불확실성을 가진 속성 값 기반 네트워크 임베딩

Attributed Network Embedding with Uncertainty

Yong-Min Shin[1], Cong Tran[1,2], Won-Yong Shin[1]
[1]Yonsei University, [2]Dankook University

jordan3414@yonsei.ac.kr, trancong208@gmail.com, wy.shin@yonsei.ac.kr

Abstract

In real-world graphs, new nodes can come in the underlying graph without any connection to existing nodes. For example, a new user who has just created an account in an online social network. In this study, we exploit the expressive capability of graph neural networks (GNNs) to generate a low-dimensional representation (i.e., embedding) of the new nodes to be used for downstream graph mining tasks. Experimental results using the Cora dataset shows that the embedding generated via our proposed method leads to high-quality link prediction performance.

Ⅰ. Introduction

Recently, graph neural networks (GNN) has gained its popularity for graph embedding (also known as graph representation learning) [1, 2]. We take advantage of both the structure of the given graph and features associated with each node, and generate a low dimensional vector representation in the Euclidian space for all nodes in the graph for graph mining and data denoising. Often in real-world scenarios, new nodes can enter the underlying graph. For example, we can easily imagine a user signing up for an online social network by creating a new account. However, in the case where the user has not made any connections, traditional GNNs cannot generate representations of this new user. In this work, we propose a method to make GNNs generate representations for such new users.

Ⅱ. Methodology

We denote the set of nodes as V_o, set of edges as E_o, and the attribute matrix as X_o, for an unweighted and undirected graph $G_o = (V_o, E_o, X_o)$. As time goes on, a set of new users V_n with its corresponding attributes X_n comes in the underlying graph, where there is no information of edges connected to existing nodes in G_o. As we aim to expand the capability of GNNs, we generate an attribute-based graph G_f from X_o so as to make the model less dependent on G_o. In this context, the embedding space using GNN shall be given by $Z = GNN_\theta(G_f, X_o)$, where θ is the parameter of the GNN model. To make use of G_o, the loss function to train the model takes the form of $\mathcal{L}(Z, G_o) = \gamma log\left(\sigma(Z_i, Z_j)\right)$, where the parameter γ is 1 when (i, j) is a positive edge in G_o and -1 otherwise. During the inference step, we create edges for V_n using the same argument for G_f and run the model once to generate the representation.

Ⅲ. Experimental results

We adopt a two-layer GraphSAGE model [1], which is a well-known GNN architecture, to generate embedding vectors with 128 dimensions. We evaluate our method using the Cora dataset, which is a citation network where the nodes are papers and edges are formed when a citation occurs. First, we mask a portion of nodes in the underlying graph to represent V_n. Then, we create an attribute-based graph G_f via k-nearest neighbor (kNN). After training the model with the proposed method, we generate links for V_n and run the model. We perform link prediction to see the superiority of our method. For comparison, we take into account the following two baseline methods: 1) a logistic regression classifier (LRC), 2) a kNN classifier (kNNC), both of which operate on the attributes and are capable of calculating the edge probabilities for V_n.

Table 1: Experimental results

Method	AUC
LR	0.7806
kNNC	0.7714
Proposed	0.8402

Table 1 shows the experimental results of the link prediction task for V_n. We observe that the proposed GNN-aided method achieves higher performance compared to the two baselines in terms of the area under the ROC curve (AUC). This is because the multi-layer of GNN can effectively aggregate the information of nodes beyond the immediate neighbors. Moreover, the learnt parameters of GNN can effectively model complex interactions within the graph, and thus embed more complex information.

ACKNOWLEDGMENT

This work was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI20C0127), and by the Yonsei University Research Fund of 2020 (2020-22-0101).

REFERENCES

[1] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, December 2017, pp. 1024-1034.

[2] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, April 2017, pp. 1-14.