Etude comparée des performances de SVM multi-classes en prédiction de la structure secondaire des protéines

Y. Guermeur

LORIA-CNRS, équipe ABC
Campus Scientifique, BP 239
54506 Vandœuvre-lès-Nancy cedex
Yann.Guermeur@loria.fr
http://www.loria.fr/~guermeur

Résumé. Les SVM bi-classes, introduites en bioinformatique à la fin des années 90, font aujourd'hui référence pour de nombreux problèmes de traitement de séquences biologiques. Les SVM multi-classes, de conception plus récente, sont progressivement appliquées à ces problèmes, singulièrement en biologie structurale prédictive. Dans cet article, nous proposons une étude comparée des performances de trois SVM multi-classes en prédiction de la structure secondaire des protéines. Les modèles impliqués sont celui de Weston et Watkins, celui de Lee et co-auteurs ainsi qu'une nouvelle machine nommée M-SVM^2. Cette étude se conçoit comme une étape dans la mise au point d'une méthode de prédiction hybride, intégrant systèmes discriminants et génératifs et s'appuyant sur une approche hiérarchique du problème.

1 Introduction

La biologie moléculaire est un domaine d'application de choix pour les modèles de l'apprentissage artificiel. Si les problèmes de discrimination qu'elle propose font intervenir des données de natures très diverses, un grand nombre d'entre eux, souvent parmi les plus importants, relèvent du traitement de séquences. C'est en particulier le cas des problèmes de biologie structurale prédictive. L'un des plus anciens, qui a résisté à quarante années de recherches intensives, est la prédiction *ab initio* du repliement des protéines globulaires. Il est ordinairement abordé par le biais d'une approche du type diviser pour régner faisant intervenir une sous-tâche nommée prédiction de la structure secondaire. Du point de vue de l'apprentissage artificiel, cette tâche de discrimination à catégories multiples est d'un intérêt majeur, ceci à plus d'un titre. Pour nous restreindre à ce qui fera l'objet de cet article, il convient d'illustrer cet intérêt en précisant que les principaux modèles connexionnistes discriminants ont été appliqués en prédiction de la structure secondaire, de même que les machines à vecteurs support (SVM). Dans ce domaine, les perceptrons multi-couches (PMC) et leurs variantes récurrentes constituent l'état de l'art depuis environ vingt ans. Ils ont été rejoints récemment par les SVM. Le choix de ce problème est donc particulièrement indiqué pour réaliser une étude comparative de SVM multi-classes (M-SVM). C'est précisément ce que propose cet article. Les trois machines considérées sont la M-SVM de Weston et Watkins (1998), celle de Lee et al. (2004)