

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-échantillonnage de la classe minoritaire et sous-échantillonnage 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 $(x_i)_{i \in \llbracket 1, n-1 \rrbracket}$.

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,