Actualisation en ligne d'un score d'ensemble

Benoît Lalloué, Jean-Marie Monnez, Eliane Albuisson

Actualisation en ligne d'un score d'ensemble

Benoît Lalloué [1,3,*], Jean-Marie Monnez [1,3,†], Éliane Albuisson [2,4,5,‡]

[1] *Université de Lorraine, CNRS, Inria*, IECL**, F-54000 Nancy, France*
**Inria, Project-Team BIGS*
***Institut Elie Cartan de Lorraine, Vandoeuvre-lès-Nancy, France*
[2] *Université de Lorraine, CNRS, IECL**, F-54000 Nancy, France*
[3] *Inserm U1116, Centre d'Investigation Clinique Plurithématique 1433, Université de Lorraine, Nancy, France*
[4] *BIOBASE, Pôle S2R, CHRU de Nancy, Vandoeuvre-lès-Nancy, France*
[5] *Faculté de Médecine, InSciDenS, Vandoeuvre-lès-Nancy, France*
∗ *benoit.lalloue@univ-lorraine.fr;* † *jean-marie.monnez@univ-lorraine.fr;*
‡ *eliane.albuisson@univ-lorraine.fr*
Financement : Programme Investissement d'Avenir ANR-15-RHU-0004

Résumé. En construisant une collection de prédicteurs (en faisant varier les échantillons utilisés, les variables retenues, les règles d'apprentissage, ...) dont les prédictions sont ensuite agrégées, les méthodes d'ensemble permettent d'obtenir de meilleurs résultats que les prédicteurs individuels. Dans un contexte en ligne où des données arrivent de façon continue, on souhaite actualiser les paramètres d'un score construit à l'aide d'une méthode d'ensemble. On considère le cas où il est impossible de conserver toutes les données obtenues précédemment et de recalculer les paramètres sur l'ensemble des données à chaque nouvelle observation. Nous proposons une méthode d'actualisation en ligne d'un score d'ensemble à l'aide de bootstrap Poisson et d'algorithmes stochastiques.

Mots-clés. Algorithmes stochastiques, apprentissage pour les données massives, médecine, méthode d'ensemble, score en ligne.

Abstract. By constructing a collection of predictors (by varying samples, selection of variables, learning rules, etc.) whose predictions are then aggregated, ensemble methods obtain better results than individual predictors. In an online setting, where data arrives continuously, we want to update the parameters of a score constructed with an ensemble method. We consider the case where it is impossible to keep all the data obtained previously and to compute again the parameters on all the data at each new observation. We propose a method for updating an ensemble score online using Poisson bootstrap and stochastic algorithms.

Keywords. Stochastic algorithms, learning for big data, medicine, ensemble method, online score.

1 Introduction

Considérons le problème de prédiction des valeurs d'une variable dépendante y continue (dans le cas de la régression) ou catégorielle (dans le cas de la classification) à partir de variables observées $x^1, ..., x^p$ continues ou catégorielles.

De nombreux prédicteurs différents peuvent être construits pour répondre à ce problème. Le principe des méthodes d'ensemble est de construire une collection de N prédicteurs "de base" (à l'aide de méthodes classiques) dont les prédictions sont ensuite agrégées par moyenne ou par vote. On s'attend à ce que le prédicteur d'ensemble soit meilleur que chacun des prédicteurs de base, si toutefois ces prédicteurs de base sont relativement bons et s'ils sont suffisamment différents les uns des autres (Genuer et Poggi 2017).

L'ensemble de prédicteurs peut être construit suivant différentes variantes, utilisées séparément ou associées :

- avec différentes types de régressions ou différentes règles de classification
- avec différents échantillons (obtenus par bootstrap, par exemple)
- avec différentes méthodes de sélection de variables (aléatoire, stepwise, pénalisée, ...)
- plus généralement, en introduisant un aléa dans la construction des prédicteurs

Le *bagging* (Breiman 1996), le *boosting* (Freund et Schapire 1996), les forêts d'arbres aléatoires (Genuer et Poggi 2017) ou les modèles linéaires généralisés aléatoires (*Random Generalized Linear Model*, RGLM)(Song *et al.* 2013) sont des exemples de méthodes d'ensemble. Duarte, Monnez et Albuisson (2018a) ont proposé une méthode d'ensemble pour la construction d'un score combinant ces différentes variantes en sept étapes :

1. Sélection de n_1 règles de classifications.
2. Pour chaque règle, génération de n_2 échantillons bootstrap. Les n_2 échantillons bootstrap sont les mêmes pour les n_2 règles.
3. Choix de n_3 modalités de sélection aléatoire de variables, identiques pour chaque échantillon bootstrap.
4. Sélection de m^* variables par une méthode de sélection (stepwise, pénalisation, ...).
5. Pour chaque règle de classification, construction des $n_2 \times n_3$ prédicteurs selon les modalités précédentes.
6. Agrégation des prédicteurs en un score synthétique pour chaque règle de classification.
7. Agrégation des scores synthétiques construits à l'étape précédente.

Dans un contexte de données en ligne, c'est à dire d'un flux de données arrivant de façon continue, on souhaite pouvoir actualiser un tel score d'ensemble sans avoir à stocker ou à réutiliser l'ensemble des données obtenues lorsque de nouvelles données sont disponibles.

Nous présentons dans la suite une méthode d'actualisation en ligne de ce score d'ensemble.

2 Actualisation en ligne d'un score d'ensemble

Afin de pouvoir actualiser en ligne le score d'ensemble, il faut pouvoir actualiser chaque échantillon bootstrap et chaque prédicteur lors de l'arrivée de nouvelles données d'apprentissage.

À partir d'un échantillon de taille n, la construction classique d'un échantillon bootstrap consiste à effectuer n tirages aléatoires *avec remise*. Dans le cas d'un flux de données, le bootstrap Poisson proposé par Oza et Russel (2001) peut être utilisé pour actualiser un échantillon bootstrap : pour toute nouvelle donnée d'apprentissage, pour chaque échantillon bootstrap b_i, on simule une réalisation k_i d'une variable aléatoire de loi de Poisson de paramètre 1 et on ajoute k_i fois la nouvelle donnée à l'échantillon b_i. Ces nouvelles données peuvent alors être utilisées pour actualiser le prédicteur défini à partir de l'échantillon b_i.

Cette actualisation peut être effectuée à l'aide d'algorithmes stochastiques récursifs prenant en compte un lot de nouvelles données à chaque étape. De tels algorithmes ont été développés pour estimer des paramètres de régression linéaire (Duarte *et al.* 2018b) ou non linéaire, ou encore pour estimer des centres de classes en classification non supervisée (Cardot *et al.* 2012) ou des composantes principales d'une analyse factorielle (Monnez et Skiredj, 2018). Ils ne nécessitent pas de stocker les données et peuvent, dans un temps fixé, traiter davantage de données que les méthodes classiques.

Une fois les prédicteurs mis à jour, les scores composites par règle de classification puis le score d'ensemble final sont obtenus en utilisant les mêmes règles d'agrégation que pour la méthode d'ensemble classique.

3 Actualisation d'un score d'ensemble basé sur l'analyse discriminante linéaire et la régression logistique

Duarte *et al.* (2018a) ont appliqué la méthode d'ensemble exposée précédemment à la construction d'un score de risque à court terme de décès ou d'hospitalisation pour les patients souffrant d'insuffisance cardiaque avec les paramètres suivants :

- $n_1 = 2$ règles de classification : l'analyse discriminante linéaire et la régression logistique.
- $n_2 = 1000$ échantillons bootstrap pour chaque règle.
- $n_3 = 3$ modalités de sélection aléatoire ont été retenues : tirage au sort d'un nombre fixé de variables ou tirage au sort d'un nombre fixé de groupes prédéfinis de variables corrélées puis tirage au sort d'une variable par groupe retenu (avec deux définitions des groupes).
- L'étape de sélection de variables stepwise ou pénalisée n'a pas été retenue car elle n'améliorait pas la précision des prédictions.

- L'agrégation des scores obtenus par les prédicteurs de base a été réalisée en prenant leur moyenne arithmétique.
- L'agrégation entre les deux scores synthétiques S_1 et S_2 a été faite par combinaison convexe : $\lambda S_1 + (1-\lambda)S_2$ avec $0 \leq \lambda \leq 1$.

Il est possible d'actualiser ce score en ligne en utilisant des processus de gradient stochastique adaptés à l'estimation des paramètres de l'analyse discriminante linéaire et de la régression logistique.

3.1 Actualisation de l'analyse discriminante linéaire

On peut noter que l'analyse discriminante linéaire est équivalente à une régression linéaire avec une variable dépendante binaire.

Soit $R(p,1)$ et $S(q,1)$ deux vecteurs aléatoires. Supposons qu'un lot de m_n nouvelles données (R_i, S_i) constituant un échantillon i.i.d de (R,S) arrive à l'étape n. On note $M_n = \sum_{i=1}^{n} m_i$ et $I_n = \{M_{n-1}+1, ..., M_n\}$

On se place dans le cadre de la régression linéaire multidimensionnelle. On cherche à déterminer $\theta(p,q)$ et $\eta(q,1)$ qui minimisent $\mathbb{E}\left[\|S - \theta' R - \eta\|^2\right]$. Pour éviter des explosions numériques, on propose dans Duarte *et al.* (2018b) d'utiliser des données centrées-réduites. Soit $\mathbb{E}[R]$ (respectivement $\mathbb{E}[S]$) le vecteur espérance-mathématique de R (respectivement S), Γ (respectivement Γ^1) la matrice diagonale des inverses des écarts-types des composantes de R (resp. S). On note $R_1 = \Gamma(R - \mathbb{E}[R])$ et $S_1 = \Gamma^1(S - \mathbb{E}[S])$ les vecteurs standardisés. On cherche alors à déterminer $\theta_1 = \Gamma^{-1}\theta\Gamma^1$ tel que $\mathbb{E}\left[\|S_1 - \theta_1' R_1\|^2\right]$ est minimale. θ_1 est solution de :

$$\nabla_{\theta_1} \mathbb{E}\left[\|S_1 - \theta_1' R_1\|^2\right] = 0 \Leftrightarrow \mathbb{E}\left[R_1 R_1'\right]\theta_1 = \mathbb{E}\left[R_1 S_1'\right]$$

On peut donc utiliser un processus de gradient stochastique pour estimer en ligne θ_1.

Soit à résoudre de façon générale un système $B\theta = F$. On définit de façon récursive le processus de gradient stochastique classique (SGD) (X_n) convergeant vers θ par :

$$X_{n+1} = X_n - a_n(B_n X_n - F_n) \text{ avec :}$$

$$\mathbb{E}[B_n|T_n] = B, \mathbb{E}[F_n|T_n] = F (T_n \text{ tribu du passé au temps } n), a_n > 0, \sum_{n=1}^{\infty} a_n = \infty, \sum_{n=1}^{\infty} a_n^2 < \infty$$

Toutefois, un mauvais choix du pas d'apprentissage a_n ou la présence de données extrêmes peuvent conduire à des problèmes d'explosion numérique lors des applications pratiques.

Dans le cas d'un flux de données, on ne connaît pas a priori les moments de R et de S. On peut les estimer en ligne pour effectuer la standardisation. On ne dispose alors plus d'un échantillon i.i.d de (R_1, S_1) et la convergence du gradient stochastique n'est pas assurée par les théorèmes usuels.

Posons :

$$B_n = \Gamma_{M_n}\left(\frac{1}{M_n}\sum_{i=1}^{n}\sum_{j\in I_i} R_j R_j' - \bar{R}_{M_n}\bar{R}'_{M_n}\right)\Gamma_{M_n}$$

$$F_n = \Gamma_{M_n}\left(\frac{1}{M_n}\sum_{i=1}^{n}\sum_{j\in I_i} R_j S_j' - \bar{R}_{M_n}\bar{S}'_{M_n}\right)\Gamma^1_{M_n}$$

avec $\bar{R}_{M_n} = \frac{1}{M_n}\sum_{i=1}^{M_n} R_i$, $\bar{S}_{M_n} = \frac{1}{M_n}\sum_{i=1}^{M_n} S_i$ et Γ_n (respectivement Γ^1_n) la matrice diagonale des inverses des estimations des écarts-types des composantes de R (resp. S) calculée récursivement à partir des données $(R_i, S_i), i \leq n$.

La convergence du processus utilisant les matrices B_n et F_n définies ci-dessus et de deux autres processus avec standardisation des données en ligne est établie dans Duarte *et al.* (2018b). Ces processus sont comparés à d'autres (avec ou sans standardisation en ligne, avec ou sans moyennisation, avec ou sans prise en compte de l'ensemble des données jusqu'à l'observation courante) sur des données réelles ou simulées. Les meilleurs résultats ont été obtenus avec le processus présenté ici.

Il est donc possible d'utiliser ce processus pour actualiser les prédicteurs par analyse discriminante linéaire dans le score d'ensemble en utilisant pour chaque prédicteur à chaque étape l'échantillon de nouvelles données généré par le bootstrap Poisson.

3.2 Actualisation de la régression logistique binaire

Plaçons nous maintenant dans le cas de la régression logistique binaire. Soit S une variable aléatoire prenant ses valeurs dans $\{0, 1\}$ et $R = (R^1, ...R^p, 1)'$ avec $R^1, ..., R^p$ des variables aléatoires réelles.

Notons R^c le vecteur R centré, σ^k l'écart-type de R^k, Γ la matrice carrée de diagonale $\frac{1}{\sigma^1}, ..., \frac{1}{\sigma^p}, 1$, $Z = \Gamma R^c$ le vecteur R standardisé (centré-réduit), $\theta(p+1, 1)$ le vecteur des paramètres et $h(u) = \frac{e^u}{1+e^u}$.

θ est solution du système d'équations $\mathbb{E}\left[\nabla_x \ln\left(\frac{e^{Z'xS}}{1+e^{Z'xS}}\right)\right] = 0$, et donc de

$$\mathbb{E}\left[Z(S - h(Z'x))\right] = 0$$

Soit $(R_n), n \geq 1$ (respectivement (S_n)) un échantillon i.i.d de R (resp. S). On note $\bar{R}^k_n$ la moyenne de l'échantillon $(R^k_1, ..., R^k_n)$ de R^k et $(V^k_n)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(R^k_i - \bar{R}^k_n\right)^2$ sa variance (toutes deux calculées récursivement), $\bar{R}_n$ le vecteur $(\bar{R}^1_n, ..., \bar{R}^p_n, 0)'$ et Γ_n la matrice de diagonale $\frac{1}{\sqrt{\frac{n}{n-1}}V^1_n}, ..., \frac{1}{\sqrt{\frac{n}{n-1}}V^p_n}, 1$.

On définit $\tilde{Z}_j = \Gamma_{M_{n-1}}\left(R_j - \bar{R}_{M_{n-1}}\right)(j \in I_n)$ le vecteur R_j standardisé à l'aide des moyennes et variances estimées à l'étape $n-1$ (M_n et I_n ont été définis dans le paragraphe précédent).

Définissons alors le processus (X_n) de manière récursive :

$$X_{n+1} = X_n - a_n \frac{1}{m_n} \sum_{j \in I_n} \tilde{Z}_j \left(h(\tilde{Z}_j' X_n) - S_j \right)$$

Monnez (2018) a montré la convergence de ce processus vers θ sous certaines conditions. On peut utiliser ce processus pour actualiser les prédicteurs par régression logistique.

4 Conclusion

En combinant le bootstrap Poisson à des processus de gradient stochastique adaptés avec données standardisées, il est possible d'effectuer une actualisation du score d'ensemble proposé par Duarte *et al.* (2018a) au fur et à mesure de l'arrivée de nouvelles données. Des résultats d'application seront présentés.

Bibliographie

Breiman, L. (1996). Bias, variance, and arcing classifiers. *Technical Report 460*, Departement of Statistics, University of California, Berkeley.

Cardot, H., Cénac, P. and Monnez, J.-M. (2012). A fast and recursive algorithm for clustering large datasets with k-medians. *Comput Stat Data Anal.* 56(6):1434-49.

Duarte, K., Monnez J.-M. and Albuisson E. (2018a). Methodology for Constructing a Short-Term Event Risk Score in Heart Failure Patients. *Appl Math*, 09(08):954-74.

Duarte, K., Monnez, J.M. and Albuisson, E. (2018b). Sequential linear regression with online standardized data, *PloS One*, 13 (1) e0191186.

Freund, Y. and Schapire, R.E. (1996). Experiments with a New Boosting Algorithm. In: *Proceedings of the 13th International Conference on Machine Learning.*

Genuer, R. and Poggi, J.-M. (2017). Arbres CART et Forêts aléatoires, Importance et sélection de variables, *hal-01387654.*

Monnez, J.-M. (2018). Online constrained binary logistic regression process with online standardized data, *Working paper.*

Monnez, J.-M. and Skiredj, A. (2018). Convergence of a normed eigenvector stochastic approximation process and application to online principal component analysis of a data stream, *hal-01844419.*

Oza, N.C. and Russell, S.J. (2001). Online Bagging and Boosting. In: *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, AISTATS 2001.*

Song, L., Langfelder, P. and Horvath, S. (2013). Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC Bioinformatics.* 14:5.