

DC-GNN : Deep clustering with Graph Neural Networks for real world data

Clustering is an unsupervised machine learning technique that involves grouping data samples. In the literature, most proposed approaches use shallow models. Recently, some works have used deep architectures such as deep K-means, graph convolutional networks based clustering and online deep clustering [1,2]. In the image domain, several solutions aim to learn deep convolutional neural networks whose loss is given by the reconstruction error and a clustering-based loss (as in Kmeans or spectral clustering). While these methods provided satisfactory results, most proposed deep solutions do not offer clustering as a direct solution. Rather, a data representation is first computed and then a clustering algorithm is applied to this representation. For shallow graph-based models or kernel-based models, it has been shown that direct solutions, also known as one-step clustering solutions, can be superior to the indirect solution. This is because data partitioning is considered in the criterion being optimized.

Learning with graph data, such as social networks and citation networks, has recently gained increasing attention. Graph Neural Network (GNN) has become the most important tool for learning representations on graphs. GNN is an extension of traditional neural networks for processing graph data. Graph Convolutional Networks (GCNs) [3,4] and their variants have been proposed for semi-supervised classification of nodes in graph structured data. The main application of GCNs is semi-supervised classification. The use of GCNs in unsupervised clusters will be explored.

The main goal of this research is to propose deep clustering algorithms based on GNN. To this end, the research work consists of:

- 1- Developing a self-supervised deep graph embedding feature that handles the interaction of high order nodes for clustering.
- 2- Learning such GNN architectures that integrate the attributes and the convolution and aggregation operator for structured information. The output can be the low dimensional embedding vectors which can be seen as the nonlinear data projection.
- 3- Using of an architecture with two heads. One head provides the nonlinear spectral representation and the other provides the cluster index matrix.
- 4- Applying proposed algorithms on image clustering.

[1]- X. Zhanga, H. Liua, X. Wuc, X. Zhang, X. Liu Spectral Embedding Network for Attributed Graph Clustering. Neural Networks, 2021.

[2]- X. Zhan, J. Xie, Z.Liu, Y. Ong, C. Loy. Online deep clustering for unsupervised learning. IEEE CVPR 2020.

[3]- T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations, 2017.

[4]- M. T. Kejani, F. Dornaika and H. Talebi. Graph Convolution Networks with Manifold Regularization for Semi-Supervised Learning. Neural Networks, volume 127, pp. 160-167, 2020.

Clustering profond par Réseaux de Neurones des Graphes

Le clustering profond est un outil important qui contribue à une meilleure compréhension, récupération, visualisation et organisation des données massives, en plus d'être un élément important des systèmes décisionnels complexes. Les méthodes de clustering traditionnelles, par exemple, k-moyennes (k-Means) reposent entièrement sur les représentations de données d'origine et peuvent alors être inefficaces lorsqu'elles se trouvent dans un espace de grande dimension - un problème communément appelé la malédiction de la dimensionnalité. Une solution intuitive largement étudiée consistait à projeter les données d'un espace de grande dimension en un espace de dimension inférieure dans lequel le clustering peut être plus facile. De telles techniques, basées sur des mesures statistiques, s'avèrent non optimales en vue de l'objectif principal du clustering qui est de découvrir la structure sous-jacente des clusters. Une autre solution consistait à utiliser les réseaux de neurones profonds (Deep Neural Network DNN) pour effectuer le clustering [1,2].

Le clustering des données graphiques, telles que les réseaux sociaux et les citations, est un sujet qui attire l'attention. Graph Neural Network (GNN) est devenu l'outil le plus important pour l'apprentissage des graphes. GNN est une extension des réseaux de neurones traditionnels pour apprendre sur les graphes. Les réseaux convolutifs de graphes (GCN) [3,4] et leurs variantes ont été proposés pour la classification semi-supervisée des nœuds dans les données structurées en graphe. L'application principale des GCN à ce jour est la classification semi-supervisée. L'utilisation des GCN pour l'apprentissage non supervisé reste à explorer.

L'objectif principal de cette thèse consiste à proposer des algorithmes de clustering profond sur graphes et les valider sur des données réelles. Le travail devra prendre en compte d'une manière originale des aspects importants :

- 1- Apprendre une architecture permettant d'intégrer les interactions aussi bien entre les nœuds voisins que ceux éloignés.
- 2- Proposer des fonctions coûts permettant d'apprendre des architectures GNN qui permettent d'extraire les caractéristiques des graphes en utilisant un opérateur de convolution approprié à ceux-ci.
- 3- Utiliser une architecture multi-têtes. Celle-là est basée sur un processus développé par plusieurs couches apprenant conjointement différentes représentations à partir de différentes positions.
- 4- Appliquer ces méthodes sur le clustering des images.

[1]- X. Zhanga, H. Liua, X. Wuc, X. Zhang, X. Liu Spectral Embedding Network for Attributed Graph Clustering. Neural Networks, 2021.

[2]- X. Zhan, J. Xie, Z.Liu, Y. Ong, C. Loy. Online deep clustering for unsupervised learning. IEEE CVPR 2020.

[3]- T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations, 2017.

[4]- M. T. Kejani, F. Dornaika and H. Talebi. Graph Convolution Networks with Manifold Regularization for Semi-Supervised Learning. Neural Networks, volume 127, pp. 160-167, 2020.