Borne "rayon-marge" sur l'erreur "leave-one-out" des SVM multi-classes

Yannick Darcy, Emmanuel Monfrini & Yann Guermeur

LORIA-CNRS-UHP, Campus scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex

Résumé : La mise en œuvre d'une machine à vecteurs support requiert la détermination des valeurs d'hyper-paramètres. Pour effectuer cette tâche, différentes procédures fondées sur la validation croisée sont actuellement disponibles. Leur principal défaut réside dans le temps de calcul qu'elles nécessitent. Afin de surmonter cette difficulté, plusieurs bornes supérieures sur l'erreur "leave-one-out" des SVM calculant des dichotomies ont été proposées. Nous présentons ici l'extension de l'une de ces bornes parmi les plus populaires, la borne "rayon-marge", au cas de la SVM multi-classe standard.

Abstract : Using a support vector machine requires to set the values of hyperparameters. To perform this task, one can use various procedures based on cross-validation. Obviously, the major drawback of such procedures rests in their time requirements. To overcome this difficulty, several upper bounds on the leave-one-out error of pattern recognition SVMs have been derived. We present here the extension of one of them among the most popular, the radius-margin bound, to the case of the standard multi-class SVM.

Mots clés : apprentissage, discrimination multi-classe, machines à vecteurs support, bornes sur le risque

1 Introduction

Pour mettre en œuvre efficacement les machines à vecteurs support (SVM), il est nécessaire d'optimiser les valeurs des hyper-paramètres : constante de marge douce et paramètres du noyau. Si de nombreuses approches ont été proposées pour résoudre ce problème de sélection de modèles (voir par exemple [6] et [7]), les procédures les plus utilisées demeurent celles basées sur la validation croisée. Leur défaut principal est leur coût en temps de calcul, en particulier lorsque l'on a recours à une procédure de type "leave-one-out". Afin de se libérer de cette contrainte, plusieurs bornes permettant de majorer l'erreur "leave-one-out" des SVM calculant des dichotomies ont été proposées dans la littérature [10, 11, 8]. A notre connaissance, une seule étude a porté sur l'extension de ces bornes au cas multi-classe. Dans [12], les auteurs, proposent deux extensions indirectes de la célèbre borne "rayon-marge" [2], soulignant qu'une extension directe n'est pas "viable" dans la mesure où le résultat de base s'appuie sur des propriétés intrinsèques de la SVM bi-classe. Cet article présente la généralisation directe de la borne "rayon-marge" à un nombre arbitraire de catégories. L'expression du théorème résultant dans le cas bi-classe produit exactement la borne originelle. Après avoir rappelé dans la section 2 ce qu'est une SVM bi-classe et la borne rayon-marge correspondante, nous introduisons et discutons le résultat multi-classe dans la section 3.

2 Borne rayon-marge pour les SVM bi-classes

On se place dans le cadre du calcul des dichotomies, consistant à associer aux éléments d'un espace de description $\mathcal{X}$ une étiquette appartenant à $\mathcal{Y} = \{-1, 1\}$. Pour déterminer, dans une famille donnée, la fonction rendant le mieux compte de cette dépendance, on dispose d'un m-échantillon $s_m = ((x_i, y_i))_{1 \leq i \leq m}$. Dans ce contexte, la famille des fonctions réalisables par une SVM [1, 3] est caractérisée à la fois par un noyau symétrique défini positif κ, de $\mathcal{X}^2$ dans $\mathbb{R}$, et par l'échantillon d'apprentissage. Il s'agit des fonctions de la forme $h(x) = \sum_{i=1}^{m} \beta_i \kappa(x_i, x) + b$, où les coefficients β_i et b sont des scalaires. En faisant intervenir l'espace de Hilbert à noyau reproduisant (RKHS) induit par κ, et en notant Φ l'une des fonctions sur $\mathcal{X}$ vérifiant $\kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle$, ceci se réécrit $h(x) = \sum_{i=1}^{m} \beta_i \langle \Phi(x_i), \Phi(x) \rangle + b = \langle w, \Phi(x) \rangle + b$. On voit ainsi apparaître l'expression d'un séparateur linéaire dans le RKHS incluant $\Phi(\mathcal{X})$. Les paramètres sont choisis de manière que le couple (w, b) définisse un *hyperplan optimal*, c'est-à-dire maximisant la *marge*. La marge est la plus petite distance euclidienne existant entre un $\Phi(x_i)$ et la frontière de décision $h(x) = 0$. La solution de ce problème est obtenue en résolvant le problème de programmation quadratique (QP) suivant :

Problème 1

$$\min_{\alpha} \left\{ \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j \kappa(x_i, x_j) - \sum_{i=1}^{m} \alpha_i \right\}$$

$$s.c. \begin{cases} 0 \leq \alpha_i \leq C, \quad (1 \leq i \leq m) \\ \sum_{i=1}^{m} y_i \alpha_i = 0 \end{cases}$$

où C, la *constante de marge douce*, permet de travailler à risque empirique non nul. L'expression des β_i en fonction des α_i est $\beta_i = y_i \alpha_i$, la valeur de b se déduisant à l'optimum des conditions de Kuhn-Tucker. Ce cadre étant posé, Vapnik a établi la borne suivante.

Théorème 1 ([9]) *Soit une SVM séparant sans erreur tous les points de l'échantillon d'apprentissage avec une marge γ et $\alpha^0 = [\alpha_i^0] \in [0, C[^m$ son vecteur de paramètres. Soit $\mathcal{L}_m$ le nombre d'erreurs commises en appliquant à cette machine une procédure de validation croisée "leave-one-out" et $\mathcal{D}_m$ le diamètre de la plus petite boule contenant les $\Phi(x_i)$ pour les indices i tels que $\alpha_i^0 > 0$. Alors,*

$$\mathcal{L}_m \leq \frac{\mathcal{D}_m^2}{\gamma^2}. \tag{1}$$

3 Borne rayon-marge pour les SVM multi-classes

La littérature propose plusieurs extensions des SVM au cas multi-classe (voir [5] pour une revue). Nous restreignons notre étude à la machine introduite dans [13, 9], que nous évoquerons dans ce qui suit comme "la" M-SVM. Toutes les démonstrations des résultats présentés dans cette section se trouvent dans [4].

3.1 SVM multi-classes

Dans le cas de la discrimination à Q catégories, l'ensemble $\mathcal{Y}$ devient $\{C_1, \ldots, C_k, \ldots, C_Q\}$. Pour simplifier les notations, on pourra identifier une catégorie à son indice. Une M-SVM réalise des fonctions vectorielles $h = (h_k)_{1 \leq k \leq Q}$ dont les fonctions composantes $h_k(.) = \sum_{i=1}^{m} \beta_{ik} \langle \Phi(x_i), \Phi(.) \rangle + b_k = \langle w_k, \Phi(.) \rangle + b_k$ sont construites comme les fonctions définissant les SVM bi-classes. Un point est affecté à la catégorie associée à la sortie la plus élevée. Ici encore, le paramétrage est obtenu par résolution d'un problème QP.

Problème 2

$$\min_{\alpha} \{J(\alpha)\}$$

$$s.c. \begin{cases} 0 \leq \alpha_{ik} < C, \quad (1 \leq i \leq m), (1 \leq k \leq Q), k \neq y_i \\ \sum_{x_i \in C_k} \sum_{l=1}^{Q} \alpha_{il} - \sum_{i=1}^{m} \alpha_{ik} = 0, \quad (1 \leq k \leq Q-1) \end{cases}$$

où la fonction objectif est donnée par :

$$J(\alpha) = \frac{1}{2} \left\{ \sum_{i \simeq j} \sum_{k=1}^{Q} \sum_{l=1}^{Q} \alpha_{ik} \alpha_{jl} \kappa(x_i, x_j) - 2 \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \alpha_{ik} \alpha_{jy_i} \kappa(x_i, x_j) \right.$$
$$\left. + \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \alpha_{ik} \alpha_{jk} \kappa(x_i, x_j) \right\} - \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik},$$

la notation $i \simeq j$ signifiant l'appartenance de x_i et x_j à une même classe.

La formulation du problème 2 fait implicitement intervenir, dans un but de simplification, des pseudo-variables α_{iy_i}, $(1 \leq i \leq m)$ toutes égales à 0. L'expression des w_k en fonction des α_{ik} et des données d'apprentissage est :

$$w_k = \sum_{x_i \in C_k} \sum_{l=1}^{Q} \alpha_{il} \Phi(x_i) - \sum_{i=1}^{m} \alpha_{ik} \Phi(x_i), \tag{2}$$

les b_k se déduisant à l'optimum des conditions de Kuhn-Tucker. La littérature propose de nombreuses notions de marges multi-classes. Nous retenons ici l'extension la plus naturelle.

Définition 1 *Considérons une M-SVM dont l'erreur en apprentissage est nulle. γ_{kl}, sa* marge *relative aux classes C_k et C_l, est la distance euclidienne minimale entre un point de s_m dans l'une ou l'autre de ces classes et l'hyperplan les séparant (d'équation $\langle w_k - w_l, \Phi(x) \rangle + b_k - b_l = 0$). En notant*

$$\delta_{kl} = \min \left[\min_{x_i \in C_k} \{ \langle w_k - w_l, \Phi(x_i) \rangle + b_k - b_l - 1 \}, \min_{x_j \in C_l} \{ \langle w_l - w_k, \Phi(x_j) \rangle + b_l - b_k - 1 \} \right], \tag{3}$$

l'expression analytique de γ_{kl} est donc :

$$\gamma_{kl} = \frac{1 + \delta_{kl}}{\|w_k - w_l\|}. \tag{4}$$

3.2 Majoration de l'erreur "leave-one-out" pour la M-SVM

Notre principal résultat, le théorème 2, est une conséquence directe d'un lemme dont la formulation nécessite l'introduction préalable de quelques notations.

Définition 2 *Soit une M-SVM séparant sans erreur tous les points de l'échantillon d'apprentissage* $s_m = ((x_i, y_i))_{1 \leq i \leq m}$ *et* $\alpha^0 = [\alpha^0_{ik}] \in [0, C[^{Qm}$ *son vecteur de paramètres. Soit* J_{VC} *la fonction qui, à un vecteur* $\nu = [\nu_{ik}] \in \mathbb{R}_+^{Qm}$ *vérifiant pour tout* i $\nu_{iy_i} = 0$, *associe* $J_{VC}(\nu) = \sum_{l=1}^{Q} \left(\sum_{j=1}^{m} \nu_{jl} \right)^2$. *Pour tout couple* (i,k) *vérifiant* $\alpha^0_{ik} > 0$, *on note respectivement* $K_{i,k}$ *et* $K'_{i,k}$ *les valeurs des fonctions objectif à l'optimum des problèmes QP suivants :*

Problème 3

$$\min_{\lambda} J_{VC}(\lambda)$$

$$s.c. \begin{cases} \forall l, \;\; \lambda_{il} = \frac{\alpha^0_{il}}{\alpha^0_{ik}} \\ \forall j \neq i, \forall l, \;\; 0 \leq \lambda_{jl} \leq \frac{\alpha^0_{jl}}{\alpha^0_{ik}} \\ \sum_{x_j \in C_l} \sum_{o=1}^{Q} \lambda_{jo} - \sum_{q=1}^{m} \lambda_{ql} = 0, \;\; (1 \leq l \leq Q-1) \end{cases}$$

Problème 4

$$\min_{\mu} J_{VC}(\mu)$$

$$s.c. \begin{cases} \begin{cases} l = k, \;\; \mu_{il} = 1 \\ l \neq k, \;\; \mu_{il} = 0 \end{cases} \\ \forall j \neq i, \forall l, \;\; \mu_{jl} \geq 0 \\ \sum_{x_j \in C_l} \sum_{o=1}^{Q} \mu_{jo} - \sum_{q=1}^{m} \mu_{ql} = 0, \;\; (1 \leq l \leq Q-1) \end{cases}$$

Lemme 1 *Soit une M-SVM séparant sans erreur tous les points de l'échantillon d'apprentissage et* $\alpha^0 = [\alpha^0_{ik}] \in [0, C[^{Qm}$ *son vecteur de paramètres. Si la même machine, entraînée sur* $s_m \setminus \{(x_p, y_p)\}$, *commet une erreur dans la classification de* x_p, *en lui attribuant indûment la classe* C_n, *alors :*

$$\alpha^0_{pn} \geq \frac{1}{K \mathcal{D}_m^2} \tag{5}$$

où $\mathcal{D}_m$ *est le diamètre de la plus petite boule contenant les* $\Phi(x_i)$ *tels qu'il existe un* α^0_{ik} *strictement positif, et* K *s'exprime en fonction des* $K_{i,k}$ *et* $K'_{i,k}$ *de la définition 2 comme :*

$$K = \sqrt{\max_{i,k \,:\, \alpha^0_{ik} > 0} (K_{i,k}.K'_{i,k})}. \tag{6}$$

Théorème 2 *Soit une M-SVM à Q catégories vérifiant les hypothèses du lemme 1. $\mathcal{L}_m$, K, $\mathcal{D}_m$, les δ_{kl} et les marges γ_{kl} étant définis comme précédemment, $\mathcal{L}_m$ est majoré par :*

$$\mathcal{L}_m \leq \frac{K\mathcal{D}_m^2}{Q} \sum_{k<l} \frac{(1+\delta_{kl})^2}{\gamma_{kl}^2}. \tag{7}$$

Dans le calcul de cette borne, la partie la plus coûteuse est l'obtention de la valeur de K. Dans la borne originelle, ce terme n'apparaît pas. Les auteurs tirent implicitement profit du fait que dans le cas bi-classe, les solutions (optimales) des problèmes 3 et 4 sont triviales, et fournissent une valeur de la fonction objectif égale à 2, donc une valeur de K égale à Q. En conséquence, la version bi-classe de notre borne n'est rien d'autre que la borne de Vapnik. Cette identité serait apparue de manière plus immédiate si nous avions pris pour terme de contrôle du problème primal d'apprentissage de la M-SVM, au lieu du terme habituel $\sum_{k=1}^{Q} \|w_k\|^2$, le terme généralisant directement celui du cas bi-classe, c'est-à-dire $\sum_{k<l} \|w_k - w_l\|^2$ (notons que $\sum_{k<l} \|w_k - w_l\|^2 = Q \sum_{k=1}^{Q} \|w_k\|^2$ puisque $\sum_{k=1}^{Q} w_k = 0$). Dans ce cas, nous aurions obtenu une expression de K qui aurait pris la valeur 1 pour $Q = 2$, et plus de coefficient $1/Q$ dans la borne sur $\mathcal{L}_m$. Etablir la pertinence de ce travail se réduit donc à en prouver l'utilité dans le cas $Q > 2$. Cette utilité s'exprime comme le fait de permettre une réalisation efficace de tâches de base, comme le choix des valeurs des hyper-paramètres, dans un temps de calcul significativement inférieur à celui de la validation croisée. Pour comprendre le gain en temps de calcul résultant de l'emploi de notre borne, il suffit d'observer que les problèmes 3 et 4 sont beaucoup plus simples que le problème 2. Il ne font en particulier pas intervenir le calcul du noyau. De plus, le nombre de problèmes à résoudre peut être petit lorsque le vecteur α^0 est creux, ce qui est généralement le cas. Une simplification supplémentaire résulterait de la démonstration de la proposition suivante, dont nous avons cherché en vain un contre-exemple : si extraire l'exemple (x_i, y_i) de la base d'apprentissage conduit à réaliser une erreur sur ce point, en le classant dans la catégorie C_k, alors on a $\alpha_{ik}^0 = \max_l \alpha_{il}^0$.

4 Conclusions et perspectives

Dans cet article, nous avons présenté une extension directe de la borne rayon-marge au cas multi-classe. Appliquée au cas bi-classe, elle correspond exactement à la borne originelle. Dans le cas multi-classe, elle doit permettre un gain significatif en temps de calcul. Cette borne est directement comparable à celles proposées dans [12]. L'étude comparative correspondante fait l'objet d'un travail en cours.

Remerciements Les auteurs souhaitent exprimer leur gratitude à Liva Ralaivola pour sa relecture et ses commentaires sur cette étude.

Références

[1] B. Boser, I. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT'92*, pages 144–152, 1992.

[2] O. Chapelle, V.N. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3) :131–159, 2002.

[3] C. Cortes and V.N. Vapnik. Support-Vector Networks. *Machine Learning*, 20 :273–297, 1995.

[4] Y. Darcy and Y. Guermeur. Radius-margin Bound on the Leave-one-out Error of Multi-class SVMs. Technical Report RR-5780, INRIA, 2005.

[5] Y. Guermeur, A. Elisseeff, and D. Zelus. A comparative study of multi-class support vector machines in the unifying framework of large margin classifiers. *Applied Stochastic Models in Business and Industry*, 21(2) :199–214, 2005.

[6] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning, Data Mining, Inference and Prediction*. Springer Series in Statistics. Springer, 2001.

[7] P. Massart. Concentrations inequalities and model selection. In *Ecole d'Eté de Probabilités de Saint-Flour XXXIII*, LNM. Springer-Verlag, 2003.

[8] M. Opper and O. Winther. Gaussian processes and SVM : Mean field and leave-one-out. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 311–326. The MIT Press, 2000.

[9] V.N. Vapnik. *Statistical learning theory.* John Wiley & Sons, Inc., N.Y., 1998.

[10] V.N. Vapnik and 0. Chapelle. Bounds on error expectation for support vector machines. *Neural Computation*, 12(9) :2013–2036, 2000.

[11] G. Wahba, Y. Lin, and H. Zhang. Generalized approximate cross-validation for support vector machines : another way to look at margin-like quantities. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 297–309. The MIT Press, 2000.

[12] L. Wang, P. Xue, and K.L. Chan. Generalized Radius-Margin Bounds for Model Selection in Multi-class SVMs. Technical report, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798, 2005.

[13] 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.