NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

URECA
Undergraduate Research Experience on CAmpus

Disease Module Detection From Machine Learning Approach

Objective

The purpose of the research is to first evaluate and cross compare the state-of-art module detection algorithm and second to introduce a new way to detecting modules especially disease modules using machine learning techniques.

Introduction

There are handful of module detection algorithms in the previous studies. Two types namely Hierarchical based and greedy based are specifically studied. A cross comparison result on their advantages and disadvantages are shown clearly in the next section.

Inspired by an integrative drug target detection method[1] based on multi-relational association mining, the method described in this poster takes all topological features, sequential features and functional features into consideration and will be tested using several mature machine learning techniques.

Methodology

Data Set

1. Mendelian Disease Dataset[2] (with labelled target): known disease genes, essential non-disease genes and non-essential non-disease genes

Feature Creation

As the conventional algorithms can be biased to detect module based on topological features only, for example, clusters with higher modularity may not function together[3], thus we use the integrative method of combining the following features.

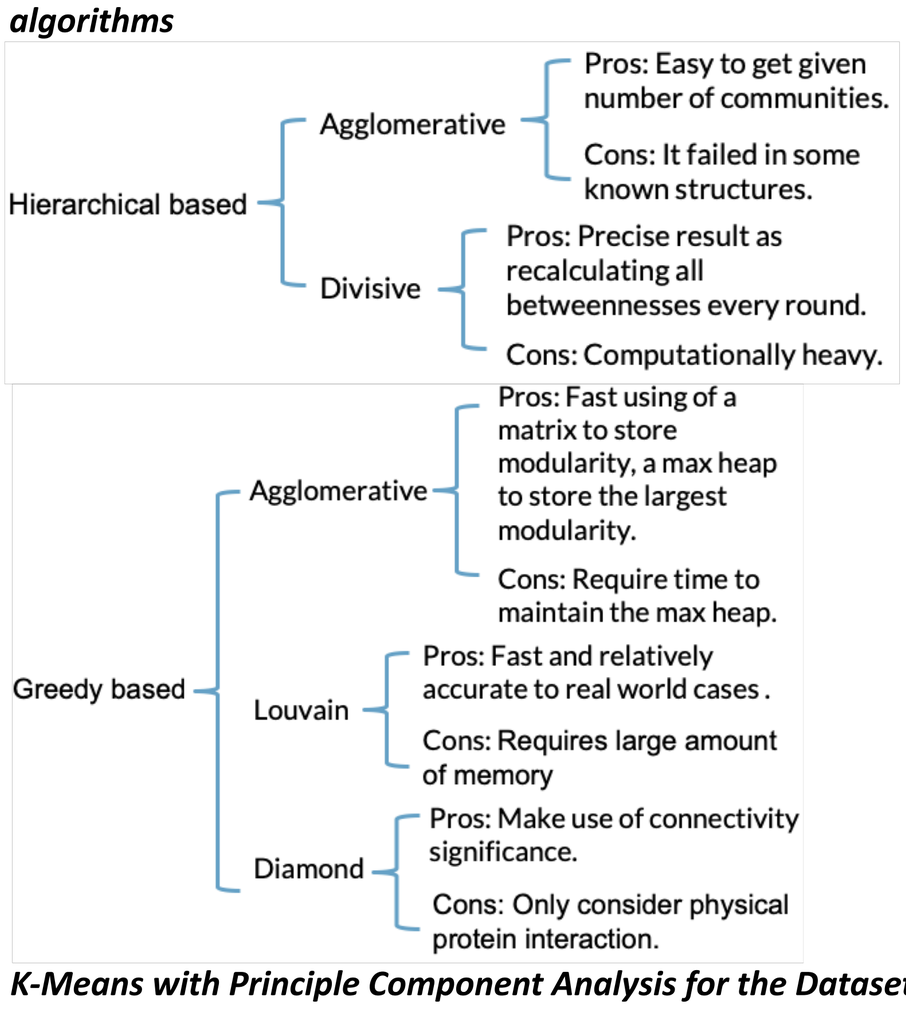
Topological Features	Sequential Features	Functional Features
Average shortest path	Amino acids percentage	Semantic Similarity among Gene Ontology terms based on Biological Processes
Local clustering coefficient	Aromaticity	Semantic Similarity among Gene Ontology terms based on Cellular Components
Centrality: degree centrality, closeness centrality, betweenness centrality, eigenvector centrality, PageRank, harmonic centrality (implemented using Python NetworkX package)	Isoelectric	Semantic Similarity among Gene Ontology terms based on molecular functions
Modularity	SS fraction: helix, turn, sheet	

Machine learning model

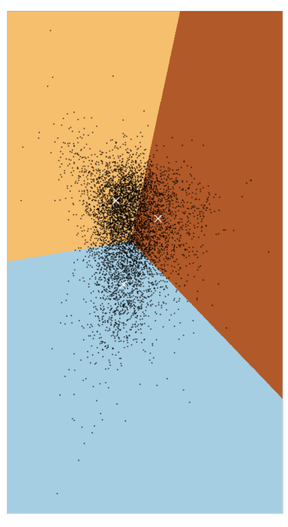
1. K-Means Clustering with Principle Component Analysis
2. Support Vector Machine with nested cross validation
3. Random Forest

Result & Discussion

Comparison between two types of module detection algorithms

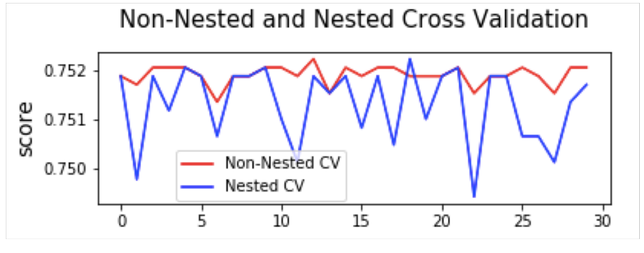


K-Means with Principle Component Analysis for the Dataset



	Disease Module 1	Disease Module 2	Disease Module 3
F1 score	0.35	0.29	0.19
Accuracy	0.54	0.63	0.58

Support Vector Machine on the Dataset



The highest accuracy score(0.75) occurs with nested cross validation.

Conclusion

Through observation, we find that Support Vector Machine provides a much higher F1 score than K-Means with PCA. Next step is to first tuning the parameters using Support Vector Machine and then study the performance by using other machine learning models.

References

[1] Nguyen, T. P., Priami, C., & Caberlotto, L. (2015). Novel drug target identification for the treatment of dementia using multi-relational association mining. *Scientific reports*, *5*, 11104.

[2] Spataro, N., Rodríguez, J. A., Navarro, A., & Bosch, E. (2017). Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. *Human molecular genetics*, *26*(3), 489-500.

[3] Ghiassian SD, Menche J, Barabási A-L (2015) A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. PLoS Comput Biol 11(4): e1004120.

Category: Engineering and Technology
School of Computer Science and Engineering
Supervisor: Prof Rajapakse Jagath Chandana

Student: Li Mengyang
Project ID: SCSE18042
Collaborators*/Co-supervisors*: Dr. Rama Kaalia