Méthodes à noyaux appliquées aux textes structurés

Sujeevan Aseervatham, Emmanuel Viennet

Université de Paris-Nord, LIPN - UMR CNRS 7030
99, avenue Jean-Baptiste Clément
93430 Villetaneuse, France
{Prénom.Nom}@lipn.univ-paris13.fr

Résumé. Cet article ébauche un état de l'art sur l'utilisation des noyaux pour le traitement des données structurées. Les applications modernes de la fouille de données sont de plus en plus confrontés à des données structurées, notamment textuelles. Les algorithmes d'apprentissage doivent donc être capables de tirer parti des informations apportées par la structure, ce qui pose d'intéressants problèmes de représentation des données. L'une des approches possibles consiste à utiliser les noyaux de Mercer. Ces noyaux permettent de calculer la similarité entre deux données de type quelconque, et peuvent être utilisés par une large gamme d'algorithmes d'apprentissage (Machines à Vecteur de Support, ACP, Analyse Discriminante, Perceptron, etc). Nous présentons dans cet article les principaux noyaux proposés ces dernières années pour le traitement des structures telles que les séquences, les arbres et les graphes.

1 Introduction

Les techniques d'apprentissage statistique sont généralement conçues pour travailler sur des données vectorielles ; chaque mesure est représentée par un ensemble de données numériques de taille fixe. Pendant plusieurs décennies, les recherches en statistique se sont centrées sur des problèmes comme la normalisation des données, le traitement des valeurs manquantes, etc. Depuis une dizaine d'années, sous la pression des applications, nous sommes confrontés à des problème dans lesquels la structure des données porte une information essentielle : textes en langage naturel, documents XML, séquences biologiques, analyse de scènes (images), analyse des réseaux sociaux. Pour attaquer ces problèmes, il est nécessaire trouver un moyen de traiter l'information structurelle, par exemple en calculant une mesure de similarité entre deux structures.

De nombreux systèmes d'apprentissage numérique des données textuelles utilisent une représentation du texte en "sac de mot". Ce type de codage, qui a l'avantage de la simplicité, n'utilise que les fréquences d'apparition des mots dans les documents et perd toute information liée à l'ordre des éléments (ordre des mots, structure en paragraphes ou sections, etc).

Depuis une petite dizaine d'années, une nouvelle famille d'algorithmes d'apprentissage basés sur la notion de noyaux, fait l'objet d'intenses recherches. Les noyaux, proposés par V. Vapnik pour les machines à vecteur de support (SVM) (Vapnik, 1995), permettent de définir des mesures de similarité non linéaires. En simplifiant, la fonction noyau calcule un produit scalaire