Modélisation de la faillite d'exploitations agricoles : comparaison de méthodes de régression non linéaire avec R

Thibault Laurent

Modélisation de la faillite d'exploitations agricoles : comparaison de méthodes de régression non linéaire avec R

Thibault Laurent

Toulouse School of Economics (GREMAQ)

Résumé

L'idée de cet article est d'essayer des méthodes de régression non paramétriques pour le "scoring" et de vérifier leur efficacité en terme de taux de mal classés. Les méthodes approchées sont le modèle additif généralisé introduit par Hastie et Tibshirani (1986), le bagging et les forêts aléatoires proposés par Breiman (1996, 2001). On essaira également les Support Vecteur Machines proposés par Vapnik (1999). Nous poursuivons les travaux de Desbois (2008) qui a modélisé les probabilités de faillite de 1260 exploitations agricoles issus de quatre départements français (l'Eure, le Nord, l'Orne et la Seine-Maritime) à partir de caractéristiques de celles-ci observées entre 1988 et 1994. Les résultats de ces méthodes en terme de mal classés sont assez proches, néanmoins le choix des "meilleurs" variables explicatives peut varier d'une méthode à une autre.

Mots clés : étude de cas, méthodes non paramétriques, modèle addititif généralisé, vecteur support machine, bagging, boostrap.

Abstract

The idea of this article is to try methods of nonparametric regression for "scoring" and to verify their effectiveness in terms of rate of misfiled. The approximate methods are the generalized additive model introduced by Hastie and Tibshirani (1986), bagging and random forests proposed by Breiman (1996, 2001). It also will try the Support Vector Machines proposed by Vapnik (1999). We continue the work of Desbois (2008) who modeled the probability of bankruptcy of 1260 farms from four French departments (Eure, Nord, Orne and Seine-Maritime) from characteristics of these observed between 1988 and 1994. The results of these methods in terms of misfiled are fairly close, but the choice of "best"variables can vary from one method to another.

Key Words : non parametric method, generalized additive model, support vector machine, bagging, boostrap.

Un des points forts du logiciel R est que ses utilisateurs peuvent mettre leurs codes à contribution de tous, sous forme de paquetages. Ainsi, on trouve la plupart des outils d'analyse statistique, aussi bien récents que classiques, appliqués à la finance, à l'écologie ou à la génétique, téléchargeables sur le site du CRAN[1]. On propose ici de comparer plusieurs méthodes de régression non paramétriques disponibles sous R, lorsque la variable à expliquer est qualitative à deux modalités (échec et réussite) et les variables explicatives, quantitatives ou qualitatives.

Nous poursuivons ici les travaux de Desbois (2008) qui a modélisé les probabilités de faillite de 1260 exploitations agricoles issus de quatre départements français (l'Eure, le Nord, l'Orne et la Seine-Maritime) à partir de caractéristiques de celles-ci observées entre 1988 et 1994. L'auteur propose de comparer les résultats en terme de taux de mal classés, de l'analyse discriminante décisionnelle d'une part et la régression logistique d'autre part. L'auteur montre ainsi que les deux méthodes fournissent des résultats relativement semblables en terme de taux de mal classés.

L'idée de cet article est d'essayer d'autres méthodes de régression non paramétriques pour le "scoring" et de vérifier leur efficacité également en terme de taux de mal classés. Les méthodes approchées sont le modèle additif généralisé introduit par Hastie et Tibshirani (1986), le bagging et les forêts aléatoires proposés par Breiman (1996, 2001). On essaira également les *Support Vecteur Machines* proposés par Vapnik (1999). L'échantillon de départ sera divisé en deux : un échantillon d'apprentissage sur lequel les modèles seront calculés et un échantillon test sur lequel nous effectuerons les prédictions. Par ailleurs, nous prédirons qu'une exploitation agricole tombe en faillite lorsque $Pr[\hat{y}_i = 1] > c$, où c sera choisi de façon à minimiser le taux de mal classé. Nous procédérons à 100 réplications pour chaque méthode, pour vérifier la robustesse des résultats obtenus.

Dans un premier temps, nous effectuerons une comparaison directe de ces méthodes, à savoir que nous utiliserons les mêmes régresseurs que ceux de Desbois (2008) pour chaque méthode. Dans un second temps, on effectuera un choix de variables en fonction de la méthode considérée et nous comparerons ainsi les meilleurs modèles de chaque méthode.

[1]The Comprehensive RArchive Network : `http://cran.r-project.org/`

Bibliographie

[1] Besse, P. (2008) *Apprentissage Statistique et Data Mining*, `http://www.math.univ-toulouse.fr/~besse/enseignement.html`.
[2] Breiman, L. (1996) Bagging predictors, *Machine Learning*, 26(2) :123-140.
[3] Breiman, L. (2001) Random Forests, *Machine Learning*, 45 :5-32.
[4] Desbois, D. (2008) Introduction to Scoring Methods : Financial Problems of Farm Holdings, *CS-BIGS*, 2(1) : 56-76.
[5] Hastie, T. et Tibshirani, R. (1986) Generalized Additive Models, *Statistical Science 1*, 297-318.
[6] Saporta, G. (2006) *Probabilités, Analyse des données et Statistique*, Technip, deuxième édition.
[7] Vapnik, V.N. (1999) *Statistical learning theory*, Wiley Inter science.
[8] Wood, S. (2006) *Generalized Additive Models : An Introduction with R.* , Chapman & Hall/CRC.