Motif-Aware Graph Embedding

Hoang Nguyen
Tokyo Institute of Technology
hoangnt@net.c.titech.ac.jp

Shun Nukui
Tokyo Institute of Techonology
nukui.s@net.c.titech.ac.jp

Tsuyoshi Murata
Tokyo Institute of Techonology
murata@c.titech.ac.jp

Abstract

Given a large complex graph, how can we have a lower dimension real-vector representations of vertices that preserve structural information? Recent advancements in graph embedding have adopted word embedding techniques and deep architectures to propose a feasible solution to this question. However, most of these former researches considers the notion of "neighborhood" by vertex adjacency only. In this paper, we propose a novel graph embedding algorithm that employs motif structures into the latent vector representation learning process. Our algorithm learns the graph latent representation by contrasting between different type of motif-biased random walk. We showed that our algorithm yields more accurate embedding results compared to other existing algorithms through various graph mining benchmark tasks.

1   Introduction

The graph (or network) data model is an useful tool for a wide range of disciplines, and it is essential to have a low dimensionality representation of a complex graph. In its simplest form, a graph is an ordered set of vertices connected by edges. Being simple and expressive, the graph-theoric approach has been applied in many scientic fields for untangling complex discrete structures. For example, the study conducted by MP Van Den Heuvel suggested that the human brain functional network provides many new discoveries about our brain's organization (**?**). The same approach for structural analysis can be found in other fields such as chemistry (**?**), physics (**?**), and sociology (**?**). However, the graph analysis process often becomes intractable due to the complexity of the data. In such case, the graph usually contains several thousands to millions of vertices and edges. Therefore, it is desirable to have a compact latent representation of the graph while its statistical properties are retained. A *high quality* latent representation of a graph can benefit machine learning algorithms in many ways. For instance, the result and runtime of machine algorithm will be improved thanks to the low dimensionality of the data. On the other hand, instead of relying only on graph mining algorithms to analyze some given data, researchers can also apply other machine learning algorithms on the learned laten

resspresentation to make predictions. Conventionally, the latent representation learning procedure on graph is called *graph embedding*.

Learning a high quality latent representation is a challenging task. Traditionally, dimensionality reduction technique such as PCA (**?**), CCA (**?**), and IsoMap (**?**) are used. Although these aforementioned techniques have a profound theoretical background (**?**), they are impractical due to computational drawbacks. To address the dimensionality problem, Peperozi et al. proposed Deepwalk - a Skipgram-based algorithm for graph embedding (**?**). Different from tradional linear algebra approach, Deepwalk learns the latent representation from probabilistic point of view. By treating a random series of vertices as if it is a sentence of words, Deepwalk adopts the operation of Skipgram model (**?**) to learn a vertex's latent representation. Following Deepwalk, other Skipgram-based graph embedding algorithms which aim to improve embedding quality were proposed (**?**; **?**; **?**; **?**). However, these algorithms do not consider the intrinsic *motif structure* of a graph. Therefore, the performance of these algorithms depend heavily on heuristic hyper-parameter tunning.

In this work, we propose an algorithm which controls both graph context and negative samples generation. The motif-aware context generation aims to emphasize the importance of local motif community, which is a strong indicator for true community in a graph. On the other hand, negative sampling is known to be the state of the art technique to estimate the normalization factor of a probabilistic model. Instead of sampling from a distorted unigram distrubution as suggested by Mikholov et al. (**?**), we propose a motif-aware method to generate negative samples from a graph. Generally, our algorithm, named Motif-Aware Graph Embedding (MAGE), has two following advantages:

- Positive samples are generated using a motif-biased walk. This motif-aware context generation procedure is backed by the hypothesis that vertices in the same motif are more likely to belong to the same community (**?**; **?**). By selecting the appropriate motif for each graph, we have a sensible way to control the context generated for embedding.

- Negative samples are also generated using a motif-biased walk. However, by choosing the *opposing* motif to the characteristic motif of the graph, we have a concrete

method to "escape" the motif community, which leads to a truely contrastive negative samples, hence better quality embedding.

Our algorithm is implemented on Keras (**?**) framework. The implementation and experimental results are available on Github. (TODO: ADD FOOTNOTE).

The remaining of this paper is divided into 4 parts. Section 2 provides additional information about related work on graph embedding and graph motif. Section 3 presents our algorithm and experimental design. Section 4 and 5 discusses results and conclusion.

2 Related work

2.1 Skipgram model

Representation learning has been one of the key to the success of machine learning algorithms (**?**). In the context of natural language processing (NPL), representation learning becomes even more important as the data has a discrete, but high-dimension nature. To address the dimensionality problem in NLP, Mikholov et al. have proposed the Skipgram model (**?**). Instead of maximizing the n-gram distribution as prior works, Skipgram maximize the occurence probability of context words given a target word. The softmax potential function for a context word given a target word is given by:

$$\Pr(v_c|v_t) = \frac{\exp\left(\langle \omega_{v_c}, \omega_{v_t} \rangle\right)}{\sum_{k \in V} \exp\left(\langle \omega_{v_k}, \omega_{v_t} \rangle\right)}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is vector dot product; v_c and v_t are the token for the context word and the target word respectively; ω_{v_t} is the embedding vector selected from the *embedding matrix* by the token v_t; ω_{v_c} is the embedding vector selected from the *context matrix* by the token v_c.

2.2 Context representation importance

The Skipgram model proposed by Mikolov et al. (**?**) is a powerful model in natural language processing. Under the Distributional Hypothesis (**?**), the Skipgram model maximizes the occurent probability of context words given a target word.