INFORMATICA, 2011, Vol. 22, No. 1, 73–96

A Quadratic Loss Multi-Class SVM for which a Radius–Margin Bound Applies

Yann GUERMEUR[1], Emmanuel MONFRINI[2]
[1]*LORIA-CNRS, Campus Scientifique, BP 239*
54506 Vandœuvre-lès-Nancy cedex, France
[2]*TELECOM SudParis, 9 rue Charles Fourier*
91011 EVRY cedex, France
e-mail: yann.guermeur@loria.fr, emmanuel.monfrini@it-sudparis.eu

Received: October 2009; accepted: December 2010

Abstract. To set the values of the hyperparameters of a support vector machine (SVM), the method of choice is cross-validation. Several upper bounds on the leave-one-out error of the pattern recognition SVM have been derived. One of the most popular is the radius–margin bound. It applies to the hard margin machine, and, by extension, to the 2-norm SVM. In this article, we introduce the first quadratic loss multi-class SVM: the M-SVM^2. It can be seen as a direct extension of the 2-norm SVM to the multi-class case, which we establish by deriving the corresponding generalized radius–margin bound.

Keywords: multi-class SVMs, model selection, leave-one-out cross-validation error, radius–margin bounds.

1. Introduction

Using an SVM (Boser *et al.*, 1992; Cortes and Vapnik, 1995) requires to set the values of two types of hyperparameters: the soft margin parameter C and the parameters of the kernel. To perform this model selection task, the solution of choice consists in applying a cross-validation procedure. Among those procedures, the leave-one-out one appears especially attractive, since it is known to produce an estimator of the generalization error which is almost unbiased (Luntz and Brailovsky, 1969). The seamy side of things is that it is highly time consuming. This is the reason why, in recent years, a number of upper bounds on the leave-one-out error of the pattern recognition SVM have been proposed (see Chapelle *et al.*, 2002, for a survey). Although the tightest one is the *span bound* (Vapnik and Chapelle, 2000), the results of Chapelle *et al.* (2002) show that when using the 2-norm SVM (see, for instance, Chapter 7 in Shawe-Taylor and Cristianini, 2004), the *radius–margin* bound (Vapnik, 1998) achieves equivalent performance for model selection while being far simpler to compute. These results are corroborated by those of several comparative studies, among which the one performed by Duan *et al.* (2003). As a consequence, this bound, with its variants (Chung *et al.*, 2003), is currently the most popular one. The first studies dealing with the use of SVMs for multi-category classification

(Schölkopf *et al.*, 1995; Vapnik, 1995) report results obtained with decomposition methods involving Vapnik's machine. A recent implementation of this approach can be found in Balys and Rudzkis (2010). Multi-class support vector machines (M-SVMs) were introduced later by Weston and Watkins (1998). Over more than a decade, many M-SVMs have been developed (see, Guermeur, 2007, for a survey), among which three have been the subject of extensive studies. However, to the best of our knowledge, literature only proposes a single multi-class extension of the radius–margin bound. Introduced by Wang *et al.* (2008), it makes use of the bi-class bound in the framework of the one-versus-one decomposition method. As such, it does not represent a direct generalization of the initial result to an M-SVM, and the authors state that "such a theoretical generalization of this bound is not that straightforward because this bound is rooted in the theoretical basis of binary SVMs."

In this article, a new M-SVM is introduced: the M-SVM^2. It can be seen either as a quadratic loss variant of the M-SVM of Lee *et al.* (2004) (LLW-M-SVM) or as a multi-class extension of the 2-norm SVM. A generalized radius–margin bound on the leave-one-out error of the hard margin version of the LLW-M-SVM is then established and assessed. This provides us with a differentiable objective function to perform model selection for the M-SVM^2. A comparative study including all four M-SVMs illustrates the generalization performance of the new machine.

The organization of this paper is as follows. Section 2 provides a general introduction to the M-SVMs and characterizes the three main models. Section 3 focuses on the LLW-M-SVM and Section 4 introduces the M-SVM^2. Section 5 is devoted to the multi-class radius–margin bound. Experimental results are given in Section 6. We draw conclusions and outline our ongoing research in Section 7.

2. Multi-Class SVMs

Like the (bi-class) SVMs, the M-SVMs are large margin classifiers which are devised in the framework of Vapnik's statistical learning theory (Vapnik, 1998).

2.1. *Formalization of the Learning Problem*

We consider the case of Q-category pattern recognition problems with $3 \leqslant Q < \infty$. Each object is represented by its description $x \in \mathcal{X}$ and the set $\mathcal{Y}$ of the categories y can be identified with the set $[\![1, Q]\!]$. We assume that the link between descriptions and categories can be described by an unknown probability measure P on $\mathcal{X} \times \mathcal{Y}$. The learning problem then consists in selecting a set $\mathcal{G}$ of functions $g = (g_k)_{1 \leqslant k \leqslant Q}$ from $\mathcal{X}$ to $\mathbb{R}^Q$, and a function g^* in that set classifying data in an optimal way. The criterion which is to be optimized must be specified. The function g assigns $x \in \mathcal{X}$ to the category l if and only if $g_l(x) > \max_{k \neq l} g_k(x)$. In case of ex æquo, x is assigned to a dummy category denoted by $*$. Let f be the decision rule (from $\mathcal{X}$ to $\mathcal{Y} \bigcup \{*\}$) associated with g and (X, Y) a random pair with values in $\mathcal{X} \times \mathcal{Y}$ distributed according to P. Ideally, the objective function to be minimized over $\mathcal{G}$ is $P(f(X) \neq Y)$. In practice, since P is unknown, other criteria are used and the optimization process, called *training*, is based on

empirical data. More precisely, we assume that what we are given to select both $\mathcal{G}$ and g^* is an m-sample $D_m = ((X_i, Y_i))_{1 \leqslant i \leqslant m}$ of independent copies of (X, Y). A realisation d_m of D_m is called a *training set*. This article focuses on the choice of $\mathcal{G}$, named *model selection*, in the particular case when the model considered is an M-SVM.

2.2. *Architecture and Training Algorithms*

M-SVMs, like all the SVMs, are *kernel machines* (Shawe-Taylor and Cristianini, 2004; Norkin and Keyzer, 2009), which means that they operate on a class of functions induced by a positive type function/kernel. This calls for the formulation of some definitions and basic results. For the sake of simplicity, we consider real-valued functions only, although the general form of these definitions and results involves complex-valued functions.

DEFINITION 1 (Positive type (positive semidefinite) function, Definition 2 in Berlinet and Thomas-Agnan, 2004). A real-valued function κ on $\mathcal{X}^2$ is called a *positive type function* (or a *positive semidefinite function*) if it is symmetric and

$$\forall n \in \mathbb{N}^*, \quad \forall (x_i)_{1 \leqslant i \leqslant n} \in \mathcal{X}^n, \quad \forall (a_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n, \quad \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \kappa(x_i, x_j) \geqslant 0.$$

DEFINITION 2 (Reproducing kernel Hilbert space, Definition 1 in Berlinet and Thomas-Agnan, 2004). Let $(\mathbf{H}, \langle \cdot, \cdot \rangle_{\mathbf{H}})$ be a Hilbert space of real-valued functions on $\mathcal{X}$. A real-valued function κ on $\mathcal{X}^2$ is a *reproducing kernel* of $\mathbf{H}$ if and only if

1. $\forall x \in \mathcal{X},\ \kappa_x = \kappa(x, \cdot) \in \mathbf{H}$;
2. $\forall x \in \mathcal{X}, \forall h \in \mathbf{H},\ \langle h, \kappa_x \rangle_{\mathbf{H}} = h(x)$ (reproducing property).

A Hilbert space of real-valued functions which possesses a reproducing kernel is called a *reproducing kernel Hilbert space* (RKHS) or a *proper Hilbert space*.

The connection between positive type functions and RKHSs is provided by the Moore-Aronszajn theorem.

Theorem 1 (Moore-Aronszajn theorem, Theorem 3 in Berlinet and Thomas-Agnan, 2004). *Let κ be a real-valued positive type function on $\mathcal{X}^2$. There exists one and only one Hilbert space $(\mathbf{H}, \langle \cdot, \cdot \rangle_{\mathbf{H}})$ of real-valued functions on $\mathcal{X}$ with κ as reproducing kernel.*

We can now define the classes of vector-valued functions at the basis of the M-SVMs as follows.

DEFINITION 3 (Classes of functions $\bar{\mathcal{H}}$ and $\mathcal{H}$). Let κ be a real-valued positive type function on $\mathcal{X}^2$ and let $(\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa})$ be the corresponding RKHS. Then, $\bar{\mathcal{H}}$ is the Hilbert space of vector-valued functions defined as follows: $\bar{\mathcal{H}} = \mathbf{H}_\kappa^Q$ and $\bar{\mathcal{H}}$ is endowed with

the inner product $\langle \cdot, \cdot \rangle_{\bar{\mathcal{H}}}$ given by:

$$\forall (\bar{h}, \bar{h}') \in \bar{\mathcal{H}}^2,\ \bar{h} = \left(\bar{h}_k\right)_{1 \leqslant k \leqslant Q},\ \bar{h}' = \left(\bar{h}'_k\right)_{1 \leqslant k \leqslant Q}, \quad \left\langle \bar{h}, \bar{h}' \right\rangle_{\bar{\mathcal{H}}} = \sum_{k=1}^{Q} \left\langle \bar{h}_k, \bar{h}'_k \right\rangle_{\mathbf{H}_\kappa}.$$

Let $\{1\}$ be the one-dimensional space of real-valued constant functions on $\mathcal{X}$.

$$\mathcal{H} = \bar{\mathcal{H}} \oplus \{1\}^Q = \left(\mathbf{H}_\kappa \oplus \{1\}\right)^Q.$$

For a given kernel κ, let Φ be the map from $\mathcal{X}$ into $\mathbf{H}_\kappa$ given by:

$$\forall x \in \mathcal{X}, \quad \Phi(x) = \kappa_x.$$

By analogy with the bi-class case, we call Φ the *reproducing kernel map* or a *feature map* and $\mathbf{H}_\kappa$ a *feature space*. It springs from Definition 3 and the reproducing property that the functions h of $\mathcal{H}$ can be written as follows:

$$h(\cdot) = \bar{h}(\cdot) + b = \left(\bar{h}_k(\cdot) + b_k\right)_{1 \leqslant k \leqslant Q} = \left(\left\langle \bar{h}_k, \Phi(\cdot) \right\rangle_{\mathbf{H}_\kappa} + b_k\right)_{1 \leqslant k \leqslant Q},$$

where $\bar{h} = (\bar{h}_k)_{1 \leqslant k \leqslant Q} \in \bar{\mathcal{H}}$ and $b = (b_k)_{1 \leqslant k \leqslant Q} \in \mathbb{R}^Q$. With these definitions and theorems at hand, a definition of the M-SVMs can be formulated as follows.

DEFINITION 4 (M-SVM, Definition 4.1 in Guermeur, 2010). Let d_m be a training set 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 functional $J_{\text{M-SVM}}$ of the form:

$$J_{\text{M-SVM}}(h) = \sum_{i=1}^{m} \ell_{\text{M-SVM}}\left(y_i, h(x_i)\right) + \lambda \left\|\bar{h}\right\|_{\bar{\mathcal{H}}}^2$$

where the data fit component involves a loss function $\ell_{\text{M-SVM}}$ which is convex.

The M-SVMs thus differ according to the nature of the function $\ell_{\text{M-SVM}}$ which corresponds to a multi-class extension of the hinge loss function.

DEFINITION 5 (Hard and soft margin M-SVM). If an M-SVM is trained subject to the constraint that $\sum_{i=1}^{m} \ell_{\text{M-SVM}}(y_i, h(x_i)) = 0$, it is called a *hard margin* M-SVM. Otherwise, it is called a *soft margin* M-SVM.

There are three main models of M-SVMs. The first one in chronological order is the model of Weston and Watkins (1998). Its loss function ℓ_{WW} is given by:

$$\ell_{\text{WW}}(y, h(x)) = \sum_{k \neq y} \left(1 - h_y(x) + h_k(x)\right)_+,$$

where $(\cdot)_+$ denotes the function $\max(0, \cdot)$. The second machine is due to Crammer and Singer (2001) and corresponds to the loss function ℓ_{CS} defined as:

$$\ell_{\mathrm{CS}}\big(y, \bar{h}(x)\big) = \big(1 - \bar{h}_y(x) + \max_{k \neq y} \bar{h}_k(x)\big)_+.$$

The most recent model is the one of Lee *et al.* (2004). Its loss function ℓ_{LLW} is given by:

$$\ell_{\mathrm{LLW}}\big(y, h(x)\big) = \sum_{k \neq y} \left(h_k(x) + \frac{1}{Q-1}\right)_+. \tag{1}$$

The LLW-M-SVM is the only model whose loss function is Fisher consistent (Lee *et al.*, 2004; Zhang, 2004; Tewari and Bartlett, 2007).

2.3. *Geometrical Margins*

Our definition of the M-SVMs locates these machines in the framework of Tikhonov's regularization theory (Tikhonov and Arsenin, 1977). This section characterizes them as large margin classifiers. From now on, we use the standard notation consisting in denoting w the vectors defining the direction of the linear discriminants in a feature space. For the sake of simplicity, the inner product of $\mathbf{H}_\kappa$ and its norm are simply denoted $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ respectively. Thus, $h(\cdot) = (\langle \bar{h}_k, \Phi(\cdot) \rangle_{\mathbf{H}_\kappa} + b_k)_{1 \leqslant k \leqslant Q}$ becomes $h(\cdot) = (\langle w_k, \Phi(\cdot) \rangle + b_k)_{1 \leqslant k \leqslant Q}$.

DEFINITION 6 (Geometrical margins, Definition 7 in Guermeur, 2007). Let $n \in \mathbb{N}^*$ and let $s_n = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}: \quad 1 \leqslant i \leqslant n\}$. If a function $h \in \mathcal{H}$ classifies these examples without error, then for any pair of distinct categories (k, l), its *margin between k and l* (computed with respect to s_n), $\gamma_{kl}(h)$, is defined as the smallest distance of a point of s_n either in k or l to the hyperplane separating those categories. Let us denote

$$d(h) = \min_{1 \leqslant k < l \leqslant Q} \Big\{ \min_{i:\ y_i \in \{k,l\}} \big|h_k(x_i) - h_l(x_i)\big| \Big\},$$

and

$$\forall (k,l): \quad 1 \leqslant k < l \leqslant Q, \quad d_{kl}(h) = \frac{1}{d(h)} \min_{i:\ y_i \in \{k,l\}} \big|h_k(x_i) - h_l(x_i)\big| - 1.$$

Then we have

$$\forall (k,l): \quad 1 \leqslant k < l \leqslant Q, \quad \gamma_{kl}(h) = \gamma_{lk}(h) = d(h) \frac{1 + d_{kl}(h)}{\|w_k - w_l\|}.$$

Since the M-SVMs satisfy the constraint $\sum_{k=1}^{Q} w_k = 0$, the connection between their geometrical margins and their penalizer is given by (2.6) in Guermeur (2007):

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

3. The M-SVM of Lee, Lin and Wahba

We now give more details regarding the LLW-M-SVM, from which the M-SVM^2 is derived. Our motivation is to establish some of the formulas that will be involved in the presentation of the new machine and the proof of the multi-class radius–margin bound.

3.1. *Training Algorithms*

The substitution in Definition 4 of $\ell_{\text{M-SVM}}$ with the expression of ℓ_{LLW} given by (1) provides us with the expressions of the quadratic programming (QP) problems corresponding to the training algorithms of the hard margin and soft margin versions of the LLW-M-SVM.

Problem 1 (Hard margin LLW-M-SVM, primal formulation).

$$\min_{h \in \mathcal{H}} J_{\text{HM}}(h)$$
$$\text{s.t.} \begin{cases} \forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, & h_k(x_i) \leqslant -\frac{1}{Q-1}, \\ \sum_{k=1}^{Q} h_k = 0, \end{cases}$$

where

$$J_{\text{HM}}(h) = \frac{1}{2} \sum_{k=1}^{Q} \|\bar{h}_k\|^2 = \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2.$$

Problem 2 (Soft margin LLW-M-SVM, primal formulation).

$$\min_{h, \xi} J_{\text{SM}}(h, \xi)$$
$$\text{s.t.} \begin{cases} \forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, & h_k(x_i) \leqslant -\frac{1}{Q-1} + \xi_{ik}, \\ \forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, & \xi_{ik} \geqslant 0, \\ \sum_{k=1}^{Q} h_k = 0, \end{cases}$$

where

$$J_{\text{SM}}(h, \xi) = \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k \neq y_i} \xi_{ik}.$$

For convenience of notation, the vector ξ of the slack variables of Problem 2 is represented as follows: $\xi = (\xi_{ik})_{1 \leqslant i \leqslant m, 1 \leqslant k \leqslant Q} \in \mathbb{R}_+^{Qm}$. ξ_{ik} is its component of index $(i-1)Q + k$ and the ξ_{iy_i} are dummy variables, all equal to 0. Using the notation e_n to designate the vector of $\mathbb{R}^n$ whose components are equal to e, we have thus $(\xi_{iy_i})_{1 \leqslant i \leqslant m} = 0_m$. The expression of the soft margin parameter C as a function of the regularization coefficient λ is: $C = (2\lambda)^{-1}$. To solve Problems 1 and 2, one usually solves their dual. We now derive the dual of Problem 2. Let $\alpha = (\alpha_{ik})$ and $\beta = (\beta_{ik})$ be

respectively the vectors of Lagrange multipliers associated with the constraints of good classification and the constraints of nonnegativity of the slack variables. These vectors are built according to the same principle as vector ξ. Let $\gamma \in \mathbf{H}_\kappa$ and $\delta \in \mathbb{R}$ be the Lagrange multipliers associated with the sum-to-0 constraints. The Lagrangian function of Problem 2 is given by:

$$\begin{aligned} &L_1(h, \xi, \alpha, \beta, \gamma, \delta) \\ &\quad = \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik} \\ &\qquad + \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik} \left(\langle w_k, \Phi(x_i) \rangle + b_k + \frac{1}{Q-1} - \xi_{ik} \right) \\ &\qquad - \sum_{i=1}^{m} \sum_{k=1}^{Q} \beta_{ik} \xi_{ik} - \left\langle \gamma, \sum_{k=1}^{Q} w_k \right\rangle - \delta \sum_{k=1}^{Q} b_k. \end{aligned} \tag{3}$$

Setting the gradient of L_1 with respect to w_k equal to the null vector provides us with Q alternative expressions for the optimal value of vector γ:

$$\forall k \in [\![1, Q]\!], \quad \gamma^* = w_k^* + \sum_{i=1}^{m} \alpha_{ik}^* \Phi(x_i). \tag{4}$$

Summing over the index k provides us with $\gamma^* = \frac{1}{Q} \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^* \Phi(x_i)$. By substitution into (4), we get the expression of the vectors w_k at the optimum:

$$\forall k \in [\![1, Q]\!], \quad w_k^* = \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \alpha_{il}^* \Phi(x_i), \tag{5}$$

where $\delta_{k,l}$ is the Kronecker symbol. Let us now set the gradient of L_1 with respect to b equal to the null vector. We get similarly

$$\forall k \in [\![1, Q]\!], \quad \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \alpha_{il}^* = 0. \tag{6}$$

Given the constraint $\sum_{k=1}^{Q} b_k = 0$,

$$\sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^* b_k^* = \sum_{k=1}^{Q} b_k^* \sum_{i=1}^{m} \alpha_{ik}^* = \delta^* \sum_{k=1}^{Q} b_k^* = 0. \tag{7}$$

Setting the gradient of L_1 with respect to ξ equal to the null vector gives:

$$\forall i \in [\![1, m]\!], \ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \quad \alpha_{ik}^* + \beta_{ik}^* = C. \tag{8}$$

By application of (5),

$$\frac{1}{2}\sum_{k=1}^{Q}\left\|w_k^*\right\|^2+\sum_{i=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}^*\left\langle w_k^*,\Phi(x_i)\right\rangle$$
$$=-\frac{1}{2}\sum_{k=1}^{Q}\left\|w_k^*\right\|^2=-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{Q}\sum_{l=1}^{Q}\left(\delta_{k,l}-\frac{1}{Q}\right)\alpha_{ik}^*\alpha_{jl}^*\kappa(x_i,x_j). \quad (9)$$

Extending to the case of matrices the double subscript notation used to designate the general terms of the vectors α, β and ξ, let $H\in\mathcal{M}_{Qm,Qm}(\mathbb{R})$ be the matrix of general term: $h_{ik,jl}=(\delta_{k,l}-\frac{1}{Q})\kappa(x_i,x_j)$. Reporting (7), (8), and (9) in (3) provides us with the following expression for the dual objective function:

$$J_{\mathrm{LLW,d}}(\alpha)=-\frac{1}{2}\alpha^T H\alpha+\frac{1}{Q-1}1_{Qm}^T\alpha.$$

Since the corresponding constraints are derived from (6) and (8), we get:

Problem 3 (Soft margin LLW-M-SVM, dual formulation).

$$\max_{\alpha} J_{\mathrm{LLW,d}}(\alpha),$$
$$\text{s.t.}\begin{cases}\forall i\in[\![1,m]\!],\ \forall k\in[\![1,Q]\!]\setminus\{y_i\}, & 0\leqslant\alpha_{ik}\leqslant C,\\ \forall k\in[\![1,Q-1]\!], & \sum_{i=1}^{m}\sum_{l=1}^{Q}\left(\frac{1}{Q}-\delta_{k,l}\right)\alpha_{il}=0,\end{cases}$$

where

$$J_{\mathrm{LLW,d}}(\alpha)=-\frac{1}{2}\alpha^T H\alpha+\frac{1}{Q-1}1_{Qm}^T\alpha,$$

with the general term of the Hessian matrix H being

$$h_{ik,jl}=\left(\delta_{k,l}-\frac{1}{Q}\right)\kappa(x_i,x_j).$$

With slight modifications, the derivation above can be adapted to express the dual of Problem 1. This leads to:

Problem 4 (Hard margin LLW-M-SVM, dual formulation).

$$\max_{\alpha} J_{\mathrm{LLW,d}}(\alpha),$$
$$\text{s.t.}\begin{cases}\forall i\in[\![1,m]\!],\ \forall k\in[\![1,Q]\!]\setminus\{y_i\}, & \alpha_{ik}\geqslant 0,\\ \forall k\in[\![1,Q-1]\!], & \sum_{i=1}^{m}\sum_{l=1}^{Q}\left(\frac{1}{Q}-\delta_{k,l}\right)\alpha_{il}=0.\end{cases}$$

3.2. *Geometrical Margins*

The geometrical margins of the hard margin Q-category LLW-M-SVM can be characterized thanks to three propositions among which the two last will prove useful to establish the radius–margin bound.

PROPOSITION 1. For a hard margin Q-category LLW-M-SVM,

$$d(h^*) \geqslant \frac{Q}{Q-1}.$$

Proof. If $h \in \mathcal{H}$ classifies the examples of the set s_n without error, then $d(h) = \min_{1 \leqslant i \leqslant n} \min_{k \neq y_i} (h_{y_i}(x_i) - h_k(x_i))$. By application of (1),

$$\forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \quad h_k^*(x_i) \leqslant -\frac{1}{Q-1}.$$

To finish the proof, it suffices to use the equation $\sum_{k=1}^{Q} h_k^* = 0$.

PROPOSITION 2. For a hard margin Q-category LLW-M-SVM trained on d_m, in the non-trivial case when $\alpha^* \neq 0$, there exists a mapping $\mathcal{I}$ from $[\![1, Q]\!]$ to $[\![1, m]\!]$ such that

$$\forall k \in [\![1, Q]\!], \quad h_k^*(x_{\mathcal{I}(k)}) = -\frac{1}{Q-1}.$$

Proof. This proposition results readily from the Karush–Kuhn–Tucker (KKT) optimality conditions and the constraints of Problem 4. Indeed, if $\alpha^* \neq 0$, then for all k in $[\![1, Q]\!]$, there exists at least one dual variable α_{ik}^* which is positive.

PROPOSITION 3. For a hard margin Q-category LLW-M-SVM, we have

$$\frac{d(h^*)^2}{Q} \sum_{k<l} \left(\frac{1 + d_{kl}(h^*)}{\gamma_{kl}(h^*)} \right)^2 = \sum_{k=1}^{Q} \|w_k^*\|^2 = {\alpha^*}^T H \alpha^* = \frac{1}{Q-1} 1_{Qm}^T \alpha^*.$$

Proof.

- $\frac{dt(h^*)^2}{Q} \sum_{k<l} (\frac{1+d_{kl}(h^*)}{\gamma_{kl}(h^*)})^2 = \sum_{k=1}^{Q} \|w_k^*\|^2$.
  This equation is a direct consequence of Definition 6 and (2).
- $\sum_{k=1}^{Q} \|w_k^*\|^2 = {\alpha^*}^T H \alpha^*$.
  This is a direct consequence of (9) and the definition of H.
- ${\alpha^*}^T H \alpha^* = \frac{1}{Q-1} 1_{Qm}^T \alpha^*$.
  By application of the KKT complementary conditions,

$$\sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^* \left(\langle w_k^*, \Phi(x_i) \rangle + b_k^* + \frac{1}{Q-1} \right) = 0.$$

Since

$$\forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \quad \left\langle w_k^*, \Phi(x_i) \right\rangle = -\left(H\alpha^*\right)_{ik},$$
$$\sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^* \left\langle w_k^*, \Phi(x_i) \right\rangle = -{\alpha^*}^T H \alpha^*.$$

Using (7), this implies that ${\alpha^*}^T H \alpha^* = \frac{1}{Q-1} 1_{Qm}^T \alpha^*$.

4. The M-SVM^2

Our new machine is a variant of the LLW-M-SVM in which the empirical contribution to the objective function is a quadratic form.

4.1. *Quadratic Loss Multi-Class SVMs: Motive and Principle*

Let ξ be the vector of slack variables of any M-SVM. In the case of the M-SVMs of Weston and Watkins and Lee, Lin and Wahba, $\xi \in \mathbb{R}_+^{Qm}$ with $(\xi_{iy_i})_{1 \leqslant i \leqslant m} = 0_m$, whereas in the case of the model of Crammer and Singer, $\xi \in \mathbb{R}_+^m$. In both cases, the empirical contribution to the objective function is $\|\xi\|_1$. The 2-norm SVM is the variant of the standard bi-class SVM obtained by replacing $\|\xi\|_1$ with $\|\xi\|_2^2$ in the objective function. Its main advantage is that its training algorithm can be expressed, after an appropriate change of kernel, as the training algorithm of a hard margin machine. Thus, its leave-one-out cross-validation error can be upper bounded thanks to the radius–margin bound. The strategy that we advocate to exhibit interesting multi-class extensions of the 2-norm SVM consists in studying the class of *quadratic loss M-SVMs*, i.e., the class of extensions of the M-SVMs such that the data fit term is $\xi^T M \xi$, where the matrix M is such that its submatrix M' obtained by suppressing the rows and columns whose indices are those of dummy slack variables is symmetric positive definite. The constraints on M correspond to necessary and sufficient conditions for $\xi^T M \xi$ to be a norm of ξ.

4.2. *The M-SVM^2 as a Multi-Class Extension of the 2-Norm SVM*

In this section, we establish that the idea introduced above provides us with a solution to the problem of interest when the M-SVM used is the LLW-M-SVM and $M = (m_{ik,jl})_{1 \leqslant i,j \leqslant m, 1 \leqslant k,l \leqslant Q}$ is the block diagonal matrix of general term

$$m_{ik,jl} = (1 - \delta_{y_i,k})(1 - \delta_{y_j,l})\delta_{i,j}(\delta_{k,l} + 1).$$

We first note that the corresponding matrix M' is actually symmetric positive definite. Indeed, it can be rewritten as follows: $M' = I_m \otimes (\delta_{k,l} + 1)_{1 \leqslant k,l \leqslant Q-1}$, where I_m designates the identity matrix of size m and $\otimes$ denotes the Kronecker product. Its spectrum is thus identical to the one of the matrix $(\delta_{k,l} + 1)_{1 \leqslant k,l \leqslant Q-1}$, i.e., made up of two positive eigenvalues: 1 and Q. The corresponding machine is named M-SVM^2. Its training algorithm is given by the following QP problem.

Problem 5 (M-SVM^2, primal formulation).

$$\min_{h,\xi} J_{\mathrm{M-SVM^2}}(h,\xi)$$
$$\text{s.t.} \begin{cases} \forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\}, \quad h_k(x_i) \leqslant -\frac{1}{Q-1} + \xi_{ik}, \\ \sum_{k=1}^{Q} h_k = 0, \end{cases}$$

where

$$J_{\mathrm{M-SVM^2}}(h,\xi) = \frac{1}{2}\sum_{k=1}^{Q} \|w_k\|^2 + C\xi^T M \xi.$$

Keeping the notations of the preceding sections, the expression of the Lagrangian function associated with this problem is:

$$\begin{aligned} &L_2(h,\xi,\alpha,\gamma,\delta) \\ &\quad = \frac{1}{2}\sum_{k=1}^{Q} \|w_k\|^2 + C\xi^T M\xi + \sum_{i=1}^{m}\sum_{k=1}^{Q} \alpha_{ik}\left(\langle w_k, \Phi(x_i)\rangle + b_k + \frac{1}{Q-1} - \xi_{ik}\right) \\ &\qquad - \left\langle \gamma, \sum_{k=1}^{Q} w_k \right\rangle - \delta \sum_{k=1}^{Q} b_k. \end{aligned} \tag{10}$$

Setting the gradient of L_2 with respect to ξ equal to the null vector gives

$$2CM\xi^* = \alpha^*. \tag{11}$$

Indeed, the coefficient $(1-\delta_{y_i,k})(1-\delta_{y_j,l})$ appears in $m_{ik,jl}$ so that:

$$\forall i \in [\![1,m]\!], \quad 2C(M\xi)_{iy_i} = \alpha_{iy_i} = 0.$$

It springs from (11) that

$$C{\xi^*}^T M\xi^* - {\alpha^*}^T\xi^* = -C{\xi^*}^T M\xi^*. \tag{12}$$

Using the same reasoning that we used to derive the objective function of Problem 3 and (12), at the optimum, (10) simplifies into

$$L_2\big(h^*,\xi^*,\alpha^*,\gamma^*,\delta^*\big) = -\frac{1}{2}{\alpha^*}^T H\alpha^* - C{\xi^*}^T M\xi^* + \frac{1}{Q-1}1_{Qm}^T\alpha^*.$$

Proving that the M-SVM^2 exhibits the same property as the 2-norm SVM amounts to exhibiting a kernel κ' such that

$$C{\xi^*}^T M\xi^* = \frac{1}{2}{\alpha^*}^T H'\alpha^* \tag{13}$$

with the general term of the matrix H' being: $h'_{ik,jl} = (\delta_{k,l} - \frac{1}{Q})\kappa'(x_i, x_j)$. Combining (11) and (13) gives:

$$\frac{1}{2}\alpha^{*T} H' \alpha^* = 2C^2 \xi^{*T} M^T H' M \xi^* = C \xi^{*T} M \xi^*.$$

After some algebra, we get the general term of the matrix $M^T H' M$, which is

$$(1 - \delta_{y_i,k})(1 - \delta_{y_j,l})(\delta_{k,l} + 1)\kappa'(x_i, x_j).$$

Thus, $2C\xi^{*T} M^T H' M \xi^* = \xi^{*T} M \xi^*$ provided that

$$\forall (i,j) \in [\![1,m]\!]^2, \quad \kappa'(x_i, x_j) = \frac{1}{2C}\delta_{i,j}.$$

This expression of the second kernel is precisely the one obtained in the case of the 2-norm SVM. With this definition of κ', we get

$$J_{\mathrm{M-SVM^2,d}}(\alpha) = -\frac{1}{2}\alpha^T \tilde{H} \alpha + \frac{1}{Q-1} 1_{Qm}^T \alpha,$$

where $\tilde{H} = H + H'$. Since $\nabla_b L_2(h, \xi, \alpha, \gamma, \delta) = \nabla_b L_1(h, \xi, \alpha, \beta, \gamma, \delta)$, the equality constraints of the dual are still given by (6). On the contrary, the only inequality constraints correspond to the nonnegativity of the Lagrange multipliers α_{ik}. Thus, the dual of Problem 5 is:

Problem 6 (M-SVM^2, dual formulation).

$$\max_{\alpha} J_{\mathrm{M-SVM^2,d}}(\alpha),$$
$$\text{s.t.} \begin{cases} \forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\}, & \alpha_{ik} \geqslant 0, \\ \forall k \in [\![1,Q-1]\!], & \sum_{i=1}^m \sum_{l=1}^Q \left(\frac{1}{Q} - \delta_{k,l}\right)\alpha_{il} = 0, \end{cases}$$

where

$$J_{\mathrm{M-SVM^2,d}}(\alpha) = -\frac{1}{2}\alpha^T \tilde{H} \alpha + \frac{1}{Q-1} 1_{Qm}^T \alpha,$$

with the general term of the Hessian matrix $\tilde{H}$ being

$$\tilde{h}_{ik,jl} = \left(\delta_{k,l} - \frac{1}{Q}\right)\left(\kappa(x_i, x_j) + \frac{1}{2C}\delta_{i,j}\right).$$

This problem is Problem 4 with $\kappa + \kappa'$ as kernel, which establishes that for the M-SVM^2, as for the 2-norm SVM, a radius–margin bound can be used to perform model

selection. By application of Proposition 3 and (13), we can check that

$$\begin{aligned} J_{\mathrm{M-SVM^2}}(h^*, \xi^*) &= \frac{1}{2}\sum_{k=1}^{Q} \|w_k^*\|^2 + C{\xi^*}^T M \xi^* \\ &= \frac{1}{2}{\alpha^*}^T H \alpha^* + \frac{1}{2}{\alpha^*}^T H' \alpha^* \\ &= \frac{1}{2}{\alpha^*}^T \tilde{H} \alpha^* = -\frac{1}{2}{\alpha^*}^T \tilde{H} \alpha^* + \frac{1}{Q-1} 1_{Qm}^T \alpha^* \\ &= J_{\mathrm{M-SVM^2,d}}\big(\alpha^*\big). \end{aligned}$$

4.3. *Properties and Implementation of the M-SVM*^{2}

Even though the training algorithm of the 2-norm SVM does not incorporate explicitly the constraints of nonnegativity of the slack variables, these constraints are satisfied by the optimal solution, for which we get:

$$\forall i \in [\![1, m]\!], \quad \xi_i^* = \frac{1}{2C}\alpha_i^*.$$

Problem 5 does not incorporate these constraints either. In that case however, this makes a significant difference since some of these variables can be negative. At the optimum, their expression can be deduced from (11), by inverting M'.

$${M'}^{-1} = I_m \otimes \big((\delta_{k,l} + 1)_{1 \leqslant k,l \leqslant Q-1}\big)^{-1} = I_m \otimes \left(\delta_{k,l} - \frac{1}{Q}\right)_{1 \leqslant k,l \leqslant Q-1}.$$

We then get

$$\forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \quad \xi_{ik}^* = (H'\alpha^*)_{ik}. \tag{14}$$

The optimal values of the slack variables are only positive on average, since applying on (14) a summation over the index k gives

$$\forall i \in [\![1, m]\!], \quad \sum_{k=1}^{Q} \xi_{ik}^* = \frac{1}{2CQ}\sum_{k=1}^{Q} \alpha_{ik}^*.$$

The relaxation of the constraints of nonnegativity of the slack variables alters the meaning of the constraints of good classification, although the global connection between a small value of the norm of ξ and a small training error is preserved. We conjecture that for any of the three M-SVMs presented in Section 2.2, no choice of the matrix M can give rise to a machine such that its dual problem is the one of a hard margin machine and its slack variables are all nonnegative.

Efficient SVM training requires to select an appropriate optimization algorithm (Bartkutė-Norkūnienė, 2009). To solve Problem 6, we developed two programs. One implements the Frank-Wolfe algorithm (Frank and Wolfe, 1956) and the other one Rosen's

gradient projection method (Rosen, 1960). The corresponding pieces of software are available from the first author's webpage. The computation of $\bar{h}$, b, and ξ as a function of the data and the dual variables calls for some explanations. At any iteration of the gradient ascent, the expression of the functions $\bar{h}_k$ is deduced from (5). Thus, in the case when x belongs to the training set,

$$\forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \quad \bar{h}_k(x_i) = -(H\alpha)_{ik}. \tag{15}$$

This formula is useful indeed, since the computation of the vector $H\alpha$ can also appear as a step in the computation of the dual objective function. The difficulty rests in the computation of the vectors b and ξ. In the case of the LLW-M-SVM, the KKT complementary conditions imply that at the optimum:

$$\begin{aligned} &\forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \\ &\quad \alpha^*_{ik} \in (0, C) \quad \Longrightarrow \quad b^*_k = -\frac{\partial}{\partial \alpha_{ik}} J_{\mathrm{LLW,d}}(\alpha^*). \end{aligned}$$

This formula can also be used before the optimum is reached, simply to obtain a "sensible" (but suboptimal) value for b. Let us define the sets $\mathcal{S}_k$ as follows: $\forall k \in [\![1, Q]\!],\ \mathcal{S}_k = \{i \in [\![1, m]\!] : \alpha^*_{ik} \in (0, C)\}$. Setting $\forall k \in [\![1, Q]\!],\ b'_k = -\frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \frac{\partial}{\partial \alpha_{ik}} J_{\mathrm{LLW,d}}(\alpha)$ and $\forall k \in [\![1, Q]\!],\ b_k = b'_k - \frac{1}{Q} \sum_{k=1}^{Q} b'_k$ provides us in turn with a value for the vector ξ thanks to the formula

$$\forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \quad \xi_{ik} = \left(\frac{\partial}{\partial \alpha_{ik}} J_{\mathrm{LLW,d}}(\alpha) + b_k \right)_+ .$$

Plugging these expressions of vectors b and ξ in the formula giving J_{SM}, one readily obtains an upper bound on the value of the primal objective function for the current step t of the gradient ascent, i.e., the current value of vector α, with

$$\lim_{t \to +\infty} J_{\mathrm{LLW,d}}(\alpha) = J_{\mathrm{LLW,d}}(\alpha^*) = J_{\mathrm{SM}}(h^*, \xi^*) = \lim_{t \to +\infty} J_{\mathrm{SM}}(h, \xi),$$

which makes it possible to specify a stopping criterion for training based on the value of the *feasibility gap*: $J_{\mathrm{SM}}(h, \xi) - J_{\mathrm{LLW,d}}(\alpha)$. Going back to the M-SVM^2, once more, the KKT complementary conditions provide us with b^*. We get

$$\begin{aligned} &\forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \\ &\quad \alpha^*_{ik} > 0 \quad \Longrightarrow \quad b^*_k = -\frac{\partial}{\partial \alpha_{ik}} J_{\mathrm{M-SVM^2,d}}(\alpha^*). \end{aligned}$$

As in the case of the LLW-M-SVM, this formula can be used to derive a value for vector b before the optimum is reached. However, since there is no analytical expression for the optimal value of vector ξ as a function of h, deriving a tight upper bound on the current value of the primal objective function requires some more work. The optimal value of ξ

is obtained by solving Problem 5 with h fixed. Then, given (14) and (15), the obvious choice for an initial feasible solution is:

$$\forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\},$$
$$\xi_{ik} = \max\left\{ -(H\alpha)_{ik} + b_k + \frac{1}{Q-1}, (H'\alpha)_{ik} \right\}.$$

5. Radius–Margin Bound on the Leave-One-Out Cross-Validation Error of the Hard Margin LLW-M-SVM

Like its bi-class counterpart, our multi-class radius–margin bound is based on a key lemma.

5.1. *Multi-Class Key Lemma*

Lemma 1 (Multi-class key lemma). *Let us consider a hard margin Q-category LLW-M-SVM trained on d_m. Consider now the same machine trained on $d_m \setminus \{(x_p, y_p)\}$. If it makes an error on (x_p, y_p), then*

$$\max_{1 \leqslant k \leqslant Q} \alpha^*_{pk} \geqslant \frac{Q}{(Q-1)^3 \mathcal{D}_m^2},$$

where $\mathcal{D}_m$ is the diameter of the smallest sphere of $\mathbf{H}_\kappa$ enclosing the set $\{\Phi(x_i)$: $1 \leqslant i \leqslant m\}$.

Proof. Let $h^p \in \mathcal{H}$ be the optimal solution when the machine is trained on $d_m \setminus \{(x_p, y_p)\}$. Let $\alpha^p = (\alpha^p_{ik}) \in \mathbb{R}_+^{Qm}$ be the corresponding vector of dual variables, with $(\alpha^p_{pk})_{1 \leqslant k \leqslant Q} = 0_Q$. This representation is used to simplify the simultaneous handling of both M-SVMs. Let us define two feasible solutions of Problem 4: λ^p and μ^p. λ^p is such that the vector $\alpha^* - \lambda^p$ is a feasible solution of Problem 4 under the additional constraint that $(\alpha^*_{pk} - \lambda^p_{pk})_{1 \leqslant k \leqslant Q} = 0_Q$, i.e., $\alpha^* - \lambda^p$ satisfies the same constraints as α^p. We have thus:

$$\begin{cases} \forall k \in [\![1, Q]\!], & \lambda^p_{pk} = \alpha^*_{pk}, \\ \forall i \in [\![1, m]\!] \setminus \{p\},\ \forall k \in [\![1, Q]\!], & 0 \leqslant \lambda^p_{ik} \leqslant \alpha^*_{ik}, \\ \forall k \in [\![1, Q-1]\!], & \sum_{i=1}^m \sum_{l=1}^Q \left(\frac{1}{Q} - \delta_{k,l}\right) \lambda^p_{il} = 0. \end{cases} \tag{16}$$

In the sequel, we write J in place of $J_{\mathrm{LLW,d}}$. By definition of μ^p, for all $K_1 \in \mathbb{R}_+^*$, $\alpha^p + K_1 \mu^p$ is a feasible solution of Problem 4. Thus, given the way λ^p has been specified, $J(\alpha^* - \lambda^p) \leqslant J(\alpha^p)$ and $J(\alpha^p + K_1 \mu^p) \leqslant J(\alpha^*)$. Hence,

$$J(\alpha^*) - J(\alpha^* - \lambda^p) \geqslant J(\alpha^*) - J(\alpha^p) \geqslant J\big(\alpha^p + K_1 \mu^p\big) - J(\alpha^p). \tag{17}$$

The value of the left-hand side of (17) is

$$J(\alpha^*) - J(\alpha^* - \lambda^p) = \frac{1}{2}{\lambda^p}^T H \lambda^p + \nabla J(\alpha^*)^T \lambda^p.$$

Since α^* and λ^p are respectively an optimal and a feasible solution of Problem 4, then necessarily, $\nabla J(\alpha^*)^T \lambda^p \leqslant 0$. This becomes obvious when one thinks about the principle of the Frank-Wolfe algorithm. As a consequence,

$$J(\alpha^*) - J(\alpha^* - \lambda^p) \leqslant \frac{1}{2}{\lambda^p}^T H \lambda^p,$$

and equivalently, in view of (5) and (9) (where α^* has been replaced with λ^p), as well as the definition of H,

$$J(\alpha^*) - J(\alpha^* - \lambda^p) \leqslant \frac{1}{2} \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \lambda_{il}^p \Phi(x_i) \right\|^2. \tag{18}$$

The line of reasoning used for the left-hand side of (17) gives:

$$\begin{aligned} &J\left(\alpha^p + K_1 \mu^p\right) - J(\alpha^p) \\ &\quad = K_1 \nabla J(\alpha^p)^T \mu^p - \frac{K_1^2}{2} \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \mu_{il}^p \Phi(x_i) \right\|^2. \end{aligned} \tag{19}$$

Since the M-SVM trained on $d_m \setminus \{(x_p, y_p)\}$ misclassifies x_p, there exists $n \in [\![1, Q]\!] \setminus \{y_p\}$ such that $h_n^p(x_p) \geqslant 0$, and α^p is not an optimal solution of Problem 4. Since μ^p is a feasible solution of the same problem, it can be built in such a way that $\nabla J(\alpha^p)^T \mu^p > 0$. These observations being made, neglecting the case $\alpha^p = 0$ as a degenerate one, we make use of Proposition 2 to build μ^p. Thus, let $\mathcal{I}$ be a mapping from $[\![1, Q]\!]$ to $[\![1, m]\!] \setminus \{p\}$ such that

$$\forall k \in [\![1, Q]\!], \quad h_k^p(x_{\mathcal{I}(k)}) = -\frac{1}{Q-1}.$$

For $K_2 \in \mathbb{R}_+^*$, let μ^p be the vector of $\mathbb{R}_+^{Qm}$ that only differs from the null vector in the following way:

$$\begin{cases} \mu_{pn}^p = K_2, \\ \forall k \in [\![1, Q]\!] \setminus \{n\}, \quad \mu_{\mathcal{I}(k)k}^p = K_2. \end{cases}$$

This definition of vector μ^p satisfies the constraints of Problem 4 and provides us with a positive lower bound for the inner product of interest.

$$\begin{aligned}\nabla J(\alpha^p)^T\mu^p &= \sum_{i=1}^{m}\sum_{k=1}^{Q}\mu_{ik}^p\left(\langle w_k^p,\Phi(x_i)\rangle+\frac{1}{Q-1}\right)\\ &= K_2\left\{\langle w_n^p,\Phi(x_p)\rangle+\frac{1}{Q-1}+\sum_{k\neq n}\left(\langle w_k^p,\Phi(x_{\mathcal{I}(k)})\rangle+\frac{1}{Q-1}\right)\right\}\\ &= K_2\left\{h_n^p(x_p)+\frac{1}{Q-1}-\sum_{k=1}^{Q}b_k^p\right\}=K_2\left\{h_n^p(x_p)+\frac{1}{Q-1}\right\}.\end{aligned}$$

As a consequence,

$$\nabla J(\alpha^p)^T\mu^p \geqslant \frac{K_2}{Q-1}.$$

Making use of this result, the combination of (17), (18), and (19) finally gives

$$\begin{aligned}&\frac{1}{2}\sum_{k=1}^{Q}\left\|\sum_{i=1}^{m}\sum_{l=1}^{Q}\left(\frac{1}{Q}-\delta_{k,l}\right)\lambda_{il}^p\Phi(x_i)\right\|^2\\ &\quad\geqslant \frac{K_1K_2}{Q-1}-\frac{K_1^2}{2}\sum_{k=1}^{Q}\left\|\sum_{i=1}^{m}\sum_{l=1}^{Q}\left(\frac{1}{Q}-\delta_{k,l}\right)\mu_{il}^p\Phi(x_i)\right\|^2.\end{aligned}\tag{20}$$

Let $\nu^p = K_2^{-1}\mu^p$. The value of $K = K_1K_2$ maximizing the right-hand side of (20) is: $K^* = \{(Q-1)\sum_{k=1}^{Q}\|\sum_{i=1}^{m}\sum_{l=1}^{Q}(\frac{1}{Q}-\delta_{k,l})\nu_{il}^p\Phi(x_i)\|^2\}^{-1}$. By substitution in (20), this implies that

$$\begin{aligned}&(Q-1)^2\sum_{k=1}^{Q}\left\|\sum_{i=1}^{m}\sum_{l=1}^{Q}\left(\frac{1}{Q}-\delta_{k,l}\right)\lambda_{il}^p\Phi(x_i)\right\|^2\\ &\quad\times\sum_{k=1}^{Q}\left\|\sum_{i=1}^{m}\sum_{l=1}^{Q}\left(\frac{1}{Q}-\delta_{k,l}\right)\nu_{il}^p\Phi(x_i)\right\|^2\geqslant 1.\end{aligned}$$

The quadratic form ${\lambda^p}^T H\lambda^p$ can be rewritten as

$$\begin{aligned}&\sum_{k=1}^{Q}\left\|\frac{1}{Q}\sum_{i=1}^{m}\sum_{l=1}^{Q}\lambda_{il}^p\Phi(x_i)-\sum_{i=1}^{m}\lambda_{ik}^p\Phi(x_i)\right\|^2\\ &\quad=\frac{1}{Q^2}\sum_{k=1}^{Q}\left\|\sum_{i=1}^{m}\sum_{l=1,l\neq k}^{Q}(\lambda_{il}^p-\lambda_{ik}^p)\Phi(x_i)\right\|^2\\ &\quad=\frac{1}{Q^2}\sum_{k=1}^{Q}\left\|\sum_{l=1,l\neq k}^{Q}\left(\sum_{i=1}^{m}\lambda_{il}^p\Phi\big(x_i\big)-\sum_{i=1}^{m}\lambda_{ik}^p\Phi\big(x_i\big)\right)\right\|^2.\end{aligned}$$

For $\eta\in\mathbb{R}^{Qm}$, let $S(\eta)=\frac{1}{Q}1_{Qm}^T\eta$. By definition of λ^p,

$$\forall k\in[\![1,Q]\!],\quad \sum_{i=1}^{m}\lambda_{ik}^p = S\big(\lambda^p\big).$$

Since $\lambda^p \in \mathbb{R}_+^{Qm}$, by construction,

$$\sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \lambda_{il}^p \Phi(x_i) \right\|^2 = \frac{S(\lambda^p)^2}{Q^2} \times \sum_{k=1}^{Q} \left\| \sum_{l=1, l \neq k}^{Q} \left(\mathrm{conv}_l\big\{\Phi(x_i)\colon 1 \leqslant i \leqslant m\big\} - \mathrm{conv}_k\big\{\Phi(x_i)\colon 1 \leqslant i \leqslant m\big\} \right) \right\|^2,$$

where the terms $\mathrm{conv}_l\{\Phi(x_i)\colon 1 \leqslant i \leqslant m\}$ are convex combinations of the $\Phi(x_i)$. As a consequence,

$$\forall (k,l) \in [\![1, Q]\!]^2, \quad \left\| \mathrm{conv}_l\big\{\Phi(x_i)\colon 1 \leqslant i \leqslant m\big\} - \mathrm{conv}_k\big\{\Phi(x_i)\colon 1 \leqslant i \leqslant m\big\} \right\| \leqslant \mathcal{D}_m$$

and applying the triangular inequality gives

$$\sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \lambda_{il}^p \Phi(x_i) \right\|^2 \leqslant \frac{(Q-1)^2}{Q} S(\lambda^p)^2 \mathcal{D}_m^2.$$

Since the same reasoning applies to ν^p, we get:

$$\frac{(Q-1)^6}{Q^2} S(\lambda^p)^2 S(\nu^p)^2 \mathcal{D}_m^4 \geqslant 1. \tag{21}$$

By construction, $S(\nu^p) = 1$. We now construct a vector λ^p minimizing the objective function S. Since $\forall k \in [\![1, Q]\!], \quad \lambda_{pk}^p = \alpha_{pk}^*$,

$$\forall k \in [\![1, Q]\!], \quad \sum_{i=1}^{m} \lambda_{ik}^p \geqslant \alpha_{pk}^*.$$

But since

$$\forall (k,l) \in [\![1, Q]\!]^2, \quad \sum_{i=1}^{m} \lambda_{ik}^p = \sum_{i=1}^{m} \lambda_{il}^p = S(\lambda^p),$$

we have further

$$\min_{\lambda^p} S(\lambda^p) \geqslant \max_{1 \leqslant l \leqslant Q} \alpha_{pl}^*.$$

Obviously, the nature of the function S calls for the choice of minimal values for the components λ_{ik}^p, which is coherent with the box constraints in (16). Thus, there exists a vector λ^{p^*} which is a minimizer of S subject to the set of constraints (16) such that

$$\forall k \in [\![1, Q]\!], \quad \sum_{i=1}^{m} \lambda_{ik}^{p^*} = \max_{1 \leqslant l \leqslant Q} \alpha_{pl}^*,$$

i.e., $S(\lambda^{p^*}) = \max_{1 \leqslant l \leqslant Q} \alpha^*_{pl}$. The substitution of the values of $S(\nu^p)$ and $S(\lambda^{p^*})$ in (21) provides us with

$$\Big(\max_{1 \leqslant k \leqslant Q} \alpha^*_{pk} \Big)^2 \geqslant \frac{Q^2}{(Q-1)^6 \mathcal{D}_m^4}.$$

Taking the square root of both sides concludes the proof of the lemma.

5.2. *Multi-Class Radius–Margin Bound*

The multi-class radius–margin bound is a direct consequence of Lemma 1.

Theorem 2 (Multi-class radius–margin bound). *Let us consider a hard margin Q-category LLW-M-SVM trained on d_m. Let $\mathcal{L}_m$ be the number of errors resulting from applying a leave-one-out cross-validation procedure to this machine and $\mathcal{D}_m$ the diameter of the smallest sphere of $\mathbf{H}_\kappa$ enclosing the set $\{\Phi(x_i)\colon 1 \leqslant i \leqslant m\}$. Then, using the notations of Definition 6, we have:*

$$\mathcal{L}_m \leqslant \frac{(Q-1)^4}{Q^2} \mathcal{D}_m^2 d(h^*)^2 \sum_{k<l} \left(\frac{1 + d_{kl}(h^*)}{\gamma_{kl}(h^*)} \right)^2. \tag{22}$$

Proof. Let $\mathcal{M}(d_m)$ be the subset of d_m made up of the examples misclassified by the cross-validation procedure ($|\mathcal{M}(d_m)| = \mathcal{L}_m$). Lemma 1 exhibits a non-trivial lower bound on $\max_{1 \leqslant k \leqslant Q} \alpha^*_{pk}$ when (x_p, y_p) belongs to $\mathcal{M}(d_m)$. As a consequence,

$$1_{Qm}^T \alpha^* \geqslant \sum_{i=1}^{m} \max_{1 \leqslant k \leqslant Q} \alpha^*_{ik} \geqslant \sum_{i\colon (x_i, y_i) \in \mathcal{M}(d_m)} \max_{1 \leqslant k \leqslant Q} \alpha^*_{ik} \geqslant \frac{Q \mathcal{L}_m}{(Q-1)^3 \mathcal{D}_m^2}. \tag{23}$$

To finish the proof, it suffices to make use of Proposition 3.

5.3. *Discussion*

When $Q = 2$, (1) implies that $d(h^*) = 1 + \frac{1}{Q-1} = \frac{Q}{Q-1} = 2$. Thus, $\frac{(Q-1)^4}{Q^2} d(h^*)^2 = 1$. Furthermore, since $d_{12}(h^*) = 0$, the sum $\sum_{k<l} (\frac{1+d_{kl}(h^*)}{\gamma_{kl}(h^*)})^2$ simplifies into $\frac{1}{\gamma^2}$. This means that the expression of the multi-class radius–margin bound simplifies into the one of the standard bi-class radius–margin bound: $\mathcal{L}_m \leqslant (\frac{\mathcal{D}_m}{\gamma})^2$. The formulation of Theorem 2 is the one involving the radius (diameter) and the geometrical margins, so that it appears clearly as a multi-class extension of the bi-class radius–margin bound. However, (23) provides us with a sharper bound, namely

$$\mathcal{L}_m \leqslant \frac{(Q-1)^3}{Q} \mathcal{D}_m^2 \sum_{i=1}^{m} \max_{1 \leqslant k \leqslant Q} \alpha^*_{ik}. \tag{24}$$

If (24) is a tighter bound, (22) could be preferable for model selection, if it can be derived simply with respect to the hyperparameters, in the same way as in the bi-class case (Chapelle *et al.*, 2002).

The comparison with the radius–margin bound introduced in Wang *et al.* (2008) is also enlightening. This bound is dedicated to the one-versus-one decomposition strategy under the rule of max wins (Hsu and Lin, 2002). It appears as a direct consequence of the application of the bi-class radius–margin bound in this framework. However, it applies to all the multi-class discriminant models based on SVMs and for which the bi-class radii and margins can be computed.

Theorem 3 (Model selection criterion I in Wang *et al.*, 2008). *Let us consider a Q-category one-versus-one decomposition method involving $\binom{Q}{2}$ hard margin bi-class SVMs. For $1 \leqslant k < l \leqslant Q$, let κ_{kl}, Φ_{kl}, and γ_{kl} be respectively the kernel, the reproducing kernel map, and the geometrical margin of the machine discriminating categories k and l. Let $\mathcal{D}_{kl}$ be the diameter of the smallest sphere of $\mathbf{H}_{\kappa_{kl}}$ enclosing the set $\{\Phi_{kl}(x_i)\colon y_i \in \{k,l\}\}$. Then, the following upper bound holds true:*

$$\mathcal{L}_m \leqslant \sum_{k<l} \left(\frac{\mathcal{D}_{kl}}{\gamma_{kl}}\right)^2. \tag{25}$$

Formulas (22) and (25) share the same structure in terms of radii and margins. An argument in favour of the use of the one-versus-one decomposition method and the second bound is that if all the machines use the same kernel κ, then

$$\forall (k,l)\colon \quad 1 \leqslant k < l \leqslant Q, \quad \frac{\mathcal{D}_{kl}}{\gamma_{kl}} \leqslant \frac{\mathcal{D}_m}{\gamma_{kl}(h^*)}.$$

However, it is no longer valid if (24) replaces (22). An argument backing the use of the M-SVM with (24) is that it requires less computational time. All in all, the most useful bound could simply correspond to the most efficient strategy to tackle the multi-class problem at hand. In that respect, it is currently admitted that no multi-class discriminant model based on SVMs is uniformly superior to the others (Hsu and Lin, 2002; Fürnkranz, 2002; Rifkin and Klautau, 2004).

6. Experimental Results

The four M-SVMs are compared on three multi-class data sets from the UCI Machine Learning Repository (`http://archive.ics.uci.edu/ml/`) and a real-world problem: protein secondary structure prediction. In each case, a multi-layer perceptron (MLP) (Anthony and Bartlett, 1999) is used to provide a performance of reference.

The three benchmarks from the UCI repository are those named “Image Segmentation”, “Landsat Satellite” and “Waveform Database Generator (Version 2)”. These bases have been devided by their authors into a training and a test set, making the reproductibility of the experiments as easy as possible. The kernel of the M-SVMs is a radial basis

Table 1
Relative prediction accuracy of the four M-SVMs on three data sets from the UCI repository

	MLP	WW	CS	LLW	M-SVM^2
Image segmentation	83.6	89.7	90.3	89.5	90.2
Landsat satellite	86.3	92.1	92.0	91.9	92.1
Waveform	85.8	86.7	86.7	86.4	86.5

function (RBF). Thus, their hyperparameters are the parameter C and the bandwidth of the kernel. As for the MLP, the capacity control is based on the choice of the size of the hidden layer. To set the values of all the hyperparameters, a cross-validation procedure was implemented on the training set. The experimental results obtained are gathered in Table 1.

Two main comments can be made regarding these initial results. First, the M-SVMs appear uniformly superior to the MLP. For the two first data sets, the gain in prediction accuracy is always statistically significant with confidence exceeding 0.95. Second, the M-SVM^2 systematically obtains slightly better results than the LLW-M-SVM. However, the difference is too small to be significant, as was confirmed by additional experiments performed on different data sets (data not shown).

Protein secondary structure prediction is an open problem of central importance in predictive structural biology. It consists in assigning to each residue (amino acid) of a protein sequence its conformational state. We consider here a three-state description of this structure ($Q = 3$), with the categories being: α-helix, β-strand and aperiodic/coil. To assess our classifiers on this problem, we used the CB513 data set of Cuff and Barton (1999). The 513 sequences of this set are made up of 84119 residues. Each sequence is represented by a position-specific scoring matrix (PSSM) produced by PSI-BLAST (Altschul *et al.*, 1997). The initial secondary structure assignment was performed by the DSSP program of Kabsch and Sander (1983), with the reduction from 8 to 3 conformational states following the CASP method, i.e., H$+$G $\rightarrow$H (α-helix), E$+$B $\rightarrow$E (β-strand), and all the other states in C (coil). To predict the conformational state of the residue of index n in a given sequence, a sliding window of size 15 is used. The vector of predictors processed by the classifiers is obtained by appending the rows of the corresponding PSSM whose indices range from $n - 7$ to $n + 7$. Since a PSSM has 20 columns, one per amino acid, this corresponds to 300 predictors. Once more, the four M-SVMs used a RBF kernel. The results obtained with the data sets from the UCI repository had highlighted a superiority of the M-SVMs over the MLP. We decided to investigate further this phenomenon by implementing two variants of the MLP. The first one combines a quadratic (Q) loss with output units using a sigmoid activation function. The second one combines a cross-entropy (CE) loss with output units using a softmax activation function. In order to perform model selection and assess the quality of the predictions, a two-level cross-validation procedure called stacked generalization (Wolpert, 1992) was implemented. In that way, the estimates of the prediction accuracy were unbiased. A secondary structure prediction method must fulfill different requirements in order to be useful for the biologist. Thus, several standard measures giving complementary indications must be used

Table 2

Relative prediction accuracy of the four M-SVMs on the 513 protein sequences (84119 residues) of the CB513 data set

	MLP (Q)	MLP (CE)	WW	CS	LLW	M-SVM^2
Q_3	72.2	72.1	76.2	76.4	75.6	76.5
C_α	0.63	0.63	0.71	0.70	0.69	0.71
C_β	0.55	0.55	0.62	0.62	0.60	0.62
C_{coil}	0.52	0.52	0.57	0.58	0.57	0.58
Sov	61.5	60.5	70.5	71.5	69.8	71.3

to assess the prediction accuracy (Baldi *et al.*, 2000). We used the three most popular ones: the recognition rate Q_3, Pearson-Matthews correlation coefficients $C_{\alpha/\beta/\mathrm{coil}}$, and the segment overlap measure (Sov) in its most recent version (Sov'99). Table 2 provides the values taken by these measures for the different classifiers.

Once more, the M-SVMs appear uniformly superior to the MLP (irrespective of the choice of its loss function). Furthermore, the difference in recognition rate between the M-SVM^2 and the LLW-M-SVM is now statistically significant with confidence exceeding 0.95. Finding the reason for this noticeable improvement could tell us more about the benefits that one can expect from using a quadratic loss M-SVM (apart from the possibility to use a radius–margin bound).

7. Conclusions and Ongoing Research

A new M-SVM has been introduced: the M-SVM^2. This quadratic loss extension of the LLW-M-SVM is the first M-SVM exhibiting the main property of the 2-norm SVM: its training algorithm can be expressed, after an appropriate change of kernel, as the training algorithm of a hard margin machine. As in the bi-class case, one can take advantage of this property by making use of a radius–margin bound as objective function for the model selection procedure. The derivation of the corresponding bound is the second main contribution of the article. At last, initial experimental results highlight the potential of the new machine, whose prediction accuracy is similar to those of the three main M-SVMs, and compares favourably with the one of the MLP. This study has highlighted different features of the M-SVMs which make their study intrinsically more difficult than the one of bi-class SVMs, like the complexity of the formula expressing the geometrical margins as a function of the vector of dual variables α^* (Proposition 3). Coming after our study of the sample complexity of classifiers taking values in $\mathbb{R}^Q$ (Guermeur, 2010), it provides us with new arguments backing our thesis that the study of multi-category classification should be tackled independently of the one of dichotomy computation.

The evaluation of the M-SVM^2 and its bound is still to be carried out in a systematic way. The aim of this study is to find a satisfactory trade-off between the prediction accuracy and the computational complexity. In that respect, the time needed to set the value of the soft margin parameter of the M-SVMs should be kept reasonable thanks to the implementation of algorithms devised to fit the entire regularization path at a cost exceeding only slightly the one of one training of the corresponding machine. The first algorithm of this kind dedicated to an M-SVM, the LLW-M-SVM, was proposed by Lee

and Cui (2006). The derivation of an algorithm dedicated to the M-SVM^2 is the subject of an ongoing research.

Acknowledgements. This work was supported by the Decrypthon program of the AFM, the CNRS, and IBM. The authors would like to thank Y. Lee for providing them with additional information on her work, F. Thomarat for generating the PSSMs, as well as the anonymous reviewers for their comments. Thanks are also due to M. Bertrand and R. Bonidal for carefully reading this manuscript.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389–3402.

Anthony, M., Bartlett, P.L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5), 412–424.

Balys, V., Rudzkis, R. (2010). Statistical classification of scientific publications. *Informatica*, 21(4), 471–486.

Bartkutė-Norkūnienė, V. (2009). Stochastic optimization algorithms for support vector machines classification. *Informatica*, 20(2), 173–186.

Berlinet, A., Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston.

Boser, B.E., Guyon, I.M., Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. In: *COLT'92*, pp. 144–152.

Chapelle, O., Vapnik, V.N., Bousquet, O., Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Mach. Learn.*, 46(1), 131–159.

Chung, K.-M., Kao, W.-C., Sun, C.-L., Wang, L.-L., Lin, C.-J. (2003). Radius margin bounds for support vector machines with the RBF kernel. *Neural Comput.*, 15(11), 2643–2681.

Cortes, C., Vapnik., V.N. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

Crammer, K., Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2, 265–292.

Cuff, J.A., Barton, G.J., (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Structure, Function, and Genetics*, 34(4), 508–519.

Duan, K., Keerthi, S.S., Poo, A.N. (2003). Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing*, 51, 41–59.

Frank, M., Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3, 95–110.

Fürnkranz, J. (2002). Round robin classification. *Journal of Machine Learning Research*, 2, 721–747.

Guermeur, Y. (2007). *SVM multiclasses, théorie et applications*. Habilitation à diriger des recherches, Université Nancy 1 (in French).

Guermeur, Y. (2010). Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs. *Communications in Statistics – Theory and Methods*, 39(3), 543–557.

Hsu, C.-W., Lin, C.-J. (2002). A comparison of methods for multi-class support vector machines. *IEEE Transaction on Neural Networks*, 13(2), 415–425.

Kabsch, W., Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577–2637.

Lee, Y., Cui, Z. (2006). Characterizing the solution path of multicategory support vector machines. *Statistica Sinica*, 16(2), 391–409.

Lee, Y., Lin, Y., Wahba, G. (2004). 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.

Luntz, A., Brailovsky, V. (1969). On estimation of characters obtained in statistical procedure of recognition. *Technicheskaya Kibernetica*, 3 (in Russian).

Norkin, V., Keyzer, M. (2009). On stochastic optimization and statistical learning in reproducing kernel Hilbert spaces by support vector machines (SVM). *Informatica*, 20(2), 273–292.
Rifkin, R., Klautau, A. (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5, 101–141.
Rosen, J.B. (1960). The gradient projection method for nonlinear programming. Part I. Linear constraints. *Journal of the Society for Industrial and Applied Mathematics*, 8(1), 181–217.
Schölkopf, B., Burges, C., Vapnik, V.N. (1995). Extracting support data for a given task. In: *KDD'95*, pp. 252–257.
Shawe-Taylor, J., Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.
Tewari, A., Bartlett, P.L. (2007). On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8, 1007–1025.
Tikhonov, A.N., Arsenin, V.Y. (1977). *Solutions of Ill-Posed Problems*. V.H. Winston & Sons, Washington.
Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.
Vapnik, V.N. (1998). *Statistical Learning Theory*. Wiley, New York.
Vapnik, V.N., Chapelle, O. (2000). Bounds on error expectation for support vector machines. *Neural Computation*, 12(9), 2013–2036.
Wang, L., Xue, P., Chan, K.L. (2008). Two criteria for model selection in multiclass support vector machines. *IEEE Transactions on Systems, Man, and Cybernetics – Part B*, 38(6), 1432–1448.
Weston, J., Watkins, C. (1998). Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science.
Wolpert, D.H. (1992). Stacked Generalization. *Neural Networks*, 5, 241–259.
Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5, 1225–1251.

Y. Guermeur received a French "diplôme d'ingénieur" from the IIE in 1991. He received a PhD in computer science from the University Paris 6 in 1997 and the "Habilitation à Diriger des Recherches" (HDR) from the University Nancy 1 in 2007. Permanent researcher at CNRS since 2003, he is currently at the head of the ABC research team in the LORIA laboratory. His research interests include machine learning and computational biology.

E. Monfrini received a PhD degree from Paris VI University, Paris, France, in 2002. He is currently associate professor at Institut Telecom and member of CITI (Communication, Image et Traitement de l'Information) Department. His research interests include statistical learning and especially multi-class SVMs, supervised or unsupervised classification and Markov models. More information can be found at:
`http://www-public.it-sudparis.eu/ monfrini/.`

Spindulio rėžio ribos taikymas kvadratinio nuostolio daugiaklasiam SVM

Yann GUERMEUR, Emmanuel MONFRINI

Atraminių vektorių klasifikavimo (SVM) metodo taikymas yra susijęs su dviejų šio metodo hiperparametrų (silpno skirtumo (soft margin) C ir branduolio parametrų) nustatymu. Parametrams įvertinti taikomas kryžminio įverčio metodas. Žinoma, kad šio metodo „palikti vieną" variantas sukuria apibendrintą paklaidos įvertį, kuris yra beveik visada nepaslinktas. Pagrindinis jo trūkumas – dideli skaičiavimo laiko ištekliai. Norint išvengti šios problemos pasiūlyti keli paklaidos „palikti vieną" viršutiniai rėžiai sprendžiant SVM vaizdų atpažinimo uždavinius. Populiariausias iš jų yra spindulinio skirtumo rėžis (radius margin bound). Jis taikomas maksimalaus atstumo SVM metodui ir išplečiamas kvadratinės normos SVM metodui. Šiame straipsnyje nagrinėjamas Lee, Lin ir Wahb daugiaklasis SVM – tai M-SVM. Šiam metodui įvedamas apibendrintas spindulinio skirtumo rėžis.