RMBI4310/COMP4332 Project 2 Report

Group 35: Jamie, Lo Sze Yuen (20619883), NANDA, Nikhil (20491384), LAI, Wai Chun (20510702), LIN Chi Wing ()

Abstract

This project aims to solve the link prediction problem on a subset of the Yelp dataset. The data consists of user_id and friends which correspond to a directed graph, whose edges serve as the labels. Two different random walk based embedding algorithms are compared - 1) DeepWalk and 2) node2vec.

1) DeepWalk

The DeepWalk approach takes fixed-length, unbiased random walks starting from each node. In order to find the values for parameters that maximize the AUC - ROC score on the given dataset, grid search is performed on 3 parameters namely - node dimension (node_dim), number of random walks (num_walks) and length of each walk (walk_length)

node_dim

Grid search has been performed on various values for node_dim including 5, 10, 20, 30 and 40. The Figure 1 heatmap visualizes the AUC - ROC score corresponding to various node_dim values. It can be seen that node_dim value of 10 results in the highest AUC - ROC score.

num_walks & walk_length

Grid search has also been performed on various values for num_walks including 5, 10, 20 & 40 and for walk_length including 10, 20 & 40. The Figure 2 heatmap visualizes the AUC - ROC score corresponding to various node_dim and walk_length values. It can be seen that num_walks value of 20 and walk_length value of 10 results in the highest AUC - ROC score.

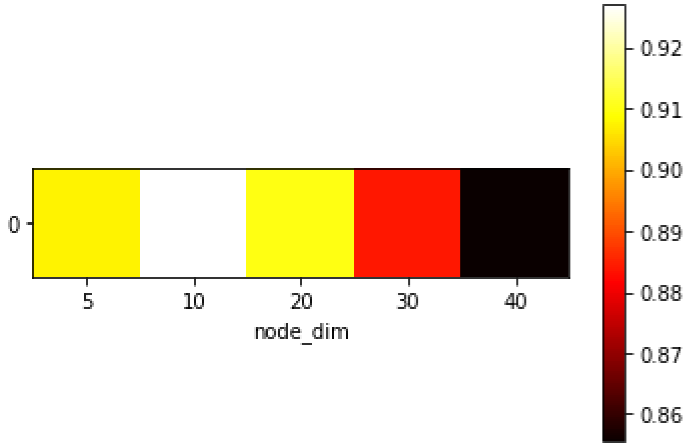

Figure 1: Grid Search on node_dim

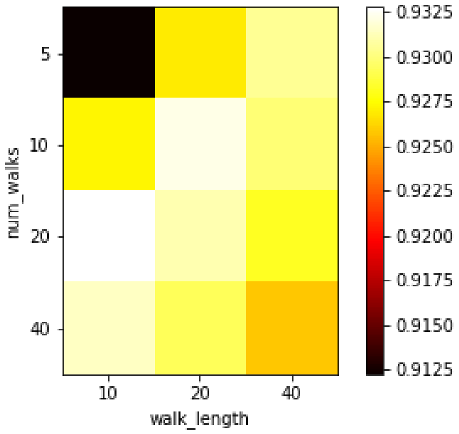

Figure 2: Grid Search on walk_length and num_walks

Final DeepWalk Model

node_dim = 10, num_walks = 20 and walk_length = 10 achieves validation accuracy of 93.23%

2) node2vec

The node2vec approach takes flexible, biased random walks that can trade off between local and global views of the directed graph which is controlled by two parameters - p and q. In order to find the values for parameters that maximize the AUC - ROC score on the given dataset, grid search is performed on 5 parameters namely - node dimension (node_dim), number of random walks (num_walks), length of each walk (walk_length), return parameter (p) and In-out parameter (q)

node_dim

Grid search has been performed on various values for node_dim including 5, 10, 20, 30 and 40. The Figure 3 heatmap visualizes the AUC - ROC score corresponding to various node_dim values. It can be seen that node_dim value of 10 results in the highest AUC - ROC score of 92.78%.

num_walks & walk_length

Grid search has also been performed on various values for num_walks including 5, 10, 20 & 40 and for walk_length including 10, 20 & 40. The Figure 4 heatmap visualizes the AUC - ROC score corresponding to various node_dim and walk_length values. It can be seen that num_walks value of 10 and walk_length value of 20 results in the highest AUC - ROC score of 93.08%.

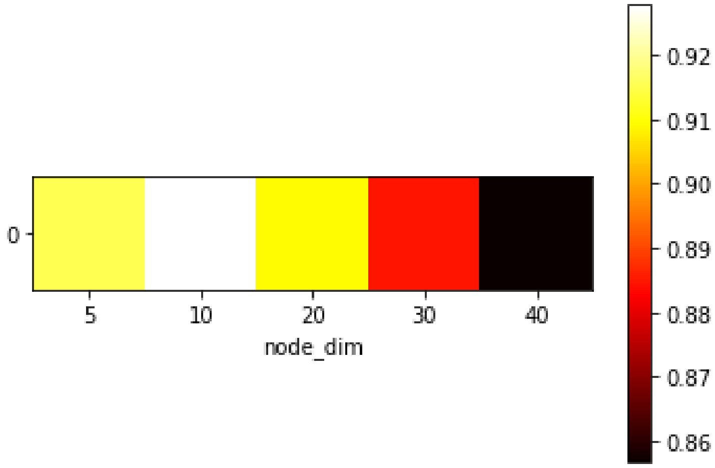

Figure 3: Grid Search on node_dim

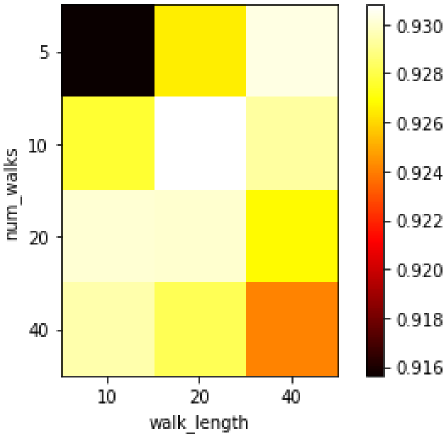

Figure 4: Grid Search on num_walks and walk_length

p and q

Lastly, grid search has been performed on various values for p including   and for q including . From the Figure 5 heatmap, it can be seen that p value of 0.5 and q value of 1.8 results in the highest AUC - ROC score of 93.02%.

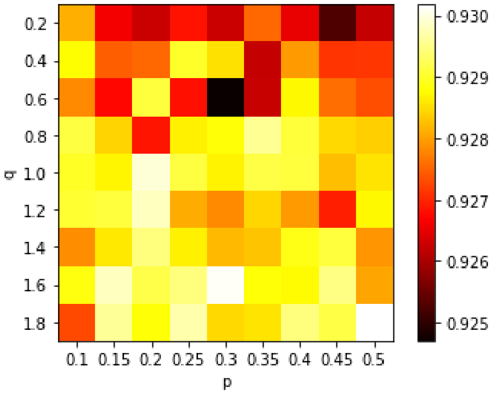

Figure 5: Grid Search on p and q

Final node2vec Model

node_dim = 10, num_walks = 10, walk_length = 20, p = 0.5 and q = 1.8 achieves validation accuracy of 93.21%