컨셉을 통한 그래프 신경망 해석

Concept-Aided Explanation of Graph Neural Networks

Sun-Woo Kim, Eun-Bi Yoon, Yong-Min Shin, Won-Yong Shin
Yonsei University

{kswoo9977, yon6286, jordan3414, wy.shin}@yonsei.ac.kr

Abstract

Despite its success in graph data, explaining the result of graph neural networks (GNNs) is challenging due to the complex structure. In this study, we develop **Graph Concept Score Evaluator** (GCSE) to evaluate the model's decision by introducing the notion of 'concept' that captures complex information of graphs in a human-interpretable fashion. Experimental results using a synthetic infection dataset show that our model can successfully evaluate the concepts of a given GNN-based model.

I. Introduction

Recently, graph representation learning using graph neural networks (GNNs) has achieved state-of-the-art performance in various graph-related tasks [1]. Along with the success of machine learning models, explainable AI (XAI) has been considered of great importance for gaining trust and achieving universal usage. Although XAI has been studied in other domains such as images, applications to GNNs remain relatively underdeveloped [2]. Due to the intrinsic complexity of the graph structure, explaining the model's decision is hard to match the human intuition, making it a much more challenging task. In this work, we propose Graph Concept Score Evaluator (GCSE), an XAI method for GNNs to intuitively explain the model decisions.

II. Methodology

We assume that a GNN model for node classification has already been trained and given. We define a ***concept*** as a human-interpretable feature of a target node (and its neighbors), which greatly influences a model's decision. For example, in an infection network, we may regard a concept for an infected person as "the contact to an already infected person". As GNN iteratively merges the node features and topological information for each layer, we consider searching for concepts in the last hidden layer. We also define ***decision*** as the classification result of a model for a given data. From this definition, we would like to provide the following conditions that an important concept must satisfy:

C1: Each node should be explained by at least one concept.

C2: Important concepts should be linearly separable from others.

C3: A single concept should be related to a single class of decision.

Given a set of user-defined concepts to be tested, GCSE evaluates whether the GNN model has learned the user-defined concepts that are important to the model decision. Denoting the set of ***K*** concepts as $C = \{c_1, c_2 \dots, c_K\}$ for each c_i, we divide nodes into two sets V_{c_i} and V'_{c_i}, where nodes in V_{c_i} have the concept c_i, and nodes in V'_{c_i} do not. For node classification, let ***y*** denote the set of node class labels. Based on this, we can calculate the followings:

$\boldsymbol{\tau_i}$ = AUROC of logistic regressor that separates V_{c_i} and V'_{c_i}

$$\boldsymbol{\delta_i} = 4\left(\frac{1}{2} - \frac{\mathrm{n}(\mathrm{c_i} \cap \mathrm{y})}{\mathrm{n}(\mathrm{c_i})}\right)^2, \quad \frac{\mathrm{n}(\mathrm{c_i} \cap \mathrm{y})}{\mathrm{n}(\mathrm{c_i})} = \frac{\text{number of nodes in } V_{c_i} \text{ with label } y}{\text{number of nodes in } V_{c_i}}$$

Finally, we calculate the concept score (***CS***) as $0.5(\tau_i + \delta_i)$ for $0 \le CS \le 1$, which evaluates whether the trained model has learned the concept. High ***CS*** of c_i indicates that the model has learned a certain concept c_i, and it plays an important role in the model's decision.

III. Experimental Results

We train a single-layer GraphSAGE model [1], a well-known GNN architecture to be used for node classification to be explained. We evaluate our GCSE using the synthetic infection dataset [2], which consists of synthetic undirected graphs of 1000 subgraphs with 10~30 nodes each. (See Figure 1). We randomly set 0 to 50% of the nodes in each subgraph as infected nodes and immune nodes each and set the rest as at-risk nodes. Edges are randomly generated; we consider an at-risk node to be infected in the next time stamp if connected to an infected node. The task is to classify whether each node is infected in the next time stamp. Ground truth concepts include 1) initially infected, 2) immune, 3) not immune and contacted no infectious node, and 4) not immune but contacted infectious node(s).

Figure 1. Node features of the infection dataset

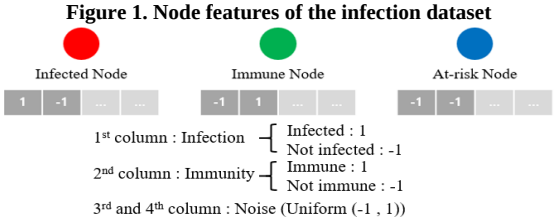

Table 1. Experimental result: CS for concepts

Concept	Ground-Truth	***CS***
Immune	True	0.9999
Infected	True	0.9998
Not immune and contacted no infected node	True	0.9962
Not immune but contacted infected node(s)	True	0.9949
Noise Feature	False	0.3849

Table 1 shows that GCSE can distinguish the ground truth from false concepts. This result implies that GCSE has the capability of distinguishing important concepts from fake concepts. Unlike existing explanation techniques, which directly assign scores to the node features and its neighbors, our method measures the overall score in terms of the user-defined concept, which makes direct comparison infeasible. However, our evaluation demonstrates that our method can successfully validate whether the model has learned the user-defined concept, along with the ***CS*** difference between user-defined concepts and noise concepts.

ACKNOWLEDGMENT

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C3004345), by the Republic of Korea's MSIT (Ministry of Science and ICT), under the High-Potential Individuals Global Training Program (No. 2020-0-01463) supervised by the IITP (Institute of Information and Communications Technology Planning Evaluation), and by the Yonsei University, Republic of Korea Research Fund of 2021 (2021-22-0083).

REFERENCES

[1] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs", In *Proc. Advances Neural Inf. Processing Syst. (NIPS)*, Long Beach, CA, Dec. 2017, pp. 1024-1034

[2] F. Baldassarre, and H. Azizpour, "Explainability techniques for graph convolutional networks." *arXiv preprint arXiv:1905.13686*, 2019.