Prédiction séquence-à-séquence de séries temporelles multivariées et déséquilibrées avec des réseaux de neurones convolutifs

Mehdi Elion*, Sonia Tabti*, Julien Budynek*

*FieldBox.ai, Quai Armand Lalande, 33300 Bordeaux
melion@fieldbox.ai, stabti@fieldbox.ai, jbudynek@fieldbox.ai
https://www.fieldbox.ai/

Les équipements industriels sont fréquemment munis de capteurs mesurant différentes grandeurs physiques. Il en résulte un historique de séries temporelles à partir desquelles des modèles prédictifs sont construits. En particulier, la prédiction d'évènements rares (ex. pannes, anomalies) est un enjeu clé des opérations industrielles. Il existe plusieurs modèles de classification adaptés à ce type de données, dites déséquilibrées. C'est par exemple le cas des méthodes dites *cost-sensitive* où la fonction objectif pénalise davantage les erreurs relatives aux cas rares, ou encore des méthodes de rééchantillonnage des données d'entraînement. Cependant, moins de travaux traitent de régression déséquilibrée. Laptev et al. (2017) propose une modélisation d'évènements rares saisonniers peu adaptée aux procédés industriels. Torgo et al. (2013) ont proposé SMOTER (Synthetic Minority Oversampling TEchnique for Regression) qui est une adaptation à la régression de SMOTE (Chawla et al., 2002), un algorithme combinant sur-échantillonage de la classe minoritaire et sous-échantillonage de la classe majoritaire dans l'ensemble d'entraînement pour la classification. SMOTER a été adapté aux séries temporelles (Moniz et al., 2017) mais uniquement dans le cas univarié et pour prédire un seul pas dans le futur. Or, les opérations industrielles exigent souvent de prédire plusieurs pas dans le futur à partir de séries temporelles multivariées. Nous avons donc développé une nouvelle technique de rééchantillonnage pour séries temporelles déséquilibrées et multivariées, SMOTEST (Synthetic Minority Oversampling TEchnique for Sequence-To-sequence). Cette méthode étend SMOTER à ces problématiques et utilise un réseau de neurones convolutif unidimensionnel pour la prédiction. Plus précisément, notre but est de prédire à chaque instant un signal de sortie, composé des $w_{out} \in \mathbb{N}^*$ valeurs futures d'une variable cible y, à partir de signaux d'entrée composés des $w_{in} \in \mathbb{N}^*$ valeurs précédentes de y et de variables $(x_i)_{i\in[\![1,n-1]\!]}$.

Soit $\mathcal{D}$ l'ensemble des cas $(X, Y) \in \mathbb{R}^{w_{in}\times n} \times \mathbb{R}^{w_{out}}$ disponibles, où X et Y représentent respectivement les signaux d'entrée et de sortie. On définit une fonction $\Phi : \mathbb{R}^{w_{out}} \mapsto \mathbb{R}$ qui caractérise chaque séquence de sortie par une grandeur. Dans cette étude, Φ renvoie l'augmentation maximale dans la séquence de sortie, mais d'autres critères sont envisageables. Étant donné un seuil $\theta \in \mathbb{R}$ défini par l'utilisateur, l'ensemble des cas rares, dits *positifs*, est alors défini par $\mathcal{D}_P = \{(X, Y) \mid \Phi(Y) \geq \theta\}$. Le sur-échantillonnage s'effectue par itération sur les cas $(X, Y) \in \mathcal{D}_P$. Pour chaque cas, ses k plus proches voisins sont isolés. Ensuite, n_g cas synthétiques sont générés comme suit : un des voisins, noté (X_{nn}, Y_{nn}), est choisi aléatoirement,

puis chaque signal d'entrée du cas synthétique courant (X_{new}, Y_{new}) est calculé par combinaison linéaire aléatoire de ceux contenus dans X_{new} et X_{nn} ; Y_{new} est alors généré comme une combinaison linéaire de Y et Y_{nn} dont les coefficients dépendent des distances entre X_{new} et X et entre X_{new} et X_{nn}. Les valeurs de k et n_g sont choisies selon le rééquilibrage souhaité.

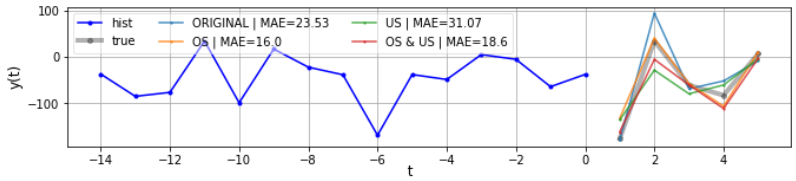

FIG. 1: Exemples de prédictions sur le jeu de données synthétique par les modèles avec différents modes de rééchantillonnage : *OS* (sur-échantillonnage), *US* (sous-échantillonnage) et *ORIGINAL* (aucun rééchantillonnage). La partie gauche (*hist*) représente l'historique de la variable cible utilisé en entrée du modèle, $w_{in} = 15$. La partie droite comprend la séquence de sortie (*true*) à prédire, $w_{out} = 5$, et les prédictions faites par les différents modèles.

Nous avons étudié sur un jeu de données synthétiques l'influence du mode de rééchantillonnage : sous-échantillonnage, sur-échantillonnage ou combinaison des deux, cf. FIG. 1. Les résultats suggèrent que le sur-échantillonnage seul présente les meilleures performances. Notons qu'il a été testé que le sur-échantillonage proposé surpasse une simple duplication de données. Enfin, une application de cette méthode sur un jeu de données industriel confidentiel de contrôle qualité a montré des résultats encourageants (cf. TAB 1), démontrant ainsi l'intérêt applicatif de cette méthode. En travaux futurs, nous prévoyons d'étudier plusieurs méthodes de sur-échantillonnage ainsi que d'autres types de modèles prédictifs.

	Cas positifs		Cas négatifs	
	MAE	RMSE	MAE	RMSE
Original	0.36 ± 0.37	0.46 ± 0.43	$\mathbf{0.16 \pm 0.21}$	$\mathbf{0.19 \pm 0.23}$
SMOTEST	$\mathbf{0.34 \pm 0.35}$	$\mathbf{0.42 \pm 0.40}$	0.19 ± 0.21	0.22 ± 0.23

TAB. 1: Résultats sur les données industrielles de test ($w_{in} = 90, w_{out} = 20$). Les valeurs à gauche et à droite du signe $\pm$ correspondent resp. à la moyenne et l'écart type de l'erreur.

Références

Laptev, N., J. Yosinski, L. E. Li, et S. Smyl (2017). Time-series Extreme Event Forecasting with Neural Networks at Uber. pp. 5.

Moniz, N., P. Branco, et L. Torgo (2017). Resampling strategies for imbalanced time series forecasting. *International Journal of Data Science and Analytics 3*(3), 161–181.

Torgo, L., R. P. Ribeiro, B. Pfahringer, et P. Branco (2013). SMOTE for Regression. In *EPIA*.