De nouvelles pondérations adaptées à la classification de petits volumes de données textuelles

Flavien Bouillot*,**, Pascal Poncelet*, Mathieu Roche*,***

* LIRMM, Univ. Montpellier 2, CNRS – France
**ITESOFT, Aimargues – France
***TETIS, Cirad, Irstea, AgroPariTech – France
Prenom.Nom@lirmm.fr

Résumé. Un des défis actuels dans le domaine de la classification supervisée de documents est de pouvoir produire un modèle fiable à partir d'un faible volume de données. Avec un volume conséquent de données, les classifieurs fournissent des résultats satisfaisants mais les performances sont dégradées lorsque celui-ci diminue. Nous proposons, dans cet article, de nouvelles méthodes de pondérations résistant à une diminution du volume de données. Leur efficacité, évaluée en utilisant des algorithmes de classification supervisés existants (Naive Bayes et Class-Feature-Centroid) sur deux corpus différents, est supérieure à celle des autres algorithmes lorsque le nombre de descripteurs diminue. Nous avons étudié en parallèle les paramètres influençant les différentes approches telles que le nombre de classes, de documents ou de descripteurs.

1 Introduction

La classification supervisée de documents vise à déterminer la ou les catégories potentielles d'un document à partir de son contenu (les termes le composant). Dans un cadre supervisé, le processus se décompose généralement en 2 phases :

1. la phase d'apprentissage, qui vise à créer un modèle à partir d'un ensemble d'exemples étiquetés (documents dont la classe est connue),
2. la phase de classification, qui va déterminer la ou les catégories d'un document dont la classe est inconnue par application du modèle.

Bien entendu, la qualité du modèle dépend de la qualité et du nombre d'exemples disponibles. Ainsi plus il y a d'exemples, plus les observations seront fiables et plus le modèle sera précis et efficace.

Il peut cependant s'avérer intéressant de pouvoir élaborer un modèle de classification fiable à partir d'un faible nombre de descripteurs (Forman et Cohen, 2004). Par exemple, le développement des réseaux sociaux, avec un nombre de plus en plus important de messages en temps réel mais d'une taille limitée (comme un tweet limité à 140 caractères), implique la mise à disposition d'outils capables de les classer rapidement avec un volume restreint de données. Dans ce contexte, l'extraction de descripteurs pertinents et discriminants représente un défi