Découvrir de nouvelles classes dans des données tabulaires

Colin Troisemaine[*,**], Joachim Flocon-Cholet[*], Stéphane Gosselin[*], Sandrine Vaton[**]
Alexandre Reiffers-Masson[**], Vincent Lemaire[*]

[*] Orange Labs, Lannion, France
[**] Département Informatique, IMT Atlantique, Brest, France

Résumé. Dans le domaine du Novel Class Discovery (NCD), le but est de trouver de nouvelles classes dans un ensemble non étiqueté lorsqu'un ensemble étiqueté de classes connues mais différentes est disponible. Bien que le NCD ait récemment attiré l'attention de la communauté scientifique, aucune solution n'a encore été proposée pour les données tabulaires, alors qu'il s'agit d'une représentation très courante des données. Dans cet article, nous proposons une nouvelle méthode pour résoudre ce problème, dans le contexte de données tabulaires contenant des variables hétérogènes. Ce processus est en partie réalisé par une nouvelle méthode de définition de pseudo-étiquettes ainsi que par une mise en œuvre des découvertes récentes de l'apprentissage multi-tâches pour optimiser une fonction objectif conjointe. Notre méthode démontre que le NCD n'est pas seulement applicable aux images mais aussi aux données tabulaires hétérogènes.

1 Introduction

Le récent succès des modèles d'apprentissage automatique a été rendu possible en partie par l'utilisation de grandes quantités de données étiquetées. De nombreuses méthodes supposent actuellement qu'une grande partie des données disponibles est étiquetée et que toutes les classes sont connues. Cependant, ces hypothèses ne sont pas toujours vraies en pratique et les chercheurs commencent à envisager des scénarios dans lesquels des données non étiquetées sont disponibles (Nodet et al., 2021). On peut distinguer dans cet apprentissage dit faiblement supervisé les méthodes qui nécessitent de connaître toutes les classes à l'avance de celles qui sont capables de gérer des classes qui ne sont jamais apparues pendant l'entraînement.

Récemment, le Novel Class Discovery (NCD) (Hsu et al., 2018) a été proposé pour combler ces lacunes et tente d'identifier de nouvelles classes dans un ensemble de données non étiquetées en exploitant un autre ensemble étiqueté de classes différentes. Plusieurs solutions ont été proposées dans le contexte de la vision par ordinateur (Han et al., 2021, 2019; Zhong et al., 2021) avec des résultats prometteurs.

Cependant, la recherche dans ce domaine est encore récente et, à notre connaissance, le NCD n'a pas été directement abordé pour les données tabulaires. Bien que les données audio et image suscitent un grand intérêt dans les publications scientifiques récentes, les données tabulaires restent une structure d'information très courante que l'on retrouve dans de nombreux problèmes du monde réel, tels que les systèmes d'information des entreprises. Dans cet article, nous nous concentrerons donc sur le NCD pour les données tabulaires. À la différence