Sélection de modèles par des méthodes à noyaux pour la classification de données séquentielles

Trinh Minh Tri Do, Thierry Artières, Patrick Gallinari
LIP6, Université Pierre et Marie Curie
{Prénom.Nom}@lip6.fr

Ce travail concerne le développement de méthodes de classification discriminantes pour des données séquentielles. Quelques techniques ont été proposées pour étendre aux séquences les méthodes discriminantes, comme les machines à vecteurs supports, par nature plus adaptées aux données en dimension fixe. Elles permettent de classifier des séquences complètes mais pas de réaliser la segmentation, qui consiste à reconnaître la séquence d'unités, phonèmes ou lettres par exemple, correspondant à un signal. En utilisant une correspondance donnée / modèle nous transformons le problème de l'apprentissage des modèles à partir de données par un problème de sélection de modèles, qui peut être attaqué via des méthodes du type machines à vecteurs supports. Nous proposons et évaluons divers noyaux pour cela et fournissons des résultats expérimentaux pour deux problèmes de classification.

1 Introduction

Cette étude concerne l'intégration d'une information discriminante dans des systèmes de classification de données reposant sur des modèles génératifs et plus spécifiquement sur des mélanges de modèles génératifs. Dans la majorité des tâches de classification, on dispose de deux possibilités principales sur la nature de l'approche à employer, l'approche discriminante et l'approche générative. On peut utiliser un modèle discriminant -- réseau de neurones, classifieur linéaire, machine à vecteurs supports (MVS) -- dont l'apprentissage est focalisé sur ce qui différencie les différentes classes. D'un point de vue probabiliste, cela correspond à apprendre les lois de probabilités a posteriori des classes. La plupart de ces techniques discriminantes sont adaptées à des données en dimension fixe et sont plus délicates à utiliser avec des données séquentielles, de taille variable, comme la parole, l'écriture, etc. Une autre approche consiste à modéliser les classes indépendamment les unes des autres, et à apprendre pour chacune un modèle correspondant à sa densité de probabilité (e.g. modèle gaussien, modèle de Markov) avec un critère du type Maximum de Vraisemblance. On utilise un modèle génératif par classe, où chaque modèle est appris indépendamment des autres avec les données de sa classe. Ensuite, via le théorème de Bayes, on peut se ramener aux probabilités a posteriori et donc construire un système de classification optimal.

En règle générale, l'approche discriminante est plus performante. Cependant, on peut avoir intérêt à employer des mélanges de modèles génératifs dans certaines conditions. Les mélanges de modèles sont particulièrement adaptés lorsque les classes sont fortement multimodales (par exemple en écriture manuscrite, un « b » peut être écrit de différentes façons, on parle d'allographes). Les modèles génératifs sont eux particulièrement intéressants lorsque les données sont de dimension variable. Ce dernier cas correspond à