Qu'est-ce que mon GNN capture vraiment ? Exploration des représentations internes d'un GNN

Luca Veyrin-Forrer*, Ataollah Kamal*, Stefan Duffner*, Marc Plantevit**, Céline Robardet*

*Université de Lyon, INSA Lyon, CNRS, LIRIS UMR5205
**EPITA Research and Development Laboratory (LRDE), France

Résumé. Nous considérons l'explication de GNN. Alors que les travaux existants expliquent la décision du modèle en s'appuyant sur la couche de sortie, nous cherchons à analyser les couches cachées pour identifier les attributs construits par le GNN. Nous extrayons d'abord des règles d'activation qui identifient des ensembles de neurones co-activés pour une classe. Ces règles définissent des représentations internes ayant un impact fort sur la classification. Ensuite, nous associons à celles-ci un graphe dont le plongement produit par le GNN est très proche de celui identifié par la règle. Des expériences sur 6 jeux de données et 3 baselines démontrent que notre méthode génère des graphes réalistes de haute qualité.

1 Introduction

Les graphes sont une structure de données puissante très utilisée pour représenter des données relationnelles. L'une de leurs spécificités est que leur structure sous-jacente n'est pas dans un espace Euclidien et n'a pas une structure en forme de grille, caractéristiques facilitant l'utilisation directe de techniques génériques d'apprentissage automatique. En effet, chaque nœud d'un graphe est caractérisé par un label, ses nœuds voisins et récursivement leurs propriétés. Ces informations intrinsèquement discrètes ne peuvent pas être directement utilisées par des méthodes d'apprentissage automatique standard pour prédire une classe associée au graphe ou à un de ses nœuds. Ainsi, les réseaux de neurones pour graphes (GNN) apprennent des vecteurs $\mathbf{h}_v \in \mathbb{R}^K$ représentant chaque nœud v dans un espace métrique afin de permettre la comparaison entre nœuds. Les GNN utilisent une stratégie de propagation de message qui agrège récursivement les informations des nœuds vers leurs voisins afin de produire des représentations vectorielles de l'ego-graphe centrées sur un nœud v – avec un rayon égal à l'indice de récursivité – de telle sorte que la tâche de classification, basée sur ces vecteurs, soit optimisée.

Bien que les GNN aient atteint des performances remarquables dans de nombreuses tâches, un inconvénient majeur est leur manque d'interprétabilité. Les cinq dernières années ont vu un énorme effort de recherche pour définir des techniques d'explication de réseaux de neurones profonds (Burkart et Huber, 2021; Molnar, 2020), en particulier pour les données sous forme d'images ou de textes. Cependant, l'explicabilité des GNN a été bien moins explorée avec seulement deux types d'approches. D'une part, les algorithmes d'explication au niveau de l'instance (Baldassarre et Azizpour, 2019; Pope et al., 2019; Schnake et al., 2020; Duval et Malliaros, 2021; Luo et al., 2020; Ying et al., 2019) visent à apprendre un masque vu