Echantillonnage adaptatif de jeux de données déséquilibrés pour les forêts aléatoires

Julien Thomas*,**, Pierre-Emmanuel Jouve*, Elie Prudhomme**

*Société Fenics Lyon, France.
**Université Lumière Lyon 2, Laboratoire ERIC.

Résumé. Dans nombre d'applications, les données présentent un déséquilibre entre les classes. La prédiction est alors souvent détériorée pour la classe minoritaire. Pour contourner cela, nous proposons un échantillonnage guidé, lors des itérations successives d'une forêt aléatoire, par les besoins de l'utilisateur.

Introduction Les jeux de données déséquilibrés constituent un problème important de l'apprentissage supervisé. Or la plupart des modèles sont conçus pour des données équilibrées. Leur utilisation sur des données déséquilibrées conduit souvent à une mauvaise prédiction de la classe minoritaire. Pourtant, cette situation se retrouve régulièrement dans la pratique (Détection de pannes (Pazzani et al., 1994), textmining, aide aux diagnostics médicaux...). Ces applications ont besoin de disposer de méthodes capables de prédire la classe minoritaire avec des performances en adéquation avec les attentes de l'utilisateur. L'évantail des solutions existantes vont de l'échantillonnage (Japkowicz, 2000; Chawla et al., 2002), à la construction d'un modèle de prédiction spécifique à la classe d'intérêt, en passant par l'utilisation de matrices de coût (Pazzani et al., 1994; Kubat et al., 1998).

FUNSS L'idée de FUNSS (*Fitting User Needs Sampling Strategy*) est de traduire le besoin en rappel pour la classe minoritaire en terme de marge de décision entre les individus de chaque classe. Les individus minoritaires (positifs) sont entourés par une quantité importante d'individus majoritaires (négatifs) qui empèchent le classifieur de les apprendre correctemment. Pour augmenter le rappel, une solution consiste à choisir des individus négatifs éloignés des individus positifs. A l'inverse, pour augmenter la précision, il suffit de garder les individus négatifs proches des individus positifs. FUNSS reprend ce principe en modifiant l'échantillonnage réalisé au cours des forêts aléatoires en un échantillonnage dirigé. A chaque tirage avec remise, le processus est le suivant : si l'individu est positif, il est intégré dans le nouvel échantillon ; sinon un groupe de n individus négatifs est tiré ainsi qu'un individu positif. L'individu négatif du groupe qui est soit le plus proche, soit le plus éloigné de l'individu positif est intégré dans le nouvel échantillon. Chaque échantillon de la forêt aléatoire est donc l'occasion d'augmenter ou de diminuer le rappel pour atteindre une valeur fixée par l'utilisateur. Pour cela, le rappel de la forêt est estimé à chaque nouvel arbre à l'aide des individus out-of-bag. S'il est en dessous du rappel désiré, l'échantillonnage suivant sélectionne des individus négatifs éloignés. Dans le cas contraire, les individus négatifs proches sont favorisés. Enfin pour déterminer l'individu le plus proche d'une cible, les individus sont ordonnés pour chaque attribut sur leur proximité à cette cible. La distance utilisée est alors la somme des rangs d'un individu.

Algorithme	R classe positive	R classe négative	P classe positive	Correction globale	Temps (s)	Indice de temps
Forêt aléatoire	52,1	98,9	83,2	94,3	828	1
BRF	71,6	99,5	68,6	94,0	1782	2,2
SMOTE900	78,4	93,8	57,6	92,3	6528	7,9
FUNSS70	70,1	96,3	67,2	93,8	2077	2,5
FUNSS80	79,6	92,1	52,0	90,9	2406	2,9
FUNSS90	88,0	85,5	39,6	85,7	2251	2,7

TAB. 1 – *Résultats de 5-CrossValidation sur le jeu Satimage.(R : Rappel ; P : Précision)*

Expérimentations Des résultats obtenus sur le jeu de données Satimage (6435 individus ; 36 variables ; 9.73% d'individus positifs) sont présentés en table 1. La version de SMOTE testée ne comporte pas de sous-échantillonnage de la classe négative. L'algorithme BRF (Balanced Random Forest) construit une forêt aléatoire à partir de bootstraps équilibrés (Chen et Liaw, 2004). La notation FUNSS*XX* signifie que l'algorithme FUNSS a été paramétré avec une barre de rappel pour la classe positive à atteindre de *XX*%.
FUNSS montre la possibilité d'apprendre de manière correcte une modalité minoritaire sans modifier la distribution des effectifs. Ses temps de calcul sont sensiblement inférieurs à SMOTE. Le rappel de la classe positive obtenu est en adéquation avec le souhait de l'utilisateur.

Références

Chawla, N., K. Bowyer, L. Hall, et P. Kegelmeyer (2002). Smote : Synthetic minority oversampling technique. *Journal of Artificial Intelligence and Research 16*, 321–357.

Chen, C. et A. Liaw (2004). Using random forest to learn imbalanced data. *Technical Report.*.

Japkowicz, N. (2000). The class imbalance problem : Significance and strategies. In *Proceedings of IC-AI'2000*, Volume 1, pp. 111–117.

Kubat, M., R. C. Holte, et S. Matwin (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning 30*(2-3), 195–215.

Pazzani, M., C. Merz, P. Murphy, K. Ali, T. Hume, et C. Brunk (1994). Reducing misclassification costs. In *Proc. of the 11th ICML*, pp. 217–225. Morgan Kaufmann.

Summary

The class imbalance problem may occur in several cases when learning algorithms are applied to real data. Prediction is therefore degraded especially for the minority class. To cope with this issue, the proposed approach consists in using an adaptative sampling scheme through the successive steps of a Random Forest, to ensure that user needs are satisfied.