ST JOSEPH'S UNIVERSITY

Classifying of blog in Facebook Graph Machine Learning

Project Report

Submitted by	
222BDA09	JITHIN JOJI
222BDA62	JOHN JEFFERSON.K
222BDA64	JOEMON PAULSON

Introduction

DeepWalk and Node2Vec are two popular graph embedding techniques used for representing nodes in a graph as low-dimensional vectors, which can then be used for various downstream tasks such as node classification, link prediction, and recommendation. To classify Facebook pages with DeepWalk or Node2Vec, you must first represent the pages and their connections as a graph. In this case, you can represent each Facebook page as a node in the graph and draw edges between nodes based on their connections, such as mutual followers or common interests.

Once the graph is built, DeepWalk or Node2Vec can be used to learn low-dimensional embeddings of the nodes, which can then be used for classification. The following are the fundamental steps for employing these techniques:

Create a network of Facebook pages and their connections.

Train a DeepWalk or Node2Vec model on the graph to learn the nodes' low-dimensional embeddings.

The embeddings can be used to categorize Facebook pages based on their content, such as topics or interests.

Literature Review

DeepWalk is a machine learning technique that learns latent representations of vertices in a network, which can then be used for various tasks such as classification and anomaly detection. It is based on the idea of treating truncated random walks as the equivalent of sentences in language modeling and using unsupervised feature learning to obtain latent representations. The resulting representations encode social relations in a continuous vector space, which can be easily used by statistical models.

DeepWalk has been evaluated on several multi-label network classification tasks for social networks such as BlogCatalog, Flickr, and YouTube, and has been shown to outperform challenging baselines which are allowed a global view of the network, especially in the presence of missing information. Its representations can provide F1 scores up to 10% higher than competing methods when labeled data is sparse, and can even outperform all baseline methods while using 60% less training data in some experiments.

One of the advantages of DeepWalk is its scalability. It is an online learning algorithm that builds useful incremental results, and is trivially parallelizable, making it suitable for a broad class of real-world applications such as network classification and anomaly detection.

The task of predicting nodes and edges in networks requires carefully engineered features for learning algorithms. However, recent advances in representation learning have made significant progress in automating prediction by learning the features themselves. Nevertheless, existing feature learning approaches are not expressive enough to capture the diverse connectivity patterns observed in networks.

To address this challenge, the authors propose node2vec, an algorithmic framework for learning continuous feature representations for nodes in networks. In node2vec, a mapping of nodes to a low-dimensional feature space is learned to maximize the likelihood of preserving network neighborhoods of nodes. The authors define a flexible notion of a node's network neighborhood and design a biased random walk procedure to efficiently explore diverse neighborhoods. This algorithm generalizes prior work that is based on rigid notions of network neighborhoods, and the authors argue that the added flexibility in exploring neighborhoods is key to learning richer representations

Methodology

DeepWalk is a method that learns node embeddings by generating random walks on the graph and treating them as sentences in a corpus. The Skip-gram model is then applied to this corpus to learn embeddings that capture the structural information of the graph.

Node2Vec is a similar method that extends the random walk approach of DeepWalk by introducing a biased random walk that balances between exploring the neighborhood of a node and staying within the local community. This allows Node2Vec to capture both the structural and community information of the graph.

Here we used packages

Networkx : NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

Umap : Uniform Manifold Approximation and Projection (UMAP) is a dimension reduction technique that can be used for visualisation similarly to t-SNE, but also for general non-linear dimension reduction.

Seaborn : Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Matplotlib : Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

Pandas : pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool,built on top of the Python programming language

Implementation

Data collection:

Collect data from Facebook pages that you want to classify. This can be done using the Facebook dataset in csv format.

Dataset

Facebook data can be downloaded from :

(https://github.com/benedekrozemberczki/datasets).

We collected data about Facebook pages (November 2017). These datasets represent blue verified Facebook page networks of different categories. Nodes represent the pages and edges are mutual likes among them. The csv files contain the edges - nodes are indexed from 0. We included 8 different distinct types of pages. These are listed below. For each dataset we listed the number of nodes an edges.

Preprocessing:

Preprocess the data by removing any irrelevant information and cleaning the text data. You may also want to extract features such as page likes, post frequency, etc.

Graph creation:

Create a graph from the preprocessed data. Each page can be represented as a node in the graph, and the edges can be defined based on similarities between pages.

Training embeddings:

Use DeepWalk and Node2vec algorithms to train embeddings on the graph. The embeddings will represent each page as a high-dimensional vector.

Model training:

Train a classification model on the embeddings. This can be done using any machine learning algorithm, such as SVM or logistic regression.

Evaluation:

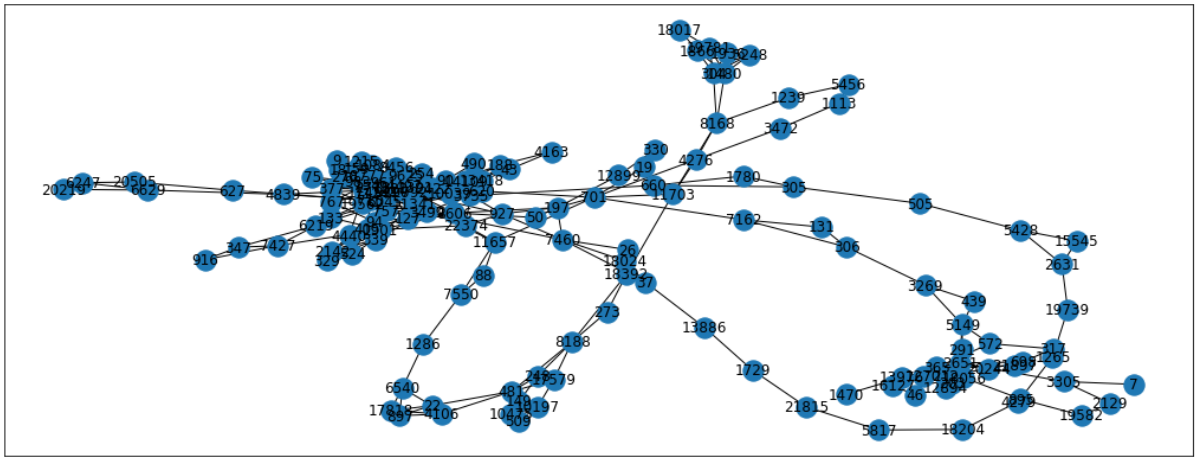
Evaluate the performance of the classification model on a validation set or through cross-validation.

Exploratory Data Analysis:

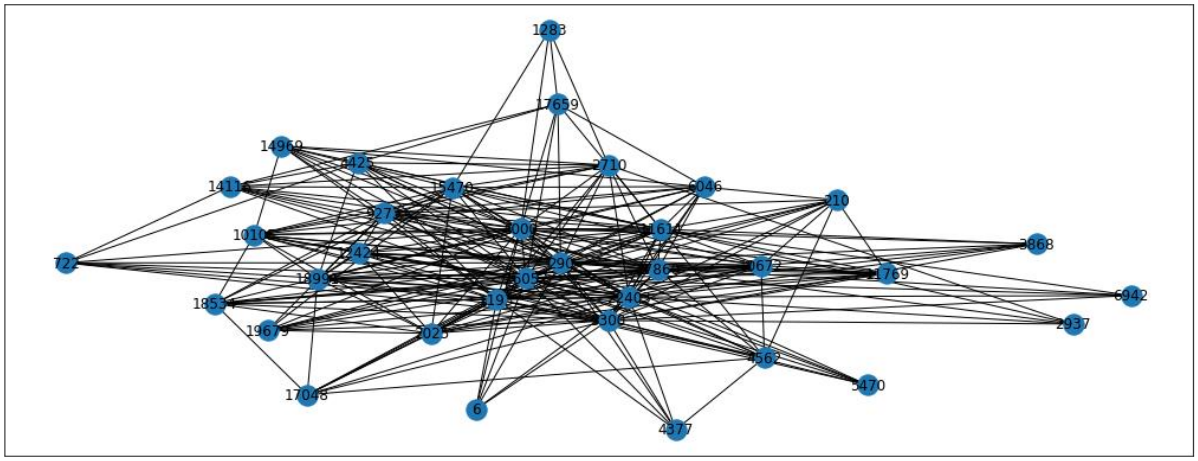
Here we used some plots scatterplot, lineplot, Umap, karateplot

Execution & Results

DeepWalk on Facebook Graph



Node2Vec on Facebook graph



Code

The link to the github repository is given below:

https://github.com/joemon-paulson/Classifying-of-blog-in-Facebook-Graph-Machine-Learning

Conclusion

DeepWalk and Node2Vec are effective graph embedding techniques for categorizing Facebook pages based on their content. These methods learn low-dimensional representations of nodes in a graph, capturing the graph's structural and community information.

We can classify Facebook pages based on their content, such as topics or interests, by using DeepWalk or Node2Vec to learn embeddings of Facebook pages and their connections. The classification can be accomplished by employing appropriate algorithms, such as logistic regression or support vector machines, which take the embeddings as input.

Reference

DeepWalk: Online Learning of Social Representations

Bryan Perozzi, Rami Al-Rfou, Steven Skiena