Supplementary Document

Recent Advances in Network-based Methods for Disease Gene Prediction

Sezin Kircali Ata [1], Min Wu [2], Yuan Fang [3], Le Ou-Yang [4], Chee Keong Kwoh [1] and Xiao-Li Li [2,*]

[1] *School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore*
[2] *Institute for Infocomm Research, 138632, Singapore*
[3] *School of Information Systems, Singapore Management University, 188065, Singapore and*
[4] *College of Information Engineering, Shenzhen University, Shenzhen 518060, China*

1) AUPR performances of the studied methods in this review.

AUPR performances of the methods presented in this review are shown in Table 1

Table 1: AUPR Performances of the studied methods.

	Disease	Alzheimer	Breast Cancer	Colon Cancer	Diabetes	Lung Cancer	Obesity	Prostate Cancer	Avg
Homogeneous Network	RWR	0.0063	0.0321	0.0140	0.0154	0.0049	0.0530	0.0123	0.0197
	node2vec	0.0411	0.0682	0.0749	0.0048	0.0046	0.0046	0.0060	0.0292
Heterogeneous Network	N2VKO	0.1592	0.0776	0.1185	0.0392	0.1083	0.0618	0.0357	0.0857
	RWRH	0.1760	0.1057	0.1411	0.0085	**0.1203**	**0.1322**	0.0042	0.0983
	IMC	0.0058	0.0094	0.0063	0.0056	0.0103	0.0073	0.0089	0.0076
	Catapult	**0.3537**	0.0718	0.0845	0.0056	0.0761	0.0109	0.0076	0.0872
	Prodige	0.3420	0.0732	0.0481	0.0154	0.1114	0.0384	**0.0532**	0.0974
	Metagraph+	0.3018	0.1308	0.1053	0.0159	0.0383	0.0679	0.0032	0.0947
	HIN2Vec	0.0165	0.0987	0.0972	0.0135	0.0517	0.0042	0.0178	0.0428
	HeGAN	0.1442	0.1167	0.1091	0.0129	0.0101	0.0047	0.0028	0.0572
Multi-view Network	MVE	0.0161	0.0622	0.0757	0.0283	0.0150	0.0016	0.0405	0.0342
	mvn2vec-r	0.0275	0.0868	**0.1570**	0.0072	0.0072	0.0794	0.0023	0.0525
	DMNE	0.0603	0.0205	0.0252	0.0123	0.0503	0.0877	0.0018	0.0369
	MANE	0.2277	**0.1889**	0.1252	**0.0435**	0.0850	0.0386	0.0084	**0.1025**

2) Average performances with rank-aware evaluation metrics.
In our dataset, the number of positive samples in each fold of 5-fold cross-validation varies between 2-6. We thus examined the ranking performances of the studied methods for the top-3 of the predictions in Table 2. These rank-aware evaluation metrics are Precison at k, Recall at k, Average Precision at k and Normalized Discounted Cumulative Gain at k. They are mostly used for recommendation systems but they are also helpful to evaluate the top-k predictions of the methods on skewed datasets.

3) RWR and RWRH performances based on restart probabilities

Both RWR and RWRH exhibit varying average AUPR performances, around 0.01 and 0.03, respectively, depending on the restart probabilities as shown in Table 3. This shows the necessity of the tuning process for these methods.

4) Clustering performances of network embedding methods

Considering the top four network embedding methods (i.e., under the same experimental settings) in AUC and AUPR evaluations, we performed K-means clustering across seven diseases (Alz, BC, CC,

Table 2: Average performances of the studied methods across seven diseases with some of the rank-aware evaluation metrics.

	Methods	AUC	AUCPR	P@3	R@3	AP@3	ndcg@3
Homogeneous Network	RWR	0.5917	0.0197	0.0190	0.0112	0.0381	0.0177
	node2vec	0.6772	0.0292	0.0095	0.0057	0.0095	0.0069
Heterogeneous Network	N2VKO	0.7558	0.0857	0.0952	0.0817	**0.2429**	0.1155
	RWRH	0.7368	0.0983	0.0857	0.0741	0.1571	0.0879
	IMC	0.7001	0.0076	0.0095	0.0143	0.0143	0.0143
	Catapult	0.7399	0.0872	0.0667	0.0762	0.1381	0.0844
	Prodige	0.7763	0.0974	0.0667	0.0748	0.1714	0.0863
	Metagraph+	0.7490	0.0947	**0.1143**	0.0929	0.1952	0.1233
	HIN2Vec	0.7593	0.0428	0.0381	0.0234	0.0714	0.0434
	HeGAN	0.7351	0.0572	0.0571	0.0386	0.1048	0.0606
Multi-view Network	MVE	0.6405	0.0342	0.0381	0.0234	0.0714	0.0434
	mvn2vec-r	0.7385	0.0525	0.0667	0.0370	0.0976	0.0680
	DMNE	0.7501	0.0369	0.0381	0.0381	0.0619	0.0376
	MANE	**0.7956**	**0.1025**	**0.1143**	**0.0949**	**0.2429**	**0.1273**

Table 3: RWR and RWRH average performances over the restart probabilities across seven diseases.

Method	Metric	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
RWR	AUC	**0.5917**	0.5909	0.5896	0.5879	0.5864	0.5842	0.5819	0.5788	0.5751
	AUPR	**0.0197**	0.0196	0.0191	0.0186	0.0185	0.0180	0.0178	0.0175	0.0177
RWRH	AUC	0.7273	0.7336	0.7368	0.7388	0.7396	**0.7402**	0.7401	0.7392	0.7374
	AUPR	0.0813	0.0947	**0.0983**	0.0963	0.0897	0.0831	0.0804	0.0758	0.0712

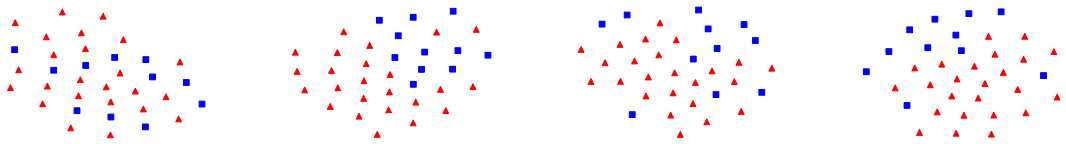
DM, LC, Ob, PC) with 100-random initialization, and evaluated the results using purity, normalized mutual information (NMI) and rand index (Table 4). The clustering performance of MANE outperforms the runner-up method, mvn2vec, by 3%, 2% and 4% in terms of purity, normalized mutual information (NMI) and rand index, respectively.

Table 4: Clustering performance comparison across seven diseases among the competitive methods.

Method	Purity	NMI	Rand
MVE	0.3167	0.1035	0.0203
mvn2vec-r	0.3561	0.1520	0.0532
DMNE	0.3238	0.1211	0.0293
MANE	**0.3662**	**0.1556**	**0.0552**

We further visualized the embeddings generated by four competitive methods in Figure 1. We showcase the separation of the Alz. disease proteins (Alz. positives) from the BC disease proteins (BC positives) based on their known disease associations obtained from OMIM database. For this purpose, t-SNE algorithm is employed with the same parameter setting over the embeddings of the positives.

Both DMNE and MVE plots are visibly inferior, where the disease proteins are mixed with each other without forming clusters. The plot of mvn2vec is relatively superior in attaining two clusters, however the location of the two BC proteins might be misleading for a classifier. MANE achieves to form a cluster of nine Alz. proteins, and this time these proteins are well separated on one side.



(a) MVE (b) mvn2vec-r (c) DMNE (d) MANE

Figure 1: Visualization of disease protein nodes (positives): Alz (square*blue), BC (triangle*red).

5) Oversampling performances of network embedding methods

We apply oversampling with SMOTE over the ratio parameter r={0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5} based on AUPR performances. AUC and AUPR results are shown in Table 5 and Table 6, respectively. We observe an overall improvement especially in AUPR for most of the methods. MANE embeddings exhibit robust performance such that the applied oversampling technique could not yield an improvement in results.

Table 5: AUC comparison on tuned oversampling results

	Disease	Alzheimer	Breast Cancer	Colon Cancer	Diabetes	Lung Cancer	Obesity	Prostate Cancer	Avg
Homogeneous Network	node2vec	0.7536	0.8951	0.8334	0.5102	0.6057	0.6051	0.5653	0.6812
Heterogeneous Network	HIN2Vec	0.9045	0.8974	0.9345	0.7716	0.6944	0.7152	0.5869	0.7864
	HeGAN	0.8388	0.9140	0.9022	0.6599	0.7404	0.5600	0.5187	0.7334
Multi-view Network	MVE	0.8293	0.8739	0.9187	0.7130	0.7097	0.7639	0.6753	0.7834
	mvn2vec-r	0.8764	0.8986	0.9194	0.6455	0.6493	0.6516	0.5302	0.7387
	DMNE	0.9457	0.7960	0.8243	0.7526	0.6670	0.7333	0.5112	0.7471
	MANE	0.9659	0.9199	0.9217	0.7040	0.6989	0.7227	0.6319	0.7950

Table 6: AUPR comparison on tuned oversampling results

	Disease	Alzheimer	Breast Cancer	Colon Cancer	Diabetes	Lung Cancer	Obesity	Prostate Cancer	Avg
Homogeneous Network	node2vec	0.0346	0.0912	0.0690	0.0050	0.0048	0.0045	0.0043	0.0305
Heterogeneous Network	HIN2Vec	0.1170	0.1795	0.1701	0.0147	0.0179	0.0164	0.0044	0.0743
	HeGAN	0.1509	0.1064	0.0952	0.0156	0.0125	0.0040	0.0036	0.0554
Multi-view Network	MVE	0.0929	0.0848	0.1108	0.0245	0.0348	0.0127	0.0294	0.0557
	mvn2vec-r	0.0226	0.0957	0.1584	0.0090	0.0081	0.0853	0.0041	0.0547
	DMNE	0.1421	0.0266	0.0175	0.0148	0.0290	0.1082	0.0023	0.0486
	MANE	0.2449	0.1618	0.1106	0.0179	0.0849	0.0261	0.0102	0.0938

6) AUC performance of four selected approaches in the constructed GO network

Table 7 shows the AUC for these methods on the GO network, respectively. Table 8 further shows the performance comparison of various methods on PPI and GO networks. Based on the results in Table 8, only RWRH achieves a better average AUC performance on the GO network. Therefore, we still report their performance on the PPI network in our main manuscript.

Table 7: AUC performance of four selected approaches in the constructed GO network.

GO	Alzheimer	Breast Cancer	Colon Cancer	Diabetes	Lung Cancer	Obesity	Prostate Cancer	Avg
RWRH	**0.9501**	**0.8722**	**0.8912**	**0.8357**	0.7968	0.7704	0.6190	**0.8193**
IMC	0.6966	0.6327	0.6500	0.6149	0.6006	0.5000	**0.6914**	0.6266
Catapult	0.5945	0.6205	0.6562	0.6578	0.5616	0.6830	0.5046	0.6112
ProDiGe	0.7705	0.7366	0.7111	0.7170	**0.7997**	**0.7901**	0.5727	0.7282

Table 8: Average AUC performance of various methods on PPI and GO networks.

	PPI network	GO network
RWRH	0.7368	**0.8193**
IMC	**0.7001**	0.6266
Catapult	**0.7399**	0.6112
ProDiGe	**0.7763**	0.7282

7) AUC and AUPR performances of GO networks across parameter K

We calculate the pairwise GO similarity between proteins based on G-SESAME[1] [1] and then construct the KNN graphs based on different values for K over $\{5, 10, 20, 50, 100\}$. For each constructed GO KNN graph, we evaluate the performance for disease gene prediction by feeding their node2vec embeddings to logistic regression model. Table 9 shows the performance in terms of AUC and AUPR when we use different K values to construct KNN graph. We set K as 10 as its GO KNN graph can lead to the best performance as shown in Table 9.

Table 9: AUC and AUPR performances of GO networks across parameter K.

	AUC	AUPR
$\text{GO}_{K=5}$	0.5982	0.0096
$\text{GO}_{K=10}$	**0.7215**	**0.0114**
$\text{GO}_{K=20}$	0.6808	0.0098
$\text{GO}_{K=50}$	0.6375	0.0083
$\text{GO}_{K=100}$	0.6877	0.0083

8) How features are extracted from the networks for machine learning

We provide details for some of the methods that we demonstrate their performance, namely, Catapult, ProDiGE and Metagraph+. Disease similarity matrix and PPI adjacency matrices are extracted for both Catapult and ProDiGe, which are feature-based methods. As described in study Catapult [2], Catapult directly utilizes them in the matrix format as inputs and then extracts paths as features by taking matrix multiplication operations. Prodige transforms these two matrices into gene-disease kernel. Metagraph+ representations, on the other hand, consist of three concatanated feature vectors, i.e., 1) annotations (keywords) of proteins 2) the number of occurrences within a subgraph type (meta-graph), 3) the number of co-occurrences with a disease protein within the same subgraph type.

[1]G-SESAME: http://bioinformatics.clemson.edu/G-SESAME/

References

[1] James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.

[2] U. Martin Singh-Blom, Nagarajan Natarajan, Ambuj Tewari, John O. Woods, Inderjit S. Dhillon, and Edward M. Marcotte. Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLOS ONE*, 8(5):1–17, 05 2013.