Multi-catégorisation de textes juridiques et retour de pertinence

Vincent Pisetta, Hakim Hacid, Djamel A. Zighed

Laboratoire ERIC – 5, av. Pierre Mendès-France- 69767 Bron- France
vpisetta@etu.univ-lyon2.fr,
hhacid@eric-univ.lyon2.fr,
zighed@univ-lyon2.fr

Résumé. La fouille de données textuelles constitue un champ majeur du traitement automatique des données. Une large variété de conférences, comme TREC, lui sont consacrées. Dans cette étude, nous nous intéressons à la fouille de textes juridiques, dans l'objectif est le classement automatique de ces textes. Nous utilisons des outils d'analyses linguistiques (extraction de terminologie) dans le but de repérer les concepts présents dans le corpus. Ces concepts permettent de construire un espace de représentation de faible dimensionnalité, ce qui nous permet d'utiliser des algorithmes d'apprentissage basés sur des mesures de similarité entre individus, comme les graphes de voisinage. Nous comparons les résultats issus du graphe et de C4.5 avec les SVM qui eux sont utilisés sans réduction de la dimensionnalité.

1 Introduction

Le cadre général de l'apprentissage automatique part d'un fichier d'apprentissage comportant n lignes et p colonnes. Les lignes représentent les individus et les colonnes les attributs, quantitatifs ou qualitatifs observés pour chaque individu ligne. Dans ce contexte, on suppose également que l'échantillon d'apprentissage est relativement conséquent par rapport au nombre d'attributs. Généralement la taille de l'échantillon est de l'ordre de 10 fois le nombre de variables pour espérer obtenir une certaine stabilité, c'est-à-dire une erreur en généralisation qui n'est pas trop loin de l'erreur en apprentissage. De plus, l'attribut à prédire est supposé à valeur unique. C'est une variable à valeurs réelles dans le cas de la régression et c'est une variable à modalités discrètes, appelées classes d'appartenance, dans le cas du classement. Ces questions relatives aux rapports entre taille d'échantillon et taille de l'espace des variables sont étudiées de façon très approfondies dans les publications relatives à l'apprentissage statistique (Vapnik, 1995). Dans ce papier nous décrivons une situation d'apprentissage qui s'écarte significativement du cadre classique tel que décrit plus haut. En effet, le contexte expérimental ne nous permet pas de disposer immédiatement d'un ensemble d'apprentissage conséquent, chaque individu peut appartenir à plusieurs classes simultanément, et chaque individu, au lieu d'être décrit par un ensemble attributs-valeurs, l'est par un texte en langage naturel en anglais.