भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad

Assignment 1

Identify clusters using embeddings (Node2Vec Embedding, Spectral, and GCN)

submitted by

Akshat Gupta – CS23MTECH11001
Ashish Emmenuel – CS23MTECH11004
Ashutosh Rajput – CS23MTECH11005
Simanta Das – CS23MTECH11018
Patnala Sai Kumar – CS23MTECH14020

CONTENTS

1. Abstract

Embedding is a technique to represent data in a way that captures the features of the original data in a more compact and meaningful form. The concept of embedding is more often used in Machine Learning to represent real world data into mathematical form, so that it is easy to do operations on it.

Here in this assignment we are studying about three such methods of finding embedding for a given graph. We use techniques such as Node2Vec, Spectral, and GCN for finding node embedding for a given graph.

2. Introduction

Problem Statement

We need to identify the clusters present in the data set given to us. We need to find these clusters using different types of embedding like Node2Vec, Spectral, Graph Convolutional Network(GCN) embedding.

Brief Description of these embeddings

1. **Node2Vec :** t continuously learns feature representations of nodes present in the network.
2. **Spectral Embedding :** It does dimensionality reduction in the data using the eigenvalues of the similarity matrix of the data.
3. **Graph Convolutional Network Embedding :** They leverages the structural information in the data to improve performance.

Description of Dataset:

We have been given the dataset **Payments.csv** file which contains the set of transactions between various senders and receivers. Each sender sends some amount of money to the receiver.

The dataset consist of 3 columns

1. Sender
2. Receiver
3. Amount

Each row of the dataset means: sender sends the amount to the receiver.
Ex of row 1 : sender number 1309 sends 123051 amount to the receiver number 1011.

3. Methodology

3.1 Creating Graph From the given data

- We were given a csv file with 3 columns consisting of payments data.
- The first second and the third column consists of Sender Receiver and Amount transacted respectively
- The graph has nodes as Sender and Receiver ids and Amount as edge weight
- For multiple edges from the same Sender to Receiver, we have averaged out the amount and added the same as the edge weight

3.2 Node2Vec

Steps for Node2Vec:

- Generate random walks
- Using the random walks create a skip gram model and get the embeddings
- Dimension for the embeddings is taken to be 8
- After getting the embeddings, using PCA we reduce the dimension to 3.

3.3 Spectral

Steps for Spectral:

To get embeddings using Spectral method, we need to create an Adjecency matrix from the given graph, with the edge weight as the value.

- Create the Adjacency Matrix
- Create the degree matrix
- Create the Laplacian matrix using the formula $Degreematrix - AdjacencyMatrix$
- Calculate the Eigenvectors of the Laplacian matrix, and create a matrixusing the largest three Eigenvectors as columns.
- Now the rows of the newly constructed matrix corresponds to the embeddings

3.3 GCN:

Steps for GCN:

- Create a Directional Graph using the given data.
- Create a GCN class and with Convolutional layers and get the embeddings by traing agains the given graph

3.4 K-means for finding clusters

As the give dataset consists of fraud and non-fraud data, we have taken the no of clusters to be 2

- Using unsupervised methods (K-Means) find the different clusters
- All the different type of embeddings will give different Clusters depending upon their hyperparameters used

4. Results and Discussion

The three figures for the clusters are given below

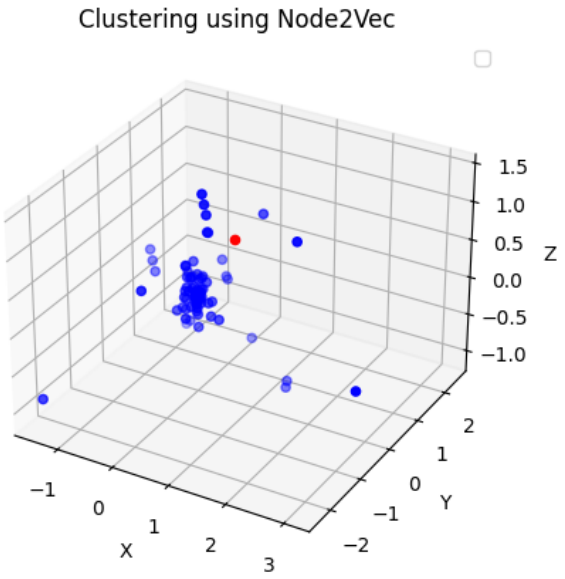

Figure 1: Node2Vec Embeedding

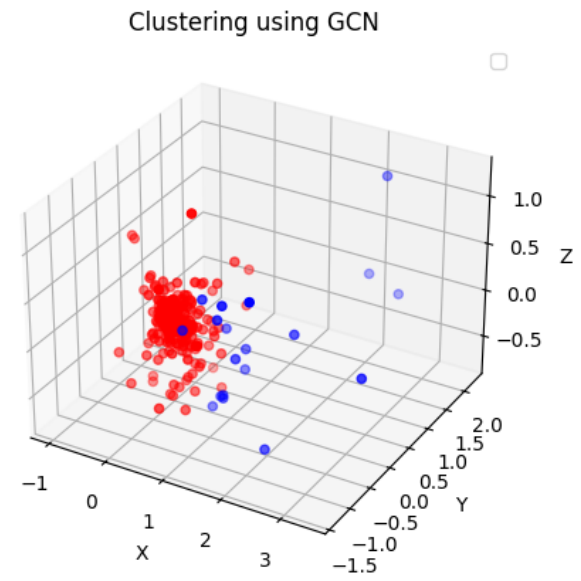

Figure 2: GCN Embedding

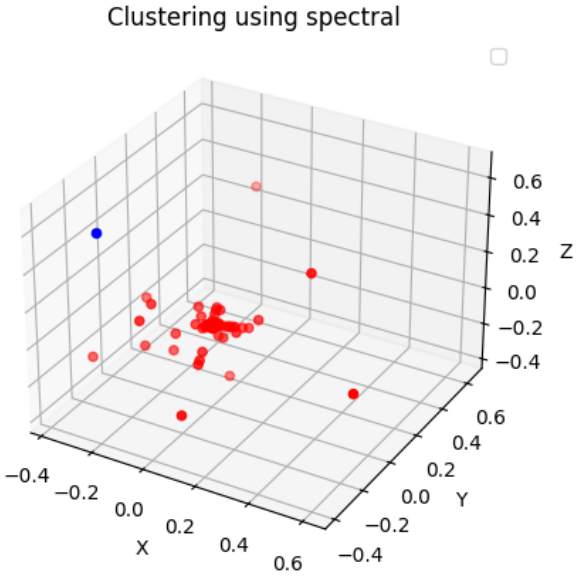

Figure 3: Spectral Embedding

5. Conclusion

After studying the three different method and implementing them to find clusters from a given data, we have come to the conclusion that, different methods capture different essence of the data. No one method is absolutely correct or accurate, but it is up-to the user, which method he uses.

Different problem require different approaches and that can only be achieved by studying the underlying data and having domain knowledge and expertise in it.

6. References

1. CS4786/5786: Machine Learning for Data Science, Spring 2015 03/05/2015: Lecture Notes: Spectral Clustering

2. Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach

3. A Tutorial on Spectral Clustering Ulrike von Luxburg Max Planck Institute for Biological Cybernetics Spemannstr. 38, 72076 T ubingen, Germany ulrike.luxburg@tuebingen.mpg.de