Ensemble prédicteur fondé sur les cartes auto-organisatrices adapté aux données volumineuses.

Elie Prudhomme*, Stéphane Lallich*

*Université Lumière Lyon 2, Laboratoire ERIC,
5 avenue Pierre Mendès-France
69676 Bron
eprudhomme@eric.univ-lyon2.fr, stephane.lallich@univ-lyon2.fr

Résumé. Le stockage massif des données noie l'information pertinente et engendre des problèmes théoriques liés à la volumétrie des données disponibles. Ces problèmes dégradent la capacité prédictive des algorithmes d'extraction des connaissances à partir des données. Dans cet article, nous proposons une méthodologie adaptée à la représentation et à la prédiction des données volumineuses. A cette fin, suite à un partitionnement des attributs, des groupes d'attributs non-corrélés sont créés qui permettent de contourner les problèmes liés aux espaces de grandes dimensions. Un Ensemble est alors mis en place, apprenant chaque groupe par une carte auto-organisatrice. Outre la prédiction, ces cartes ont pour objectif une représentation pertinente des données. Enfin, la prédiction est réalisée par un vote des différentes cartes. Une expérimentation est menée qui confirme le bien-fondé de cette approche.

1 Espaces de grandes dimensions

Les systèmes d'information tendent vers le stockage et l'analyse d'une quantité croissante d'information. En apprentissage, cette évolution a pour conséquence une augmentation massive du nombre d'individus et du nombre d'attributs. Elle est due à la fois à un faible coût de stockage et à une collecte plus facile des informations. Les individus d'intérêt sont ainsi devenus des objets de plus en plus complexes. C'est le cas par exemple des puces à ADN qui décrivent des individus à travers l'expression de milliers de gènes, des banques d'images dont chacune nécessite des centaines d'attributs ou encore de l'internet et des documents qu'il contient. Les outils classiques de l'apprentissage automatique, notamment ceux relatif à la prédiction, perdent une partie de leur efficacité pour traiter ces données (Verleysen, 2003). En effet, que ce soit du point de vue des individus ou des attributs, leur présence en un nombre important pose deux catégories de problème.

La première catégorie est relative à la qualité des données. Celle-ci dépend de la pertinence des informations récoltées. Elle est améliorée lors d'une phase de prétraitement. Du point de vue des individus, il s'agit de détecter ceux qui sont mal étiquetés comme ceux qui présentent des valeurs atypiques, ces deux types d'individus faussant fortement l'apprentissage et nuisant