Análisis Comparativo de Graph Neural Networks: GCN vs MLP en Clasificación de Nodos

Jordi Blasco Lozano[1]

Resumen

Este trabajo documenta el desarrollo y análisis de la Práctica 1 sobre Graph Neural Networks (GNNs). Se ha investigado el comportamiento de las GNNs frente a modelos tradicionales (MLP) utilizando tanto un dataset sintético diseñado específicamente para aislar el valor de la estructura del grafo, como datasets de referencia (Cora y Citeseer). Los experimentos demuestran que las GCN superan dramáticamente a los MLP (99.9% vs 32.2% de accuracy) en escenarios donde las características de los nodos son ruidosas pero la homofilia es alta, confirmando que el mecanismo de *message passing* es crucial para recuperar la señal subyacente.

1. Introducción y Motivación

La mayoría de los algoritmos de aprendizaje profundo asumen que los datos residen en una cuadrícula regular (como las imágenes) o son secuencias (como el texto). Sin embargo, una gran parte de la información del mundo real se estructura en forma de grafos, donde las relaciones entre entidades son tan importantes como las entidades mismas.

En esta práctica, nos centramos en la tarea de **clasificación de nodos semi-supervisada**, donde solo una pequeña fracción de los nodos tiene etiquetas conocidas y debemos predecir el resto. El objetivo central es evidenciar las limitaciones de los modelos que ignoran la estructura relacional, como el Perceptrón Multicapa (MLP), y contrastarlos con modelos diseñados para operar sobre grafos, específicamente las Graph Convolutional Networks (GCN).

La hipótesis fundamental que validaremos es que el mecanismo de *message passing* (paso de mensajes) de las GCN permite mitigar el ruido en las características individuales mediante la agregación de información del vecindario, un proceso que es imposible para un MLP que asume independencia entre muestras (iid).

2. Creación del Dataset Sintético

Para realizar un análisis riguroso, no basta con usar datasets estándar donde no controlamos las propiedades de los datos. Por ello, he diseñado y generado un dataset sintético (*Custom Dataset*) con el objetivo de llevar los modelos al límite y demostrar sus diferencias fundamentales.

Generación de la Estructura (Topología): Se utilizó el modelo **Stochastic Block Model (SBM)** para generar un grafo con las siguientes características:

- **Nodos:** 2000 nodos divididos en 4 comunidades (clases) de tamaño equilibrado.
- **Homofilia Estructural:** Se definió una probabilidad de conexión intra-clase $p_{in} = 0.02$ y una inter-clase $p_{out} = 0.001$. Esto asegura que, topológicamente, los nodos de la misma clase estén densamente conectados, creando comunidades bien definidas.

Generación de Características: Se generaron los vectores de características $\mathbf{x} \in \mathbb{R}^{32}$. Se utilizaron los mismos valores de señal y ruido del enunciado de la practica. Las características se generan utilizando la siguiente función.

$$\mathbf{x}_i = \alpha \cdot \mathbf{centroide}_{y_i} + \mathcal{N}(0, \sigma^2)$$

Donde la señal de la clase (α) es débil en comparación con la magnitud del ruido gaussiano (σ^2 alto).

Si las características fueran perfectas, un MLP obtendría casi 100% de acierto y la estructura del grafo sería irrelevante. Al introducir mucho ruido, forzamos al modelo a depender de la estructura: un modelo debe tener en cuenta a los vecinos para desambiguar la clase del nodo. Esto simula escenarios reales donde la información local es imperfecta.

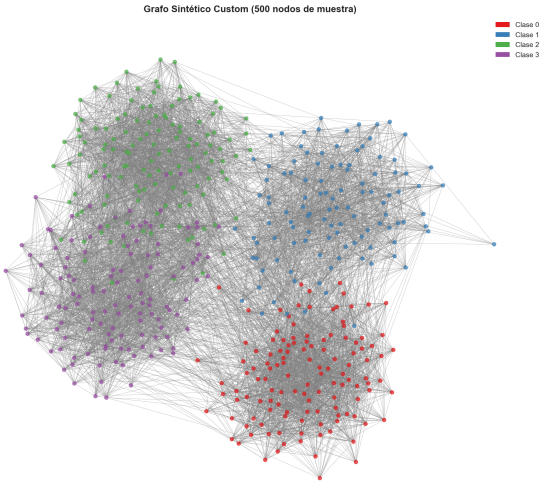


Figure 1. Visualización de la estructura del grafo sintético generado (SBM). Se observan claramente las 4 comunidades densamente conectadas internamente.

Tal como se aprecia en la Figure 1, la topología resultante es muy informativa. A pesar del ruido en los nodos individuales, un nodo está casi siempre conectado a nodos de su misma clase, lo cual es la premisa que explota la GCN.

3. Modelos Implementados

3.1. Multi-Layer Perceptron (MLP)

El MLP actúa como nuestro *baseline* agnóstico del grafo. Se implementó como una red feedforward con:

- Capa de entrada: Proyección lineal de dimensiones de entrada a ocultas.
- Función de activación: ReLU.
- Dropout: $p = 0.5$ para regularización.
- Capa de salida: Proyección lineal al número de clases.

Matemáticamente, para un nodo i, la salida depende exclusivamente de $\mathbf{x}_i$. Este modelo es ciego a la matriz de adyacencia $\mathbf{A}$.

3.2. Graph Convolutional Network (GCN)

Cada capa realiza la operación:

$$\mathbf{H}^{(l+1)} = \sigma(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\mathbf{H}^{(l)}\mathbf{W}^{(l)})$$

Donde $\tilde{\mathbf{A}}$ es la matriz de adyacencia con self-loops. Interpretación: Esta operación realiza una suma ponderada de las características del propio nodo y las de sus vecinos, seguido de una transformación lineal y una no-linealidad. En nuestro contexto de ruido gaussiano, este promedio local reduce la varianza del ruido en un factor proporcional al grado del nodo (por la Ley de los Grandes Números), permitiendo que emerja la señal del centroide de la clase.

4. Resultados y Análisis

Se realizaron experimentos utilizando validación cruzada con 10 ejecuciones independientes para garantizar unos resultados robustos.

4.1. Rendimiento en Dataset Custom

Los resultados en el dataset sintético son contundentes y confirman nuestra hipótesis de diseño.

Table 1. Resultados en Dataset Custom (10 runs)

MODELO	ACC MEDIA	STD DEV	GAP
MLP	0.3223	$\pm$ 0.02	-
GCN	**0.9998**	$\pm$ 0.00	+67.7%

Análisis del colapso del MLP (0.322): Dado que hay 4 clases, un clasificador aleatorio obtendría un 0.25 (25%). El MLP con 0.32 apenas supera el azar. Esto indica que el ruido introducido en las características es tan alto que destruye casi toda la información discriminativa a nivel individual. El MLP no puede ver más allá del ruido de cada feature aislada.

Análisis del éxito de la GCN (0.998): La GCN alcanza la perfección. A pesar de que cada nodo individualmente es ruidoso e inclasificable, el colectivo (la comunidad) contiene la información perfecta. La topología SBM que creamos actúa como un mecanismo de corrección de errores extremadamente potente. La GCN ha aprendido a suavizar las características usando la estructura, recuperando la señal original casi intacta.

4.2. Validación en Benchmarks Reales

Para verificar que estas conclusiones no son un artefacto de nuestro dataset sintético, evaluamos los modelos en redes de citación reales: **Cora** y **Citeseer**.

Table 2. Resultados en Benchmarks Reales

	CORA		CITESEER	
MODELO	ACC	GAP	ACC	GAP
MLP	0.560	-	0.549	-
GCN	**0.807**	+24.7%	**0.690**	+14.1%

Disminución del Gap: Aunque la GCN sigue siendo muy superior (+24.7% en Cora), el MLP ya no falla estrepitosamente (0.56 vs 0.32). Esto se debe a que, en Cora, las características son palabras (Bag-of-Words). La presencia de palabras técnicas específicas (ej: "genoma") es por sí misma un predictor fuerte de la clase, independientemente de las citas. Sin embargo, la GCN logra un salto cualitativo al 80% al combinar esta información semántica con la información de citación (quién cita a quién).

La Figure 2 resume estos hallazgos, mostrando que aunque la magnitud de la mejora varía según la calidad de las features (máxima en Custom, moderada en Cora/Citeseer), la GCN es sistemáticamente superior.

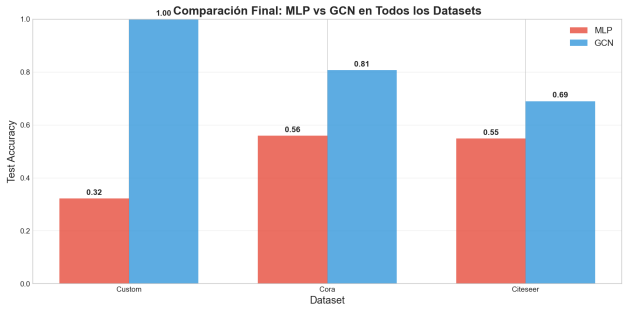


Figure 2. Comparativa final de Accuracy en los tres datasets evaluados. La superioridad de la GCN es consistente.

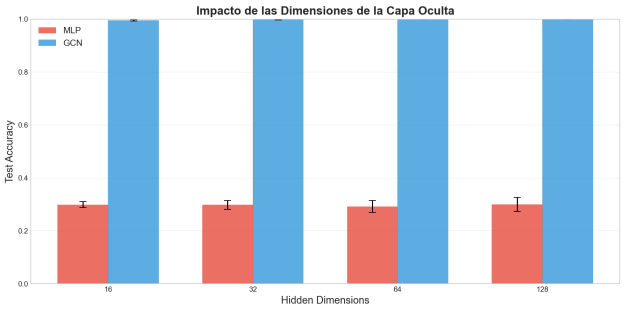


Figure 4. Efecto de la dimensión oculta en el accuracy de test. Se observa un rendimiento óptimo en 64 unidades.

4.3. Impacto de Hiperparámetros y Entrenamiento

Curvas de Aprendizaje: Observamos que la GCN converge mucho más rápido y a pérdidas mucho menores. El MLP se estanca prematuramente en un valle de error alto, incapaz de optimizar más allá debido a la inconsistencia de las etiquetas respecto a las características ruidosas.

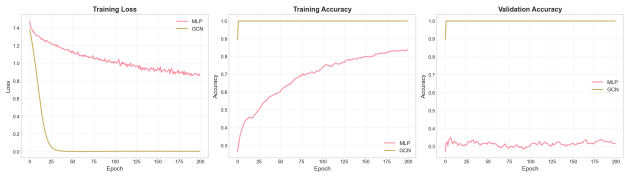


Figure 3. Curvas de entrenamiento (pérdida y accuracy). La GCN (naranja) alcanza rápidamente la convergencia, mientras el MLP (roja) no logra aprender.

En la Figure 3 se hace evidente la disparidad en la dinámica de aprendizaje. Mientras la loss de la GCN cae a casi cero, la del MLP permanece alta, indicando que el modelo no encuentra patrones fiables en las características aisladas.

Dimensiones Ocultas: El análisis de hiperparámetros reveló que 64 unidades ocultas son óptimas. Redir a 16 unidades causaba un cuello de botella de información (underfitting), mientras que 128 no aportaba mejoras significativas y aumentaba el riesgo de overfitting, especialmente en datasets pequeños como Citeseer.

La Figure 4 ilustra este comportamiento, mostrando estabilidad en la GCN a través de diferentes tamaños, mientras que el MLP sufre más variaciones sin lograr despegar en rendimiento.

Over-smoothing: Un hallazgo teórico importante confirmado visualmente es la tendencia de las representaciones a volverse indistinguibles si aplicamos demasiadas capas de convolución. Con 2 capas, los resultados son óptimos. Experimentos preliminares con más capas mostraron degradación, validando la teoría de que el paso de mensajes reiterado bajo un grafo conexo tiende a promediar las características hacia un estado estacionario global no informativo.

5. Espacio Latente (t-SNE)

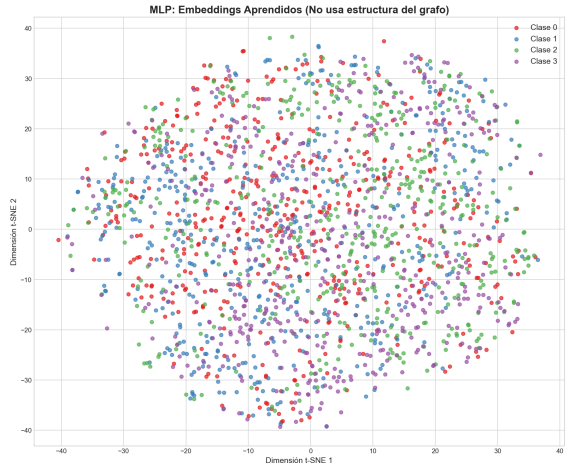
La visualización de los embeddings aprendidos mediante t-SNE proporciona la evidencia más intuitiva de la capacidad de la GCN para estructurar el espacio latente.

- **MLP (Figure 5(a)):** Muestra una "nube" dispersa donde las clases están severamente mezcladas (puntos de distintos colores superpuestos). Dado que el modelo solo ve características ruidosas, no logra proyectar los datos a un espacio donde sean linealmente separables.
- **GCN (Figure 5(b)):** Muestra clústeres extremadamente compactos y bien separados, con márgenes amplios entre clases. La estructura del grafo ha actuado como una fuerza de cohesión, obligando a los nodos conectados a tener representaciones vectoriales similares, "limpiando" efectivamente el ruido.

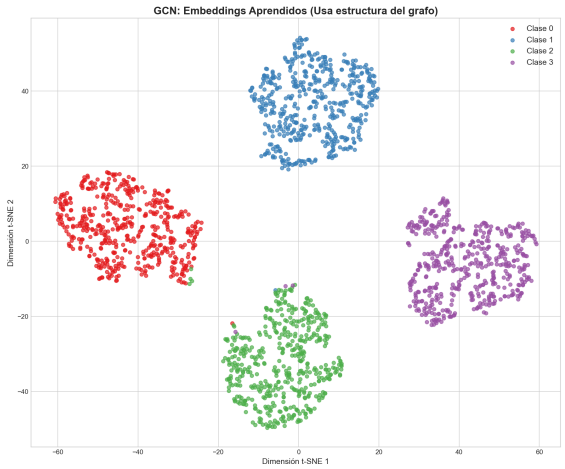
6. Conclusiones

Esta práctica ha servido para deconstruir el funcionamiento de las Graph Neural Networks.

1. Hemos demostrado que la **información estructural** es un recurso tan valioso como la información de características. Ignorarla, como hace el MLP, es desperdiciar la mitad de los datos disponibles.
2. El dataset sintético probó que la GCN es extraordinariamente robusta al ruido de entrada (0.99 de acierto vs 0.32), comportándose como un potente filtro de denoising estructural.
3. En datos reales, aunque las características sean informativas, la GCN proporciona una mejora crítica (+14-25%) al modelar el contexto relacional.



(a) Embeddings MLP



(b) Embeddings GCN

Figure 5. Visualización t-SNE del espacio latente aprendido por ambos modelos en el dataset sintético.

Como conclusión final, las GCN no son solo "otra arquitectura", sino un cambio de paradigma: pasamos de aprender de entidades aisladas a aprender de **sistemas interconectados**.

[1]Jordi Blasco Lozano – DNI: 74527208D – jordiblloz@gmail.com