Your Original and Relevant Course Project Title

Social Network Analysis for Computer Scientists — Course paper

Chenyu Shi
s3500063@umail.leidenuniv.nl
LIACS, Leiden University
Leiden, Netherlands

Shupei Li
s3430863@umail.leidenuniv.nl
LIACS, Leiden University
Leiden, Netherlands

ABSTRACT

KEYWORDS

node2vec, GCN, graph embeddings, social network analysis, network science

ACM Reference Format:
Chenyu Shi and Shupei Li. 2022. Your Original and Relevant Course Project Title: Social Network Analysis for Computer Scientists — Course paper. In *Proceedings of Social Network Analysis for Computer Scientists Course 2022 (SNACS '22).* ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

Graphs are mathematical objects that can model complex relationships on non-Eucildean space. They are widely used in multiple domains such as molecular structure modelling, social network analysis, recommender systems, etc. To leverage the information contained in graphs, it is essential to develop efficient techniques for representing graph-structured data numerically.

Traditional statistical and machine learning methods are designed for extracting features from structured data on Eucildean space. For example, principal component analysis(PCA), uniform manifold approximation and projection (UMAP), and t-distributed stochastic neighbor embedding (T-SNE) are common techniques to reduce dimensions and capture features of data. Although these methods have achieved satisfactory performance on structured-data, they are hard to be generalized to graph-structured data, because they highly depend on properties of Eucildean space.

This challenge led to the development of techniques specifically for graph-based representation. There are two main types of these techniques: shallow embedding method and deep embedding method [5]. Shallow embedding methods use shallow encoder functions to map the original graph structure onto a Euclidean space and obtain the embedding matrix. If data is labelled, we can apply supervised learning algorithms, e.g. label propagation, to extract embeddings later used in a supervised task. However, labels are not available or only partly available in most cases, where we need unsupervised learning or semi-supervised learning to distill information about graph structure. These methods can be divided into distance-based method and outer product method further [5]. Generally, distance-based methods select a metric function that indicates distances between any pairs of nodes and optimize the function to generate embeddings. Representative distance-based methods include multi-dimensional scaling and laplacian eigenmaps. Outer product-based methods use matrix operations to evaluate the similarity between nodes. Most of early studies in graph embedding field adopt matrix factorization to reduce dimensionality of data while preserve the structure information [1]. Another mainstream outer product-based method is inspired by the development in natural language processing. Existing research generalizes the skip-gram word embedding framework to capture the graph embeddings, which has been proved to be efficient on many graph related tasks [6][7]. Node2vec [3], one of algorithms addressed in this paper, is also a variation of skip-gram-based method.

Node2vec is a semi-supervised algorithm whose goal is learning features from networks [3]. It transforms the graph embedding learning into a maximum likelihood optimization problem in a similar way to skip-gram architecture of word embedding learning. Likelihood calculation requires a clear definition of the neighborhood. Textual data has the intrinsic semantic order that can be naturally employed as word neighborhoods. However, graph-structured data has no explicit neighborhoods. Node2vec introduces the idea of the second-order random walk into graph neighborhood sampling strategy. The emphasis of node2vec model is easy to switch between breadth-first sampling (BFS) and depth-first sampling (DFS) by adjusting hyperparameters. Moreover, its computational complexity is less than classical BFS and DFS strategies. Because of its efficiency and great performance on graph embedding learning task, node2vec is an ideal choice among shallow embedding methods.

In recent years, a lot of studies have focused more on deep embedding method rather than the shallow one. Deep embedding method usually refers to algorithms that learn graph features via graph neural networks (GNN). The GNN is a class of artificial neural networks constructed for graph-structured data. Inspired by the success of convolutional neural networks (CNN) on grid data, many GNN architectures have been proposed to generalize the convolution operation on graphs. In this paper, we mainly focus on a method called graph convolutional networks (GCN) [4]. GCNs defines the graph convolution based on the graph Laplacian spectrum. It has achieved state-of-the-art performance on common graph related tasks, such as node classification, link prediction, etc.

We propose a novel method to extract embeddings from graphs in this paper. Our method is based on node2vec and GCNs. Motivated by the concept of meta learning, we regard the embeddings returned by node2vec as the meta information for GCNs. This prior knowledge helps to improve the quality of final graph embeddings, and therefore enhances the model performance in various tasks.

The rest of the paper is organized as follows. Section 2 reviews related works on graph embedding learning methods. We illustrate

SNACS '22, Master CS, Fall 2022, Leiden, the Netherlands

necessary notations, formula, and assumptions in Section 3. And then we describe approaches in detail in Section 4. Section 5 is an introduction to five open source data sets we use in the project. After that, we present our experimental set-up and results in Section 6. The paper ends with a conclusion section.

2 RELATED WORK

Our project draws inspiration both from the skip-gram-based shallow embedding method and the deep embedding method. The following will briefly review existing works related to these methods.

Skip-gram-based methods optimize graph embeddings to predict nodes in the defined context. Actually, this kind of methods leverages the matrix factorization technique implicitly [5]. Compared to early works that explicitly use matrix factorization, skip-gram-based methods are usually more computationally efficient. Deepwalk [6], an algorithm proposed in 2014, is one of pioneering studies that generalizes the idea of skip-gram model in language processing to graph embedding learning field. Deepwalk models the context of nodes by truncated random walks, which is analogous to sentences in textual data. It utilizes the local information obtained from random walks and learns a latent space that corresponds to features of vertices. Following the Deepwalk, LINE algorithm [7] is proposed to address the preservation of network properties and applicability on large-scale networks. The neighborhood sampling strategy of LINE is firstly simulating a BFS-style search for half of the feature dimensions and then a DFS-style search for remaining dimensions. Both Deepwalk and LINE are restricted to a specific sampling strategy of neighborhoods. Node2vec [3] algorithm provides a more flexible option. It introduces a return parameter and an in-out parameter during second-order random walks that enables users to adjust the style of sampling. Its workflow is similar to Deepwalk and can be considered as a generalization.

GNNs have been extensively researched these years. Learning graph embeddings is one of the important applications of GNNs. According to [5], GNN models used to devise graph embeddings are called deep embedding methods. A main challenge in designing GNNs is finding an efficient and easy-to-train filter to process graph signals. One solution is applying graph Fourier transform to define the convolution operation on graphs. However, using a filter directly based on graph Laplacian matrix is computationally expensive. Defferrard et al. [2] suggest to approximate the filter by the k-order Chebyshev polynomial, which has a practical computational complexity. Kipf and Welling [4] propose GCN model that simplifies the approximation further. GCNs limit the order Chebyshev polynomial to one and assume the coefficients in polynomial are equal. Besides, GCNs consider the self-loops in the graph and intrduce the renormalization trick, which improves the quality of learned graph embeddings. So far, GCNs have achieved the great performance on different graph related tasks.

Based on these works, we make the following contributions in the project.

(1) We propose a novel method to learn graph embeddings, inspired by meta learning concept. Our method fully utilizes the information from node2vec and the power of GCN architecture. We test our model on node classification and link prediction task.
(2) We evaluate node2vec, GCN, and our proposed method on five real-word data sets. In node classification task, we use additional metrics to assess the model performance, i.e. accuracy, recall, and weighted f1-score, while the original paper of node2vec only reports macro f1-score.

3 PRELIMINARIES

4 APPROACH

5 DATA

6 EXPERIMENTS

7 CONCLUSION

ACKNOWLEDGMENTS

REFERENCES

[1] HongYun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang. 2018. A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. *IEEE Transactions on Knowledge and Data Engineering* 30, 9 (2018), 1616–1637. https://doi.org/10.1109/TKDE.2018.2807452
[2] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*.
[3] Aditya Grover and Jure Leskovec. 2016. Node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*.
[4] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.
[5] Kevin P. Murphy. 2022. *Probabilistic Machine Learning: An Introduction*. MIT Press.
[6] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*.
[7] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-Scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*.