Apport de la prise en compte du contexte structurel dans les modèles bayésiens de classification de documents semi-structurés

Pierre-François Marteau, Gildas Ménier, Leopold Ekamby
VALORIA, Université de Bretagne Sud, Campus de Tohannic, 56 000 Vannes
{Pierre-François.Marteau, Gildas.Menier}@univ-ubs.fr

Résumé. Nous nous intéressons dans cet article au problème de la classification supervisée de documents semi-structurés. Un modèle formel basé sur des hypothèses simples et originales à notre connaissance est proposé. Ce modèle puise ses fondements dans les modèles de classification bayésiens, en ciblant la prise en compte de la structure des documents dans les tâches de classification. Ce modèle permet d'envisager la fusion de données numériques ou symboliques structurées et de données non structurées qui peuvent faire l'objet d'une modélisation spécifique. Des versions simplifiées de ce modèle sont implémentées pour évaluer de manière comparée à d'autres approches l'impact de la prise en compte de la structure documentaire dans des tâches de classification de documents textuels. Les premiers résultats, qui confirment ceux déjà obtenu dans le cadre de travaux similaires, montrent que la prise en compte du contexte structurel d'occurrence des mots améliore de manière significative les performances d'un classifieur bayésien naïf multinomial. Cette implémentation conduit à des performances comparables à celles atteintes par les classifieurs SVM sur la tâche considérée. Une implémentation plus complète de ce modèle doit permettre d'envisager des expérimentations ou des applications plus complexes et plus riches. Ces résultats ouvrent des perspectives autour de l'exploitation d'heuristiques de pondération des estimateurs associés aux composantes structurelles des documents.

1 Introduction

Les volumes croissants d'information stockée tant dans les bases documentaires traditionnelles que sur le World Wide Web constituent une motivation de plus en plus pressante pour le développement de méthodes de classification et de filtrage automatique de documents performantes, ceci afin de mieux organiser et structurer ces données et en faciliter l'accès. Dans cet article, nous nous intéressons à la classification de documents textuels semi-structurés au format XML. La recherche sur les méthodes de classification de documents textuels a conduit à une production impressionnante d'articles depuis les années 1990 (Cf. [Yang, 1999] pour une synthèse récente de ces travaux). Une liste non exhaustive d'approches basées sur les techniques d'apprentissage automatique inclut le Classifieur Bayésien Naif (CBN) [Duda & Hart, 1973], [Lewis et Ringuette 1994], [McCallum et Nigam 1998], les k plus proches voisins (k-NN) [Yang, 1999], les Machine à support vectoriel (SVM) [Vapnik, 1995,98], [Joachims, 1999], [Dumais et al., 1998], le boosting appliqué à la