INDIANA UNIVERSITY
BLOOMINGTON

ENGR-E536 Graph Analytics

Studying the impact of Spare Embedding on Node Classification and Link Prediction

Project Proposal

by

Anirudh Penmatcha
Nischal Bangalore Krupashankar
Sathya N C

Luddy School of Informatics, Computing, and Engineering
Indiana University Bloomington

1 Objective of the project

The primary objective of our project is to study the impact of Sparse embeddings on graph based machine learning tasks like node classification, clustering and node classification. Identifying specific parameters to study the effects of such embeddings will shed light on this aspect of graph analytics, aiding researchers in developing insights into various ways graphs can be represented and manipulated. We hypothesize that using such embeddings would result in a moderate decrease in accuracy, balanced by a corresponding increase in execution speed.

Such an outcome would be favorable in application where researchers value the increase in speed over the moderate to minimal decline in accuracy.

2 Method

The project can be divided into two distinct parts: a) Creation of sparse embeddings b) Quantifying the effects of these embeddings for various ML tasks.

The creation of embeddings can further be broken down into two techniques. In the first technique we propose creating shallow embeddings using standard algorithms like DeepWalk and Node2Vec. The second technique will focus on generating shallow embeddings by using Graph Neural Networks. A new sparseGNN algorithm can be modelled which would sparsify the embeddings in between layers while generating the computational graphs.

Once such embeddings are obtained, they can be tested on a range of different tasks. A few important things to consider here would be the size of the embeddings, the performance on various tasks and lastly the efficiency (quantified by execution speed). All three parameters can be carefully iterated over to understand the effects of shallow embeddings, which can in turn be used to understand their significance. To validate any gains we will be using dense embeddings as the baseline.

3 Expected results

Due to the exploratory nature of the project, it is difficult to accurately ascertain the exact outcomes. That said, we do have some deliverables in mind:

1) Understand how sparse embeddings affect various ML tasks.
2) Develop a way to balance the theoretical speed improvements versus the decline in accuracy.
3) A novel sparse graph neural network that can be leveraged to work with sparse embeddings.

4 Dataset

Our project involves experimentation on the efficacy of a technique that can be replicated across different types of graphs. Due to this reason, it is not feasible to list down a particular dataset that would be optimal to use. The effects of sparse embeddings can be tested across the different types of graphs including synthetic graphs and real world graphs.

We could employ techniques such as Erdos-Renyi graph generation, Kronecker graph product etc. for generation of synthetic graphs of desired sizes. As for real world graphs, we will be using the Stanford Network Analysis Project. (Leskovec and Krevl, 2014)

References

Leskovec, J. and Krevl, A. (2014). SNAP Datasets: Stanford large network dataset collection http://snap.stanford.edu/data.