Traitement des exemples atypiques en apprentissage par la régression

Stéphane Lallich — Fabrice Muhlenbach — Djamel A. Zighed

E.R.I.C. Lyon
Université Lumière Lyon 2
5, avenue Pierre Mendès-France
69676 BRON Cedex – France

lallich@univ-lyon2.fr — muhlenba@eric.univ-lyon2.fr — zighed@univ-lyon2.fr

RÉSUMÉ. *Nos travaux concernent la qualité de la représentation et la détection des points atypiques en apprentissage supervisé. Dans le cas où la variable à prédire est numérique – on parle alors d'apprentissage par la régression –, nous avons proposé d'évaluer la qualité de la représentation associée à un graphe de voisinage issu des prédicteurs à partir d'un coefficient d'autocorrélation de voisinage. Ce coefficient est construit sur le modèle du coefficient d'autocorrélation spatiale de Moran. Poursuivant l'analogie avec l'analyse spatiale, nous proposons dans ce papier de décomposer ce coefficient en une somme des coefficients locaux associés à chaque exemple et de tracer le diagramme de dispersion de Moran afin de repérer les exemples pour lesquels la valeur de la variable à prédire est atypique. L'expérimentation conduite sur diverses bases du site UCI Machine Learning a donné des résultats satisfaisants.*

ABSTRACT. *Our research thematic deals with the representation quality and the outlier detection in supervised learning. The prediction of a continuous value is referred to as regression learning. In this case, once constructed the neighbourhood graph resulting from the predictors, we proposed in a recent work to evaluate the representation quality using neighbourhood autocorrelation coefficient, like Moran spatial coefficient. Extending the analogy with spatial analysis, we suggest in this paper to detect abnormal values of the variable to predict from local components of global neighbourhood autocorrelation coefficient and Moran scatterplot. Experimentation led on various bases from UCI Machine Learning repository has given interesting results.*

MOTS-CLÉS : *apprentissage supervisé, variable à prédire numérique, graphe de voisinage, exemples atypiques, filtrage.*

KEYWORDS: *Supervised Learning, Continuous Class, Neighborhood Graph, Mislabeled Examples, Filtering.*

RSTI – 17/2003. EGC 2003, pages 399 à 410