Prédiction d'évènements rares : une revue de la littérature

Mathieu Berthe [1] & Pierre Druilhet [2] & Stéphanie Léger [3] & Olivier Brachet [4]

[1,2,3] *Université Clermont Auvergne*
Laboratoire de Mathématiques Blaise Pascal UMR 6620 - CNRS
Campus des Cézeaux
3, Place Vasarely
TSA 60026 CS 60026
63178 Aubière Cedex
[1] *Mathieu.Berthe@math.univ-bpclermont.fr*
[2] *Pierre.Druilhet@math.univ-bpclermont.fr*
[3] *Stephanie.Leger@math.univ-bpclermont.fr*
[4] *Olivier@playsharp.pro*

Résumé. La modélisation et la prédiction d'évènements rares ou peu fréquents est un sujet délicat. En effet, les techniques de modélisation habituelles pour données binaires sont souvent peu efficaces. Pour pallier ces problèmes, de nombreuses méthodes ont été développées, que ce soit des méthodes dites de sampling (Oversampling, undersampling, SMOTE...), des améliorations et corrections de biais sur des méthodes existantes (régression logistique pour événements rares) ou encore des méthodes dites d'agrégation (Bagging, boosting). Nous proposons de présenter, grâce à une revue de la littérature, les avantages et les inconvénients de chaque méthode. À l'aide de simulations, nous comparerons les méthodes afin de mettre en évidence les meilleures d'entre elles. Ceci nous permettra également de proposer des outils de comparaison plus adaptés aux jeux de données déséquilibrées, par rapport à ceux couramment utilisés.

Mots-clés. Évènements rares; Donnés déséquilibrées ; Boosting ; Bagging; SVM; Régression logistique, Arbre de classification;

Abstract. Predicting unusual events is one of today's challenges in data science. Imbalanced datasets degrade the performance of common modeling techniques and make the tools used to compare different models inefficient and biased. In order to overcome such issues, many methods have been developed whether it be sampling methods (Oversampling, undersampling, SMOTE), upgrading and correcting bias to adjust existing methods (regression logistic for unusual events), or aggregating methods (Bagging, Boosting). Advantages and drawbacks of each method will be assessed thanks to a literature review. Every method will be compared through simulations in order to highlight the best ones. This will also allow more suitable comparison tools to be suggested in comparison to those commonly used.

Keywords. Rare event; Sparse data; Boosting; Bagging; SVM; Logistic regression; Classification tree

1 Introduction

On appellera évènement rare tout évènement se produisant avec une probabilité p petite (voire très petite). On peut rencontrer ce type de situation dans le milieu médical notamment dans l'apparition de maladies rares, dans le milieu météorologique avec la survenue de catastrophes naturelles, ou bien dans des domaines plus communs tel que le sport avec l'apparition de blessures. Un évènement est considéré comme rare ou peu fréquent lorsque $p \leq 0.05$. Les méthodes habituelles, comme la régression logistique, les arbres de classification ou les SVM, ont des difficultés à modéliser correctement les jeux de données déséquilibrées. De nombreuses améliorations ont été apportées à ces modèles afin d'améliorer la qualité prédictive [1]. D'autres types de méthodes consistent à modifier de façon aléatoire ou non le jeu de données, le plus souvent en rajoutant ou en supprimant des individus. Enfin, des méthodes dites d'agrégation montrent une amélioration notable de la prédiction des évènements rares. L'agrégation consiste à construire plusieurs "classifieurs" et à les agréger en un seul. Ces trois groupes de techniques peuvent être utilisées conjointement pour obtenir des résultats intéressants, ce qui engendre un nombre de méthodes possibles très important. Nous présentons ici les principales méthodes et les confrontons entre elles, afin d'en mesurer les avantages et les inconvénients.

2 Echantillonnage des données

Les méthodes d'échantillonnage consistent à créer, à partir d'un jeu de données Z déséquilibrées de taille N, un jeu de données $\tilde{Z}$ modifiées plus équilibrées. Les deux méthodes les plus simples utilisées sont "l'undersampling" [2], qui consiste à supprimer aléatoirement des individus de la classe majoritaire pour rééquilibrer l'échantillon (on a alors $N_Z > N_{\tilde{Z}}$), et "l'oversampling" [2] qui consiste à dupliquer de manière aléatoire des individus de la classe minoritaire pour rééquilibrer l'échantillon (on a alors $N_Z < N_{\tilde{Z}}$). De nombreuses autres méthodes plus complexes sont également disponibles. Parmi les plus abouties, SMOTE [3] est une méthode d'oversampling qui consiste à créer selon certains critères un individu synthétique de la classe minoritaire (c'est-à-dire qui n'existe pas réellement dans la base de données), qui se situe entre chaque individu de la classe minoritaire et ses k-plus proches voisins de la classe majoritaire.

3 Modèle

Les modèles habituellement utilisés pour la prédiction de variables binaires sont pour la plupart biaisés lorsque les données sont déséquilibrées. Cependant, il existe des corrections, comme pour la régression logistique[1], notamment sur la constante de la régression logistique β_0 lorsqu'elle associée à des méthodes d'échantillonnages. Notons τ la fraction d'individus présentant l'événement dans la population, et $\bar{y}$ la fraction observée dans

l'échantillon. On peut alors calculer $\tilde{\beta}_0$ non biaisé:

$$\tilde{\beta}_0 = \hat{\beta}_0 - \log\left[\left\{\frac{1-\tau}{\tau}\right\}\left\{\frac{\bar{y}}{1-\bar{y}}\right\}\right]. \quad (1)$$

Il existe également plusieurs techniques de corrections dans le cas des Support Vector Machine (SVM) [4][5]. On peut par exemple citer les corrections apportées à la constante b de l'hyperplan séparateur [6]. Soit $Z = \{(x_1, y_1), ..., (x_N, y_N)\}$ la base d'apprentissage, avec $x_i \in X$ et $y_i \in Y = \{-1, 1\}$ et soit Z_1 les données de la classe positive (+) et Z_2 les données de la classe négative (-), alors une correction $\tilde{b}$ de la constante peut être donnée par :

$$\tilde{b} = b - \frac{\alpha + \beta}{2}. \quad (2)$$

Avec: α la valeur maximale de l'hyperplan appliquée aux données négatives Z_2 et β la valeur minimale de l'hyperplan appliquée aux données positives Z_1.

D'autres corrections existent et permettent d'améliorer la prédiction et la classification.

4 Méthode d'Agrégation

Les méthodes d'agrégation consistent à construire plusieurs "classifieurs" à partir de modèles choisis arbitrairement (arbre de classification/régression, SVM, régression logistique, etc.), puis à les regrouper par vote ou moyenne, pour obtenir la prédiction ou la classification. Prenons la méthode de Bagging (Bootstrap aggregating)[7]; elle consiste à construire m modèles à partir de m échantillons obtenus par bootsrap et à les regrouper par vote pour la classification et par moyenne pour la régression. La seconde méthode que l'on peut citer est le boosting [8], et plus particulièrement ADABOOST [8] qui repose sur la construction itérative de "classifieurs". A chaque itération, les exemples mal classés par le "classifieur" courant sont pondérés et leurs importances augmente pour la construction du "classifieur" suivant. Finalement, les "classifieurs" sont regroupés et pondérés en fonction de leur taux de bonne classification, pour obtenir une prédiction finale. Les méthodes d'agrégation sont très réputées pour augmenter la performance des prédictions lorsque les données sont rares.

5 Outils de mesure

Les outils de mesure utilisés habituellement (sensibilité, spécificité et taux de bien classés) pour comparer l'efficacité des modèles, sont moins performants lorsque les données sont déséquilibrées ($p \leq 0.05$). En effet, prenons le cas d'un échantillon déséquilibré de taille $N = 1000$, de sorte que la taille de classe majoritaire est $N_+ = 950$ et la taille de

la classe minoritaire est $N_- = 50$. Alors, un modèle dont le taux de bien classés= 0.95 (considéré comme un bon modèle pour des données déséquilibrées) et dont la sensibilité=0 et la spécificité=1, est en réalité un mauvais modèle. Cependant, il existe des indices permettant de comparer efficacement des modèles. Comme les courbes roc qui permettent d'évaluer visuellement un modèles et numériquement à l'aide d'indices connus tel que l'AUC (Area Under Curve) et d'autres moins utilisés, en particulier l'indice de Peirce[9][10].

6 Application

Les méthodes seront également utilisées sur des données réelles de blessure sans contact chez les footballeurs professionnels. La blessure sans contact peut être considérée comme évènement rare puisque sa probabilité d'apparition est inférieure à 0.05. Sur ces données réelles nous présenterons les méthodes permettant de donner la meilleure probabilité qu'un joueur se blesse durant un match professionnel. La figure 1 présente les courbe ROC de la méthode régression logistique (AUC=0.642), régression logistique utilisée avec oversampling (AUC=0.671) et la régression logistique utilisé avec BAGGING (AUC=0.787).

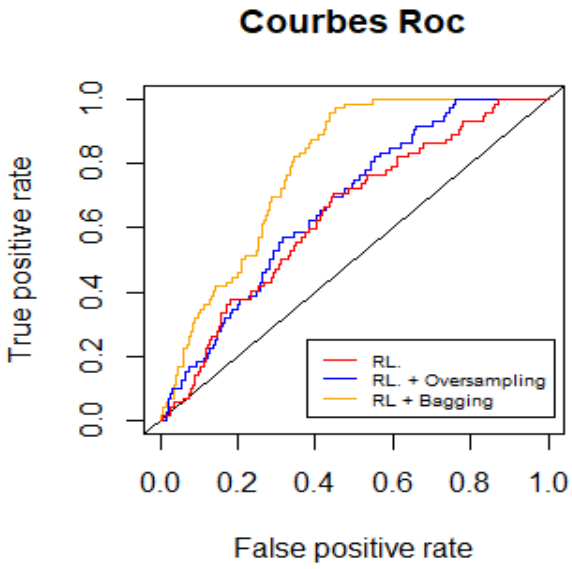

Figure 1: Courbes ROC des méthodes de prédiction de blessure

7 Conclusion

Les évènements rares sont présents dans différents milieux tel que la médecine, le sport ou encore la géologie. Il existe une multitude de méthodes pouvant être utilisées que ce soit par le choix du modèle (régression logistique, arbre, svm), par les méthodes d'échantillonnage (Undersampling, Oversampling, SMOTE), ou encore par des méthodes d'agrégation (Baggin, Boosting). Dans cette perspective, il semble intéressant d'évaluer les meilleures stratégies pour choisir les combinaisons les plus optimales et de mettre en avant leurs défauts et leurs qualités à l'aide des outils de comparaison adaptés. Il convient également d'évaluer les différents outils de comparaison de modèle et d'évaluer leur comportement selon différentes simulations et d'étudier leur efficacité sur des données rélles.

Bibliographie

[1] King, Gary et Zeng, L. (2001). Logistic Regression in Rare Event Data, *Political Analysis*, 9, pp. 137-163.
[2] Drummond, C. et Holte, R.C. , D. J. (2003). Class Imbalance , and Cost Sensitivity : Why Under-Sampling beats OverSampling.
[3] Chawla, N.V. et Bowyer, K.C. et Hall, L.O. et Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, pp. 321-357.
[4] Cortes, C. et Vapnik, V. (1995). Support vector networks. *Machine Learning,*, 20, pp. 1-25.
[5] Platt, J.C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in large margin classifiers*, pp. 61-74.
[6] Haydemar, N. et Gonzalez-Abril, L. et Angulo, C. (2017). Improving SVM Classification on Imbalanced Datasets by Introducing a New Bias. *Journal of Classification.*
[7] Breiman, L. (1996). Bagging predictors. *Machine Learning*,24, pp. 123-140.
[8] Freund, Y. et Schapire, R.E. (1999). A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*,14, pp. 771-780.
[9] Peirce, C.S. (1884). The numerical measure of the success of predictions. *Science*,4, pp. 453-454.
[10] Mason, I.B. (2003). *Binary events. Forecast Verification: A Practitioners Guide in Atmospheric Science*, pp. 37-76.