Sélection non paramétrique et régularisée d'instances et de variables

Sylvain Ferrandiz[1,2], Marc Boullé[1,]

France Télécom R&D,
2, avenue Pierre Marzin, 22300 Lannion
sylvain.ferrandiz@francetelecom.com et
marc.boulle@francetelecom.com

Résumé : La classification suivant le plus proche voisin est une règle simple et souvent performante. Sa mise en oeuvre pratique nécessite de sélectionner les instances ainsi que les variables descriptives. Dans cet article, on introduit un critère d'évaluation supervisée d'un ensemble de variables et d'un ensemble d'instances. Ce critère est non-paramétrique et prévient le sur-apprentissage. Une heuristique d'optimisation étant adoptée, la méthode obtenue est évaluée sur une vingtaine de jeux de données de l'UCI.

Mots-clés : Classification supervisée, plus proche voisin, sélection de variables, sélection d'instances.

1 Evaluation d'un ensemble d'instances et de variables

La classification par le plus proche voisin consiste à attribuer à une instance l'étiquette de l'instance la plus proche parmi celles constituant l'échantillon. La qualité de cette règle de classification repose sur la qualité de l'échantillon constitué. La sélection des variables et instances pertinentes est ainsi une étape cruciale. Si on dispose initialement d'un échantillon de N instances étiquetées représentées par D variables descriptives, on propose le critère suivant d'évaluation de la qualité d'un sous-échantillon résultant de la sélection de L variables et K instances :

$$\log(D+1) + \log \binom{D+L-1}{L} + \log N + \log \binom{N+K-1}{K}$$
$$+ \sum_{k=1}^{K} \log \binom{N_k+J-1}{J-1} + \sum_{k=1}^{K} \log \frac{N_k!}{N_{k1}! \ldots N_{kJ}!}.$$

Ce critère résulte de l'instanciation du principe de longueur de description minimum. Le premier terme correspond à la description du nombre L, le second à la spécification des L variables à prendre en compte parmi les D descripteurs ; le troisième correspond à la description du nombre K, le quatrième à la spécification des K instances sélectionnées.

A la partition de Voronoi de l'échantillon en K groupes induite par ces K instances sont attachés les cardinaux $N_1, \ldots, N_K$ des groupes et les cardinaux $N_{11}, \ldots, N_{KJ}$ respectifs des J classes cibles dans chaque groupe. Le cinquième terme correspond à la description des distributions de fréquences des cibles dans chaque groupe et le dernier spécifie l'attribution des étiquettes à chacune des instances.

2 Expérimentation

On dispose d'un critère d'évaluation d'un sous-échantillon. On propose une heuristique de recherche du meilleur sous-échantillon procédant en deux temps. Un ensemble d'instances étant fixé aléatoirement, on considère l'ensemble des variables descriptives et on supprime une variable lorsque cela conduit à une meilleure solution. Les variables sont considérées dans un ordre aléatoire. Cette phase décrémentale est suivie d'une phase incrémentale analogue. Les variables sélectionnées sont alors fixées et on applique l'heuristique de sélection des instances décrite dans (Ferrandiz & Boullé, 2006). Une validation croisée à 10 niveaux est appliquée pour évaluer la méthode. La méthode est évaluée sur des jeux de données de l'UCI (Blake & Merz, 1996).

En plus des 7 variables descriptives du jeu LED, le jeu LED17 contient 17 variables indépendantes de la variable cible. La méthode trouve les 7 variables significatives 7 fois sur 10 et sélectionne 6 variables dans les trois autre cas.

Sur le jeu de données Spam, qui contient 57 variables descriptives, la sélection d'instances seule conduit à une légère perte de performance en prédiction vis-à-vis de la classification sans sélection (82.7% contre 85.2%). En pratiquant une sélection de variables avant la sélection d'instances, on obtient une performance de 88.9%. Travailler suivant un axe de sélection (ici la sélection d'instances), aussi bien qu'on puisse le faire, peut parfois être inutile si le problème est mal situé sur l'autre axe (ici la sélection de variables). Le critère permet de travailler sur les deux axes simultanément.

On évalue la performance de la méthode sur 22 jeux de données de l'UCI. La prédiction après sélection (75.6%) est statistiquement au même niveau qu'avant sélection (75.9%), avec 8 victoires significatives contre 5 défaites significatives). La réduction de l'échantillon ne s'accompagne donc pas d'une perte de performance prédictive.

En pratique, l'ensemble de données est divisé en trois parties : une pour construire le modèle (l'échantillon d'apprentissage), une pour ajuster les éventuels paramètres (l'échantillon de validation) et une pour évaluer la performance en généralisation du modèle (l'échantillon de test). Le critère étant non-paramétrique, l'échantillon de validation est inutile. Plus de données peuvent ainsi être utilisées pour l'apprentissage.

Références

BLAKE C. & MERZ C. (1996). Uci repository of machine learning databases. http ://www.ics.uci.edu/mlearn/MLRepository.html.

FERRANDIZ S. & BOULLÉ M. (2006). Sélection supervisée d'instances : une approche descriptive. In G. V. D.A. ZIGHED, Ed., *Actes de la conférence extraction et gestion des connaissances*, volume 2, p. 421–432 : Cépaduès.