Validation statistique des cartes de Kohonen en apprentissage supervisé

Elie Prudhomme, Stéphane Lallich
Laboratoire E.R.I.C, Université Lumière Lyon 2
5, avenue Pierre Mendès-France, 69676 BRON Cedex – France
elie.prudhomme@etu.univ-lyon2.fr, stephane.lallich@univ-lyon2.fr

Résumé. En apprentissage supervisé, la prédiction de la classe est le but ultime. Plus largement, on attend d'une bonne méthodologie d'apprentissage qu'elle permette une représentation des données susceptible de faciliter la navigation de l'utilisateur dans la base d'exemples et d'aider au choix des exemples et des variables pertinents tout en assurant une prédiction de qualité dont on comprenne les ressorts. Différents travaux ont montré l'aptitude des graphes de voisinage issus des prédicteurs à fonder une telle méthodologie, ainsi le graphe des voisins relatifs de Toussaint. Cependant, la complexité de leur construction, en $O(n^3)$, reste élevée.

Dans le cas de données volumineuses, nous proposons de substituer aux graphes de voisinage les cartes de Kohonen construites sur les prédicteurs. Après un bref rappel du principe des cartes de Kohonen en apprentissage non supervisé, nous montrons comment celles-ci peuvent fonder une stratégie d'apprentissage optimisée. Nous proposons ensuite d'évaluer la qualité de cette stratégie par une statistique originale qui est étroitement corrélée au taux d'erreur en généralisation. Différentes expérimentations montrent la faisabilité de cette approche. On dispose alors d'un critère fiable pour sélectionner les individus et les attributs pertinents.

Mots-clefs : apprentissage supervisé, cartes de Kohonen, validation statistique

1 Position du problème

Les méthodes d'apprentissage supervisé d'une variable catégorielle ont pour objet *in fine* la prédiction de la classe d'appartenance d'un nouvel exemple à partir d'un échantillon d'exemples étiquetés. En fait, la prédiction n'est qu'une étape de la procédure d'apprentissage, qui est enrichie par l'analyse exploratoire des données tout à la fois pour les préparer au mieux et pour leur donner du sens en intégrant d'éventuelles informations contextuelles.

Dans une telle perspective, le recours aux graphes de voisinage apporte une solution efficace. On construit le graphe de voisinage issu des prédicteurs, par exemple le graphe des voisins relatifs de Toussaint (Toussaint et Menard, 1980), puis l'on colorie les sommets du graphe en fonction de leur classe d'appartenance. Pour trouver la classe d'un nouvel exemple, on insère celui-ci dans le graphe de voisinage et on lui attribue la classe majoritaire parmi ses voisins dans le graphe. Divers travaux ont proposé une statistique (le poids des arêtes coupées) qui évalue la capacité prédictive d'un graphe de voisinage et permet la sélection de variables pertinentes ou la détection