Community Detection In Social Networks Using K-Means

Group Members

Shahzaib Ahmed (22K-4390)

Ali Wasif (22K-4511) - Group Leader

Introduction

Community detection is crucial in analyzing social networks, revealing hidden structures, and helping in tasks like recommendation systems and information diffusion. This project focuses on detecting communities using a combination of DeepWalk graph embeddings and K-Means clustering. The aim is to represent nodes in a high-dimensional space and cluster them based on structural similarity. DeepWalk, inspired by the research on adaptations of k-Means for community detection in parallel environments, was employed for node embedding, while K-Means clustering grouped nodes into communities based on their embeddings.

Methodology

Graph Input:
A sample edge list representing relationships between nodes was used as input. Each edge signifies a connection between two nodes, similar to a social network's interactions.

DeepWalk:
DeepWalk generates node embeddings by simulating random walks on the graph, capturing structural relationships. Each random walk is treated as a sentence, and Word2Vec is applied to learn the embeddings. These embeddings capture the nodes' neighborhood information in a high-dimensional vector space.

Clustering with K-Means:
The generated embeddings were clustered using the K-Means algorithm. The number of clusters is optimized based on the silhouette score, which measures the quality of clustering. The optimal cluster count is the one that maximizes this score, ensuring that communities are cohesive and well-separated.

Visualization:
The clusters were visualized using static plots, where each community was assigned a distinct color for easy identification. This visualization helps in interpreting the structure of the network and evaluating the clustering results.

Tools and Libraries

- **Python:** Used for algorithm implementation and data analysis.
- **NetworkX:** For graph operations such as graph creation and edge list handling.
- **Word2Vec (Gensim):** Used to generate DeepWalk embeddings.
- **Scikit-learn:** For K-Means clustering and silhouette score calculation.

Results

The clustering analysis revealed clear community structures within the graph. By optimizing the number of clusters based on the silhouette score, the best clustering configuration was determined. The silhouette score of 0.37 for 8 clusters indicates that the nodes were effectively grouped based on structural similarities.

Visualization:
The graph was visualized with each community in a distinct color, offering an intuitive representation of how nodes were grouped based on their embeddings. This helped to confirm that the identified clusters reflected meaningful groupings within the network.

Performance Evaluation Metrics

The silhouette score was used to evaluate the clustering performance, providing insight into how well nodes within the same cluster were grouped. A high score indicates strong cohesion within communities and clear separation between them. The optimal silhouette score for the selected clustering configuration was 0.3823, confirming the quality of the clustering.

Discussion

This project demonstrated the effective use of DeepWalk and K-Means for community detection in social networks. DeepWalk's random walk-based embeddings successfully captured the graph's structural properties, while K-Means clustering effectively grouped nodes into distinct communities. Maximizing the silhouette score ensured that the identified clusters were both cohesive and well-separated. This approach, inspired by adaptations of the k-Means algorithm for parallel environments (as discussed in ResearchGate), showed strong potential for large-scale community detection tasks.

Challenges

Key challenges included tuning hyperparameters for DeepWalk, such as the walk length and number of walks, to ensure high-quality embeddings. Additionally, computational efficiency was a concern for larger graphs, as generating embeddings and running K-Means clustering can be time-consuming.

Conclusion

This project successfully applied DeepWalk and K-Means clustering for community detection in social networks. The method provided accurate clustering results, validated by silhouette scores, and visualized community structures effectively. The approach can be adapted to larger networks and real-world applications with minimal adjustments. Future work could involve optimizing the algorithm for larger graphs and exploring alternative clustering methods.