Representation Learning and its Applications in a Heterogeneous Interactions Network

Arunav Saikia, arsaikia@iu.edu
M.S. Data Science *(Residential)*
Indiana University, Bloomington

1 Abstract

In this work, we plan to implement metapath2vec, a meta-path based representation learning technique that uses a modified skip-gram model to learn latent d-dimensional representation of nodes in a user-music heterogeneous interactions network. We will show that metapath2vec embedding can be used for heterogeneous network mining tasks like node classification, similarity search and it outperforms the traditional state of the art representation learning technique like Node2vec which is designed specifically for homogeneous networks.

2 Introduction

Representation learning can be leveraged to encode the structure and semantics of the rich and complex data present in social and information networks. Advances in natural language processing (NLP), specifically a group of models called word2vec [1] has lead to word2vec inspired network representation learning frameworks such as DeepWalk [2], LINE [3], and node2vec [4]. These methods enable automated discovery of low dimensional meaningful features from raw networks by using a random walk based neighborhood sampling strategy followed by a skip-gram model to encode structurally and semantically similar nodes. While these frameworks work well for homogeneous networks, most social and information networks are heterogeneous in nature with diverse node types and edges between them. An example of such a network is an interactions graph representing users, songs they listen to and artists/bands who wrote these songs.

metapath2vec [5] is a scalable representation learning framework that tries to solve the issues arising from the homogeneous treatment of diverse node types and edges in heterogeneous networks. It learns latent feature representation of nodes by generating meta-paths based biased neighborhoods and leveraging the skip-gram model that maximizes the probability of having a heterogeneous context. The learned features can be then used for network mining tasks such as node classification, community detection, and similarity search.

3 Data and Methods

For this work, we plan to acquire user-music interactions data from Taste Profile Dataset [6]. It is a collection of 1,019,318 unique users, 384,546 unique songs for a total of 48,373,586 user-song interactions. For computational efficiency, we plan to subset this data to a manageable size ($\leq$100k nodes and edges). We also plan to augment the nodes with metadata which can be pulled from Echo Nest API. From a network standpoint, the heterogeneous nodes in this network are users(U), songs(S) and artists(A).

The different experiments we plan to run to test the efficacy of metapath2vec are as follows -

- What is the right meta-path for this network?
- Decompose the heterogeneous network of users-songs-artists into individual songs/artists network(s), investigate centrality measures and perform community detection.
- Use the metapath2vec embeddings to perform multiclass node classification and try to predict the artist for each song. Gauge performance of these embeddings across different classifiers. Compare results with node2vec embeddings. Does adding metadata like danceability, tempo, etc as additional features to the latent vector enhance the performance of the classification task?
- Perform clustering on artist nodes and validate clusters based on 3rd party labels such as genre. Is metapath2vec able to cluster artists of the same genre?
- Use embeddings to perform similarity search and recommend similar songs to users. Validate recommendations by looking at historical user behavior in the network.

- Look at the sensitivity of these results by varying hyperparameters of the algorithm like walks per node, walk length, dimensions and neighborhood size
- Use tensorflow embedding projector [7] to visualize if the embedding vectors can implicitly learn semantic similarities between nodes and internal relationships between different types of nodes.

References

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781 (2013). http://arxiv.org/abs/1301.3781

[2] Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 701-710. 2014

[3] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding.. In WWW '15. ACM.

[4] Aditya Grover and Jure Leskovec. 2016. Node2Vec: Scalable Feature Learning for Networks. In KDD '16. ACM, 855–864.

[5] Dong, Yuxiao, Nitesh V. Chawla, and Ananthram Swami. "metapath2vec: Scalable representation learning for heterogeneous networks." In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 135-144. 2017.

[6] The Echo Nest Taste profile subset, the official user data collection for the Million Song Dataset, available at: http://millionsongdataset.com/tasteprofile

[7] Smilkov, Daniel, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg. "Embedding projector: Interactive visualization and interpretation of embeddings." arXiv preprint arXiv:1611.05469 (2016).