Méthodologies de Recherche

Projet : Non-linear Attributed Graph Clustering by Symmetric NMF with PU Learning

Réalisé par:

- Sana OUBENYAHYA
- Chaimae HILAL
- Kenza OUAZZANI CHAHDI
- Mouad ET-TALI

I- Introduction

Les graphes sont un moyen flexible et puissant de représenter les données. Les bases de données relationnelles traditionnelles, avec leurs schémas fixes, rendent difficile le stockage des connexions entre différentes entités, alors que ces connexions sont une partie abondante et vitale pour bien analyser les données et les classifier.

Or, Le clustering de graphes est le processus qui consiste à regrouper les nœuds du graphe en clusters, en tenant compte de la structure des arêtes du graphe de manière à ce qu'il y ait plusieurs arêtes au sein de chaque cluster et très peu entre les clusters. Le clustering de graphe vise à partitionner les nœuds du graphe en groupes disjoints.

Le projet est basé sur l'article scientifique « ***Non-linear Attributed Graph Clustering by Symmetric NMF with PU Learning*** » de Maekawa, Takeuchi et Onizuka. L'objectif de l'article consiste à concevoir une méthode de clustering efficace et efficiente qui capture précisément la relation cachée entre la topologie et les attributs dans les graphes du monde réel.

La méthode NAGC se divise en trois parties principales, premièrement, elle apprend une projection non linéaire entre les deux espaces latents embedding de la topologie et des attributs des graphes, deuxièmement, elle tire parti de l'apprentissage positif non labelisé pour prendre en compte l'effet des bords positifs partiellement observés, finalement, elle atteint une complexité de calcul efficace, O((n2 + mn)kt) pour l'apprentissage de l'affectation des clusters.

II- Méthodes et approches utilisées

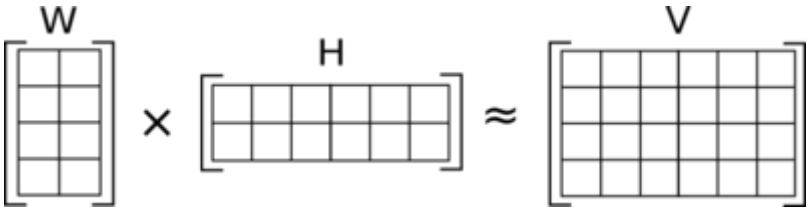
1. Non negative Matrix Factorization

La factorisation matricielle non négative (NMF) est un groupe d'algorithmes en analyse multivariée et en algèbre linéaire où une matrice V est factorisée en (généralement) deux matrices W et H, avec la propriété que les trois matrices n'ont pas d'éléments négatifs tel que :

$$\mathbf{V} \simeq \mathbf{WH} \quad (1)$$

Cette non-négativité rend les matrices résultantes plus faciles à inspecter. En effet, c'est une technique permettant d'obtenir une représentation de bas rang des matrices avec des éléments non négatifs ou positifs. De telles matrices sont courantes dans une variété d'applications d'intérêt. Par exemple, les images sont un ensemble des matrices de nombres entiers positifs représentant des intensités de pixels. Dans la recherche d'informations et l'exploration de texte, nous nous appuyons sur des matrices de documents terminologiques pour représenter les collections de documents. Dans les systèmes de recommandation, nous avons des matrices d'utilité montrant les préférences des clients pour les articles.

Étant donné une matrice de données V de m lignes et n colonnes avec chaque élément aij ≥ 0, NMF recherche les matrices W et H de taille m lignes et k colonnes, et k lignes et n colonnes, respectivement, telles que **:**



La valeur k est définie par l'utilisateur et doit être égale ou inférieure au plus petit de m et n. La matrice W est généralement appelée dictionnaire ou matrice de base, et H est appelée matrice d'expansion ou de coefficient. L'idée sous-jacente de cette terminologie est qu'une matrice de données V donnée peut être exprimée en termes de sommation de k vecteurs de base (colonnes de W) multipliés par les coefficients correspondants (colonnes de H).

Pour réaliser cette factorisation, on doit résoudre le problème de minimisation suivant :

(2) **Minimize $||V - W.H||^2$**

en cherchant un minimum local qui satisfait l'approximation voulu. Par conséquent, l'algorithme Multiplicative Update introduit par Lee et Seung en 1999 qui est basé sur la descente du gradient, est utilisé pour aboutir au minimum local.

2. Symmetric Non negative Matrix Factorization

Symmetric NMF (SymNMF) est une méthode de clustering de graphes, qui hérite des avantages de NMF en appliquant la non-négativité sur la matrice d'affectation de clustering. Contrairement à NMF, SymNMF est basé sur une mesure de similarité entre des points de données et factorise une matrice symétrique contenant des valeurs de similarité par paires (pas nécessairement non négatives).

La plupart des succès de NMF dans le clustering ont été autour du clustering de documents. L'une des raisons est que chaque vecteur de base représente la distribution de mots d'un sujet, et les documents avec des distributions de mots similaires doivent être classés dans le même groupe. Cette propriété n'est pas valable dans tous les types de données. C'est pour cela que SymNMF était introduit par Kuang, Ding et Park en 2012.

La formulation standard de NMF dans (2) a été appliquée à de nombreuses tâches de clustering où les n points de données sont explicitement disponibles dans V et sont directement utilisés comme input. Cependant, dans de nombreux cas, comme lorsque des points de données sont intégrés dans une variété non linéaire, il est préférable de décrire la relation entre les points de données sous la forme d'un graphe. Dans le modèle de graphe, chaque nœud correspond à un point de données, et une matrice de similarité $A_{n\times n}$ contient des valeurs de similarité entre chaque paire de nœuds, c'est-à-dire que la (i, j)-ième entrée de A représente la similarité entre x_i et x_j. En SymNMF, une variation symétrique de NMF qui utilise A directement en entrée est appliquée. Lorsque A est correctement construit, la

factorisation de A générera une matrice d'affectation de cluster qui est non négative et capture bien la structure de cluster inhérente à la représentation graphique.

3. PU Learning

Positive Unlabeled Learning c'est d'apprendre à partir d'exemples positifs et non labelisés (ou apprentissage PU) peut être considéré comme un problème de classification à deux classes (positives et négatives), où il n'y a que des données d'entraînement positives labelisés, mais pas de données d'entraînement négatives labelisés.

La méthode originale du PU est :

Étant donné un ensemble d'apprentissage contenant uniquement des classes positives (P) et Unlabeled (U), on suit les étapes suivantes :

1. En traitant tous les U comme négatifs (N), on forme un classificateur P vs U.
2. À l'aide du classificateur, on note la classe inconnue et isolez l'ensemble des négatifs « fiables » (RN).
3. Former un nouveau classificateur sur P vs. RN, l'utiliser pour marquer le U restant, isoler RN supplémentaire et agrandir RN.
4. Répétez l'étape 3, en agrandissant itérativement l'ensemble de RN jusqu'à ce que la condition d'arrêt soit remplie.

III- Dataset

Quatre datasets réels ont été utilisés dans les expériences réalisés dans l'article :

- WebKB : le graphe web de quatre universités. Le label d'un sommet indique l'université propriétaire de la page, et les attributs d'un sommet représentent les mots apparaissant dans la page.
- Citeseer et Cora : les réseaux de citation. Le label d'un sommet correspond à un domaine de recherche de l'article, et les attributs d'un sommet sont constitués des mots apparaissant dans l'article.
- Polblog4 : est un réseau d'hyperliens entre des blogs sur la politique américaine. Le label d'un sommet indique si le blog est libéral ou conservateur, et les attributs d'un sommet représentent les sources des blogs.

Références

- D. Kuang, C. Ding and H.Park. Symmetric Nonnegative Matrix Factorization for Graph Clustering, 1:3, 2012.

- D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization, 1:2, 1999.

- https://medium.com/octavian-ai/how-to-get-started-with-machine-learning-on-graphs-7f0795c83763

- https://iksinc.online/2016/03/21/what-is-nmf-and-what-can-you-do-with-it /