

Pondération et classification simultanée de données binaires et continues

Nicoleta Rogovschi*, Mustapha Lebbah**, Nistor Grozavu**

*LIPADE, Université Paris Descartes
45 rue des Saints Pères
75270 Paris Cedex 06, France
prenom.nom@parisdescartes.fr

**LIPN-UMR 7030 Université Paris 13 - CNRS
99, av. J-B Clément - F-93430 Villetaneuse France.
prenom.nom@lipn.univ-paris13.fr

Résumé. Dans cet article, nous proposons une nouvelle approche de classification topologique et de pondération des variables mixtes (qualitatives et quantitatives codées en binaire) durant un processus d'apprentissage non supervisé. Cette approche est basée sur le modèle des cartes auto-organisatrices. L'apprentissage est combiné à un mécanisme de pondération des différentes variables sous forme de poids d'influence sur la pertinence des variables. L'apprentissage des pondérations et des prototypes est réalisé d'une manière simultanée en favorisant une classification optimisée des données. L'approche proposée a été validée sur des données qualitatives codées en binaire et plusieurs bases de données mixtes.

1 Introduction

La carte topologique proposée par (Kohonen, 2001) utilise un algorithme d'auto-organisation (SOM) qui fournit la quantification et la classification de l'espace des observations. Récemment, de nouveaux modèles de cartes auto-organisatrices dédiés à des données spécifiques ont été proposés dans (Bishop et al., 1998; Lebbah et al., 2008). Quelques-uns de ces modèles sont basés sur un formalisme probabiliste et d'autres sont des méthodes de quantification. Dans la littérature on trouve des approches basées sur la pondération comme les travaux de (Huang et al., 2005; Blanche et al., 2006; Grozavu et al., 2009). Pour les données continues, un modèle de cartes auto-organisatrices a été déjà proposé pour la pondération locale des variables appelé *lw*-SOM, (Grozavu et al., 2009). Cet algorithme se présente comme l'adaptation aux cartes SOM de l'approche de pondération proposée pour les *K*-moyennes par (Huang et al., 2005). Le modèle *lw*-SOM est dédié uniquement au cas des variables continues et n'est pas directement applicable aux données catégorielles. A notre connaissance, parmi les approches de pondération qui existent nous n'avons pas rencontré des travaux qui portent sur la classification non supervisée pondérée basée sur les cartes auto-organisatrices qui traitent des données mixtes (qualitatives codées en binaire et quantitatives). Nous voulons dans ce papier présenter une version déterministe qui tient compte de la nature des données sans utiliser des versions