GENDESC : Vers une nouvelle représentation des données textuelles

Guillaume Tisserant *, Violaine Prince *, Mathieu Roche *,**

* LIRMM, CNRS, Université Montpellier 2
161 rue Ada, 34095 Montpellier Cedex 5, France
tisserant@lirmm.fr, prince@lirmm.fr, mroche@lirmm.fr
** TETIS, Cirad, Irstea, AgroParisTech
500 rue Jean-François Breton, 34093 Montpellier Cedex 5, France
mathieu.roche@cirad.fr

Résumé. Dans cet article, nous nous intéressons à la classification automatique de données textuelles par des algorithmes d'apprentissage supervisé. L'objectif est de montrer comment l'amélioration de la représentation des données textuelles influe sur les performances des algorithmes d'apprentissage. Partant du postulat qu'un mot n'a pas un sens bien établi sans son contexte, nous proposerons des descripteurs donnant le plus d'information possible sur le contexte des mots. Pour cela, nous avons mis au point une méthode, nommée GENDESC, qui consiste à "généraliser" les mots les moins pertinents pour la classification, c'est-à-dire, à éviter le bruit sémantique (souvent dû à la polysémie) provoqué par ces termes non ou peu pertinents. Cette généralisation s'appuie sur des informations grammaticales, telles que la catégorie et la position dans la structure. La méthode GENDESC a été évaluée et adaptée à la problématique de classification de textes selon une opinion ou une thématique.

1 Introduction

La problématique à laquelle cet article se confronte, est liée à la tâche de classification de données textuelles. Les données textuelles sont extrêmement difficiles à analyser et classifier d'après Witten et Frank (2005). Les algorithmes d'apprentissage supervisé que nous nous proposons d'utiliser dans cette étude nécessitent de connaître la classe (e.g. thème, sentiment, etc.) à associer à chaque document. Les entrées de ces algorithmes sont des "paquets" de descripteurs linguistiques (c'est-à-dire des critères de classification issus des propriétés du matériau langagier, comme la catégorie grammaticale, la fonction lexicale, le rôle syntaxique, etc. mais aussi des critères terminologiques) représentant le document à classifier. Une fois la phase d'apprentissage effectuée, le modèle appris peut attribuer une classe à des "paquets de descripteurs" non étiquetés qui lui sont donnés. Un résumé ce cette approche est donné en Figure 1. La qualité de la classification proposée par l'algorithme va donc dépendre à la fois de la qualité de l'algorithme d'apprentissage, mais aussi de la façon dont les données qui lui sont transmises sont représentées, comme le montre Béchet (2009).