Link Prediction using Node2Vec

Manas Mittal

1 Introduction

Link prediction is a fundamental problem in network analysis, aiming to predict the likelihood of an edge existing between two nodes in a graph. In this task, we apply the Node2Vec algorithm to learn node embeddings and train a classifier for link prediction.

2 Methodology

Node2Vec is a framework for learning continuous feature representations of nodes in a graph. It uses second order biased random walks to explore node neighborhoods and captures both local and global structural patterns. We generate embeddings for each node using Node2Vec and train a RandomForestClassifier for link prediction. Additionally, we experiment with incorporating node character features.

To train the model, we masked 15 percent of the edges before training, ensuring that the classifier learns in a realistic setting where some links are hidden.

We perform two passes of the algorithm:

- First, we train a classifier using only Node2Vec embeddings.
- Second, we introduce two additional features: the first and last letter of each country's name, encoded numerically and scaled appropriately.

3 Challenges and Observations

Initially, the values of the additional character features ranged between 0 and 25, resulting in minimal performance gain. However, after scaling these features by a factor of 10, the model exhibited a noticeable improvement in performance.

4 Results (Homophilic setting)

The performance metrics for the two experiments are as follows:

- Using only Node2Vec embeddings:

  - Link Prediction Accuracy: 0.8366
  - Link Prediction AUC-ROC: 0.8379

- Using Node2Vec embeddings + scaled character features:
  - Link Prediction Accuracy: 0.8718
  - Link Prediction AUC-ROC: 0.8730

  The hyperparameters used for Node2Vec are as follows:

  - $p = 1$, $q = 0.5$
  - Number of walks per node: 50
  - Walk length: 20

  The hyperparameters used by the SkipGram are as follows :

  - vector size = 16
  - window=5
  - min count=1

5 Evaluation Metrics

Accuracy: This metric represents the proportion of correctly classified link predictions over the total predictions made. Higher accuracy indicates better overall performance of the classifier.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve): This metric evaluates the classifier's ability to distinguish between positive and negative links. A higher AUC-ROC score indicates better discriminative power, as it measures the trade-off between true positive and false positive rates.

6 Visualization

To better understand the learned embeddings and the structural roles of nodes, we use two visualization techniques:

- **t-SNE Visualization of Node Embeddings**: We reduce the dimensionality of embeddings using PCA and then apply t-SNE for a 2D projection. Nodes are clustered using K-Means and plotted using different colors.

- **Structural Role Analysis**: We combine embeddings with graph structural properties such as degree, clustering coefficient, betweenness, and eigenvector centrality. After scaling, we apply t-SNE for visualization and color the nodes based on clustering.

Both visualizations provide insights into the latent structure of the graph and how embeddings capture meaningful patterns.

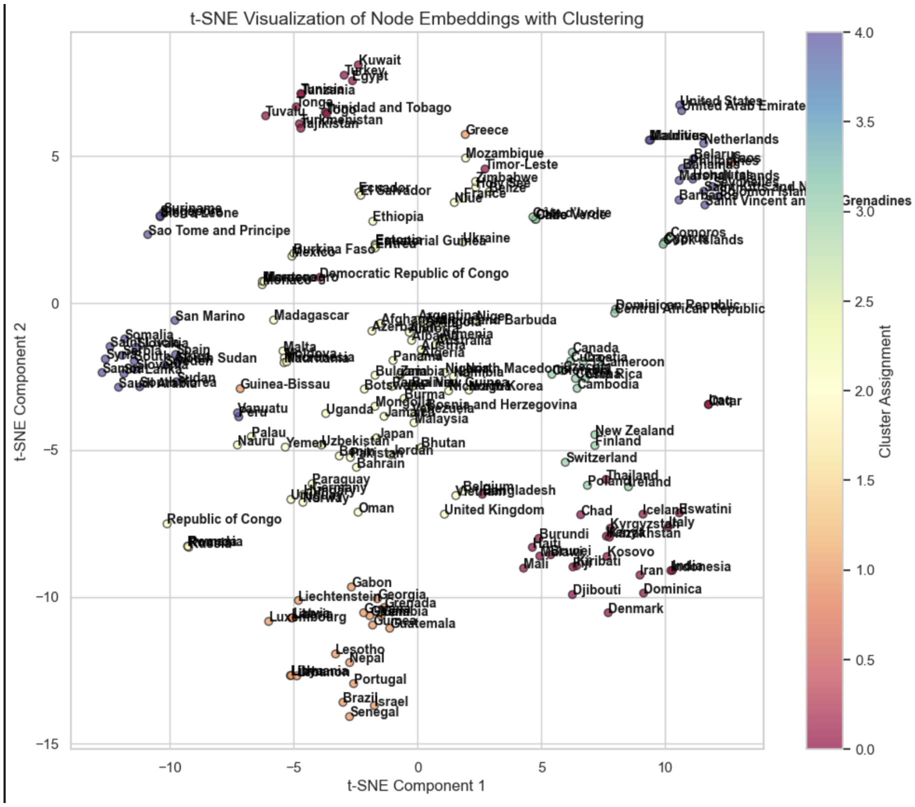

Figure 1: t-SNE visualization of node embeddings

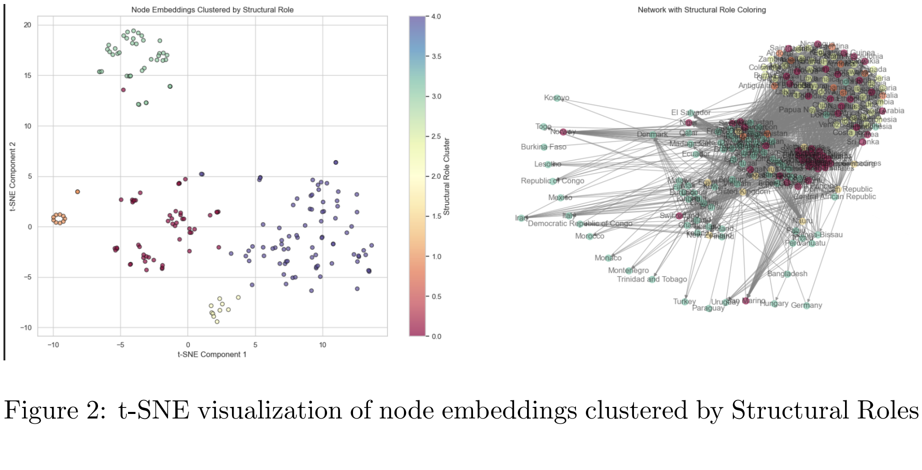

Figure 2: t-SNE visualization of node embeddings clustered by Structural Roles

7 Conclusion

Our experiments demonstrate that incorporating scaled character-based features alongside Node2Vec embeddings significantly improves link prediction performance. This highlights the importance of feature engineering and data preprocessing in machine learning applications.