Towards Gene Function Prediction via Multi-Networks Representation Learning

Hansheng Xue,[1,2] **Jiajie Peng,**[1*] **Xuequn Shang**[1]
[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
xhs1892@gmail.com, jiajiepeng@nwpu.edu.cn, shang@nwpu.edu.cn

Abstract

Multi-networks integration methods have achieved prominent performance on many network-based tasks, but these approaches often incur information loss problem. In this paper, we propose a novel multi-networks representation learning method based on semi-supervised autoencoder, termed as DeepMNE, which captures complex topological structures of each network and takes the correlation among multi-networks into account. The experimental results on two real-world datasets indicate that DeepMNE outperforms the existing state-of-the-art algorithms.

Introduction

Annotating gene function is an important and challenging problem in biological area, which aims to assign an unknown gene to the correct functional categories. Because of the complementary nature of different data sources, multi-networks based function prediction always performs better than other single-network based methods. Thus, many approaches have been proposed for gene function prediction by integrating multiple biological networks (Cho, Berger, and Jian 2016; Sara and Quaid 2010; Cao and et al. 2014).

Gene function prediction can be considered as a node classcification problem. It has been proved that better feature representation of nodes can enhance the performance of models for node classification (Cho, Berger, and Jian 2016). However, current network representation learning methods mainly focus on single-network embedding and represent nodes with topological structure information (Grover and Leskovec 2016; Perozzi and et al. 2014). Besides, existing multi-networks integration methods are linear and shallow approaches which cannot capture complex and highly non-linear structure across all networks.

In this work, we propose a novel multi-networks based feature learning algorithm, named DeepMNE. Considering correlation between multiple networks, DeepMNE applies stacked semi-supervised autoencoder to map input multi-networks into a low-dimension and non-linear space. We evaluate DeepMNE for the task of gene function prediction. The experimental results show that DeepMNE performs better than exisiting state-of-the-art algorithms.

Our Proposed Approach

Learning global structure information In order to learn single network topological information, we run random walk with restart on each network and capture feature representations for each node. The adjacency matrix cannot describe the global structure of the whole network. RWR can makeup this drawback, and represent nodes using these high-dimensional network structural information.

Prior Constraints Extraction Given pairs of nodes, we calculate pairwise pearson correlation coefficient of all pairs of nodes based on their feature vectors. We then set two thresholds for must-link and cannot-link to extract prior constraints. After extracting the constraints from the previous layer (*i* layer), we can apply these constraints to the next layer (*i*+1 layer) as the prior information.

Multi-networks representations Integration using SemiAE The core of DeepMNE is to integrate prior constraints into the network representation. Here we revise the original autoencoder and propose a novel variant of autoencoder, termed as Semi-Supervised AutoEncoder (semiAE). Let x_i be the *i*-th input vector or node representation of network, and f and g be the activations of the hidden layer and the output layer respectively. We have $h_i = f(Wx_i + b)$ and $y_i = g(Mh_i + d)$, where $\Theta = \{\theta_1, \theta_2\} = \{W, b, M, d\}$ are the parameters to be learned, f and g are the non-linear operators such as the sigmoid function ($sigmoid(z) = 1/(1 + exp(-z))$). Assuming M and C are must-link and cannot-link pairwise constraints extracted at the previous step.

The hypothesis is that x_i and x_j should also close based on the low-dimensional space if there is a must-link constraint between them in previous layer. Ideally, after encoding, two must-link nodes should be closer, and two cannot-link nodes may be more distant. Mathematically, if the pair (x_i, x_j) belongs to must-link constraints, we add a penalty on the loss function. If the pair (x_i, x_j) belongs to cannot-link constraints, we add a reward on the loss function. The loss function of semiAE can be defined as follows:

$$loss = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \parallel y_i - x_i \parallel^2 + \lambda L_{mc} \qquad (1)$$

where the first part of Equation 1 measures the squared error between input and output feature vectors, and the second

[*]Corresponding author

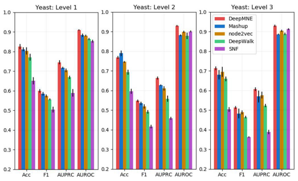

Figure 1: Performance comparison of different metrics on the task of predicting functional labels for yeast genes.

part measures error score of constraints in hidden layer. The loss function for modeling constraints is defined as follows:

$$\begin{aligned} L_{mc} &= \lambda_1 \sum_{(x_i,x_j)\in M} d(h(x_i),h(x_j)) - \lambda_2 \sum_{(x_i,x_j)\in C} d(h(x_i),h(x_j)) \\ &= \lambda_1 \sum_{i,j=1}^{n} M_{i,j}||h(x_i),h(x_j)||_2^2 - \lambda_2 \sum_{i,j=1}^{n} C_{i,j}||h(x_i),h(x_j)||_2^2 \end{aligned} \tag{2}$$

After several iterations of multi-networks integration, we can obtain low-dimensional feature representations of nodes in each network. We then run support vector machine on the output of DeepMNE to annotate unlabeled gene functions.

Experiments

Datasets and Baseline Algorithms

Gene function prediction can be treated as multi-label classification problem. We validate the performance of DeepMNE on datasets of Yeast and Human, which all consisted of six networks with 6,400 and 18,362 genes respectively.

We compare DeepMNE with the four state-of-the-art network representation learning and multi-networks integration algorithms, including Mashup (Cho, Berger, and Jian 2016), SNF (Wang and et al. 2014), node2vec (Grover and Leskovec 2016) and DeepWalk (Perozzi and et al. 2014). We then run support vector machine on the outputs of these methods to predict gene function.

Parameter Settings

We adopt 5-fold cross-validation to evaluate the performance of DeepMNE. The parameters vary with different datasets. The whole iteration model contains five layers. The restart probability of RWR is 0.5, and the final dimension of feature vectors are 500 for yeast and 800 for yeast. The DeepMNE algorithm is optimized using stochastic gradient descent. The batch size is 128, the initial learning rate is 0.1 for yeast and 0.2 for human, and the epochs are 200 and 400 respectively.

Experimental Results on Yeast and Human

Compared with other approaches, DeepMNE achieves better performance on yeast dataset at all three functional classification levels. At level 1, DeepMNE achieves the highest accuracy score that is 0.8248, in constract to 0.8117 for Mashup, 0.6511 for SNF, 0.8036 for node2vec and 0.7707 for DeepWalk. The micro-F1 score, micro-average AUPRC and AUROC achieved by DeepMNE on level 1 of yeast are 0.5994, 0.7452 and 0.9097 respectively, which are significantly higher than other four methods (see Figure 1).

Table 1: The Accuracy, AUPRC for gene function prediction on Molecular Function category of human dataset.

	MF:101-300		**MF:31-100**		**MF:11-30**	
	Acc	AUPRC	Acc	AUPRC	Acc	AUPRC
Mashup	0.5761	0.5236	0.4717	0.3666	0.4486	0.3836
SNF	0.4248	0.3473	0.2689	0.1546	0.3006	0.1662
node2vec	0.5291	0.4959	0.4355	0.3456	0.4482	0.3742
DeepWalk	0.5365	0.5011	0.4488	0.3654	0.4466	0.3762
DeepMNE	**0.5882**	**0.5406**	**0.4936**	**0.4002**	**0.4751**	**0.3897**

DeepMNE also achieves great performance on Molecular Function category of human dataset(see Table 1). The accuracy of DeepMNE on human MF-11-30 is 0.4751, which is higher than Mashup, SNF, node2vec and DeepWalk (0.4486, 0.3006, 0.4482 and 0.4466 respectively). The AUPRC scores of DeepMNE implemented on the human MF-101-300 are 0.5406, which are all higher than other four methods.

Conclusions

In this paper, we propose a novel multi-networks representation learning algorithm based on semi-supervised autoencoder, named DeepMNE. The experimental results on gene function prediction demonstrate the superior performance of DeepMNE over four state-of-the-art algorithms.

Acknowledgments This work was supported by National Natural Science Foundation of China (No. 61702421, 61332014, 61772426), China Postdoctoral Science Foundation (No. 2017M610651), Fundamental Research Funds for the Central Universities (No. 3102018zy033).

References

Cao, M., and et al. 2014. New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics* 30(12):i219.

Cho, H.; Berger, B.; and Jian, P. 2016. Compact integration of multi-network topology for functional analysis of genes. *Cell Systems* 3(6):540.

Grover, A., and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proc. of the 22nd ACM SIGKDD Inter. Conf. on Know. Disc. & Data Min.*, 855–864.

Perozzi, B. a., and et al. 2014. Deepwalk: online learning of social representations. In *Proc. of the 20th ACM SIGKDD Inter. Conf. on Know. Disc. & Data Min.*, 701–710.

Sara, M., and Quaid, M. 2010. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics* 26(14):1759–1765.

Wang, B., and et al. 2014. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* 11(3):333–337.