Comparing node embedding methods and classifiers for predicting disease genes

1 Introduction

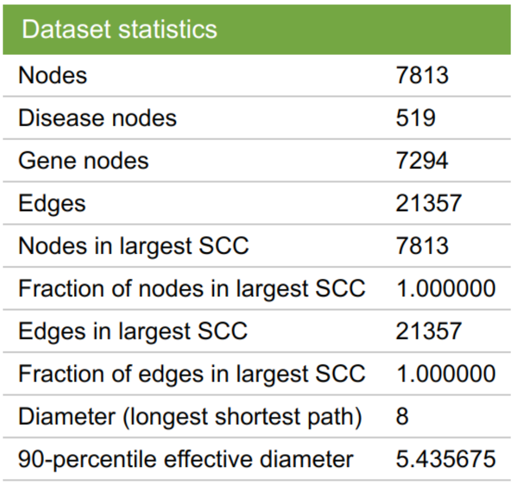
The human genome is the entire set of DNA instructions found in a cell. The term genome refers collectively to the DNA and associated protein molecules contained in an organism or a cell. The genome consists of 23 pairs of chromosomes. A gene is a specific sequence of DNA and is actually the functional unit of inheritance. Most genes contain the information needed to make a protein, or molecules that carry out all of a cell's vital activities. Therefore, slight variations in genes lead to slight changes in a protein. Although some human diseases are explained by alterations in a single gene or of a single chromosome, most are complex and may involve multiple genes and protein pathways. For these reasons, one of the most difficult problems is to find out how genes contribute to diseases.

2 Motivation

Understanding and predicting gene-disease associations is crucial for the prevention, diagnosis, and treatment of diseases. Thanks to the research and studies conducted on the human genome, a lot of data are now easily accessible and interpretable and are widely used for discovering unknown interactions between genes as well as the relationships of these with chemicals and diseases. In order to facilitate the prediction task, several techniques have been used to encode the input graph into data structures that preserve the original local and global properties while lowering its dimensionality. The key feature of graph embedding is the consistency of node embedding distances with respect to the original node ones, thus resulting in the node distances themselves embedded. Graph embedding methods can be mainly divided into three categories: random walk-based, factorization-based, and deep learning-based. The prediction module instead, typically involves machine learning classifiers such as Random Forests, Support Vector Machines, or an ensemble of these. In state-of-the-art algorithms, deep learning classifiers such as Convolutional Neural Networks (CNNs) can also be found. Some Deep learning-based graph embedding such as Graph Neural Networks (GNNs) overcome the need for a classifier by both embedding and classifying the input graph in a standalone manner. In this work, we evaluate several combinations of node embedding techniques along with different classifiers to compare the results in the disease genes predictions in terms of multiple metrics.

2.1 Datasets

Although the initial dataset intended for this study was DisGeNet, a graph comprising more than 15 million gene-disease associations as edges, we switched to a smaller version of this dataset due to the huge time involved in computing the embeddings and training the classifiers. The smaller version of DisGeNet turned out to be easier to manage, comprising less than 22 thousand edges, but still informative in terms of the quality of the predictions.



Dataset statistics	
Nodes	7813
Disease nodes	519
Gene nodes	7294
Edges	21357
Nodes in largest SCC	7813
Fraction of nodes in largest SCC	1.000000
Edges in largest SCC	21357
Fraction of edges in largest SCC	1.000000
Diameter (longest shortest path)	8
90-percentile effective diameter	5.435675

Figure 1: DisGeNet (smaller version), source: `https://snap.stanford.edu/biodata/datasets/10012/10012-DG-AssocMiner.html`.