Une Approche Filtre pour la Sélection de Variables en Apprentissage Non Supervisé

Pierre-Emmanuel JOUVE *, Nicolas NICOLOYANNIS *
*LABORATOIRE ERIC, Université Lumière - Lyon2, http://eric.univ-lyon2.fr
Bâtiment L, 5 av. Pierre Mendès-France
69 676 BRON cedex FRANCE
pierre.jouve@eric.univ-lyon2.fr, nicoloyannis@univ-lyon2.fr

Résumé. La Sélection de Variable (SV) constitue une technique efficace pour réduire la dimension des espaces d'apprentissage et s'avère être une méthode essentielle pour le pré-traitement de données afin de supprimer les variables bruitées et/ou inutiles. Peu de méthodes de SV ont été proposées dans le cadre de l'apprentissage non supervisé, et, la plupart d'entre elles, sont des méthodes dites "enveloppes" nécessitant l'utilisation d'un algorithme d'apprentissage pour évaluer les sous ensembles de variables. Or, l'approche "enveloppe" est largement mal adaptée à une utilisation lors de cas "réels". En effet, d'une part ces méthodes ne sont pas indépendantes vis à vis des algorithmes d'apprentissage non supervisé qui nécessitent le plus souvent de fixer un certain nombre de paramètres ; mais surtout, il n'existe pas de critères bien adaptés à l'évaluation de la qualité d'apprentissage non supervisé dans des sous espaces différents. Nous proposons et évaluons dans ce papier une méthode "filtre" et donc indépendante des algorithmes d'apprentissage non supervisé. Cette méthode s'appuie sur deux indices permettant d'évaluer l'adéquation entre deux ensembles de variables (entre deux sous espaces).

1 Introduction

La grande dimensionnalité de l'espace de représentation des données est un problème commun en apprentissage. La Sélection de Variables (SV) permet de déterminer quelles sont les variables pertinentes et constitue ainsi une technique efficiente pour la réduction de la dimension. Une variable pertinente pour une tâche d'apprentissage peut être définie comme une variable dont la suppression dégrade de manière significative la qualité de l'apprentissage réalisé. La suppression des variables non pertinentes permet donc la réduction de dimensionnalité, et, peut simultanément impliquer un accroissement de la précision et de la compréhensibilité des modèles bâtis. Il existe deux contextes principaux pour l'apprentissage : l'apprentissage supervisé et l'apprentissage non supervisé (clustering). S'il existe nombre de méthodes pour la SV dans le contexte supervisé (Dash et al. 1997), il n'existe que peu de méthodes (la plupart étant récentes) pour le contexte non supervisé. Cela peut être expliqué par le fait qu'il est plus aisé de sélectionner des variables pour l'apprentissage supervisé que pour le clustering. Dans le cadre supervisé, ce qui doit être appris est "connu a priori" alors que cela n'est pas le cas pour le clustering, dès lors, déterminer les variables pertinentes pour cette tâche peut être ardu. Le processus de SV pour le clustering peut être vu comme le processus de