Supplementary for Heterogeneous Networks Integration for Disease Gene Prioritization with Node Kernels

December 20, 2019

1 Model Selection

DiGI is parametrized by four integers and a real valued hyper-parameter. The network decomposition has two parameters: the degree threshold D and the clique size threshold C. The CDNK node kernel has two parameters: maximal radius r^* and maximal distance d^*. Finally the SVM classifier has a regularization parameter c. To identify the best DiGI model we performed model selection using a 3-fold cross validation strategy on the training set (i.e. only information about the training set is used for fitting on two thirds and estimate the best parameters on the remaining third). Given the 3-fold strategy, for each disease, we then perform a grid search in the following parameter space: $c \in \{10^{-4}, 10^{-3}, 10^{-2},\ 10^{-1}, 1,\ 10,\ 10^2,\ 10^3,\ 10^4\}$, $D \in \{10, 15, 20\}$, $C \in \{4, 5\}$, $r^* \in \{1, 2\}$ and $d^* \in \{2, 3, 4\}$

2 Runtime

We report empirical runtimes in hours for both evaluation settings. In all cases we have degree threshold $D = 15$, clique threshold $C = 4$, maximal distance $d^* = 2$, maximal radius $r^* = 2$, SVM regularization trade off factor $c = 0.1$. In the cross-validation setup we use data from HPRD, BioGPS and Pathways. As an example we report results for the disease-gene association *Skeletal* which includes 35 positive and 17 negative genes. The runtime for the creation of graphs is 0.25, for the graph decomposition and union is 1.96, the feature computation takes 0.07, training the model and computing the predictions using leave-one-out cross validation takes 0.004. For the unbiased evaluation, we use String and Phenotype. For the 42 novel associations, the runtime for the creation of graphs is 0.53, graph decomposition and union is 0.12, kernel computation is 0.01, training models and predicting is 0.002.

3 Graph Node Labeling

In [2], the authors propose two labeling functions that employ gene ontology (GO) [3] on to label for nodes in gene networks in both cases: discrete and real vector labels.

Discrete labels: Using a gene ontology (GO) [3], first we construct a binary vector representing a bag-of-terms for each gene. Next the k-mean clustering is performed on the resulting vectors to cluster genes into a pre-defined number of clusters. Finally, the cluster identifier is used as the discrete label for all genes belonging to that cluster.

Real vector labels: Similar to the procedure used to generate discrete labels, here we first cluster genes into C clusters based on their binary vectors. For each gene, we calculate its similarity with each of the central points of all C clusters. Then the obtained similarity vector is considered as the real vector for that gene. Formally, given a vector $v \in \mathbb{R}^k$ we compute a similarity vector $S(v) = s_1, s_2, \ldots, s_C$ with entries $s_i = \frac{1}{1+\ell(v,c_i)}$ where $\ell(v, c_i)$ is the Euclidean distance of v from the center of the i^{th} cluster $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$.

4 Experimental results

In Table 1, we report the detailed results obtained for each of the 12 disease-gene classes: Ophthamological, Connective, Endocrine, Skeletal, Metabolic, Cardiovascular, Dermatological, Renal, Hematological, Immunological, Muscular, Developmental.

Apart from the common used metrics, auc-roc and auc-pr, we also report the performances of different methods using following metrics:

- fmax: $fmax = \max_s \{2\frac{pr(s).rc(s)}{pr(s).rc(s)}\}$, in which pr and rc are precision and recall, respectively.
- p-qscores: the average rank of positive genes over all candidates
- rank: the order of a method in the descending rank list sorted based on methods' performances, 1 for the best.

Table 1: Predictive performance comparison for a leave-one-gene-out cross validation setting for disease gene prioritization on 12 disease-gene associations problems with 2 modern information sources. For each case, we report not only the performance, but also the rank of each method comparing with others. In bold the best performance. (p-qscores is the average rank of positive genes over all candidates)

Disease class	Measure	Scuba [4]	SmuDGE [5]	NSPDK [6]	DiGI
Ophthamological	fmax	71.19/1	68.94/3	68.66/4	70.97/2
	auc-roc	72.19/1	71.76/2	68.23/4	70.39/3
	aupr	69.98/4	73.86/1	70.09/3	73.01/2
	p-qscores	41.56/4	29.5/1	40.49/2	41.18/3
Connective	fmax	69.8/3	66.67/4	70.40/2	71.19/1
	auc-roc	69.42/2	51.17/4	68.34/3	74.52/1
	aupr	73.95/2	52.41/4	69.93/3	79.92/1
	p-qscores	32.34/3	40.95/4	31.22/2	30.67/1
Endocrine	fmax	70.87/3	69.00/4	70.93/2	75.76/1
	auc-roc	73.78/2	63.84/4	70.46/3	78.40/1
	aupr	77.93/2	63.89/4	72.39/3	80.72/1
	p-qscores	34.09/4	26.64/2	32.25/3	26.28/1
Skeletal	fmax	77.50/2	69.82/4	76.19/3	81.13/1
	auc-roc	77.20/3	68.97/4	80.35/2	84.24/1
	aupr	80.00/3	68.39/4	80.46/2	83.81/1
	p-qscores	29.07/4	19.31/2	28.90/3	17.21/1
Metabolic	fmax	77.29/3	69.96/4	80.41/2	83.24/1
	auc-roc	83.57/3	70.73/4	85.85/2	87.01/1
	aupr	83.67/3	69.44/4	86.51/2	87.68/1
	p-qscores	23.38/3	30.89/4	18.97/2	17.96/1
Cardiovascular	fmax	66.95/4	67.59/3	70.42/2	71.14/1
	auc-roc	63.47/3	63.42/4	74.75/2	77.19/1
	aupr	69.83/3	60.74/4	78.52/2	80.76/1
	p-qscores	37.79/4	27.61/2	29.60/3	26.81/1
Dermatological	fmax	75.36/3	80.73/1	67.80/4	79.19/2
	auc-roc	77.80/3	84.30/2	64.82/4	85.44/1
	aupr	80.69/3	81.25/2	65.73/4	82.79/1
	p-qscores	27.24/3	14.95/1	33.06/4	21.05/2
Renal	fmax	75.51/2	82.17/1	72.55/3	72.07/4
	auc-roc	79.52/2	88.58/1	72.93/3	71.96/4
	aupr	77.10/2	88.33/1	75.56/3	71.49/4
	p-qscores	28.13/3	14.32/1	33.76/4	25.84/2
Hematological	fmax	68.63/3	69.15/2	67.30/4	70.24/1
	auc-roc	65.68/2	63.48/3	61.17/4	74.73/1
	aupr	69.64/2	64.31/3	62.87/4	78.71/1
	p-qscores	31.55/2	37.21/3	37.43/4	26.36/1
Immunological	fmax	75.16/3	71.43/4	81.43/2	86.33/1
	auc-roc	80.63/3	70.11/4	86.70/2	92.27/1
	aupr	82.00/3	70.39/4	86.44/2	92.85/1
	p-qscores	15.82/3	26.31/4	13.20/2	9.85/1
Muscular	fmax	72.18/4	73.10/3	76.52/1	74.34/2
	auc-roc	77.29/3	75.00/4	81.39/1	80.96/2
	aupr	80.47/3	74.95/4	84.11/2	84.45/1
	p-qscores	22.97/3	23.77/4	18.41/1	20.85/2
Developmental	fmax	70.91/1	67.13/4	70.09/2	68.80/3
	auc-roc	64.36/1	33.98/4	62.52/3	63.87/2
	aupr	60.84/2	41.04/4	61.71/1	60.24/3
	p-qscores	35.80/3	58.30/4	30.32/1	33.38/2
Average	$\overline{\text{fmax}}$	72.6±3.5	71.3±5.1	72.7±4.8	75.4±5.8
	$\overline{\text{auc-roc}}$	73.7±6.7	67.1±14.3	73.1±8.8	78.4±8.0
	$\overline{\text{aupr}}$	75.5±6.8	67.4±12.5	74.5±8.8	79.7±8.4
	$\overline{\text{p-qscores}}$	30.0±7.1	29.2±12.1	29.0±8.1	24.8±8.2
	$\overline{\text{rank}}$	2.73±0.8	3.08±1.1	2.62±0.9	1.56±0.9

References

[1] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource *Oxford University Press*, 2004.

[2] Dinh Tran Van et al. The conjunctive disjunctive graph node kernel for disease gene prioritization. In *Neurocomputing. pages 255–262, 2018.*

[3] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. In *Nucleic acids research. pages D258–D261, 2004.*

[4] Zampieri G et al. Scuba: scalable kernel-based gene prioritization. *BMC Bioinformatics*, 2018.

[5] Alshahrani, Mona et al. Semantic Disease Gene Embeddings (SmuDGE): phenotype-based disease gene prioritization without phenotypes. *Bioinformatics*, 34(17), i901–i907, 2018.

[6] Fabrizio Costa and Kurt De Grave. Fast neighborhood subgraph pairwise distance kernel. In *Proceedings of the 26th International Conference on Machine Learning*, pages 255–262. Omnipress, 2010.