Accélération de la méthode des K plus proches voisins pour la catégorisation de textes

Fatiha Barigou*, Baghdad Atmani**
Youcef Bouziane***, Naouel Barigou****

Département d'Informatique, Université d'Oran
BP 1524, El M'Naouer, Es Senia, 31 000 Oran, Algérie.
*,** Laboratoire d'informatique d'Oran
(fatbarigou, atmani.baghdad)@gamil.com,
,* (youcefbouzianemi, barigounaouel)@gmail.com

Résumé. Parmi la panoplie de classificateurs utilisés dans la catégorisation de textes, nous nous intéressons à l'algorithme des k-voisins les plus proches. Ces performances le situent parmi les meilleures méthodes de catégorisation de textes. Toutefois, il présente certaines limites: (i) coût mémoire car il faut stocker l'ensemble d'apprentissage en entier et (ii) coût élevé de calcul car il doit explorer l'ensemble d'apprentissage pour classer un nouveau document. Dans ce papier, nous proposons une nouvelle démarche pour réduire ce temps de classification sans dégrader les performances de classification.

1 Introduction

La Catégorisation de textes joue un rôle très important dans la recherche d'information et la fouille de textes. Cette tâche a été couronnée de succès en faisant face à une grande variété d'applications. Ce succès est du principalement à la participation croissante de la communauté d'apprentissage machine. Dans ce travail, nous nous intéressons à l'algorithme des K-plus proches voisins (Cover et Hart, 1967). Ce dernier développé tout d'abord par (Fix et Hodges, 1989) est devenu l'un des algorithmes les plus populaires dans la catégorisation de textes. Il est robuste et placé parmi les meilleurs algorithmes (Sebastiani, 2002). Toutefois, il présente certaines limites, (i) stockage mémoire énorme car il faut stocker l'ensemble complet d'apprentissage et (ii) coût élevé de calcul car il doit explorer l'ensemble d'apprentissage en entier pour pouvoir classer un nouveau document. Une solution intéressante à base d'automate cellulaire appelée CAkNN (Cellular Automaton combined with k-NN) a été proposée dans (Barigou et al., 2012) pour réduire le temps de classification, dans le cadre du filtrage de spam. Les expériences réalisées sur le corpus LingSpam ont montré que la méthode CAkNN permet d'atteindre de meilleures performances de classification comparée à d'autres travaux publiés dans le domaine de filtrage de spam. Dans ce papier, nous allons reprendre cette solution pour la catégorisation de textes et nous allons montrer à travers un ensemble d'expériences que CAkNN permet de réduire le temps de classification par une sélection d'un minimum d'instances d'apprentissage pour la classification d'un nouveau document et ceci sans que la performance prédictive n'en soit affectée. Ce papier est organisé comme suit :la section 2 est dédiée