

Vers un échantillonnage de flux de données transformé

Olivier Parisot, Thomas Tamisier

Luxembourg Institute of Science and Technology, Belvaux, Luxembourg
olivier.parisot@list.lu

De nombreuses techniques ont été mises au point récemment afin d'extraire des modèles prédictifs depuis des flux de données (Nguyen et al., 2015). Dans ce domaine, le calcul de l'exactitude des résultats est crucial pour évaluer la performance des modèles obtenus. Avec des flux potentiellement infinis, il faut donc maintenir un échantillonnage, ce dernier étant utilisé – à intervalle régulier ou à la demande – pour calculer le taux d'erreur courant sur un ensemble représentatif du flux (par ex., un mélange bien balancé entre éléments récents/anciens).

Dans ce papier, nous appliquons l'échantillonnage durant l'entraînement d'un modèle prédictif et nous le transformons afin de générer à la demande un arbre de décision simplifié montrant quand ce modèle se trompe. Pour se faire, notre méthode se décline de manière classique en une phase *online* et en une phase *offline* (Fig. 1).

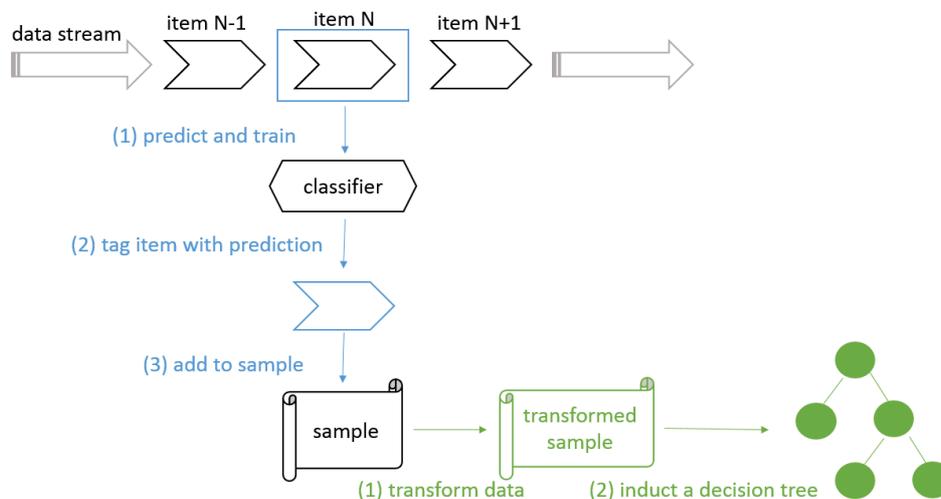


FIG. 1 – Approche pour créer à la demande un arbre de décision montrant les erreurs d'un modèle de prédiction entraîné sur un flux (phase online en bleue, phase offline en vert).

Durant la phase *online*, chaque élément du flux est analysé : il est évalué par le modèle prédictif en cours, tagué en fonction de la justesse de la prédiction ('*bien classé / mal classé*'), puis inséré dans un échantillonnage (de type réservoir, par exemple – Vitter (1985)).