Project Proposal DS 3002 Data Mining

GitHub Social Networks

Name **Student ID**

Problem and data set description

In this project, we will focus on the problem of binary classification of nodes in a GitHub Social Network. For this purpose, we will work on a dataset of GitHub developers collected in 2019 [1]. This dataset contains an undirected graph. Nodes are basically the developers who have starred at least 10 repositories and edges are mutual follower relationships between nodes. Each node of the graph has multiple features which are extracted based on the location of the developer, repositories starred, employer, and email address of the node. Each feature is coded into an integer list with different lengths.

The aim of this project is to predict if a user is a web developer or a machine learning developer, based on the graph structure as well as the features corresponding to the nodes. The dataset also contains the ground truth, which may help in supervised learning algorithms.

Preliminary ideas on how you plan to address it (models/algorithms/techniques)

Since the features of different nodes can have different lengths in the dataset, the first step of this work is to construct another feature representation of each node that has the same length (dimension) that can facilitate the node embedding process afterward. After analyzing the features, we note that each feature is a subset of the set of integers [0:4044]. Therefore, a direct approach to constructing a new feature vector is to construct a 4005-dimensional vector with elements 0/1 representing the existence of a corresponding element. This will generate a 4005xNv feature matrix. We may then consider using SVD to reduce the dimension of the feature matrix.

After obtaining the feature matrix we will explore multiple machine-learning algorithms which can use structural information for node classification. One of the famous algorithms for node classification in graphs is Graph neural networks. We will explore Graph neural networks for solving our problem. For this purpose, we plan to use specialized TensorFlow-based libraries such as Spectral, StellarGraph, and GraphNets. Additionally, we also plan to explore Graph Convolution networks. The GCN model is based on the graph convolution layer. This layer is like a dense convolution layer that incorporates the adjacency matrix of the graph to use information about the connections of nodes. We will also study and use Graph ATtention Network (GAT) for solving classification problem [2]. Besides this, we will explore GraphSAGE techniques, Inductive GraphSAGE and directive GraphSAGE.

One of the common problems while working on a graph-based dataset is that traditional machine learning algorithms use numerical data. Hence, transforming the structural information of a network into numerical representation is an important task. In this situation, the Node2Vec algorithm helps to translate the nodes of a Graph to an embedding space. This algorithm preserves the structural information. Node2Vec is a famous algorithm for classification problems, and we also plan to explore it [3].

Software tools

We plan to use python for coding. Specifically, we will use libraries including but not limited to NetworkX, Spectral, StellarGraph, GraphNets, and node2vec.

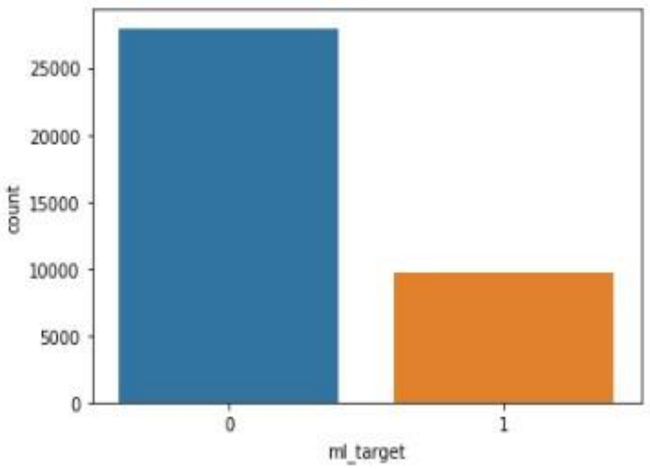
Expected Results and Evaluation

Because this project is a binary classification task, we expect to get the classification results of the GitHub developers and compare the results with the corresponding labels from the dataset. We will evaluate the result of our project through metrics including accuracy, sensitivity, specificity, and F1 scores.

Preliminary Results and Data set explored.

For this project, we have explored the data set used by Benedek Rozemberczki et.al [4] in their research paper related to Multi-scale Attributed Node Embedding. This dataset is based on a homogenous graph where nodes represent the subset of people who use GitHub. The dataset has 37,700 nodes and 289,003 edges. The transitivity of the graph is 0.013 and the density is 0.01. The nodes are binary labeled 0 and 1. The Ids of nodes and the names of users corresponding to different Id numbers are also mentioned.

We performed statistical analysis of data. The targets in the 25th and 50th percentile is 0 and in the 75th percentile are 1. To be exact 27961 targets are denoted as 0 while 9739 are denoted as 1. Hence, this is clearly a class imbalance problem. That is why we will use an evaluation matrix that is suitable for class imbalance problems i.e., sensitivity and specificity.



After analyzing the features for each node, we realized that each feature is a subset of the set of integers [0:4044]. We plan to find an alternative representation of features.

Outline of the work-to-do

1. Construct a 4005-dimensional vector with elements 0/1 representing the existence of a corresponding element.
2. Use SVD to reduce the dimension of the feature matrix (tentative).
3. Apply the graph convolutional neural network to learn the embeddings. After getting the node embeddings, we will use supervised learning to do the binary classification. Specifically, we will use a subset of the nodes and the ground truth to train the classifier. Then, we test our classifier in the rest of the nodes.
4. After that, we will compute the correctness of the classifier by building a confusion matrix.
5. Repeat the same procedure for the rest of the algorithms.
6. We will try Girvan Newman's method as well as some spectral methods mentioned earlier in graph bisection to derive baselines. We then compare our results with these baselines to see if the information beyond structure will help in the classification problem.

Reference

[1] Benedekrozemberczki/datasets: A repository of pretty cool datasets that I collected for
Network Science and Machine Learning Research., GitHub. https://github.com/benedekrozemberczki/datasets#github-social-network (Accessed: February 12, 2023).

[2] Veličković, P. et al. (2018) Graph attention networks, arXiv.org. Available at: https://arxiv.org/abs/1710.10903 (Accessed: February 12, 2023).

[3] Scalable feature learning for networks, node2vec. Available at: https://snap.stanford.edu/node2vec/ (Accessed: February 12, 2023).

[4] Rozemberczki, Benedek, Carl Allen, and Rik Sarkar. "Multi-scale attributed node embedding." Journal of Complex Networks 9.2 (2021): cnab014.

[5] Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016.

[6] Girvan M. and Newman M. E. J., Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)

[7] Newman, M. E. J. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America, 103(23), 8577–82. https://doi.org/10.1073/pnas.0601602103'

[8] M. Fielder, “A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory,” Czechoslovak Mathematical Journal, vol. 25, no. 4, pp. 619– 633, 1975.

[9] L. Hagen and A. B. Kahng, "New spectral methods for ratio cut partitioning and clustering," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 11, no. 9, pp. 1074-1085, Sept. 1992, doi: 10.1109/43.159993.

[10] Jianbo Shi and J. Malik, "Normalized cuts and image segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888-905, Aug. 2000, doi: 10.1109/34.868688.

[11] Chunaev, Petr. "Community detection in node-attributed social networks: a survey." Computer Science Review 37 (2020): 100286.

[12] Ahmed, Amr, et al. "Distributed large-scale natural graph factorization." Proceedings of the 22nd international conference on World Wide Web. 2013.

[13] Yang, H., Pan, S., Zhang, P., Chen, L., Lian, D. & Zhang, C. (2018) Binarized Attributed Network Embedding. In IEEE International Conference on Data Mining.

[14] Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.

[15] Wang, Daixin, Peng Cui, and Wenwu Zhu. "Structural deep network embedding." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016.

[16] Cao, Shaosheng, Wei Lu, and Qiongkai Xu. "Deep neural networks for learning graph representations." Proceedings of the AAAI conference on artificial intelligence. Vol. 30. No. 1. 2016.

[17] Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K. & Tang, J. (2018) Network Embedding as Matrix Factoriza- tion: Unifying Deepwalk, LINE, PTE, and Node2Vec. In International Conference on Web Search and Data Mining.