Calibration des modèles d'apprentissage pour l'amélioration des détecteurs automatiques d'exemples mal-étiquetés

Ilies Chibane*, Thomas George*,
Pierre Nodet*, Vincent Lemaire*

*Orange Innovation
prénom.nom@orange.com,
https://hellofuture.orange.com

Résumé. Les données mal-étiquetées sont un problème répandu qui dégrade la performance des modèles d'apprentissage automatique en contexte industriel. Les méthodes qui permettent de détecter les exemples mal-étiquetés reposent la plupart du temps sur l'introspection d'un modèle d'apprentissage, qui est entraîné puis sondé pour chaque exemple afin d'obtenir un score de confiance indiquant si l'étiquette fournie est bonne ou mauvaise. Dans cet article, nous étudions la calibration de ce modèle sous-jacent. Nous montrons empiriquement que l'emploi de méthodes de calibration améliore la précision et la robustesse de la détection d'exemples mal-étiquetés, ce qui permet d'obtenir une solution pratique et efficace pour des applications industrielles.

1 Introduction

La qualité d'un modèle d'apprentissage automatique est conditionnée par la qualité des données sur lesquelles il est entraîné. Or, la collection d'exemples d'apprentissage et leur annotation peuvent se révéler coûteuse pour les jeux de données les plus volumineux. Obtenir des étiquettes de qualité reste donc un problème majeur : on constate que même des jeux de données publics populaires tels que MNIST ou CIFAR-10/CIFAR-100, présentent des problèmes d'étiquetage (Northcutt et al., 2021a). Pour éviter un examen manuel de chaque étiquette, des détecteurs d'exemples mal-étiquetés visent à identifier ces exemples de manière automatique (Frénay et Verleysen, 2013). Ces détecteurs fournissent un *score de confiance* pour chaque exemple du jeu données sur lequel ils sont appliqués, indiquant si l'étiquette fournie peut être considérée comme bonne ou mauvaise. Parmi ceux-ci, les détecteurs basés sur l'*introspection* examinent s'il existe une différence de traitement entre les exemples bien et mal-étiquetés lors de l'apprentissage, mesurée à l'aide de *sondes*. Dans cet article, nous explorons une piste d'amélioration de ces détecteurs : nous proposons d'ajouter une étape intermédiaire de *calibration* du modèle d'apprentissage avant de le sonder (figure 1 et section 2.1).

Travaux connexes La détection automatique d'exemples mal étiquetées a fait l'objet de nombreux travaux de recherche. Ces travaux ont permis notamment de mettre en lumière une des limites de ces détecteurs, notamment qu'ils aient du mal à distinguer les exemples difficiles