Sélection de prototypes en vue d'une catégorisation de textes avec les K plus proches voisins : étude comparative

Fatiha Barigou*, Baya Naouel Barigou**
Baghdad Atmani***, Bouziane Beldjilali****

Département d'Informatique, Université d'Oran
BP 1524, El M'Naouer, Es Senia, 31 000 Oran, Algérie.
*,**,***,**** Laboratoire d'informatique d'Oran
Équipe de simulation,intégration et fouille de données (SIF)
(fatbarigou, barigounaouel, baghdad.atmani)@gmail.com, bouzianebeldjilali@yahoo.fr

Résumé. La technique des K plus proches voisins (KNN) est une méthode d'apprentissage à base d'instances, elle a été appliquée dans la catégorisation de textes depuis de nombreuses années. En contraste avec ses performances de classification, il est reconnu que cet algorithme est lent pendant la classification d'un nouveau document. Les Techniques de sélection de prototypes sont apparues comme des méthodes très compétitives pour améliorer le KNN grâce à la réduction des données. L'étude contenue dans ce papier a pour objectif d'analyser l'impact de ces méthodes sur la performance de la classification de textes avec l'algorithme KNN.

1 Introduction

En termes de performance de classification de textes, KNN se classe parmi les classifieurs les plus performants, un résultat obtenu d'une multitude de tests de comparaison effectués sur le corpus Reuters Yang (1999). En contraste avec ses performances de classification, il est reconnu que cet algorithme est lent puisqu'il requiert qu'une mesure de similarité soit calculée entre tous les documents d'apprentissage et le nouveau document. Il est caractérisé par un apprentissage très rapide, il est facile à apprendre, il est robuste aux ensembles d'apprentissage bruités et il est efficace si le corpus est grand Bhatia et Vandana (2010). Un inconvénient majeur du KNN reste le temps qu'il met pour classer un nouveau document. Différentes solutions ont été proposées pour réduire la complexité de calcul. Nous nous intéressons, dans ce papier, aux méthodes de sélection de prototypes. Plus précisément, nous étudions l'impact de différentes méthodes de sélection de prototypes sur la performance de la catégorisation de textes avec le classifieur KNN. Essentiellement, voici comment se structure la suite du papier, la section 2 présente une série de méthodes de sélection de prototypes, en décrivant leurs principales caractéristiques. La section 3 présente les différentes expérimentations effectuées sur les différents corpus de textes pour comparer les différentes méthodes de sélection de prototypes. La conclusion générale résume le travail effectué et les résultats obtenus.