Sujet de thèse
proposé pour une allocation de recherche du ministère

Ecole doctorale IAEM, n° 0077

LORIA UMR 7503, équipe ABC

Localisation : LORIA, équipe ABC, Campus Scientifique, BP 239, 54506 Vandœuvre-lès-Nancy cedex
Directeur de thèse : Yann Guermeur, Yann.Guermeur@loria.fr, 03 83 59 30 18
Titre de la thèse : Sélection de modèle pour les systèmes discriminants multi-classes à grande marge

Résumé du sujet

L'apprentissage statistique [21] apporte depuis plusieurs décennies une contribution majeure à de nombreux thèmes de la reconnaissance des formes. Citons parmi ceux-ci le traitement de la parole, des images, de l'écriture cursive ou le thème de prédilection de l'équipe "Apprentissage et Biologie Computationnelle" (ABC) du LORIA, le traitement des séquences biologiques. Ce domaine de l'apprentissage automatique a récemment connu des développements importants, du fait de l'apparition des machines à noyau [18], dont l'exemple le plus connu est celui des machines à vecteurs support (SVM) [21], et des nouvelles techniques introduites pour l'étude des capacités de généralisation des modèles [3, 19]. Ces progrès sont d'autant plus intéressants qu'ils contribuent à résorber le fossé existant entre théorie et pratique. A présent, les inégalités de concentration les plus puissantes sont incorporées dans des algorithmes réalisant des tâches de sélection de fonction ou sélection de modèle [11, 14], avec pour conséquence directe l'amélioration des performances des méthodes de prédiction. Néanmoins, durant de nombreuses années, cette théorie n'a considéré la discrimination qu'à travers sa version bi-classe. Ce choix était d'autant moins satisfaisant que la théorie statistique de la discrimination multi-classe ne peut être considérée comme une extension triviale de celle du cas bi-classe (voir par exemple [7]). Ceci apparaît immédiatement lorsque l'on étudie les propriétés de généralisation des SVM multi-classes (M-SVM) [22, 20, 7, 8, 6].

L'objet de cette thèse est de développer et mettre en œuvre des méthodes de sélection de modèle dédiées aux systèmes discriminants multi-classes à grande marge, avec comme modèles de prédilection les M-SVM. Ces méthodes s'appuieront sur des majorations du risque (risques garantis) et de l'erreur de validation croisée "leave-one-out" déjà établies ou à établir. Les premières bornes multi-classes qui pourront être évaluées dans ce cadre sont celles démontrées ces dernières années dans l'équipe [8, 6, 16]. Leur utilisation comme fonctions objectif sera facilitée par le développement concomitant d'algorithmes permettant d'emprunter le chemin de régularisation sur toute sa longueur avec un temps de calcul réduit, tels ceux déjà proposés pour les SVM ou les M-SVM [10, 13]. Les principales idées sur lesquelles il sera possible de s'appuyer pour établir de nouvelles bornes multi-classes sont celles qui sous-tendent les résultats les plus prometteurs du cas bi-classe, comme les bornes locales, celles faisant intervenir une mesure de capacité empirique ou les bornes relevant de l'apprentissage PAC-bayésien [15, 12, 1]. De ce point de vue, les travaux de Peter Bartlett, d'Olivier Bousquet et de leurs co-auteurs [2, 5, 4] constitueront une source d'inspiration très utile. Naturellement, la pierre de touche des méthodes de sélection de modèle est l'évaluation sur des données du monde réel. Les réalisations produites par cette thèse seront évaluées sur des problèmes relevant d'un domaine unique : le traitement de données biologiques. Ces problèmes seront ceux sur lesquels l'équipe travaillera avec l'Institut de Biologie et Chimie des Protéines (IBCP), dirigé par Monsieur Deléage. Nous poursuivrons ainsi une collaboration établie de longue date et ayant déjà produit des résultats significatifs [9, 17].

Le sujet de thèse décrit ci-dessus trouverait tout naturellement sa place dans un projet de M-SVM "clef en main" que l'équipe ABC propose comme "projet pilote" pour le centre de compétences et de transfert (CCT G-BioModeL) du thème "Modélisation des Biomolécules et de leurs Interactions" (MBI) du projet "Modélisations, Informations et Systèmes Numériques" (MISN) du CPER 2007-2013. Notre objectif consiste à mettre à la disposition des biologistes et bioinformaticiens des logiciels implémentant des M-SVM en fixant automatiquement les valeurs des hyperparamètres. De cette manière, les utilisateurs de ces machines pourront les appliquer aux problèmes de leur choix sans que cela nécessite la moindre connaissance en statistique inférentielle. Ces logiciels s'organiseront autour des programmes de M-SVM que nous avons déjà développés et qui sont diffusés par le site des "kernel machines", à l'adresse suivante : `http://www.kernel-machines.org`.

Connaissances et compétences requises

Il est attendu du candidat qu'il possède des connaissances de base en statistique et probabilités, ainsi qu'une expérience minimale de programmation en C. A l'inverse, le sujet n'appelle aucun pré-requis en bioinformatique. Les connaissances utiles de biologie moléculaire pourront être acquises auprès des membres d'ABC et de ses partenaires scientifiques.

Références

[1] A. Ambroladze, E. Parrado-Hernandez, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In *NIPS 19*, 2007. (à paraître).

[2] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities : Risk bounds and structural results. *Journal of Machine Learning Research*, 3 :463–482, 2002.

[3] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification : A survey of some recent advances. *ESAIM : Probability and Statistics*, 9 :323–375, 2005.

[4] S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. In O. Bousquet, U. von Luxburg, and G. Rätsch, editors, *Advanced Lectures on Machine Learning*, pages 208–240. Springer, 2004.

[5] O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. Thèse de doctorat, Ecole Polytechnique, 2002.

[6] Y. Guermeur. Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs. *Communications in Statistics*, 2007. (soumis).

[7] Y. Guermeur. *SVM multiclasses, théorie et applications*. Habilitation à diriger des recherches, UHP, 2007.

[8] Y. Guermeur. VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 8 :2551–2594, 2007.

[9] Y. Guermeur, C. Geourjon, P. Gallinari, and G. Deléage. Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics*, 15(5) :413–421, 1999.

[10] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5 :1391–1415, 2004.

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer, New York, 2001.

[12] J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. In *NIPS 15*, pages 423–430, 2003.

[13] Y. Lee and Z. Cui. Characterizing the solution path of multicategory support vector machines. *Statistica Sinica*, 16 :391–409, 2006.

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

[15] D.A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3) :355–363, 1999.

[16] E. Monfrini and Y. Guermeur. A quadratic loss multi-class SVM. Technical report, LORIA, 2008. hal-00276700.

[17] N. Sapay, Y. Guermeur, and G. Deléage. Prediction of amphipathic in-plane membrane anchors in monotopic proteins using a SVM classifier. *BMC Bioinformatics*, 7(255), 2006.

[18] B. Schölkopf and A.J. Smola. *Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, MA, 2002.

[19] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 35(2) :575–607, 2007.

[20] A. Tewari and P.L. Bartlett. On the consistency of multiclass classification methods. In *COLT'05*, pages 143–157, 2005.

[21] V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., New York, 1998.

[22] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5 :1225–1251, 2004.