Research Progress

Nicolás López

July 17th, 2020

PIPPL3:
Protein-Protein Interaction Predictor using Length-3 paths information

1. Models of interest are:
   (a) L3 results
   (b) A3 Results
   (c) CN Results
   (d) L3+`Node2Vec` results
   (e) A3+`Node2Vec` results
   (f) `Node2Vec` results
2. In the original paper (Barabasi et al.), the Human Interactome is used. It is recommended to test the model on both Human and Rice Interactomes.
3. The following sampling procedure was established:
   - `r = 0.90` (10% edge removal from the graph $G = (V, E)$)
   - if R is the set of edges removed from E, then R' edges are chosen randomly such that $R' \in E$, $R \cap R' = \emptyset$ and $|R| = |R'|$ (balanced dataset for training and validation).
   - `valid_percent = 20` (20% of dataset for validation)
   - `test_percent = 20` (20% of dataset for test inside training, after extracting `valid_percent`)
   - the remaining dataset after extracting `validation` and `test` is used for `training`.

Work done:

- (1.) Progress
    1. L3 results (DONE but *weird*)
    2. A3 Results (DONE but *weird*)
    3. CN Results (DONE but *weird*)
    4. L3+`Node2Vec` results (DONE)
    5. A3+`Node2Vec` results (DONE)
    6. `Node2Vec` results (**pending**)
- (2.) **Pending** calculations for the Human Interactome for the models with Node2Vec
- (3.) The problem with that procedure is that a lot of random pairs seem to be VERY unlikely and therefore, the ML algorithm predicts very well such outliers. I propose a new scheme for finding R' (to be tested)
  After executing **L3**, **A3** and **A2** predictions, R' are the edges in those predictions that don't belong to G.