Ensemble Methods of Appropriate Capacity for Multi-Class Support Vector Machines

Yann Guermeur

LORIA-CNRS
Campus Scientifique, BP 239
54506 Vandœuvre-lès-Nancy Cedex, France
(e-mail: Yann.Guermeur@loria.fr)

Abstract. Roughly speaking, there is one single model of pattern recognition support vector machine (SVM), with variants of lower popularity. On the contrary, among the different multi-class SVMs (M-SVMs) published, none is clearly favoured. Although several comparative studies between M-SVMs and decomposition methods have been reported, no attention had been paid so far to the combination of those models. We investigate the combination of M-SVMs with low capacity linear ensemble methods that estimate the class posterior probabilities.
Keywords: Ensemble methods, M-SVMs, Capacity control.

1 Introduction

Most of the statistical models developed for pattern recognition are based on a principle that does not change fundamentally with the number of categories. Things are more complex in the case of SVMs. Those machines were initially devised to compute dichotomies [2], and the first articles dealing with their use for polytomy computation report results obtained with decomposition methods [10]. M-SVMs were introduced later [11]. Since then, a few M-SVMs have been proposed and evaluated, with the attention of the community focusing on four models exhibiting distinct properties. Several comparative studies between M-SVMs and decomposition methods have established that in practice, each model has its advantages and drawbacks (see for instance [7]). The behaviours observed are different, in accordance with what was predicted by the theory. To the best of our knowledge, nobody has tried so far to take benefit of that phenomenon by combining different M-SVMs. To fill this void, we deal with the combination of M-SVMs subject to two constraints: the sample complexity of the *combiners* must be low, to avoid overfitting, and the outputs must be class posterior probability estimates.

We propose to combine the post-processed outputs of M-SVMs with linear ensemble methods which differ with respect to their objective function. They satisfy the aforementioned constraints and experimental results illustrate their potential. The organization of the paper is as follows. Section 2 provides a general introduction to the M-SVMs and characterizes the four main models. Section 3 deals with the description and statistical analysis of the linear combiners. Experimental results are exposed in Section 4, and we draw conclusions in Section 5. For lack of space, simple proofs are omitted.

2 Multi-class SVMs

We consider discrimination problems where $\mathcal{X}$ is the description space and $\mathcal{Y} = [\![1, Q]\!]$ is the set of categories. M-SVMs are kernel machines: they operate on a class of functions induced by a positive semidefinite function/kernel [1]. Let κ be a kernel on $\mathcal{X}^2$ and let $\left(\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa}\right)$ be the RKHS spanned by κ [1]. Let $\bar{\mathcal{H}} = \left(\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa}\right)^Q$ and $\mathcal{H} = \left(\left(\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa}\right) + \{1\}\right)^Q$. By construction, $\mathcal{H}$ is the class of vector-valued functions $h = (h_k)_{1 \leqslant k \leqslant Q}$ on $\mathcal{X}$ such that:

$$\forall k \in [\![1, Q]\!], \quad h_k(\cdot) = \sum_{i=1}^{m_k} \beta_{ik} \kappa\left(x_{ik}, \cdot\right) + b_k$$

where the x_{ik} are elements of $\mathcal{X}$ (the β_{ik} and b_k are scalars), as well as the limits of these functions as the sets $\{x_{ik} : 1 \leqslant i \leqslant m_k\}$ become dense in $\mathcal{X}$, in the norm induced by $\langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa}$. It springs from Mercer's theorem [1] that there exists a map Φ from $\mathcal{X}$ into a Hilbert space $\left(E_{\Phi(\mathcal{X})}, \langle \cdot, \cdot \rangle\right)$ such that $\mathcal{H}$ defines a multivariate affine model on $\Phi\left(\mathcal{X}\right)$. Functions h can then be rewritten as

$$h(\cdot) = \left(\langle w_k, \cdot \rangle + b_k\right)_{1 \leqslant k \leqslant Q}$$

where the vectors w_k belong to $E_{\Phi(\mathcal{X})}$. $\bar{\mathcal{H}}$ can then be seen as a multivariate linear model on $\Phi\left(\mathcal{X}\right)$, endowed with a norm $\| \cdot \|_{\bar{\mathcal{H}}}$ given by:

$$\forall \bar{h} \in \bar{\mathcal{H}}, \quad \left\|\bar{h}\right\|_{\bar{\mathcal{H}}} = \sqrt{\sum_{k=1}^{Q} \left\|\bar{h}_k\right\|_{\mathbf{H}_\kappa}^2} = \sqrt{\sum_{k=1}^{Q} \|w_k\|^2} = \sqrt{\sum_{k=1}^{Q} \langle w_k, w_k \rangle}.$$

Definition 1 (M-SVM). Let $\left((x_i, y_i)\right)_{1 \leqslant i \leqslant m} \in \left(\mathcal{X} \times [\![1, Q]\!]\right)^m$ and $\lambda \in \mathbb{R}_+^*$. A *Q-category M-SVM* is a classifier obtained by minimizing over the hyperplane $\sum_{k=1}^{Q} h_k = 0$ of $\mathcal{H}$ a penalized convexified empirical risk of the form:

$$J_{\text{M-SVM}}\left(h\right) = \|\xi_{\text{M-SVM}}\|_{\text{M-SVM}}^p + \lambda \left\|\bar{h}\right\|_{\bar{\mathcal{H}}}^2$$

where $\xi_{\text{M-SVM}}$ is a vector of slack variables associated with the constraints of good classification, which are linear, and $\|\cdot\|_{\text{M-SVM}}$ is either the ℓ_1 norm ($p = 1$) or the norm induced by a symmetric positive definite matrix ($p = 2$).

In chronological order, the four main M-SVMs are the machines of Weston and Watkins (WW) [11], Crammer and Singer (CS) [3], and Lee, Lin and Wahba (LLW) [8], and the M-SVM^2 [6]. Their characteristics are summarized in Table 1 (in the sequel, when no confusion is possible, the subscript identifying the machine, i.e., instantiating M-SVM, is omitted).

M-SVM	$\xi_{\text{M-SVM}}$ (constraints of good classification)	$\|\cdot\|_{\text{M-SVM}}$	p
WW	$\forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \begin{cases} h_{y_i}(x_i) - h_k(x_i) \geqslant 1 - \xi_{ik} \\ \xi_{ik} \geqslant 0 \end{cases}$	ℓ_1	1
CS	$\forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \ h_{y_i}(x_i) - h_k(x_i) \geqslant 1 - \xi_i$ $\forall k \in [\![1,Q]\!],\ b_k = 0,\ \forall i \in [\![1,m]\!],\ \ \xi_i \geqslant 0$	ℓ_1	1
LLW	$\forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \begin{cases} h_k(x_i) \leqslant -\frac{1}{Q-1} + \xi_{ik} \\ \xi_{ik} \geqslant 0 \end{cases}$	ℓ_1	1
M-SVM^2	$\forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \ h_k(x_i) \leqslant -\frac{1}{Q-1} + \xi_{ik}$	M	2

Table 1. Specifications of the four main M-SVMs

While the CS-M-SVM has one slack variable per training example, the other three have $Q-1$. In that second case, ξ is the vector of $\mathbb{R}^{Qm}$ whose component of index $(i-1)Q+k$ is ξ_{ik}, with the ξ_{iy_i} being equal to 0. The matrix M is such that the quadratic form $\xi^T M \xi$ defining $\|\xi_{\text{M-SVM}^2}\|^2_{\text{M-SVM}^2}$ is given by $\xi^T M \xi = \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \sum_{l=1}^{Q} \delta_{i,j}\left(\delta_{k,l}+1\right)\xi_{ik}\xi_{jl}$, where δ is the Kronecker symbol. The following observations illustrate the differences between these machines. The implementation of the training algorithm of the CS-M-SVM is the easiest one, the LLW-M-SVM was the first M-SVM with a Fisher consistent loss and the M-SVM^2 is the first soft margin M-SVM for which a generalized radius-margin bound applies.

3 Linear ensemble methods

We make the hypothesis that N classifiers $g^{(j)} = \left(g_k^{(j)}\right)_{1 \leqslant k \leqslant Q}$, $(1 \leqslant j \leqslant N)$, are available to perform the classification task of interest. For all n in $\mathbb{N}^*$, let U_n be the polytope given by: $U_n = \left\{u = (u_p)_{1 \leqslant p \leqslant n} \in \mathbb{R}_+^n : \sum_{p=1}^{n} u_p = 1\right\}$. The outputs of the classifiers are supposed to be nonnegative and sum to one, i.e., to belong to U_Q. We first describe the ensemble methods considered, and then characterize their sample complexity as a function of N and Q.

3.1 Class of functions and training algorithms

Let $\tilde{g}$ be the function from $\mathcal{X}$ to U_Q^N obtained by appending the component functions of the N classifiers $g^{(j)}$: $g_k^{(j)}$ is its component function of index $(j-1)Q+k$.

Definition 2 (multivariate linear model). We consider the multivariate linear model parameterized by the matrix $B \in \mathcal{M}_{Q,NQ}(\mathbb{R})$ such that

$$\forall x \in \mathcal{X},\ \ \bar{g}(x) = \left(\bar{g}_k(x)\right)_{1 \leqslant k \leqslant Q} = B\tilde{g}(x)$$

$$s.t.\ \ \forall u \in U_Q^N,\ \ Bu \in U_Q.$$

The transposes of the rows of B are denoted β_k, so that $\bar{g}_k(x) = \beta_k^T \tilde{g}(x)$, and $\beta = (\beta_k)_{1 \leqslant k \leqslant Q} \in \mathbb{R}^{NQ^2}$. The general term of B is written with three indices, i.e., β_{kjl} (β_{kjl} is the component of β_k of index $(j-1)Q + l$). Let $d_m = \{(x_i, y_i) : 1 \leqslant i \leqslant m\}$ be a training set. Let t_k denote the *one of Q coding* of category k: $t_k = (\delta_{k,l})_{1 \leqslant l \leqslant Q}$. We consider combiners obtained by solving convex programming problems of the form:

Problem 1 (Linear ensemble methods).

$$\min_{B} \sum_{i=1}^{m} \ell_{\mathrm{LEM}} \left(t_{y_i}, B\tilde{g}\left(x_i\right) \right)$$

$$s.t.\ \forall u \in U_Q^N, \ \ Bu \in U_Q$$

where the loss function ℓ_{LEM} is convex.

Proposition 1 makes the optimization computationally tractable.

Proposition 1. *Irrespective of the nature of ℓ_{LEM}, there is an optimal solution of Problem 1 which belongs to the polytope $V_{N,Q}$ given by:*

$$\begin{cases} \beta \in \mathbb{R}_+^{NQ^2} \\ \forall j \in [\![1, N]\!], \ \forall l \in [\![1, Q-1]\!], \ \sum_{k=1}^{Q} \left(\beta_{kjl} - \beta_{kjQ}\right) = 0 \\ \sum_{k=1}^{Q} \sum_{j=1}^{N} \beta_{kjQ} = 1 \end{cases} .$$

We focus on two natural choices for ℓ_{LEM} that give rise to class posterior probability estimates: the quadratic loss and the cross-entropy loss. Let $\tilde{G}$ be the matrix of $\mathcal{M}_{m,NQ}(\mathbb{R})$ whose rows are the vectors $\tilde{g}\left(x_i\right)^T$. Let I_Q denote the identity matrix of size Q and $\otimes$ the Kronecker product. For all k in $[\![1, Q]\!]$, let $Y_k = \left(\delta_{y_i,k}\right)_{1 \leqslant i \leqslant m}$ and let $Y = (Y_k)_{1 \leqslant k \leqslant Q} \in \{0,1\}^{Qm}$. The objective function (empirical risk) corresponding to the quadratic loss is:

$$J_{\mathrm{Quad}}\left(\bar{g}\right) = \frac{1}{2} \beta^T \left\{ I_Q \otimes \left(\tilde{G}^T \tilde{G} \right) \right\} \beta - \left\{ Y^T \left(I_Q \otimes \tilde{G} \right) \right\} \beta.$$

The standard expression of the cross-entropy loss ℓ_{CE} is:

$$\forall (x,y) \in \mathcal{X} \times \mathcal{Y}, \ \ \ell_{\mathrm{CE}}\left(t_y, \bar{g}\left(x\right)\right) = -\sum_{k=1}^{Q} \left\{ \delta_{y,k} \ln\left(\bar{g}_k\left(x\right)\right) + \left(1 - \delta_{y,k}\right) \ln\left(1 - \bar{g}_k\left(x\right)\right) \right\}.$$

This loss function can be used here since $U_Q \subset [0,1]^Q$. We take benefit of the fact that the component functions sum to one to substitute to ℓ_{CE} a simplified expression, so that the objective function becomes

$$J_{\mathrm{CE}}\left(\bar{g}\right) = -\sum_{i=1}^{m} \sum_{k=1}^{Q} \delta_{y_i,k} \ln\left(\frac{\beta_k^T \tilde{g}_k\left(x_i\right)}{\delta_{y_i,k}} \right).$$

It is well known that the combination of the one of Q coding of the desired outputs with these two loss functions leads to the selection of a function that generates estimates of the class posterior probabilities (see [9] for a proof).

3.2 Sample complexity of the linear ensemble methods

For $\beta \in V_{N,Q}$, let $g_\beta = (g_{\beta,k})_{1 \leqslant k \leqslant Q}$ be the function from U_Q^N to U_Q such that $g_\beta(u) = \left(\beta_k^T u\right)_{1 \leqslant k \leqslant Q}$. Let $\mathcal{G}_\beta = \{g_\beta : \beta \in V_{N,Q}\}$. We identify the capacity of the combiners with that of $\mathcal{G}_\beta$. In [4], we prooved that for large margin multi-category classifiers, the appropriate generalizations of the Vapnik-Chervonenkis dimension are the γ-Ψ-dimensions. Their use involves the application of margin operators. Here, a suitable γ-Ψ-dimension is an extension of the Natarajan dimension and the operator needed is the Δ one.

Definition 3 (Δ operator). Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}^Q$.

$$\forall g \in \mathcal{G},\ \forall x \in \mathcal{X},\ \ \Delta g(x) = \left(\Delta g_k(x)\right)_{1 \leq k \leq Q} = \frac{1}{2}\left(g_k(x) - \max_{l \neq k} g_l(x)\right)_{1 \leq k \leq Q}.$$

For the sake of simplicity, Δg_k is used in place of $(\Delta g)_k$. Let $\Delta\mathcal{G} = \{\Delta g : g \in \mathcal{G}\}$.

Definition 4 (Natarajan dimension with margin γ). Let $\mathcal{G}$ be defined as above. For $\gamma \in \mathbb{R}_+^*$, $s_n = \{x_i : 1 \leqslant i \leqslant n\} \subset \mathcal{X}$ is said to be *γ-N-shattered* by $\Delta\mathcal{G}$ if there is a set $I(s_n) = \{(i_1(x_i), i_2(x_i)) : 1 \leqslant i \leqslant n\}$ of couples of integers satisfying $1 \leqslant i_1(x_i) < i_2(x_i) \leqslant Q$ and a vector $v_b = (b_i)$ in $\mathbb{R}^n$ such that, for each vector $v_y = (y_i)$ in $\{-1, 1\}^n$, there is g_y in $\mathcal{G}$ satisfying

$$\forall i \in [\![1, n]\!],\ \begin{cases} \text{if } y_i = \ \ \ 1,\ \Delta g_{y,i_1(x_i)}(x_i) - b_i \geqslant \gamma \\ \text{if } y_i = -1,\ \Delta g_{y,i_2(x_i)}(x_i) + b_i \geqslant \gamma \end{cases}.$$

The *Natarajan dimension with margin* γ of $\Delta\mathcal{G}$, N-dim$(\Delta\mathcal{G}, \gamma)$, is the maximal cardinality of a subset of $\mathcal{X}$ γ-N-shattered by $\Delta\mathcal{G}$, if this maximum exists, and $+\infty$ otherwise.

An upper bound on N-dim$(\Delta\mathcal{G}_\beta, \gamma)$ is provided by Theorem 1.

Theorem 1.

$$\forall \gamma \in \mathbb{R}_+^*,\ \ \textit{N-dim}\left(\Delta\mathcal{G}_\beta, \gamma\right) \leqslant \binom{Q}{2}\frac{NQ}{4\gamma^2}. \tag{1}$$

The proof of Theorem 1 is based on two lemmas.

Lemma 1. *Let $\gamma \in \mathbb{R}_+^*$ and $n \in \mathbb{N}^*$. If $s_n = \{u_i : 1 \leqslant i \leqslant n\} \subset U_Q^N$ is γ-N-shattered by $\Delta\mathcal{G}_\beta$, then there exists a subset s_p of s_n of cardinality $p = \left\lceil \frac{n}{\binom{Q}{2}} \right\rceil$ such that for every partition of s_p into two subsets $s_{p,1}$ and $s_{p,2}$,*

$$\left\| \sum_{u_i \in s_{p,1}} u_i - \sum_{u_i \in s_{p,2}} u_i \right\|_2 \geqslant \frac{2p}{\sqrt{Q}}\gamma. \tag{2}$$

Proof. Let $(I(s_n), v_b)$ witness the γ-N-shattering of s_n by $\Delta\mathcal{G}_\beta$. According to the pigeonhole principle, there is at least one couple of indices (k_1, k_2) such that there are at least p points in s_n for which $(i_1(u_i), i_2(u_i))$ is (k_1, k_2). For the sake of simplicity, the points in s_n are reordered in such a way that the p first of them exhibit this property. The corresponding subset of s_n is denoted s_p. This means that for all vector $v_y = (y_i)$ in $\{-1, 1\}^n$, there is a function g_{β_y} in $\mathcal{G}_\beta$ characterized by the vector $\beta_y = (\beta_{y,k})_{1 \leqslant k \leqslant Q} \in V_{N,Q}$ such that:

$$\forall i \in [\![1, p]\!], \quad \begin{cases} \text{if } y_i = 1, \; \Delta g_{\beta_y,k_1}(u_i) - b_i \geqslant \gamma \\ \text{if } y_i = -1, \; \Delta g_{\beta_y,k_2}(u_i) + b_i \geqslant \gamma \end{cases} .$$

By definition of $\mathcal{G}_\beta$ and the margin operator Δ, this is implies:

$$\forall i \in [\![1, p]\!], \quad \begin{cases} \text{if } y_i = 1, \; \frac{1}{2}\left(\beta_{y,k_1}^T u_i - \beta_{y,k_2}^T u_i\right) - b_i \geqslant \gamma \\ \text{if } y_i = -1, \; \frac{1}{2}\left(\beta_{y,k_2}^T u_i - \beta_{y,k_1}^T u_i\right) + b_i \geqslant \gamma \end{cases} . \tag{3}$$

Consider now any partition of s_p into two subsets $s_{p,1}$ and $s_{p,2}$. Consider any vector v_y in $\{-1, 1\}^n$ such that $y_i = 1$ if $u_i \in s_{p,1}$ and $y_i = -1$ if $u_i \in s_{p,2}$. It results from (3) that there exists g_{β_y} in $\mathcal{G}_\beta$ such that:

$$\frac{1}{2}\left(\beta_{y,k_1} - \beta_{y,k_2}\right)^T \left(\sum_{u_i \in s_{p,1}} u_i - \sum_{u_i \in s_{p,2}} u_i \right) - \sum_{u_i \in s_{p,1}} b_i + \sum_{u_i \in s_{p,2}} b_i \geqslant p\gamma.$$

Conversely, consider any vector v_y such that $y_i = -1$ if $u_i \in s_{p,1}$ and $y_i = -1$ if $u_i \in s_{p,2}$. There exists g_{β_y} in $\mathcal{G}_\beta$ such that:

$$\frac{1}{2}\left(\beta_{y,k_2} - \beta_{y,k_1}\right)^T \left(\sum_{u_i \in s_{p,1}} u_i - \sum_{u_i \in s_{p,2}} u_i \right) + \sum_{u_i \in s_{p,1}} b_i - \sum_{u_i \in s_{p,2}} b_i \geqslant p\gamma.$$

Thus, whatever the sign of $\sum_{u_i \in s_{p,1}} b_i - \sum_{u_i \in s_{p,2}} b_i$ is, by application of the Cauchy-Schwarz inequality, there is a vector β_y in $V_{N,Q}$ such that:

$$\frac{1}{2}\left\|\beta_{y,k_1} - \beta_{y,k_2}\right\|_2 \left\| \sum_{u_i \in s_{p,1}} u_i - \sum_{u_i \in s_{p,2}} u_i \right\|_2 \geqslant p\gamma. \tag{4}$$

For $\beta \in V_{N,Q}$, $\max_{1 \leqslant k \neq l \leqslant Q} \|\beta_k - \beta_l\|_2$ is reached when one of the vectors is the null vector and the other one concentrates all the mass on as few components as possible. A situation of this kind is obtained by choosing any couple (k_0, j_0) in $[\![1, Q]\!] \times [\![1, N]\!]$ and defining the vector β as follows

$$\forall k \in [\![1, Q]\!], \; \forall j \in [\![1, N]\!], \; \forall l \in [\![1, Q]\!], \quad \beta_{kjl} = \delta_{k_0,k}\delta_{j_0,j}.$$

In that case, for all k in $[\![1, Q]\!] \backslash \{k_0\}$, $\|\beta_{k_0} - \beta_k\|_2 = \sqrt{Q}$. Thus, $\|\beta_{y,k_1} - \beta_{y,k_2}\|_2 \leqslant \sqrt{Q}$, and a substitution in (4) concludes the proof.

Lemma 2. *For all $n \in \mathbb{N}^*$, all subset $s_n = \{u_i : 1 \leqslant i \leqslant n\}$ of U_Q^N can be partitioned into two subsets s_1 and s_2 satisfying*

$$\left\| \sum_{u_i \in s_1} u_i - \sum_{u_i \in s_2} u_i \right\|_2 \leqslant \sqrt{Nn}. \tag{5}$$

Proof. Let $\sigma = (\sigma_i)_{1 \leqslant i \leqslant n}$ be a Rademacher sequence: the σ_i are i.i.d. Bernoulli random variables with parameter $p = \frac{1}{2}$. $\forall (i,j) \in [\![1,n]\!]^2, \mathbb{E}_\sigma [\sigma_i \sigma_j] = \delta_{i,j}$.

$$\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i u_i \right\|_2^2 = \sum_{i=1}^n \sum_{j=1}^n u_i^T u_j \mathbb{E}_\sigma [\sigma_i \sigma_j] = \sum_{i=1}^n \|u_i\|_2^2 \leqslant n \max_{u \in U_Q^N} \|u\|_2^2 .$$

The points of U_Q^N whose ℓ_2-norm is maximum are its vertices. The value of their ℓ_2-norm is $\sqrt{N}$. Thus, $\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i u_i \right\|_2^2 \leqslant Nn$. This implies that there exists a vector $v = (v_i)_{1 \leqslant i \leqslant n} \in \{-1,1\}^n$ such that $\left\| \sum_{i=1}^n v_i u_i \right\|_2 \leqslant \sqrt{Nn}$. Setting $s_1 = \{u_i \in s_n : v_i = 1\}$ and $s_2 = s_n \setminus s_1$ then concludes the proof.

With Lemmas 1 and 2 at hand, the proof of Theorem 1 is straightforward.

Proof. Let $s_n = \{u_i : 1 \leqslant i \leqslant n\}$ be a subset of U_Q^N γ-N-shattered by $\Delta\mathcal{G}_\beta$. According to Lemma 1, there is at least a subset s_p of s_n of cardinality $p = \left\lceil \frac{n}{\binom{Q}{2}} \right\rceil$ satisfying (2) for all its partitions into two subsets $s_{p,1}$ and $s_{p,2}$. Since, according to Lemma 2, there is at least one of these partitions for which (5) holds true, $\frac{2p}{\sqrt{Q}} \gamma \leqslant \sqrt{Np}$, which implies (1) since $n \leqslant \binom{Q}{2} p$.

4 Experimental results

The problem considered is protein secondary structure prediction. It consists in assigning to each residue of a protein sequence its conformational state: α-helix, β-strand or coil ($Q = 3$). The four main M-SVMs and the two combiners resulting from using the quadratic and cross-entropy losses are assessed on the P1096 data set [5]. The experimental protocol differs from the one used in [5] in two respects. The outputs of the M-SVMs are normalized:

$$\forall j \in [\![1,4]\!],\ \forall k \in [\![1,3]\!],\ \ g_k^{(j)}(\cdot) = \frac{\exp\left(h_k^{(j)}(\cdot)\right)}{\sum_{l=1}^3 \exp\left(h_l^{(j)}(\cdot)\right)}$$

and an additional level of cross-validation is introduced so as to train the M-SVMs and the combiners on different data. Table 2 summarizes the results obtained. Prediction accuracy is described by means of three standard measures giving complementary indications: the recognition rate Q_3, Matthews' correlation coefficients $C_{\alpha/\beta/\mathrm{coil}}$, and the segment overlap measure Sov.

The comparison of the performance of the M-SVMs considered individually and in the framework of a combination shows a gain induced by the combination which is statistically significant with confidence at least 0.95.

	WW	CS	LLW	M-SVM^2	Combiner Quad	Combiner CE
Q_3	66.9	66.5	66.7	66.7	67.7	67.6
C_α	0.52	0.50	0.51	0.51	0.54	0.54
C_β	0.42	0.40	0.40	0.42	0.44	0.43
C_{coil}	0.46	0.44	0.46	0.44	0.47	0.48
Sov	56.0	55.7	56.2	56.0	58.1	57.9

Table 2. Relative prediction accuracy of the M-SVMs and the linear combiners on the 1096 sequences (268575 residues) of the P1096 data set

5 Conclusions and ongoing research

We have introduced linear combiners for M-SVMs. Their low sample complexity should prevent them from overfitting and they provide estimates of the class posterior probabilities. We are currently performing a large scale study of their performance, focusing on the quality of these estimates, used to derive emission probabilities in a hidden Markov model.

References

1. A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Kluwer Academic Publishers, Boston, 2004.
2. B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT'92*, pages 144–152, 1992.
3. K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.
4. Y. Guermeur. Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs. *Communications in Statistics*, 39(3):543–557, 2010.
5. Y. Guermeur, A. Lifchitz, and R. Vert. A kernel for protein secondary structure prediction. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, chapter 9, pages 193–206. The MIT Press, 2004.
6. Y. Guermeur and E. Monfrini. A quadratic loss multi-class SVM for which a radius-margin bound applies. *Informatica.* (conditionally accepted).
7. C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.
8. Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.
9. M.D. Richard and R.P. Lippmann. Neural network classifiers estimate Bayesian *a posteriori* probabilities. *Neural Computation*, 3(4):461–483, 1991.
10. B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In *KDD'95*, pages 252–257, 1995.
11. J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science, 1998.