INFORME PROYECTO

Overlapping community detection with graph neural networks

Estructura del repositorio

La estructura del repositorio es bastante simple:

1. 01-community_detection_gnn.ipynb es el notebook que contiene todo lo necesario para cargar datos, crear el modelo, entrenarlo y evaluarlo.
2. data/ es una carpeta con los datos necesarios.
    a. colombia_openalex_adj.npz matriz dispersa de scipy con la matriz de adjacencia.
    b. colombia_openalex_att.npz matriz dispersa de scipy con los atributos de los nodos de la red.
    c. colombia_openalex_classes.npz matriz diospersa de scipy con las clases a las que pertenecen los nodos de la red.
    d. colombia_openalex_mappings.pkl archivo pickle binarizado con el mapeo de los números de las matrices de atributos y clases a palabras (No es necesario para la ejecución del notebook).
3. ENTREGA1.PDF es el documento con el anteproyecto
4. INFORME_PROYECTO.PDF es este archivo.

Estructura de la solución

Modifiqué en parios aspectos la solución propuesta en el artículo Shchur and Günnemann (2019):

- El conjunto de datos no parte de Microsoft Academics sino de su sucesor, OpenAlex.
- El conjunto de datos fue extraído con los autores colombianos de los trabajos en OpenAlex.

Presento un resumen de lo que logré implementar del artículo:

- El cálculo de la métrica NMI (McDaid et al., 2011).
- La función de loss de Bernoulli-Poisson pero sin garantizar distribuciones uniformes en sus partes.
- Las capas convolucionales con la operación **A** @ (**X** @ **W**) + **b**. Dónde **A** es la matriz de adyacencia, **X** es la matriz de caracteristicas, **W** son los pesos de la capa y **b** es el sesgo.
- El modelo con las capas escondidas variables, la salida al número de posibles comunidades a las que puede pertenecer el autor con relu como función de activación.
- Un extractor de subgrafos para ir alimentando el entrenamiento por lotes. Cada subgrafo es un lote.

El modelo podría mejorar si:

- Se mejoran las variables descriptoras de los investigadores.
- Se hace normalización de dichas variables.
- Se implementa una normalización por lotes.
- Se mejora la función de loss para compensar el desbalance.
- Se mejora el extractor de subgrafos.

Iteraciones

References

McDaid, A. F., Greene, D., & Hurley, N. (2011). Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*.

Shchur, O., & Günnemann, S. (2019). Overlapping community detection with graph neural networks. *arXiv preprint arXiv:1909.12201*.