Application de GNN sur des graphes non-attribués : benchmark de performances.

Ikram Boukharouba[1, 2], Florence Sèdes[1], Christophe Bortolaso[2], Florent Mouysset[2].

[1] *IRIT, Université Toulouse III Paul Sabatier, Toulouse/France*
[2] *Berger Levrault, Labège/France*

ikram.boukharouba@irit.fr

MOTS-CLES : Traces d'activité utilisateurs, Logs, Application industrielle, Graphe de navigation, Graph Neural Network (GNN), Graphe non-attribué.

KEYWORDS : Traces of user activity, Logs, Industrial application, Navigation graph, Graph Neural Network (GNN), Non-attributed graph.

1. Introduction

Les traces utilisateurs peuvent être modélisées sous forme de graphes qui traduisent la navigation de l'utilisateur au sein d'un logiciel. Analyser ces graphes afin d'éliciter les comportements des utilisateurs nécessite l'utilisation de méthodes de machine learning traditionnelles. Or, les graphes de navigation générés à partir des ensembles de traces ne conviennent pas à l'application directe de ces méthodes : graphes incomplets, non labellisés,... Les Graph Neural Networks (GNNs) représentent des algorithmes de machine learning non traditionnels dédiés aux graphes qui permettent de pallier ces insuffisances en rendant possible la prédiction des liens et des nœuds manquants. L'avantage des GNNs réside dans leur capacité à prendre en considération la structure de graphe ainsi que les features de ses nœuds pour la tâche d'apprentissage**.** Néanmoins, cet avantage entrave aussi l'applicabilité des GNNs sur certains graphes « du monde réel », **lorsque peu de features sur les nœuds sont disponibles.** Plus précisément, les graphes de terrain manipulés peuvent s'avérer incomplets ; par exemple, en raison de problèmes de confidentialité : ces graphes sont qualifiés de non-attribués.

Dans cet article nous abordons le problème d'application de GNN sur les graphes non-attribués et montrons **l'importance du nombre de features** des nœuds pour le bon fonctionnement de GNN (section 1). Dans la section 2, un état de l'art sur les défis d'application de GNN sur des graphes non-attribués est présenté. Ensuite dans la section 3, nous présentons notre étude de cas sur la classification des nœuds sur notre graphe de terrain non-attribué ; ainsi qu'une comparaison entre les résultats sur

notre graphe de navigation et ceux obtenus sur des jeux de données standards du domaine dans la section 4. Enfin, dans la section 5. Nous concluons cet article et ouvrons par quelques perspectives.

2. État de l'art

L'extraction des informations pertinentes dans les graphes est devenue un problème important pour la communauté de l'exploration des données. Différents types de réseaux de neurones pour graphes (GNNs) ont prouvé leur efficacité pour la classification des nœuds (Wu *et al.*, 2019a) et la prédiction des liens (Wu *et al.*, 2019b). À notre connaissance, il existe très peu de travaux qui abordent l'importance de features des nœuds dans le fonctionnement des GNNs et qui proposent des solutions pour remédier au problème de manque de features dans les graphes. Duong *et al.* (Duong *et al.*, 2019) montrent la sensibilité des GNNs aux modifications des features des nœuds. Leur protocole expérimental consiste à intervertir les features des nœuds du graphes sans changer les étiquettes. Une revue sur les techniques d'initialisation de features artificielles pour l'application des GNNs sur les graphes non attribués est présentée aussi. Ces techniques peuvent être classées en 2 catégories : les techniques basées sur la centralité comme Degree (Rossi *et al.*, 2017) et Egonet (Henderson *et al.*, 2012). Et les techniques basées sur l'apprentissage : DeepWalk (Perozzi *et al.*, 2014), HOPE(*Ou et al.*, 2016)). Les résultats prouvent que les features artificielles donnent les mêmes résultats que les features originales voir mieux dans certains cas. Cependant, ce travail a été validé que sur des datasets standards.

Dans la continuité de Duong, Cui *et al* (Cui *et al.*, 2021) proposent une revue sur les techniques de générations des features artificielles des nœuds pour les graphes non-attribuées. Ils regroupent les features de nœuds en deux grandes familles : les features de nœuds positionnelles, basées sur la position dans le graphe. Ainsi que les features de nœuds structurels qui capturent les informations structurelles des nœuds. Néanmoins, leur travail n'est testé et validé que sur des datasets standards. Zhou *et al.* (Zhou *et al.*, 2021), proposent une approche pour la classification des nœuds dans les graphes multi-étiquetées. Leur contribution se base sur l'exploitation simultanée des informations de structure dans le graphe, ainsi que les features et les étiquettes de nœuds. Cette méthode prouve son efficacité aussi bien sur les graphes attribués que non attribués où les résultats étaient fortement affectés par l'absence de features.

3. Cas d'utilisation et expérimentations

Lorsqu'un utilisateur navigue d'une page à une autre, il crée un lien entre ces pages et ça génère un graphe de navigation. Dans notre cas d'étude, nous analysons les traces utilisateurs de logiciel de gestion Sedit. Ce que nous cherchons à faire est de trouver la distribution des nœuds par module de l'application ; un module regroupant des pages métier. Nous considérons 3 datasets standards avec des features de nœuds: Cora (Sen *et al.*, 2008), Citesser (Sen *et al.*, 2008), et Pubmed (Namata *et al.*, 2012). Également nous considérons notre ensemble de données

Sedit. Les statistiques des données sont présentées dans le Tableau 1 Afin de montrer l'importance des features des nœuds, notre protocole d'évaluation est le suivant : tester les GNNs avec l'algorithme GraphSAGE (Hamilton et al., 2017) sur les 4 datasets avec les mêmes paramètres. Les deux méthodes d'agrégation : Sum et Mean ont été utilisées (voir Tableau 1).

Tableau 1. Jeux de données et résultats de la classification de nœuds avec les GNNs.

Jeux de données	Nb nœuds	Nb liens	Nb features	Nb classes	Taux de classification	
					Mean	Sum
Cora	2.708	8.278	1.433	7	80,15%	70,32%
Citesser	3.312	4.614	3.703	6	73,81%	59,01%
PubMed	19.717	44.325	500	3	77,88%	75,15%
Sedit	**108**	**17.192**	**6**	**6**	**19,45%**	**16,25%**

4. Observations et discussion

Méthodes d'agrégation : la méthode Mean présente de meilleur taux que la méthode Sum sur les 4 datasets. En effet, la méthode d'agrégation Sum peut filtrer efficacement l'influence de la structure du voisinage, ce qui contribue peu à la performance de la classification des nœuds dans les datasets positionnels comme les datasets de citations. De manière similaire aux précédents datasets Sedit semble sensible à la position. En effet compte tenue de peu d'informations sur les transitions métiers sur les pages ces résultats sont prévisibles

Les performances : le Tableau1 montre des résultats similaires entre les datasets attribués. Les performances de classification des nœuds avec les GNNs sur les datasets de graphes attribués ont largement surpassés ceux sur le jeu de données non-attribué. Par exemple, sur l'ensemble de données Cora, le taux de précision des classifications le plus élevé, parmi les méthodes de centralité, est de 80.15%, soit 60% de plus que le taux le plus élevé de classification sur Sedit. Ce qui est prévisible. Étant donnée le peu de features sur Sedit. La comparaison entre les performances obtenues avec les GNNs sur les datasets des graphes attribués et notre jeu de données industriel Sedit prouvent l'importance des features pour le bon fonctionnement de GNN.

5. Conclusion et futurs travaux

Dans cet article, nous avons abordé l'application des GNNs sur des graphes non attribués. Notre état de l'art montre l'importance des features de nœuds dans le fonctionnement des GNNs et la dégradation de ces performances en leurs absence. Notre hypothèse est que les GNN peuvent classifier les écrans d'une application par

module métier. . Quelques travaux existants montrent que le faible nombre de features est un inconvénient pour les GNNs. Cet article reproduit des conditions d'expérimentation et conduit à confirmer ces résultats. Enfin, notre originalité est dans la vérification de ces limites avec un corpus de terrains orienté « traces d'activités utilisateurs ».

Dans la continuité de notre travail, nous envisageons de générer des features artificielles sur nos corpus de données afin de valider les techniques proposées dans l'état de l'art sur un jeu de données de terrain.

6. Bibliographie

Cui H., Lu Z., Li P., Yang C., « On positional and structural node features for graph neural networks on non-attributed graphs », arXiv preprint arXiv :2107.01495, 2021.

Duong C. T., Hoang T. D., Dang H. T. H., Nguyen Q. V. H., Aberer K., « On node features for graph neural networks », arXiv preprint arXiv :1911.08795, 2019.

Hamilton W., Ying Z., Leskovec J., « Inductive representation learning on large graphs », Advances in neural information processing systems, 2017.

Henderson K., Gallagher B., Eliassi-Rad T., Tong H., Basu S., Akoglu L., Koutra D., Faloutsos C., Li L., « Rolx: structural role extraction & mining in large graphs », Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, p. 1231-1239, 2012.

Namata G., London B., Getoor L., Huang B., EDU U., « Query-driven active surveying for collective classification », 10th International Workshop on Mining and Learning with Graphs, vol. 8, p. 1, 2012.

Ou M., Cui P., Pei J., Zhang Z., Zhu W., « Asymmetric transitivity preserving graph embedding », Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, p. 1105-1114, 2016.

Perozzi B., Al-Rfou R., Skiena S., « Deepwalk: Online learning of social representations », Proceedings of the 20th ACM SIGKDD international conference on Knowledge disco very and data mining, p. 701-710, 2014.

Rossi R. A., Zhou R., Ahmed N. K., « Deep feature learning for graphs », arXiv preprint arXiv :1704.08829, 2017.

Sen P., Namata G., Bilgic M., Getoor L., Galligher B., Eliassi-Rad T., « Collective classification in network data », AI magazine, vol. 29, no 3, p. 93-93, 2008.

Wu J., He J., Xu J., « Net: Degree-specific graph neural networks for node and graph classification », Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, p. 406-415, 2019a.

Wu Y., Lian D., Jin S., Chen E., « Graph Convolutional Networks on User Mobility Heterogeneous Graphs for Social Relationship Inference. », IJCAI, p. 3898-3904, 2019b.

Zhou C., Chen H., Zhang J., Li Q., Hu D., Sheng V. S., « Multi-label graph node classification with label attentive neighborhood convolution », Expert Systems with Applications, vol. 180, p. 115063, 2021.