Proposition de sujet de thèse

Yann Guermeur

27 avril 2007

Titre

Etude des capacités de généralisation des systèmes discriminants à catégories multiples, application au traitement de séquences biologiques.

Encadrants

Jean-Paul Haton et Yann Guermeur

Résumé du sujet

L'apprentissage statistique [8, 25] 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 [20], dont l'exemple le plus connu est celui des machines à vecteurs support (SVM) [2, 6], et des nouvelles techniques introduites pour l'étude des capacités de généralisation des modèles [3, 22]. 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 [17], 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 (calcul des dichotomies). 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 [12, 11]). Ceci apparaît immédiatement lorsque l'on étudie les propriétés des SVM multi-classes [27, 23, 15, 11]. L'objet de cette thèse est donc de développer la théorie de la discrimination multi-classe, en étendant en particulier les résultats les plus prometteurs du cas bi-classe, comme les bornes locales, ou celles faisant intervenir une mesure de capacité empirique. De ce point de vue, les travaux d'Olivier Bousquet et ses co-auteurs [5, 4, 1] constitueront une source d'inspiration très utile. L'étude se concentrera sur les classifieurs à grande marge [21] construits autour d'un noyau. Une application privilégiée de ces travaux en théorie des bornes sera la spécification de méthodes de sélection de modèle dédiées aux SVM multi-classes. On pourra pour ce faire partir des résultats déjà établis dans l'équipe, comme ceux exposés dans [7].

Le traitement des séquences biologiques soulève plusieurs types de problèmes particulièrement intéressants pour l'apprenticien. Il permet en particulier d'étudier la prise en compte de dépendances à distance, parfois fortement distribuées [9], la gestion de masses de données [24], incomplètes [26, 18] ou entachées d'erreurs, ainsi que la gestion de données ne satisfaisant pas aux hypothèses habituelles concernant l'échantillonnage. Les travaux théoriques réalisés au cours de cette thèse seront appliqués dans le

cadre du développement de la méthode de prédiction hiérarchique et modulaire conçue dans l'équipe [10, 14]. Le problème pourrait être celui de la prédiction de la structure secondaire des protéines, dont ABC s'est fait une spécialité [13, 16, 19], mais d'autres options peuvent également être envisagées, en fonction de la nature des résultats produits.

Le sujet de thèse décrit ci-dessus trouverait tout naturellement sa place dans un projet portant sur la prédiction de la structure secondaire et tertiaire des protéines qui devrait être prochainement soumis au thème "Modélisation des Biomolécules et de leurs Interactions" (MBI) du projet "Modélisations, Informations et Systèmes Numériques" (MISN) du CPER 2007-2013.

Références

[1] P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher Complexities. *Annals of Statistics*, 33(4) :1497–1537, 2005.

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

[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*. PhD thesis, Ecole Polytechnique, 2002.

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

[7] Y. Darcy, E. Monfrini, and Y. Guermeur. Borne "rayon-marge" sur l'erreur "leave-one-out" des SVM multi-classes. In *JdS'06*, 2006.

[8] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New-York, 1996.

[9] Y. Guermeur. *Combinaison de classifieurs statistiques, application à la prédiction de la structure secondaire des protéines*. PhD thesis, Université Paris 6, 1997. (in French).

[10] Y. Guermeur. Combining discriminant models with new multi-class SVMs. *Pattern Analysis and Applications*, 5(2) :168–179, 2002.

[11] Y. Guermeur. VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 2007. (accepté).

[12] Y. Guermeur, A. Elisseeff, and H. Paugam-Moisy. Estimating the sample complexity of a multi-class discriminant model. In *ICANN'99*, pages 310–315. IEE, 1999.

[13] 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.

[14] Y. Guermeur, A. Lifchitz, and R. Vert. A kernel for protein secondary structure prediction. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, chapter 9, pages 193–206. The MIT Press, 2004.

[15] Y. Guermeur, M. Maumy, and F. Sur. Model selection for multi-class SVMs. In *ASMDA'05*, pages 507–516, 2005.

[16] Y. Guermeur, G. Pollastri, A. Elisseeff, D. Zelus, H. Paugam-Moisy, and P. Baldi. Combining protein secondary structure prediction models with ensemble methods of optimal complexity. *Neurocomputing*, 56C :305–327, 2004.

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

[18] H. Philippe, E.A. Snell, E. Bapteste, P. Lopez, P.W.H. Holland, and D. Casane. Phylogenomics of eukaryotes : Impact of missing data on large alignments. *Molecular Biology and Evolution*, 21(9) :1740–1752, 2004.

[19] 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.

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

[21] A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors. *Advances in Large Margin Classifiers*. The MIT Press, 2000.

[22] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 35(2), 2007. (à paraître).

[23] A. Tewari and P.L. Bartlett. On the consistency of multiclass classification methods. In *COLT'05*, pages 147–153, 2005.

[24] F. Thomarat, C.P. Vivares, and M. Gouy. Phylogenetic analysis of the complete genome sequence of *Encephalitozoon cuniculi* supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes. *J Mol Evol.*, 59(6) :780–791, 2004.

[25] V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., N.Y., 1998.

[26] J.J. Wiens. Missing data, incomplete taxa, and phylogenetic accuracy. *Syst. Biol.*, 52(4) :528–538, 2003.

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