Proposition de thèse en informatique

Titre : Extraction de motifs dans les graphes à l'aide de Graph Neural Networks : Vers une formalisation des règles de classification

Institution de rattachement et encadrement

Etablissement de rattachement

Institut des Sciences Exactes et Appliquées (ISEA), Université de la Nouvelle-Calédonie, Nouméa

Direction et encadrement

- Nazha Selmaoui Professeur des universités nazha.selmaoui@unc.nc
- Loïc Salmon Maître de conférences loic.salmon@unc.nc
- Mustapha Lebbah, Professeur à l'université Paris-Saclay UVSQ Campus mustapha.lebbah@uvsq.fr
- Adrien Sales, Chef de Bureau/Division Manager at OPT New Caledonia, Adrien.SALES@opt.nc

Contexte général

Les graphes dynamiques attribués sont des structures de données omniprésentes dans de nombreux domaines, et ont récemment fait l'objet d'une grande attention [6]. Les structures de graphes fournissent une représentation riche des phénomènes du monde réel. Ils ont été largement utilisés pour décrire de nombreux ensembles de données complexes (par exemple, des données spatio-temporelles, des données de santé, des données biologiques et chimiques les systèmes de recommandation, des données sociales, etc.) [3, 4, 8]. L'extraction de motifs (ou sous-structures) dans ces graphes est une tâche fondamentale pour comprendre leur structure et leur dynamique. Les Graph Neural Networks (GNN) ont émergé comme une méthode puissante pour

apprendre des représentations des graphes [9], mais leur utilisation pour l'extraction de motifs reste un défi, en particulier pour en tirer des règles de classification interprétables et formellement vérifiables.

Contexte scientifique

Les GNN sont des modèles de réseaux de neurones conçus pour traiter des données structurées en graphes. Ils opèrent par propagation de messages entre les nœuds, permettant de capturer à la fois les informations locales et globales du graphe [7]. Cependant, bien qu'ils soient efficaces pour des tâches comme la classification de nœuds ou la prédiction de liens, leur capacité à extraire des motifs significatifs et à en déduire des règles de classification reste peu explorée.

L'extraction de motifs dans les graphes est un problème complexe, car elle nécessite de combiner des aspects structurels (topologie du graphe) et sémantiques (attributs des nœuds et des arêtes). De plus, les motifs extraits doivent être explicables et généralisables, ce qui pose des défis tant sur le plan algorithmique que sur celui de la formalisation mathématique [11].

Travaux visés

Cette thèse vise à explorer l'utilisation des GNN pour l'extraction de motifs dans les graphes, avec un focus sur la génération de règles de classification formelles et explicables. Les travaux seront structurés en quatre axes :

1. *Modélisation des motifs* : Définir une représentation formelle des motifs dans les graphes, en s'appuyant sur les capacités des GNN à capturer les dépendances structurelles et sémantiques [2].
2. *Apprentissage des motifs* : Développer des algorithmes basés sur les GNN pour extraire des motifs significatifs à partir de graphes, en garantissant leur explicabilité [12].
3. *Génération de règles de classification* : Utiliser les motifs extraits pour dériver des règles de classification formelles, permettant de prédire des propriétés sur les nœuds ou les arêtes du graphe [13].
4. *Validation formelle* : Proposer une formalisation mathématique des règles de classification générées, en s'appuyant sur des outils de vérification formelle pour garantir leur robustesse et leur généralisation

Aspect novateur

L'originalité de cette thèse réside dans la combinaison des GNN avec des techniques d'extraction de motifs et de génération de règles de classification. Alors que les GNN sont principalement utilisés pour des tâches prédictives [9], cette thèse propose de les exploiter pour extraire des connaissances interprétables sous forme de règles. De plus, la formalisation mathématique de ces règles ouvre la voie à leur utilisation dans des systèmes critiques, où la vérification formelle est essentielle [1].

Les méthodes développées dans cette thèse sont conçues pour être génériques et applicables à différents cas d'application, tels que la biologie moléculaire [2], les réseaux sociaux [12], ou les systèmes de recommandation [13]. Cette généralité permettra d'adapter les approches proposées à une variété de domaines, tout en garantissant leur robustesse et leur interprétabilité

Datasets

Notre équipe informatique collabore sur plusieurs problématiques scientifiques de domaines variés, et les données recueillies dans ces contextes incluent :

- Des séries d'images satellitaires et des données terrains pour suivre et prédire l'érosion des sols (en lien avec un projet CNRT en cours de démarrage) (voir l'article que nous avons publié sur l'érosion du sol en calédonie [10, 5]).
- Des données issues de l'Office des Postes et Télécommunications de Nouvelle-Calédonie (OPT-NC) qui possède divers systèmes d'informations nécessitant la création d'un méta-graphe permettant d'identifier des failles de vulnérabilité, comme l'optimisation du flux applicatif de l'OPT.
- Des données de molécules dont l'analyse et l'extraction de motifs pourront permettre des sous-structures clés dans la découverte de nouvelles propriétés ou l'optimisation de composés chimiques

Profil recherché

- Bac+5 ou équivalent en informatique, mathématiques appliquées, ou domaine connexe.
- Compétences en apprentissage automatique, en particulier sur les réseaux de neurones.

- Connaissances en théorie des graphes et en extraction de motifs et GNN.
- Bon niveau d’anglais (lu, écrit, parlé).

Modalités de candidatures

Les candidats sont invités à envoyer les documents suivants par mail :

- Un CV détaillé.
- Une lettre de motivation
- Les relevés de notes du Master (ou diplôme équivalent).
- Une ou plusieurs lettres de recommandation, ou les coordonnées de référents académiques ou industriels.
- Tout autre document jugé pertinent pour la candidature.

Références

- [1] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv :1806.01261*, 2018.
- [2] Christian Borgelt and Michael R. Berthold. Mining molecular fragments : Finding relevant substructures of molecules. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM)*, 2002.
- [3] Elise Desmier, Marc Plantevit, Céline Robardet, and Jean-François Boulicaut. Cohesive co-evolution patterns in dynamic attributed graphs. In *International Conference on Discovery Science*, pages 110–124. Springer, 2012.
- [4] Elise Desmier, Marc Plantevit, Céline Robardet, and Jean-François Boulicaut. Trend mining in dynamic attributed graphs. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 654–669. Springer, 2013.
- [5] Frédéric Flouvat, Nazha Selmaoui-Folcher, Jérémy Sanhes, Chengcheng Mu, Claude Pasquier, and Jean-François Boulicaut. Mining evolutions of complex spatial objects using a single-attributed directed acyclic graph. *Knowledge and Information Systems*, 62(10) :3931–3971, 2020.
- [6] P. Fournier-Viger, G. He, C. C. Cheng, J. Li, M. Zhou, J. C. Lin, and U. Yun. A survey of pattern mining in dynamic graphs. *WIREs Data Mining and Knowledge Discovery*, 10, 2020.

- [7] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [8] Mehdi Kaytoue, Yoann Pitarch, Marc Plantevit, and Celine Robardet. Triggering patterns of topology changes in dynamic graphs. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 158–165. IEEE, 2014.
- [9] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*, 2017.
- [10] Nazha Selmaoui-Folcher, Frédéric Flouvat, Chengcheng Mu, Jérémy Sanhes, and Jean-François Boulicaut. Extraction complète efficace de chemins pondérés dans un a-dag. In Benoît Otjacques, Jérôme Daramont, and Thomas Tamisier, editors, *15èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2015, 27-30 Janvier 2015, Luxembourg*, volume E-28 of *Revue des Nouvelles Technologies de l'Information*, pages 179–190. Hermann-Éditions, 2015.
- [11] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *International Conference on Learning Representations (ICLR)*, 2019.
- [12] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [13] Ziwei Zhang, Peng Cui, and Wenwu Zhu. Deep learning on graphs : A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2020.