Predicting Drug-Gene Interactions via Graph Structure

Kayla Bennett, Zachary Abrams, and Jeff Kinne

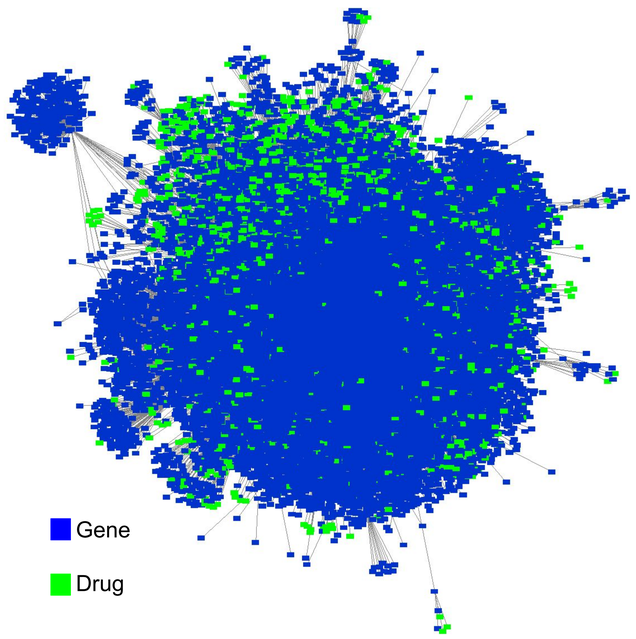
Indiana State University

Motivation

The goal of this project is to predict unknown interactions between known genes and druglike molecules. Predicted interactions can then be explored for validity and usefulness by other researchers.

Network Data

- Gene-gene interactions from BioGRID
- Drug-gene interactions from dgiDB
- Removed interactions without Entrez/ChemBL identifiers
- Treated resulting data as network edgelist
- Removed nodes not connected to the main graph



Above: interaction network displayed by Cytoscape

The node2vec Algorithm

- Represents network nodes as lower dimensional vectors
- Generalized from word2vec model for chain graphs
- Randomly walks graph to approximate similarity between nodes
- Encodes similarity between nodes as cosine between vectors
- Embeddings can be tweaked by several hyperparameters:
  - d - number of dimensions
    [16, 32, 64, 128]
  - p - return to previous node
    [0.25, 0.5, 1, 2, 4]
  - q -Walk away from (DFS) or around (BFS) source
    [.25, 0.5, 1, 2, 4]

Generating Edge Embeddings

- Embed nodes with node2vec
- Convert pairs of node vectors to edge vectors using binary operator
- 4 binary operators tested:
  - L1
  - L2
  - Hadamard
  - Average
- Best performance: L2 operator

Optimizing Performance: Grid Search

Grid Search: Best Hyperparameters			
d	**p**	**q**	**AUC score**
16	0.50	0.25	0.877738
16	1.00	0.50	0.877185
16	0.25	0.25	0.875204
16	4.00	0.50	0.874753
16	0.50	0.50	0.873145

Evaluating Performance

Using the best set of parameters above, I attained the following confusion matrix and predictions

Model Evaluation: Confusion Matrix		
	Actual True	**Actual False**
Predicted True	35626	12545
Predicted False	7217	40954

Interactions Predicted by Classifier				
CHEMBL ID	**Entrez ID**	**Drug Name**	**Gene name**	**Probability**
CHEMBL1200790	57624	METHYPRYLON	NYAP2	0.9999999745
CHEMBL861	729085	MEPHENYTOIN	FAM198A	0.9999999476
CHEMBL452	3034	CLONAZEPAM	HAL	0.9999999441
CHEMBL1522	3680	ESZOPICLONE	ITGA9	0.9999999168
CHEMBL591	83990	DAP000163	BRIP1	0.9999999155
CHEMBL591	14873	DAP000163	Gsto1	0.9999999032
CHEMBL452	26532	CLONAZEPAM	OR10H3	0.9999998378
CHEMBL1213252	1748	CLORAZEPATE	DLX4	0.9999998086
CHEMBL1213252	29785	CLORAZEPATE	CYP2S1	0.9999997669
CHEMBL285674	55862	ESTAZOLAM	ECHDC1	0.9999997263

Possible Improvements

- Different performance metrics
- K-Fold cross-validation
- Concatenating Additional Features

Conclusion

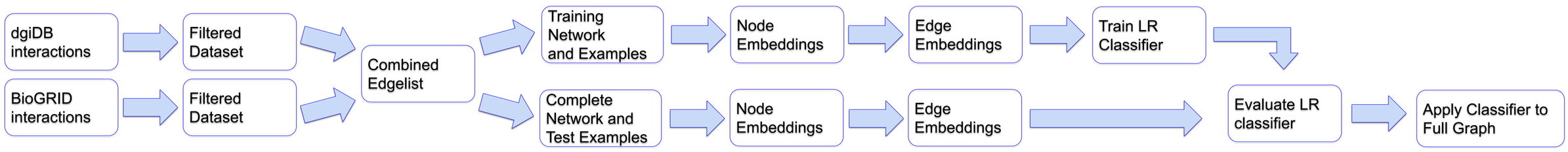
While this data does not achieve the performance detailed in the original node2vec paper, it indicates performance better than chance. The results of this project corroborate the potential of link prediction in biological networks using node embeddings generated by node2vec.

References

- Grover, A. & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (p./pp. 855--864).
- Stanford Network Analysis Package (http://snap.stanford.edu)

Acknowledgements

- NIH R25-MD011712-04: Big Data for Indiana State University (BD4ISU))
- Kevin Coombes (Ohio State University Department of Biomedical Informatics), Yan Zhang (United States Food and Drug Administration)



Indiana State University