit implementation of the SRM inductive principle. Precisely, the training procedure amounts to minimising an expression of the guaranteed risk derived from a uniform convergence result specifically established for Q-category discriminant models. This bound is tighter than those derived so far for models with multiple outputs, which should make the implementation of the SRM principle better justified in the context of multi-category discriminant analysis. Moreover, an appealing feature of these M-SVMs is the fact that they exhibit the properties which represent the main advantages of the bi-class SVM. Indeed, our models appear as natural generalisations of Vapnik's, since their definitions are compatible with the extension of some of the main theorems regarding the generalisation capacities [75]. Two of them have been implemented to combine protein secondary structure prediction methods. These combinations appear to give better performance than those resulting from the implementation of standard ensemble methods, the gain becoming statistically significant when the outputs are post-processed with a DP algorithm. The recognition rate of the overall system highlights the benefits that one could expect from generalising the use of M-SVMs in the discriminant models performing tasks in biocomputing. So far, only bi-class SVMs, or variants of them, had been implemented in biology, for protein homology detection [76–78] or to process gene expression data [79].

Since we started this work, new prediction methods with superior accuracy have become available [80–82]. We have begun to assess the influence of their inclusion in different combinations [83]. The rudimentary hierarchical approach represented by the combination of the base classifiers and the DP algorithm can be developed in various ways. Currently, we are studying the use of *N-Best algorithms* [84] to provide the practitioner with alternative predictions among which he will be able to make his own choice, based on his expertise. Furthermore, we are implementing a system with multiple sliding windows, inspired by what has been done in Krogh and Riis [85]. Our long-term goal is to incorporate in our prediction systems the symbolic knowledge currently available for the task. This is the subject of collaborations with biologists. Concomitantly, we intend to derive new theoretical results, and specifically study the asymptotical behaviour of the different multi-class SVMs developed so far. In this respect, we see another benefit bestowed upon us by the use of models the capacity of which can be estimated precisely. It must be borne in mind that with the rapid strides made in molecular biology, especially in the field of genome sequencing, huge quantities of data will soon become available to underly the main predictive tasks in bioinformatics. As a consequence, implementing cross-validation to estimate the generalisation error will become prohibitive, particularly in the context of hierarchical systems such as ours [11], trained with variants of stacked generalisation [68]. Bounds similar to those presented in this article, and specifically distribution-dependent bounds, should then represent an efficient alternative, which could make it possible to save both in terms of CPU time and training data. These bounds could naturally be adapted, to make a better use of the specificities of model combination. Finally, *concentration inequalities* [86] could provide us with new tools to meet all these goals.

APPENDIX

This appendix contains details and illustrations regarding the computations of Section 2.

A.1. Illustration of Theorem 2

In the case of the $d_{l_\infty, l_\infty(s_{2N})}$ pseudo-metric, a particular case of inequality (2) can be simply derived form the following proposition:

Proposition 2. $\forall (g, h) \in \mathcal{H}^2$

$$d_{l_\infty, l_\infty(s_{2N})}(\Delta g^\gamma, \Delta h^\gamma) \leq d_{l_\infty, l_\infty(s_{2N})}(\Delta g, \Delta h)$$
$$\leq \frac{1}{2} \|T_{s_{2N}}(g) - T_{s_{2N}}(h)\|_{l_\infty, l_\infty} \quad (9)$$

The left inequality directly springs from the fact that π_γ satisfies a Lipschitz condition with constant 1. Furthermore, from an exhaustive study of the different cases:

- $(k = M_1(g, x_i) \wedge k = M_1(h, x_i))$;
- $(k = M_1(g, x_i) \wedge k \neq M_1(h, x_i))$;
- $(k \neq M_1(g, x_i) \wedge k \neq M_1(h, x_i))$.

It appears that we have

$$\forall (i, k) \in \{1, \ldots, 2N\} \times \{1, \ldots, Q\}, \exists l(i, k) \in \{1, \ldots, Q\} \backslash \{k\} /$$
$$|\Delta g_k(x_i) - \Delta h_k(x_i)| \leq \frac{1}{2} |g_{l(i,k)}(x_i) - g_k(x_i) - h_{l(i,k)}(x_i) + h_k(x_i)|$$

Let $T_{s_{2N}, i, j}(g, h)$ be the general term of $T_{s_{2N}}(g - h) = T_{s_{2N}}(g) - T_{s_{2N}}(h)$. We thus have

$$\forall (i, k) \in \{1, \ldots, 2N\} \times \{1, \ldots, Q\}, \exists j_0 \in \{1, \ldots, Q(Q-1)/2\} /$$
$$|\Delta g_k(x_i) - \Delta h_k(x_i)| \leq \frac{1}{2} |T_{s_{2N}, i, j_0}(g, h)|$$

By way of consequence,

$$\max_{x_i \in s_{2N}} \max_k |\Delta g_k(x_i) - \Delta h_k(x_i)|$$
$$\leq \frac{1}{2} \max_i \max_j |T_{s_{2N}, i, j}(g, h)|$$

from which Proposition 2 directly springs, due to the definitions of $d_{l_\infty, l_\infty(s_{2N})}$ and $\|.\|_{l_{\infty,\infty}}$.

A.2. Proof of Theorem 6

$$\forall s_{2N} \in \mathcal{X}^{2N}, \forall h \in \mathcal{H}, \|T_{s_{2N}}(h)\|_{l_\infty, l_1} = \max_{1 \leq i \leq 2N} \sum_{k<l} |(w_k - w_l)^T x_i|$$

Applying the Cauchy–Schwarz inequality, it becomes

$$\|T_{s_{2N}}(h)\|_{l_\infty, l_1} \leq \Lambda_{\mathcal{X}} \sum_{k<l} \|w_k - w_l\|$$

and consequently

$$\|T_{s_{2N}}(h)\|_{l_\infty, l_1} \leq \Lambda_{\mathcal{X}} \sqrt{\frac{Q(Q-1)}{2}} \sqrt{\sum_{k<1} \|w_k - w_l\|^2}$$

which is precisely (7).

A.3. Equivalence of the M-SVMs of the Literature

Computing the gradient of the Lagrangian function of the M-SVM proposed by Bredensteiner and Bennett and setting it equal to the null vector, at the optimum we get $\Sigma_{k=1}^{Q} W_k = 0_d$. This equality is also satisfied by the other multi-category SVMs [21,31]. As a consequence, at the optimum we get

$$\sum_{k<l}^{Q} \|w_k - w_l\|^2 = Q \sum_{k=1}^{Q} \|w_k\|^2$$

from which it springs that all the objective functions appearing in Table 1 are identical (modulo a multiplicative constant). A more formal proof can be obtained by studying the Kuhn–Tucker conditions associated with the different quadratic programming problems considered, conditions which are satisfied by a common set of primal variables W, b (with different Lagrange multipliers). The extension of the proof of equivalence to the non-separable case is also straightforward.

Acknowledgements

The author gratefully acknowledges the support of the ESPRIT funded Working Group N. 27150 'Neural Networks and Computational Learning Theory'. Most of the theory grounding the M-SVMs described in this paper has been developed in collaboration with H. Paugam-Moisy and A. Elisseeff. Experimental results could also be checked thanks to André's software. The author would like to thank G. Deléage, J. Garnier, C. Geourjon, J.-F. Gibrat and J.-M. Levin for the availability of the software and predictions of the SOPMA, GOR IV and SIMPA96 methods. Thanks are also due to A. Smola, K. Bennett, J. Weston and C. Watkins for interesting discussions on capacity measures and multi-class SVMs, and to A. Brun for carefully reading this manuscript.

References

1. Bates JM, Granger CWJ. The combination of forecasts. Opl Res Q 1969; 20:451–468
2. Granger CWJ. Combining forecasts – twenty years later. J Forecasting 1989; 8:167–173
3. Dickinson JP. Some statistical results in the combination of forecasts. Opl Res Q 1973; 24:253–260
4. Dickinson JP. Some comments on the combination of forecasts. Opl Res Q 1975; 26:205–210
5. Perrone MP. Improving Regression Estimation: Averaging Methods for Variance Reduction with Extensions to General Convex Measure Optimisation. PhD thesis, Department of Physics at Brown University, 1993
6. LeBlanc M, Tibshirani R. Combining estimates in regression and classification. Technical Report 9318, Department of Preventive Medicine and Biostatistics and Department of Statistics, Univeristy of Toronto, Toronto, 1993
7. Jordan MI, Jacobs RA. Hierarchical mixtures of experts and the EM algorithm. Neural Computation 1994; 6(2):181–214
8. Peng F, Jacobs RA, Tanner MA. Bayesian Inference in Mixture-of-Experts and Hierarchical Mixture-of-Experts Architectures. Technical report, Department of Biostatistics, University of Rochester, June 1994
9. Jacobs RA. Methods for combining experts' probability assessments. Neural Computation 1995; 7:867–888
10. Tumer K, Ghosh J. Estimating the Bayes error rate through classifier combining. ICPR'96, 1996; II:695–699

11. Guermeur Y. Combinaison de classifieurs statistiques, application à la prédiction de la structure secondaire des protéines. PhD thesis, Université Paris 6, 1997 (in French)
12. Tumer K, Ghosh J. Theoretical foundations of linear and order statistics combiners for neural pattern classifiers. Technical Report 95–02–98, The Computer and Vision Research Center, University of Texas, Austin, 1995
13. Tumer K, Ghosh J. Error correlation and error reduction in ensemble classifiers. Connection Science 1996; 8(3 & 4):385–404
14. Clemen RT, Winkler RL. Limits for the precision and value of information from dependent sources. Operations Res 1985; 33(2):427–442
15. Breiman L. Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California, Berkeley, 1996
16. Breiman L. Bagging predictors. Machine Learning 1996; 24:123–140
17. Schapire RE, Freund Y, Bartlett P, Lee WS. Boosting the margin: A new explanation for the effectiveness of voting methods. Ann Statistics 1998; 26(5):1651–1686
18. Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. COLT'92 1992; 144–152
19. Cortes C, Vapnik VN. Support-Vector Networks. Machine Learning 1995; 20:273–297
20. Vapnik VN. Estimation of Dependences Based on Empirical Data. Springer-Verlag, NY, 1982
21. Vapnik VN. Statistical Learning Theory. Wiley, NY, 1998
22. Burges CJC. A tutorial on Support Vector Machines for pattern recognition. Data Mining and Knowledge Discovery 1998; 2(2):121–167
23. Guermeur Y, Paugam-Moisy H. Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines. In: Sebban M, Venturini G (eds), Apprentissage Automatique. Hermès, 1999; 109–138 (in French)
24. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000
25. Vapnik VN, Chervonenkis, Ya A. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications 1971; 16:264–280
26. Haussler D. Decision theoretic generalizations of the PAC model for neural net and other learning applications. Information and Computation 1992; 100:78–150
27. Anthony M. Probabilistic analysis of learning in artificial neural networks: the PAC model and its variants. Neural Computing Surveys 1997; 1:1–47
28. Schölkopf B, Burges C, Vapnik V. Extracting support data for a given task. ICKDDM'95 1995; 252–257
29. Vapnik VN. The Nature of Statistical Learning Theory. Springer-Verlag, NY, 1995
30. Mayoraz E, Alpaydin E. Support Vector Machines for Multi-Class Classification. Technical Report 98–06, IDIAP, 1998
31. Weston J, Watkins C. Multi-class Support Vector Machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science, 1998
32. Bredensteiner EJ and Bennett KP. Multicategory classification by Support Vector Machines. Computational Optimization and Applications 1999; 12(1/3):53–79
33. 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
34. Eisenhaber F, Persson B, Argos P. Protein structure prediction: recognition of primary, secondary and tertiary structural features from amino acid sequence. Crit Rev Biochem Mol Biol 1995; 30:1–94
35. Rost B, O'Donoghue S. Sisyphus and prediction of protein structure. CABIOS 1997; 13:345–356
36. Baldi P, Brunak S. Bioinformatics: The machine learning approach. MIT Press, Cambridge, MA, 2nd edition, 2001
37. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993; 232:584–599
38. Zhang X, Mesirov JP, Waltz DL. Hybrid system for protein secondary structure prediction. J Mol Biol 1992; 225:1049–1063
39. Riis S, Krogh A. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. J Comput Biol 1996; 3:163–183
40. Biou V, Gibrat J-F, Levin J-M, Robson B, Garnier J. Secondary structure prediction: combination of three different methods. Prot Eng 1988; 2:185–191
41. Guermeur Y, Geourjon C, Gallinari P, Deléage G. Improved performance in protein secondary structure prediction by inhomogeneous score combination. Bioinformatics 1999; 15(5):413–421
42. Richard MD, Lippmann RP. Neural network classifiers estimate bayesian a posteriori probabilities. Neural Computation 1991; 3:461–483
43. Bishop CM. Neural Networks for Pattern Recognition. Oxford University Press, 1995
44. Kolmogorov AN, Tihomirov VM. ϵ-entropy and ϵ-capacity of sets in function spaces. Am Math Soc Translations (2) 1961; 17:277–364
45. Carl B, Stephani I. Entropy, Compactness, and the Approximation of Operators. Cambridge University Press, 1990
46. Barlett 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 Trans Information Theory 1998; 44(2):525–536
47. Shawe-Taylor J, Bartlett PL, Williamson RC, Anthony M. Structural Risk Minimization over Data-Dependent Hierarchies. Technical Report NC-TR-96-053, NeuroCOLT, 1996
48. Shawe-Taylor J, Cristianini N. Robust Bounds on Generalization from the Margin Distribution. Technical Report NC2-TR-1998-029, NeuroCOLT2, 1998
49. Guermeur Y, Elisseeff A, Zelus D. Bounding the capacity measure of multi-class discriminant models. Technical report, NC2-TR-2002-123, NeuroCOLT2, 2002
50. Natarajan BK. On learning sets and functions. Machine Learning 1989; 4:67–97
51. Guermeur Y, Elisseeff A, Paugam-Moisy H. Estimating the sample complexity of a multi-class discriminant model. ICANN'99. IEE, 1999; 310–315
52. Haussler D, Long PM. A generalization of Sauer's lemma. J Combinatorial Theory, Series A 1995; 71:219–240
53. Alon N, Ben-David S, Cesa-Bianchi N, Haussler D. Scale-sensitive dimensions, uniform convergence, and learnability. J ACM 1997; 44:615–631
54. Williamson RC, Smola AJ, Schölkopf B. Generalization performance of regularization networks and Support Vector Machines via entropy numbers of compact operators. IEEE Trans Information Theory 2001; 47(6):2516–2532
55. Smola AJ. Learning with Kernels. PhD thesis, Technische Universität Berlin, 1998
56. Kearns MJ, Schapire RE. Efficient distribution-free learning of probabilistic concepts. Proceedings 31st Annual Symposium on Foundations of Computer Science, IEEE Press, 1990; 1:382–391
57. Girosi F, Jones MJ, Poggio T. Priors, Stabilizers and Basis Functions: from regularization to radial, tensor and additive splines. Technical Report A.I. Memo N. 1430, C.B.C.L. Paper N. 75, MIT – AI laboratory, 1993
58. Williamson RC, Smola AJ, Schölkopf B. Entropy numbers of linear function classes. COLT'00, 2000
59. 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

60. Fletcher R. Practical Methods of Optimization. Wiley, 1987
61. Elisseeff A. Etude de la complexité et contrôle de la capacité des systèmes d'apprentissage: SVM multi-classe, réseaux de régularisation et réseaux de neurones multicouches. PhD thesis, ENS Lyon, 2000 (in French)
62. Geourjon C, Deléage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. CABIOS 1995; 11(6):681–684
63. Garnier J, Gibrat J-F, Robson B. GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence. Methods Enzymol 1996; 266:540–553
64. Levin J-M. Exploring the limits of nearest neighbour secondary structure prediction. Protein Eng 1997; 10(7):771–776
65. Platt J-C, Cristianini N, Shawe-Taylor J. Large margin DAGs for multiclass classification. NIPS'12 2000; 547–553
66. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994; 3:522–524
67. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983; 22(12):2577–2637
68. Wolpert DH. Stacked generalization. Neural Networks 1992; 5:241–259
69. Breiman L. Stacked regressions. Machine Learning 1996; 24:49–64
70. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975; 405:442–451
71. Rost B, Sander C, Schneiden R. Redefining the goals of protein secondary structure prediction. J Mol Biol 1994; 235:13–26
72. Zemla A, Venclovas Č, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins: Structure, Function, and Genetics 1999; 34:220–223
73. Geourjon C, Deléage G. SOPM: a self-optimized method for protein secondary structure prediction. Protein Engineering 1994; 7(2):157–164
74. Ramesh P, Wilpon JG. Modeling state durations in Hidden Markov Models for automatic speech recognition. ICASSP-92 1992; I:381–384
75. Guermeur Y, Elisseeff A, Paugam-Moisy H. A new multi-class SVM based on a uniform convergence result. IJCNN'00 2000; IV:183–188
76. Jaakola TS, Haussler D. Exploiting generative models in discriminative classifiers. Advances in Neural Information Processing Systems 11 1998
77. Jaakola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. J Computational Biology 2000; 7:95–114
78. Jaakola T, Diekhans M, Haussler D. Using the Fisher kernel method to detect remote protein homologies. ISMB'99 1999; 149–158
79. Brown M, Grundy W, Lin D, Cristianini N, Sugnet C, Furey T, Ares M, Haussler D. Knowledge-based analysis of microarray gene expression data using support vector machines. Technical report, University of California, Santa Cruz, 1999 (submitted for publication)
80. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999; 292:195–202
81. Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 1999; 15(11):937–946
82. Petersen TN, Lundegaard C, Nielsen M, Bohr H, Bohr J, Brunak S, Gippert GP, Lund O. Prediction of protein secondary structure at 80% accuracy. PROTEINS: Structure, Function, and Genetics 2000; 41(1):17–20
83. Guermeur Y, Zelus D. Combining protein secondary structure prediction models with ensemble methods of optimal complexity. JOBIM'01 2001; 97–104
84. Steinbiss V. Sentence hypotheses generation in a continuous-speech recognition system. Eurospeech-89 1989; 051–054
85. Krogh A, Riis S. Prediction of beta sheets in proteins. NIPS 8 1996; 917–923
86. Boucheron S, Lugosi G, Massart P. A sharp concentration inequality with applications. Technical Report NC2-TR-1999-057, NeuroCOLT2, 1999

Yann Guermeur received a 'diplôme d'ingénieur' from the IIE in 1991. He then worked in the industry for two years. Four years later, he obtained a PhD in computer science from the University Paris 6. He has worked at the ENS of Lyon and the University Paris 6. At present, he is assistant professor at the University Henri Poincaré, Nancy 1. He carries out his research at the LORIA. His main research topics are statistical learning theory and its applications to bioinformatics.

Correspondence and offprint requests to: Y. Guermeur, LORIA, Campus Scientifique, BP 239, 54 506 Vandoeuvre-lès-Nancy cedex, France. E-mail: Yann. Guermeur@loria.fr