

Construction de variables et arbre de décision

Gaëlle Legrand*, Nicolas Nicoloyannis*

*Laboratoire ERIC Université Lumière - Lyon2 Bâtiment L 5 av. Pierre
Mendès-France 69 676 BRON cedex FRANCE
gaelle.legrand@eric.univ-lyon2.fr ; nicolas.nicoloyannis@univ-lyon2.fr,

Dans ce papier, nous discutons du problème de la construction de nouvelles variables dans le but d'améliorer la qualité d'apprentissage supervisé. L'utilisation de la construction de variables lors de la phase de pré-traitement des données du processus d'extraction des connaissances à partir des données, permet d'obtenir des variables synthétiques pour re-décrire les données d'entrée du problème d'apprentissage.

La méthode proposée construit de nouvelles variables en se basant sur l'analyse des arbres d'induction et appartient ainsi à l'ensemble des méthodes de construction de combinaison de variables par analyse topologique des arbres, [Pagallo et Haussler, 1990], [Oliveira et Vincentelli, 1993]. Les nouvelles variables créées sont sous la forme de conjonction de modalités des variables initiales.

Notre méthode se décompose en trois étapes :

1. Génération de l'ensemble de toutes les règles d'apprentissage issues de l'arbre de décision.
2. Classement des règles ainsi obtenues : elles sont regroupées en fonction de la classe de la variable endogène qui leur est associée. Il y aura donc autant de regroupements de règles que de classes de la variable endogène.
3. Construction des nouvelles variables : il y a une variable construite par regroupement de règles. Donc, le nombre de variables construites est égal au nombre de classes de la variable endogène. Chaque règle est une conjonction de modalités de variables et chaque nouvelle variable est la disjonction des règles appartenant à un même regroupement. Les nouvelles variables sont de type booléen.

Références

- [Pagallo et Haussler, 1990] G. Pagallo et D. Haussler. Boolean feature discovery in empirical learning. *Machine Learning*, 5(1):71–100, 1990.
- [Oliveira et Vincentelli, 1993] A. Oliveira et A. Vincentelli. Learning complex boolean functions : Algorithms and applications. *Advances in Neural Information Processing Systems*, 5:7–8, 1993.