Amélioration d'une classification par filtrage des exemples mal étiquetés

Fabrice Muhlenbach — Stéphane Lallich — Djamel A. Zighed

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

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

RÉSUMÉ. *Ce papier propose une nouvelle méthode qui améliore la performance en généralisation d'un algorithme d'apprentissage supervisé lorsque certains exemples de la base d'apprentissage ont été mal étiquetés. Notre méthode procède en deux temps. Une première étape de filtrage permet d'identifier et de retirer les exemples mal étiquetés. Ensuite, au cours de l'étape d'apprentissage, l'algorithme choisi – ici celui du plus proche voisin – opère sur l'ensemble d'apprentissage résultant du filtrage. La nouveauté de notre méthode réside dans la stratégie de filtrage qui repose sur la statistique des arêtes pondérées coupées, dans le cadre conceptuel des graphes de voisinage géométrique – ici le Graphe des Voisins Relatifs de Toussaint. L'expérimentation conduite sur 10 bases du serveur UCI Machine Learning a montré que lorsque l'on augmente le niveau de bruit sur la classe, de 0 à 20 %, notre méthode permet de maintenir le taux d'erreur en généralisation associé aux données non bruitées, plus particulièrement si les classes sont bien séparables.*

ABSTRACT. *This paper presents a new approach to improve classification accuracies produced by learning algorithms when there are mislabeled training instances. Our approach operates in two steps. First, a new filtering procedure identifies and removes mislabelled instances. Secondly, the filtered set of training instances is provided as input to the learning algorithm. Our filtering procedure is based on geometrical neighbourhood concept (the Relative Neighbourhood Graph of Toussaint) and uses weighted cut edges statistic. Final learning algorithm is 1-NN. Our experiments on ten benchmarks of UCI Machine Learning Repository show that our method allows to maintain the generalisation error rate when we introduce from 0 to 20 percent of noise on the class, especially when classes are well separable.*

MOTS-CLÉS : *Apprentissage Supervisé, Graphes de Voisinage Géométrique, Exemples mal étiquetés, Filtrage*

KEYWORDS: *Supervised Learning, Geometrical Neighbourhood Graph, Mislabelled Examples, Filtering*