A Hybrid Kernel Machine for Protein Secondary Structure Prediction

Yann Guermeur
LORIA
Université Nancy I

Alain Lifchitz
LIP6
CNRS

Régis Vert
LRI
Université Paris XI

http://www.loria.fr/~guermeur/

Overview

Multi-class support vector machines (M-SVMs)

- One single architecture
- Several training algorithms
- M-SVMs and the SRM inductive principle

M-SVMs for protein secondary structure prediction

- Improving the generalization performance of SSpro2
- Hybrid and modular architecture : M-SVM + IHMM
- Experimental results

Multi-class pattern recognition

Hypotheses : empirical data characterizing a joint probability distribution

- Q-category discrimination problem
- $Z = (X, Y)$: random variable on a probability space $(\Omega, \mathcal{B}, P)$
- $X(\Omega) = \mathcal{X}$: input space (set of descriptions), $Y(\Omega) = \mathcal{Y}$: finite set of categories
- P : joint probability distribution function on $\mathcal{X} \times \mathcal{Y}$, fixed but unknown
- $s = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset (\mathcal{X} \times \mathcal{Y})^m$, learning set : observations i.i.d. according to P

- $\mathcal{H}$: **family of vector-valued functions** $h = [h_k]$, $(1 \leq k \leq Q)$, from $\mathcal{X}$ into $\mathbb{R}^Q$

Goal : for a given pattern, find its category

Find in $\mathcal{H}$ a function associated with the lowest expected risk (generalization error)

$$R(h) = R(f) = \int_{\mathcal{X} \times \mathcal{Y}} \mathbb{1}_{\{f(x) \neq y\}} dP(x, y)$$

f **: discriminant function** corresponding to h, obtained by choosing the category associated with the **index of the highest output**

Multi-class Support Vector Machines

Architecture

Functions $h = [h_k]$ computed by the architecture are defined by :

$$\forall x \in \mathcal{X},\ \forall k \in \{1, \ldots, Q\},\ h_k(x) = \tilde{h}_k(x) + b_k = \langle w_k, \Phi(x) \rangle + b_k$$

where

- Φ is either a nonlinear map into a *feature space* or identity
- $\tilde{\mathcal{H}} = \left\{ \tilde{h} = \left[\tilde{h}_k \right] \right\}$ is the product of Q *reproducing kernel Hilbert spaces* (RKHS)
- K, the *kernel* associated with $\tilde{\mathcal{H}}$, is related to Φ through :

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

Training algorithm of M-SVMs : general principle

Let $s_m = \{(x_1, C(x_1)), \ldots, (x_m, C(x_m))\}$ be the training set

Primal formulation

Problem 1

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \|\tilde{h}\|^2_{\tilde{\mathcal{H}}} + C \sum_{i=1}^{m} l\left(h(x_i), C(x_i)\right) \right\}$$

Dual formulation

Representer theorems establish that training amounts to finding the values of the coefficients β_{ik} in :

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

Training algorithms of M-SVMs : primal formulation

Problem 2 (M-SVM1 (Vapnik & Blanz 98, Weston & Watkins 98,...))

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

$$s.t.\begin{cases}\langle w_{C(x_i)}-w_k,\Phi(x_i)\rangle+b_{C(x_i)}-b_k\geq 1-\xi_{ik}, & (1\leq i\leq m),(1\leq k\neq C(x_i)\leq Q)\\ \xi_{ik}\geq 0, & (1\leq i\leq m),(1\leq k\neq C(x_i)\leq Q)\end{cases}$$

Problem 3 (M-SVM2 (Guermeur 02))

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

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

Training algorithms of M-SVMs : primal formulation

Problem 4 (M-SVM3 (Crammer & Singer 01))

$$\min_{h\in\mathcal{H}}\left\{\frac{1}{2}\sum_{k=1}^{Q}\|w_k\|^2+C\sum_{i=1}^{m}\xi_i\right\}$$

$$s.t.\ \langle w_{C(x_i)}-w_k,\Phi(x_i)\rangle+b_{C(x_i)}-b_k+\delta_{C(x_i),k}\geq 1-\xi_i,\ (1\leq i\leq m),(1\leq k\leq Q)$$

Problem 5 (M-SVM4 (Lee, Lin & Wahba 01))

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

$$s.t.\begin{cases}\sum_{k=1}^{Q}w_k=0,\ \ \sum_{k=1}^{Q}b_k=0 & \\ \langle w_k,\Phi(x_i)\rangle+b_k\leq -1/(Q-1)+\xi_{ik}, & (1\leq i\leq m),(1\leq k\neq C(x_i)\leq Q)\\ \xi_{ik}\geq 0, & (1\leq i\leq m),(1\leq k\neq C(x_i)\leq Q)\end{cases}$$

Behaviour of a M-SVM

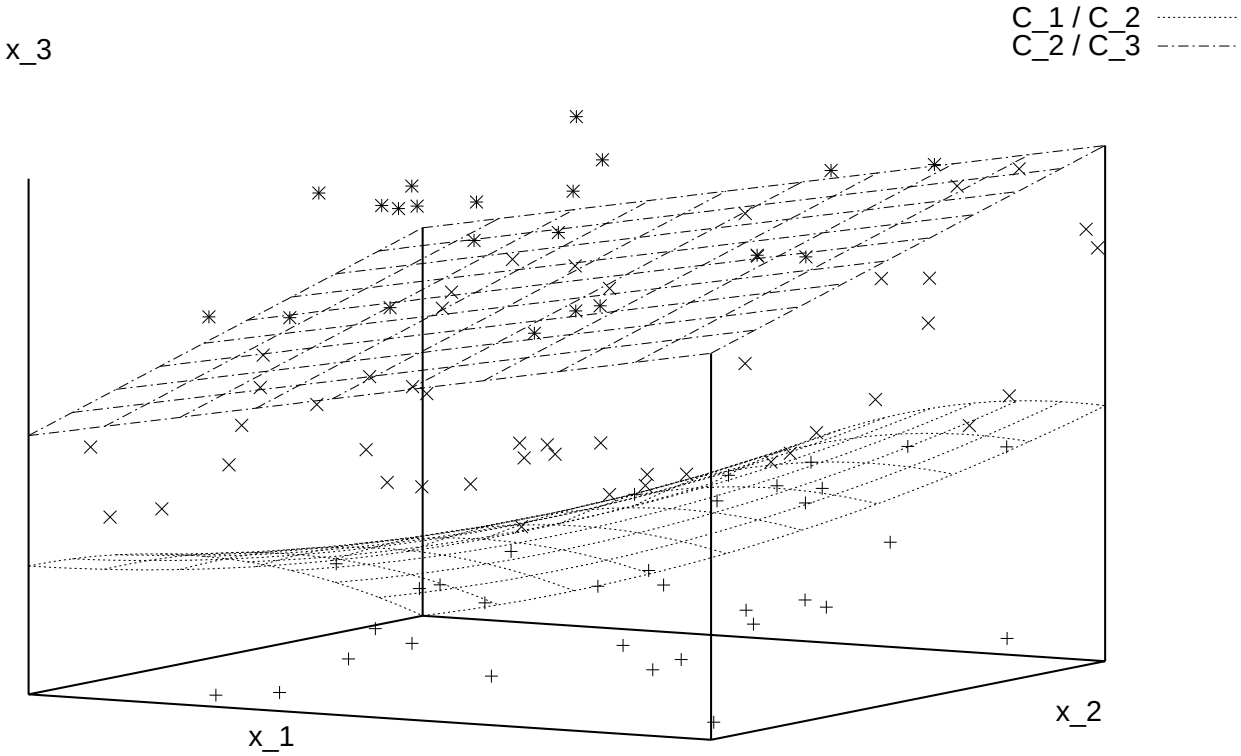


FIG. 1 – 3 categories non-linearly separable in 3D

Behaviour of a M-SVM

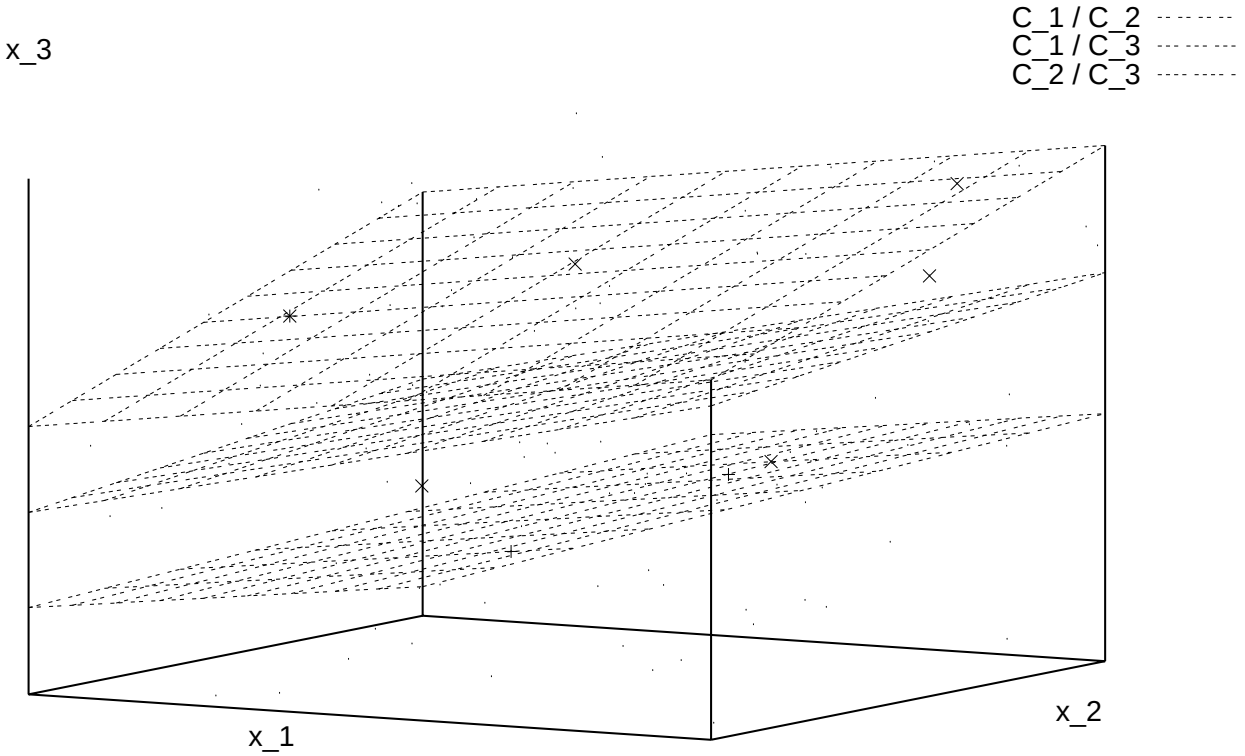


FIG. 2 – Hyperplanes and support vectors of a linear M-SVM

Empirical margin risk and uniform convergence result : multi-class case

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

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

Definition 1 (Canonical function) *Let $h = [h_k]$ be a fuction from $\mathcal{H}$. Its canonical function $\Delta h = [\Delta h_k]$, $(1 \leq k \leq Q)$, is the function from $\mathcal{X}$ into $\mathbb{R}^Q$ satisfying*

$$\Delta h_k(x) = \frac{1}{2}\left\{ h_k(x) - \max_{l \neq k} h_l(x) \right\}$$

Definition 2 (Empirical margin risk (Elisseeff & al. 99)) *The empirical risk with margin $\gamma \in (0,1]$ of h on a set $s_m = \{(x_1, C(x_1)), \ldots, (x_m, C(x_m))\}$ of size m is*

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

Empirical margin risk and uniform convergence result : multi-class case

Theorem 1 (Elisseeff & al. 99) *Let s_m be a m-sample of examples drawn independently from P. With probability at least $1-\delta$, for every value of γ in $(0,1]$, the risk $R(h)$ of a function h computed by a numerical Q-class discriminant model $\mathcal{H}$ is bounded above by :*

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

where $\Delta h^{\gamma} = [\pi_{\gamma} \circ \Delta h_k]$, $(1 \leq k \leq Q)$, $\Delta\mathcal{H}^{\gamma} = \{\Delta h^{\gamma} \ / \ h \in \mathcal{H}\}$

$$\forall s_m \in \mathcal{X}^m,\ \forall (h^{(1)}, h^{(2)}) \in \mathcal{H}^2,\ d_{l_\infty, l_\infty(s_m)}(h^{(1)}, h^{(2)}) = \max_{x_i \in s_m} \max_{k \in \{1,\ldots,Q\}} \left| h^{(1)}_k(x_i) - h^{(2)}_k(x_i) \right|$$

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

Extended notions of VC dimension

Definition 3 (Fat-shattering dimension (Kearns & Schapire 90)) *Let $\mathcal{H}$ be a set of real-valued functions on a set $\mathcal{X}$. For $\gamma > 0$, a subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ is said to be* γ-shattered *by $\mathcal{H}$ if there is a vector $v_b = [b_i] \in \mathbb{R}^m$ such that, for each binary vector $v_y = [y_i] \in \{-1, 1\}^m$, there is a function $h_y \in \mathcal{H}$ satisfying*

$$(h_y(x_i) - b_i)\, y_i \geq \gamma, \ (1 \leq i \leq m)$$

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

Definition 4 (Graph dimension (Dudley 87, Natarajan 89)) *Let $\mathcal{H}$ be a set of functions on a set $\mathcal{X}$ taking their values in a countable set. For any $h \in \mathcal{H}$, the* graph $\mathcal{G}$ *of h is $\mathcal{G}(h) = \{(x, h(x)) \ / \ x \in \mathcal{X}\}$ and the* graph space *of $\mathcal{H}$ is $\mathcal{G}(\mathcal{H}) = \{\mathcal{G}(h) \ / \ h \in \mathcal{H}\}$. Then the* graph dimension *of $\mathcal{H}$ is defined to be the VC dimension of the space $\mathcal{G}(\mathcal{H})$.*

M-fat-shattering dimension

Definition 5 (M-fat-shattering dimension) *Let $\mathcal{H}$ be a set of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$. For $\gamma > 0$, a subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ is said to be* M-γ-shattered *by $\mathcal{H}$ if there is a vector $v_b = [b_i] \in \mathbb{R}^m$ and a vector $v_c = [c_i] \in \{1, \ldots, Q\}^m$ such that, for each binary vector $v_y = [y_i] \in \{-1, 1\}^m$, there is a function $h_y = [h_{yk}]$, $(1 \leq k \leq Q) \in \mathcal{H}$ satisfying*

$$(h_{yc_i}(x_i) - b_i)\, y_i \geq \gamma, \ (1 \leq i \leq m)$$

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

M-fat-shattering dimension : extension of the fat-shattering dimension to the multivariate case and scale-sensitive version of the graph dimension

Generalization of Sauer's lemma : multi-class case

Theorem 2 *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For every value of γ in $(0,1]$ and every value of m in $\mathbb{N}^*$, the following bound is true :*

$$\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m) \leq 2\left(2mQ9^Q\right)^{d\log_2(18emQ/d)}$$

where $d = M\text{-}fat_{\Delta\mathcal{H}^\gamma}(\gamma/8)$.

Theorem 3 *Let $\mathcal{H}$ be a set of vector-valued functions $h = [h_k]$, $(1 \leq k \leq Q)$, from a set $\mathcal{X}$ into $\mathbb{R}^Q$. Let $\mathcal{H}_k$, $(1 \leq k \leq Q)$, be the sets of real-valued functions h_k corresponding to the different components of the functions h. Then the following bound holds true for all positive value of ϵ :*

$$M\text{-}fat_{\mathcal{H}}(\epsilon) \leq \sum_{k=1}^{Q} fat_{\mathcal{H}_k}(\epsilon)$$

Main difficulty : M-fat$_{\Delta\mathcal{H}}$ cannot be bounded in terms of M-fat$_{\mathcal{H}}$

Dependence of the capacity on the control term

Theorem 4 *Let $\mathcal{H}$ be the set of vector-valued functions $h = [h_k]$, $(1 \leq k \leq Q)$, computed by a M-SVM. For k in $\{1, \ldots, Q\}$, let $\Delta\mathcal{H}_k$ be the set of real-valued functions Δh_k corresponding to the k^{th} component of the functions Δh :*

$$\Delta h_k(x) = \frac{1}{2} \min_{l \neq k} \{\langle w_k - w_l, \Phi(x) \rangle + b_k - b_l\}$$

Suppose that $\Phi(\mathcal{X})$ is included in a ball of radius $\Lambda_{\Phi(\mathcal{X})}$ and that the vectors w_k satisfy $\max_{1 \leq k < l \leq Q} \|w_k - w_l\|_2 \leq \Lambda_w$. Then

$$\forall k \in \{1, \ldots, Q\}, \quad fat_{\Delta\mathcal{H}_k}(\epsilon) = O\left(\left(\frac{\Lambda_{\Phi(\mathcal{X})}\Lambda_w}{\epsilon}\right)^2\right)$$

Simple pathway to bound the covering numbers

Theorem 5 *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For every value of ϵ and γ satisfying $0 < \epsilon < \gamma$, the following bounds hold true :*

$$\mathcal{N}_{\infty,\infty}(\epsilon, \Delta\mathcal{H}^\gamma, m) \leq \mathcal{N}_{\infty,\infty}(\epsilon, \Delta\mathcal{H}, m) \leq \mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m)$$

Theorem 6 *Let $\mathcal{H}$ be a set of functions computed by a M-SVM. Suppose that the biases b_k all belong to the interval $[-b, b]$ Then :*

$$\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m) \leq \left\lceil \frac{b}{\epsilon} \right\rceil^Q \mathcal{N}_{\infty,\infty}(2\epsilon, \tilde{\mathcal{H}}, m)$$

Covering numbers and entropy numbers

Definition 6 (Entropy numbers) *Let (E, ρ) be a pseudo-metric space and H a subset of E. Then, for n in $\mathbb{N}^*$, the n^{th} entropy number of H, $\epsilon_n(H)$, is :*

$$\epsilon_n(H) = \inf \{\epsilon > 0 \ / \ \mathcal{N}(\epsilon, H, \rho) \leq n\}$$

Definition 7 (Entropy numbers of a bounded linear operator) *Let E_H and F_H be Hilbert spaces endowed with the norms $\|.\|_{E_H}$ and $\|.\|_{F_H}$. Let $\mathcal{L}(E_H, F_H)$ be the set of all bounded linear operators between $(E_H, \|.\|_{E_H})$ and $(F_H, \|.\|_{F_H})$. Let $T \in \mathcal{L}(E_H, F_H)$.*

$$\|T\| = \sup_{x \in U_{E_H}} \|Tx\|_F \quad \epsilon_n(T) = \epsilon_n(T(U_{E_H}))$$

Theorem 7 (Maurey) *Let $T \in \mathcal{L}(E_H, \mathbb{R}^m)$. Then, there exists a constant c such that, for all $n \in \mathbb{N}^*$,*

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

Covering numbers and entropy numbers

Theorem 8 *Let* $\tilde{\mathcal{H}} = \left\{\tilde{h}\right\}$ *be defined as above, with the additional constraint that each function* $\tilde{h} = [w_k]$, $(1 \leq k \leq Q)$, *satisfies :* $\max_k \|w_k\|_2 \leq \Lambda_w$

Let T_{s_m} *be the linear operator given by* $T_{s_m} = \Lambda_w S_{s_m}$ *with :*

$$S_{s_m} : \begin{array}{l} l_2 \longrightarrow l_\infty^{Qm} \\ w \mapsto [\langle w_1, \Phi(x_1)\rangle, \ldots \langle w_Q, \Phi(x_1)\rangle, \ldots, \langle w_1, \Phi(x_m)\rangle, \ldots \langle w_Q, \Phi(x_m)\rangle] \end{array}$$

Then

$$\epsilon_n\left(T_{s_m}\right) \leq \epsilon_0 \Longrightarrow \mathcal{N}_{\infty,\infty}(\epsilon_0, \tilde{\mathcal{H}}, m) \leq n$$

Constraints on the hyperplanes : $\|w_k\|_2^2$ or $\|w_k - w_l\|_2^2$?

Multi-class SVM	Objective function	Add. const.
Vapnik & Blanz 98	$J_1(w,b,\xi) = \sum_{k=1}^{Q} \|w_k\|^2 + C_1 1^T \xi$	-
Weston & Watkins 98	$J_1(w,b,\xi) = \sum_{k=1}^{Q} \|w_k\|^2 + C_1 1^T \xi$	-
Bredensteiner & al. 99	$J_2(w,b,\xi) = \sum_{k<l}^{Q} \|w_k - w_l\|^2 + \sum_{k=1}^{Q} \|w_k\|^2 + C_2 1^T \xi$	-
Guermeur & al. 00	$J_3(w,b,\xi) = \sum_{k<l}^{Q} \|w_k - w_l\|^2 + C_3 1^T \xi$	$\sum_{k=1}^{Q} w_k = 0_d$

Objective function	Add. const.	C	Solution
$J_1(w,b,\xi)$	-	C_1	$\left(w^{(1)}, b^{(1)}, \xi^{(1)}, \alpha^{(1)}, \beta^{(1)}\right)$
$J_2(w,b,\xi)$	-	$(Q+1)C_1$	$\left(w^{(1)}, b^{(1)}, \xi^{(1)}, (Q+1)\alpha^{(1)}, (Q+1)\beta^{(1)}\right)$
$J_3(w,b,\xi)$	$\sum_{k=1}^{Q} w_k = 0_d$	QC_1	$\left(w^{(1)}, b^{(1)}, \xi^{(1)}, Q\alpha^{(1)}, Q\beta^{(1)}, 0_d\right)$

The same set of primal variables generates solutions for the three problems

$\Longrightarrow$ All these multi-class SVMs are equivalent

Protein secondary structure prediction

Basic notions about proteins

– Proteins : macromolecules made up of amino acids
– 20 amino acids, each of them represented by a letter (A, R, N, D, C, E, . . .)

Hierarchical description of the conformation

– Primary structure (sequence of amino acids) $\Longleftarrow$ sequencing
– Secondary structure (sequence of structural elements) $\Longleftarrow$ circular dichroism
– Tertiary structure (three-dimensional structure) $\Longleftarrow$ X-ray, NMR
– . . .

Importance of structure prediction

– Knowing the structure is a prerequisite to gain a thorough understanding of the function
– Difficulty to determine the structure experimentally

Modular and hierarchical approach of the prediction

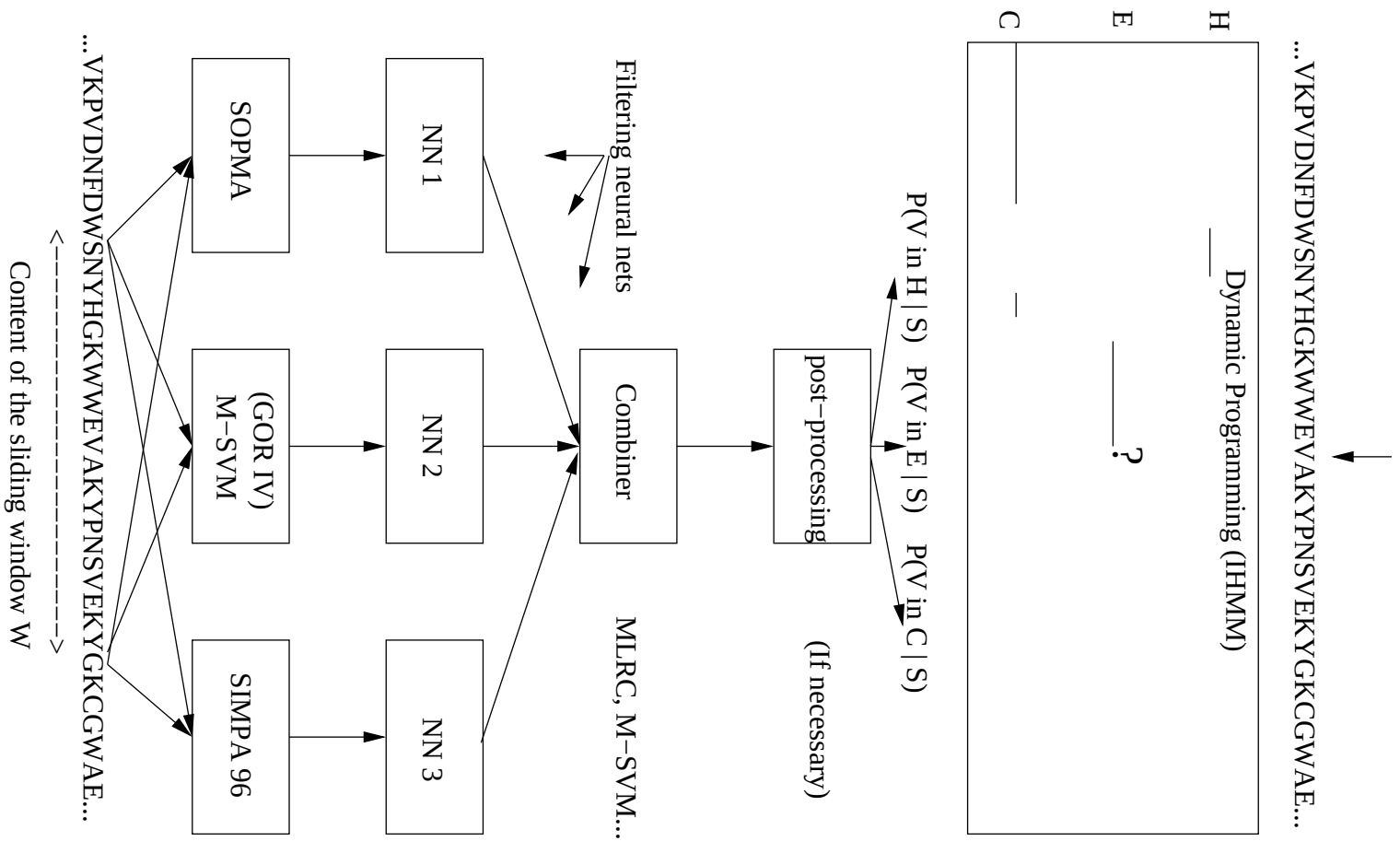


FIG. 3 – Hierarchical architecture for protein secondary structure prediction

Combination of the BRNNs of SSpro2

Collaboration with the biocomputing group of Prof. P. Baldi at the UCI

SSpro2 (Pollastri & al. 02)

- Data : profiles of alignments derived from PSI-BLAST
- Experts : 11 "Bidirectional Recurrent Neural Networks (BRNNs)
- SSpro2 : unweighted average on the outputs of the BRNNs

Goal

Improve performance by implementing a superior ensemble method $\Longrightarrow$ M-SVMs !

Experimental protocol

- New database of **1096 protein sequences (255551 amino acids)** exhibiting no homology with the training set of the BRNNs
- Training of the M-SVM + post-processing (perceptron with *softmax* units) + IHMM : *stacked generalization*

Different measures of prediction accuracy

Q_3 : recognition rate at the residue level

Pearson's/Matthews' correlation coefficients

$$C_i = \frac{p_i n_i - u_i o_i}{\sqrt{(p_i + u_i)(p_i + o_i)(n_i + u_i)(n_i + o_i)}}$$

Root mean square deviation (r.m.s.d.)

$$\sigma_i = \sqrt{\frac{1}{n_s} \sum_{j=1}^{j=n_s} (obs_{ij} - pred_{ij})^2}$$

Sov coefficients (Rost & al. 94, Zemla & al. 99)

$$Sov(\delta) = \frac{1}{n} \sum_{S_1} \left\{ \frac{1}{n_{S_2}} \sum_{S_2 / S_1 \cap S_2 \neq \emptyset} \frac{\min(end(S_1), end(S_2)) - \max(beg(S_1), beg(S_2)) + 1 + \delta}{\max(end(S_1), end(S_2)) - \min(beg(S_1), beg(S_2)) + 1} len(S_1) \right\}$$

Relative prediction accuracy of combiners

	Av.	MLP	$SVM_{1v.r}$	M-SVM1	M-SVM2	M-SVM3
Q_3	**76.94**	76.91	**77.01**	**77.09**	**77.12**	**77.03**
Q_α	86.7	86.7	86.8	86.8	86.9	86.8
Q_β	87.7	87.6	87.3	87.8	87.9	87.6
Q_c	79.5	79.6	79.6	79.6	79.6	79.5
C_α	0.72	0.72	0.71	0.72	0.72	0.72
C_β	0.62	0.63	0.62	0.63	0.63	0.63
C_c	0.58	0.58	0.58	0.58	0.58	0.58
Sov	72.2	72.2	72.0	72.4	72.5	72.3
Sov_α	75.6	76.1	76.0	76.1	76.0	76.0
Sov_β	67.1	69.0	67.3	68.9	69.0	68.6
Sov_c	69.0	67.6	68.1	68.5	68.5	68.4

TAB. 1 – Combination of the 11 BRNNs of SSpro2 with two-class and multi-class SVMs

Taking the PSI-BLAST profiles into account improves performance

	MLP	$SVM_{\alpha+\beta+c}$	M-SVM1
Q_3	**77.02**	**77.06**	**77.26**
Q_α	86.9	86.9	87.1
Q_β	87.6	87.3	87.8
Q_c	79.5	79.7	79.6
C_α	0.72	0.73	0.73
C_β	0.63	0.62	0.64
C_c	0.58	0.58	0.58
Sov	72.3	72.2	72.5
Sov_α	74.8	74.8	74.6
Sov_β	69.0	68.7	69.4
Sov_c	68.8	68.5	68.7

TAB. 2 – Combination of the 11 BRNNs of SSpro2 and PSI-BLAST derived profiles

M-SVM as basic classifier - Shortcomings of standard kernels

Canonical coding of the content of the sliding window

- $x = [x_{-n}, \ldots, x_i, \ldots, x_n]^T \in \{0,1\}^{(2n+1).22}$ ($|W| = 2n+1$)
- x_i : canonical coding of the amino acid in position i (binary vector)

Function computed by the kernel

$$K(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) = \exp\left(-\frac{(2n+1) - \sum_{i=-n}^{n} \delta_{x_i,x'_i}}{2\sigma^2}\right)$$

The kernel only computes the Hamming distance !

The possibility of insertions/deletions, the nature of the substitutions and the position in the window are not taken into account

Taking into account the substitutions

	G	P	D	E	A	N	Q	S	T	K	R	H	V	I	M	C	L	F	Y	W
G	2																			
P	1	3																		
D	0	0	2																	
E	0	-1	1	2																
A	0	-1	0	1	2															
N	0	0	1	0	0	3														
Q	0	0	0	1	0	1	2													
S	0	0	0	0	1	0	0	2												
T	0	0	0	0	0	0	0	0	2											
K	0	0	0	0	0	1	0	0	0	2										
R	0	0	0	0	0	0	0	0	0	1	2									
H	0	0	0	0	0	0	0	0	0	0	0	2								
V	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	2							
I	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	2						
M	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	0	0	2					
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2				
L	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	0	2	0	2			
F	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	0	1	0	-1	0	2		
Y	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	-1	0	1	2	
W	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	0	0	0	0	-1	0	0	0	2

FIG. 4 – Secondary structure similarity matrix **(Levin & al. 86)**

M-SVM as basic classifier - Optimization of the kernel

Kernel target alignment : multi-class case (Vert 02)

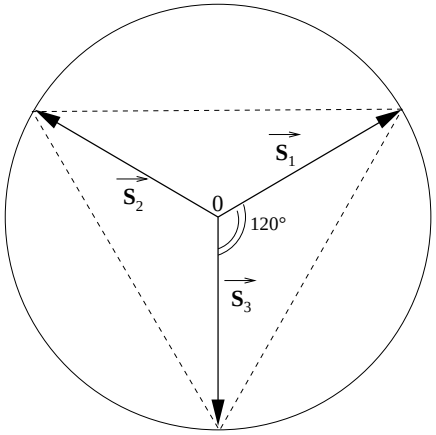


FIG. 5 – Optimal choice of the representatives of the categories

$$K_t(x, x') = \begin{cases} 1 & \text{if } C(x) = C(x') \\ -1/(Q-1) & \text{otherwise} \end{cases}$$

Training algorithm

Stochastic gradient descent

Dedicated kernel

Analytical expression

$$K_\theta(x, x') = \exp\left(-\frac{\sum_{i=-n}^{n} \theta_i^2 (1 - \langle \tilde{x}_i, \tilde{x}'_i \rangle)}{2\sigma^2}\right)$$

Influence of the position in the window

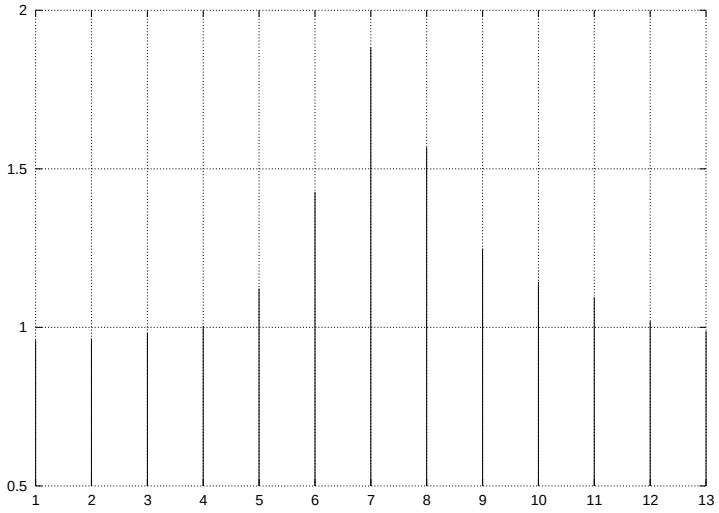


FIG. 6 – θ^* maximizing the alignment of K_θ

Prediction accuracy of a M-SVM without post-processing

	$SVM_{1v.r}$	M-SVM1
Q_3	68.4	69.0
C_α	0.53	0.54
C_β	0.50	0.51
C_c	0.50	0.52

TAB. 3 – Prediction accuracy on a set of **1096** non-homologous protein sequences

Post-processing of the posterior probability estimates with a Dynamic Programming algorithm

Underlying Hidden Markov Model : IHMM (Ramesh & Wilpon 92)

- One state for each structural state
- Parameters : $\lambda = (A, B, \Pi)$

$A = \{a_{kl}(d)\}$, $(1 \leq k, l \leq 3)$, $(1 \leq d \leq D_k)$

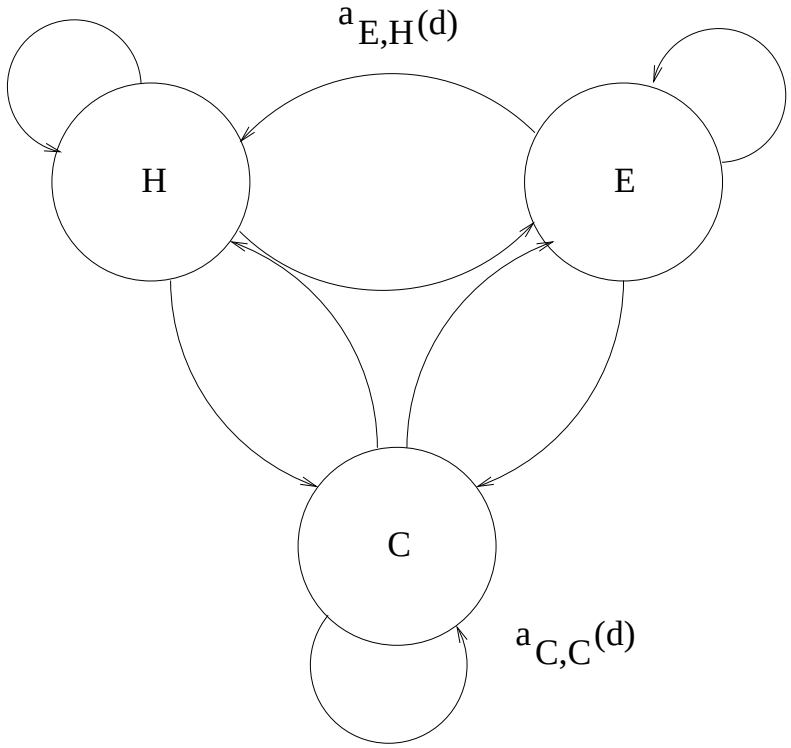
$a_{kl}(d)$: transition probability including a *duration model*

Prediction

"Best" sequence of states obtained with a variation of Viterbi's algorithm

Implementation of *N-Best* algorithms

Top of the hierarchy : higher-level treatments



Training algorithm

Transition probabilities $a_{kl}(d)$: iteratively adapted with a **Monte-Carlo and simulated annealing like algorithm** so that the corresponding frequences before and after post-processing become similar

Lengths of the conformational segments, observed and predicted

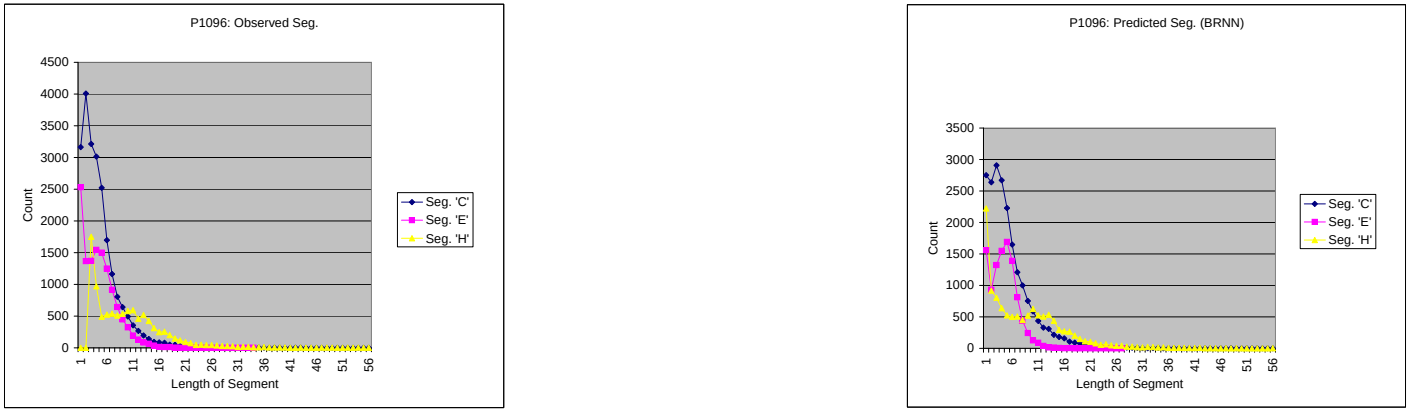


FIG. 7 – Distributions for the observed (left) and predicted (right) structures

Conclusions

Theory of M-SVMs

- New pathway to bound the risk of multi-class discriminant models
- New justifications of the training algorithms of all the M-SVMs proposed so far
- Possibility to develop new machines

M-SVMs for secondary structure prediction

- As ensemble methods, provide good generalization performance
- Little difference in generalization performance
- As basic classifiers, should prove superior to standard connectionist architectures

Work in progress

Theory and implementation of M-SVMs

- Comparison with studies involving *data dependent capacity measures* **(Boucheron & al. 99, Bartlett & al. 02, Bousquet 02)**
- Specification of optimization methods devoted to the new machines

M-SVMs for secondary structure prediction / biological sequence processing

- Additional work on the design and (efficient) implementation of specific kernels
- Fusion of additional knowledge sources provided by the biologists
- Global optimization of all the components of the hierarchical architecture