第47卷 第6期 计 算 机 工 程 2021年6月
Vol.47 No.6 Computer Engineering June 2021

·开发研究与工程应用· 文章编号：1000-3428(2021)06-0299-06 文献标志码：A 中图分类号：TP391

基于GAN异质网络表示学习的疾病关联预测算法

郭梦洁，熊 赟

（复旦大学 计算机科学技术学院 上海市数据科学重点实验室，上海 201203）

摘 要：分析疾病与基因、miRNA等生物实体之间的关联是生物研究领域的重要目标，然而利用海量的数据进行生物学实验成本过高。提出一种基于网络表示学习的关联预测算法，通过多源数据集构建生物异质网络，并给出基于生成式对抗网络的异质网络表示学习算法学习鲁棒的向量表示，算法中的判别器和生成器考虑网络中的关系来捕获丰富的异质语义信息，并通过对抗学习进行训练，在此基础上通过衡量实体向量的相似性预测疾病和基因、miRNA之间的关联。实验结果表明，与HSSVM、GAN等算法相比，该算法在两个关联预测任务上均取得了最高的AUC值，具有更好的预测结果，并且通过引入更多异质数据进行训练，有效提升了算法性能。

关键词：异质网络；网络表示学习；疾病关联预测；生成式对抗网络；对抗学习

开放科学（资源服务）标志码（OSID）：

中文引用格式：郭梦洁，熊赟. 基于GAN异质网络表示学习的疾病关联预测算法[J]. 计算机工程，2021，47(6)：299-304.

英文引用格式：GUO Mengjie，XIONG Yun. Disease association prediction algorithm using GAN-Based heterogeneous network representation learning[J]. Computer Engineering，2021，47(6)：299-304.

Disease Association Prediction Algorithm Using GAN-Based Heterogeneous Network Representation Learning

GUO Mengjie，XIONG Yun

（Shanghai Key Laboratory of Data Science，School of Computer Science，Fudan University，Shanghai 201203，China）

【Abstract】 Analyzing the relationship between diseases and biological entities such as genes and miRNAs is an important goal in the field of biological research. However，the cost of biological experiments based on massive data is too high. This paper proposes a correlation prediction algorithm based on network representation learning. A biological heterogeneous network is constructed by integrating multi-source datasets，and this basis a heterogeneous network representation learning algorithm based on Generative Adversarial Network（GAN） is proposed to learn robust vector representations. The discriminator and generator in the proposed method both consider relations in the network to capture rich heterogeneous semantic information and are trained by adversarial learning. On this basis，associations between diseases and genes as well as diseases and miRNAs are predicted by measuring the similarity between entity vectors. Experimental results show that compared with HSSVM，GAN and other algorithms，the algorithm has better prediction performance，achieving the highest AUC value on the two related prediction tasks，and demonstrates that the introduction of more heterogeneous data for training can improve the performance of the algorithm.

【Key words】 heterogeneous network；network representation learning；disease association prediction；Generative Adversarial Network（GAN）；adversarial learning

DOI：10. 19678/j. issn. 1000-3428. 0057626

0 概述

随着各种高通量生物技术的迅速发展，生物学领域产生了海量数据，使研究人员能够收集和研究大量数据，以更好地阐释复杂疾病的潜在生物学机制[1]。科研机构在生物医学数据的研究上取得了重要进展，但是由于利用海量的数据进行生物学实验需要耗费大量的时间和资源，大部分数据在最初的获取和分析后被搁置[2]，因此，数据的生成和整合分

基金项目：国家自然科学基金（U1936213，U1636207）；上海市科委发展基金（19511121204，19DZ1200802）；上海市科技创新行动计划项目（18511107800）。

作者简介：郭梦洁（1994—），女，硕士研究生，主研方向为数据挖掘；熊 赟，教授、博士生导师。

收稿日期：2020-03-09 **修回日期：**2020-05-13 **E-mail：**mjguo17@fudan.edu.cn

析数据的能力之间的差距越来越大。

很多疾病关联数据可以表示成网络，其中节点代表生物实体，如疾病、基因等，节点间的边指代它们之间的关系。这些网络往往都包含多种类型的实体和关系，被称作异质网络[3]。疾病或其他生物实体在异质网络中是相似的，则它们有很大可能性存在关联。例如一种miRNA在一种疾病起关键作用，则很有可能在相似疾病中起到相似的作用[2]。

为充分利用网络中的信息，研究人员采用网络表示学习算法[3]，将网络映射到低维向量空间，同时保留原有的网络结构、节点内容等。近年来，兴起了异质网络表示学习算法的研究，一类是基于随机游走采样正负节点的训练，代表性的算法包括Metapath2vec[4]、HeteWalk[2]，但它们都依赖合适的元路径[2]，元路径的选择需要人工经验，另一类是将异质网络分解成子网络表示学习后进行信息融合，例如PTE[5]、AspEm[6]，但在分解和融合过程中容易丢失网络中的重要信息。此外，上述算法都忽视了节点的数据分布，因此学习的向量表示缺乏鲁棒性。

本文提出一种基于生成式对抗网络(Generative Adversarial Network，GAN)[7]的异质网络表示学习算法DisGAN。该算法中的判别器和生成器设计通过异质网络中的关系区分不同关系链接的节点对，一对节点被认为是真实的必须满足基于网络拓扑结构的真实节点被正确的关系链接。DisGAN算法考虑了网络中的关系以捕获丰富的异质信息，并通过对抗学习得到鲁棒的向量表示，同时为实现关联预测的目标并验证DisGAN算法性能，本文整合6个公开数据集构建一个生物异质网络，进行基因-疾病关联预测和miRNA-疾病关联预测。

1 问题定义

异质网络定义为$\mathcal{G}=(\mathcal{V},\mathcal{E})$，$\mathcal{V}$和$\mathcal{E}$分别代表节点集合和边集合。该网络也关联一个节点类型映射函数$\phi:\mathcal{V}\rightarrow\mathcal{A}$和一个边类型映射函数$\varphi:\mathcal{E}\rightarrow\mathcal{R}$，其中$\mathcal{A}$和$\mathcal{R}$分别代表节点类型和边类型集合且$|\mathcal{A}|+|\mathcal{R}|>2$[3]。

异质网络表示学习[3]的目标是学习一个映射函数，将网络中每个节点$v\in\mathcal{V}$映射到一个低维向量空间$\mathbb{R}^d$，其中$d\ll|\mathcal{V}|$，尽可能保留网络原有信息。

生成式对抗网络[6]公式定义如下：

$$\min_{\theta^G}\max_{\theta^D} E_{x\sim P_{\text{data}}}\left[\ln D\left(x;\ \boldsymbol{\theta}^D\right)\right]+ E_{z\sim P_z}\left[\ln\left(1-D\left(G\left(z;\ \boldsymbol{\theta}^G\right);\ \boldsymbol{\theta}^D\right)\right)\right] \tag{1}$$

生成器G使用来自预定义分布P_z的噪声z，生成尽可能接近真实数据的伪样本，判别器D旨在区分来自P_{data}的真实数据和生成器生成的伪数据，$\boldsymbol{\theta}^G$、$\boldsymbol{\theta}^D$表示参数。

2 DisGAN算法

本节介绍DisGAN算法，DisGAN包括判别器Discriminator和生成器Generator两部分。网络中真实存在的节点对且通过正确的关系链接是正样本，其他均为负样本，判别器需要进行区分，而生成器需要生成和给定节点通过给定关系相连的伪节点。DisGAN模型框架如图1所示。

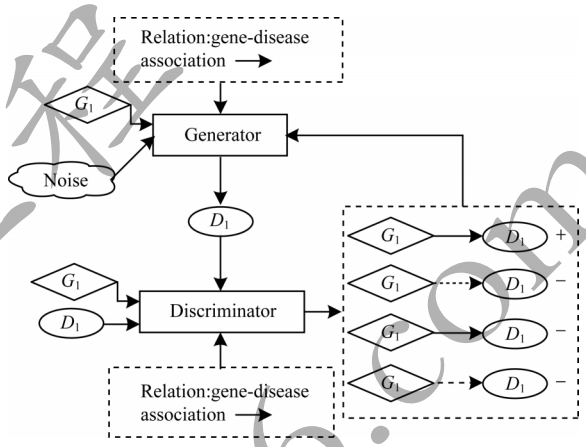

图1 DisGAN模型框架

Fig.1 Framework of DisGAN model

2.1 DisGAN中的判别器和生成器

2.1.1 DisGAN中的判别器

在异质网络中必须区分给定关系下的真实和虚假节点，因此判别器需要评估一对节点在给定关系下的链接性。给定异质网络$\mathcal{G}$中一个节点$i\in\mathcal{V}$和关系$r\in\mathcal{R}$，参数为$\boldsymbol{\theta}^D$的判别器D给出一个样本j是否通过关系r和节点i相连的概率。样本j可以是真实节点，也可以是生成器生成的伪节点。

判别器D公式定义如下：

$$D(j|i,r)=\frac{1}{1+\exp\left(-\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{M}_r\boldsymbol{v}_j\right)} \tag{2}$$

其中，$\boldsymbol{v}_i,\boldsymbol{v}_j\in\mathbb{R}^d$是节点$i$和$j$的$d$维表示向量，$\boldsymbol{M}_r\in\mathbb{R}^{d\times d}$是关系$r$的关系矩阵，参数$\boldsymbol{\theta}^D$是判别器$D$学习的所有节点的表示向量和关系的表示矩阵。

如果样本j是通过关系r和节点i相连的真实节点，判别器给出的概率值应该较高，而对伪样本应该较低。通常，样本j与给定的i和r组成一个三元组$<i,j,r>$，就其正负极性而言，每个三元组均属于以下4种情况之一，每种情况也构成了判别器损失函数的一部分。

1）通过正确关系链接的真实节点

节点i和j是异质网络$\mathcal{G}$中的真实节点，并通过真实关系r连接，这样的三元组$<i,j,r>$是正样本，希

望判别器将其标记为真,因此损失函数定义如下:

$$L_1 = E_{(<i,j,r>\sim\mathcal{G})} - \ln\left(D\left(j|i,r\right)\right) \quad (3)$$

从网络$\mathcal{G}$中提取上述三元组,即$<i,j,r>\sim\mathcal{G}$。

2)通过错误关系链接的真实节点

异质网络中的节点i和j通过一个错误的关系$r'(r'\neq r)$链接。由于它们的链接性与给定关系r携带的期望语义信息不匹配,因此判别器希望将其判定为负样本:

$$L_2 = E_{(<i,j>\sim G,\, r'\sim\mathcal{R}')} - \ln\left(1-D\left(j|i,r'\right)\right) \quad (4)$$

节点对(i,j)从网络$\mathcal{G}$提取,关系r'从$\mathcal{R}'=\mathcal{R}-r$获得。

3)通过正确关系链接的伪节点

给定异质网络中一个节点i和其关系r,然后通过生成器$G(i,r)$生成节点j'。该三元组$<i,j',r>$也应该被判别为负,所以将损失函数定义如下:

$$L_3 = E_{(<i,r>\sim\mathcal{G},\, j'\sim G(i,r))} - \ln\left(1-D\left(j'|i,r\right)\right) \quad (5)$$

伪节点j'的表示向量是从生成器G学习到的分布中提取的,与生成器G的模型参数$\boldsymbol{\theta}^G$不同。

4)通过错误关系链接的伪节点

给定节点i和一个i中不存在的关系r^*,输入生成器$G(i,r^*)$生成一个伪样本j'。判别器的目标是将该三元组也判定为负:

$$L_4 = E_{(i\sim\mathcal{G},\, r^*\sim\mathcal{R}^*,\, j'\sim G(i,r^*))} - \ln\left(1-D\left(j'|i,r^*\right)\right) \quad (6)$$

其中,$\mathcal{R}^*$代表网络$\mathcal{G}$的关系集合$\mathcal{R}$和节点i拥有的关系子集之间的差集。

整合上述4个部分作为损失函数训练判别器:

$$L_D = L_1 + L_2 + L_3 + L_4 + \lambda^D \| \boldsymbol{\theta}^D \| \quad (7)$$

其中,$\lambda^D\| \boldsymbol{\theta}^D\|$是正则化项用来防止过拟合,通过最小化$L_D$来优化判别器参数$\boldsymbol{\theta}^D$。

2.1.2 DisGAN中的生成器

生成器同样考虑到网络的异质性,即给定来自异质网络$\mathcal{G}$的节点$i\in\mathcal{V}$和关系$r\in\mathcal{R}$,参数为$\boldsymbol{\theta}^G$的生成器G希望生成尽可能和节点i通过关系r链接的节点。

生成器原始输入定义为$\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{M}^r$,其中,$\boldsymbol{v}_i\in\mathbb{R}^d$是节点$i$的表示向量,$\boldsymbol{M}_r\in\mathbb{R}^{d\times d}$是关系$r$的表示矩阵,该输入代表一个可能和节点$i$通过关系$r$链接的伪节点。然后从标准正态分布$\mathcal{N}(0,1)$中采样噪声,将噪声和原始输入相加作为最终输入。生成器被定义为多层感知器(Multilayer Perceptron,MLP),接收上述输入后生成伪样本。生成器参数$\boldsymbol{\theta}^G$包含所有节点向量、关系矩阵和MLP的参数。

生成器希望通过生成接近真实节点的伪样本来欺骗判别器,使判别器给伪样本赋予高分:

$$L_G = E_{(<i,r>\sim\mathcal{G},\, j'\sim G(i,r))} - \ln\left(D\left(j'|i,r\right)\right) + \lambda^G \| \boldsymbol{\theta}^G \| \quad (8)$$

其中,$\lambda^G\| \boldsymbol{\theta}^G\|$是正则化项,通过最小化公式$L_G$训练生成器。

2.2 DisGAN模型训练与分析

DisGAN模型使用迭代的数值计算方法[8]进行训练。首先初始化模型参数$\boldsymbol{\theta}^G$和$\boldsymbol{\theta}^D$,然后迭代训练判别器和生成器直到模型收敛。DisGAN模型训练过程如算法1所示。

算法1 DisGAN模型训练

输入 异质网络$\mathcal{G}$,生成器G、判别器D每轮训练次数n_G、n_D,样本数目n_s

1.分别初始化判别器参数θ^D和生成器参数θ^G

2.while没有收敛do

3.for n = 0; n < n_D do //训练判别器

4.采样一批三元组,即< i,j,r >~ $\mathcal{G}$

5.对每个< i,r >,生成器G生成n_s个伪节点,即j'~G(i,r)

6.对每个< i,j>采样n_s个关系r'~R'

7.根据式(7)更新参数θ^D

8.end for

9.for n = 0; n < n_G do //训练生成器

10.采样一批三元组,即< i,j,r >~ $\mathcal{G}$

11.对每个< i,r >,生成器G生成n_s个伪节点,即j'~G(i,r)

12.根据式(8)更新参数θ^G

13.end for

14.end while

15.return θ^D和θ^G

DisGAN模型生成器和判别器每次更新主要涉及节点向量和关系矩阵的更新,每轮迭代时间复杂度为$O((n_G+n_D)\cdot n_s\cdot|\mathcal{V}|\cdot d^2)$,其中,$n_G$和$n_D$是训练次数,$n_s$是样本数目,$|\mathcal{V}|$是节点数目,$d$是维度。本文将$n_G$、$n_D$、$n_s$和$d$视为常数,所以DisGAN每轮迭代的时间复杂度是$O(|\mathcal{V}|)$。

DisGAN模型的生成器是使用Leaky ReLU[9]激活函数的两层感知机,将最后一层输出当作伪节点无需softmax计算采样伪节点,所以对于整个网络每轮迭代生成器采样伪节点的时间复杂度为$O(|\mathcal{V}|)$,$|\mathcal{V}|$是网络中节点数目;而对每个节点每次softmax计算需要遍历网络中所有节点,因此每轮迭代时间复杂度为$O(|\mathcal{V}|^2)$,计算代价非常高。

DisGAN模型中判别器的参数$\boldsymbol{\theta}^G$是网络节点的表示向量和关系的表示矩阵,通过对抗学习优化模型参数。判别器和生成器在对抗学习过程中迭代训

练：首先固定生成器G的参数$\boldsymbol{\theta}^G$，然后从网络中采样真实节点关系三元组，生成器G对每个给定的节点和关系生成n_s个伪节点，最后通过最小化式(7)定义的损失函数优化参数$\boldsymbol{\theta}^D$从而训练判别器；固定判别器的参数$\boldsymbol{\theta}^D$，然后同样采样真实样本和伪样本，最后根据式(8)优化生成器参数$\boldsymbol{\theta}^G$以生成更好的伪节点。上述迭代过程直到模型收敛时停止。

DisGAN相对于GAN[7]的改进主要在于将其扩展应用到网络表示学习：GAN仅仅区分真伪节点无法捕获网络节点间的关系信息，而DisGAN区分不同关系链接的节点对，从而捕获网络的结构和语义信息；GAN中生成器输入为随机噪声，DisGAN加上网络中的节点和关系，从而生成和真实节点更相似的伪节点进行训练提升模型表现。

3 实验结果与分析

3.1 实验数据集

本文实验所用数据集如下：

1）基因相互作用网络：从HPRD数据库[10]中获得的39 240条记录。

2）miRNA相似性网络：从MISIM数据库[11]中提取的56 289条数据。

3）疾病相似性网络：从MimMiner[12]中提取的3 162 016条数据。

4）基因-疾病关联网络：从DisGeNET数据库[13]中提取的19 714条记录。

5）基因-miRNA关联网络：从miRTarBase数据库[14]中提取的21 259条记录。

6）miRNA-疾病关联网络：从文献[15]提供的数据集和miRNet[16]中提取的878条数据。

通过共同节点链接上述6个网络来构建一个生物异质网络。

3.2 对比算法

本文的实验对比算法主要包括：

1）HSSVM[17]：基于HeteSim得分[18]衡量节点相关性，使用监督学习算法进行疾病关联预测。

2）GAN[7]：生成器输入从正态分布中采样的噪声生成伪节点，判别器区分网络节点和生成器产生的伪节点，将网络节点表示作为模型参数训练。

3）DeepWalk[19]：使用随机游走得到节点序列基于skip-gram[20]模型学习表示向量。

4）AspEm[6]：通过将异质网络分解成语义子图，分别学习每个子图中节点向量表示后进行拼接得到最终节点向量表示。

5）HeteWalk[2]：使用元路径和链接权重指导的随机游走并基于异质skip-gram模型进行表示学习。

3.3 实验结果

本文分别进行基因-疾病关联和miRNA-疾病关联实验。每次实验将已知的关联数据随机划分为训练集和测试集，训练集所占比例(R)从50%变化到90%。在进行测试时，已知的关联作为正样本，随机选择相同数目且相同类型但没有关联的节点对作为负样本，通过算法得到节点表示向量的余弦相似度(归一化后的点积)得分作为预测值。不同算法在不同训练比例下的AUC得分[21]如表1和表2所示。

表1 基因-疾病关联预测实验的AUC得分

Table 1 AUC score of gene-disease association prediction experiment

算法	AUC得分				
	R=50	R=60	R=70	R=80	R=90
HSSVM	0.609	0.653	0.693	0.734	0.779
GAN	0.587	0.492	0.565	0.610	0.582
DeepWalk	0.454	0.461	0.481	0.433	0.477
AspEM	0.639	0.667	0.659	0.657	0.681
HeteWalk	0.638	0.674	0.723	0.759	0.798
DisGAN	0.844	0.866	0.859	0.869	0.875

表2 miRNA-疾病关联预测实验的AUC得分

Table 2 AUC score of miRNA-disease association prediction experiment

算法	AUC得分				
	R=50	R=60	R=70	R=80	R=90
HSSVM	0.841	0.877	0.902	0.922	0.932
GAN	0.532	0.581	0.546	0.554	0.550
DeepWalk	0.498	0.511	0.534	0.611	0.677
AspEM	0.765	0.819	0.761	0.849	0.819
HeteWalk	0.937	0.951	0.953	0.946	0.969
DisGAN	0.956	0.955	0.972	0.962	0.985

从表1和表2可以发现，DisGAN算法在两个预测任务所有训练比例上的表现一直都超过所有对比算法。HSSVM没有采用网络表示学习，只提取沿路径的两个节点之间可访问性的简单特征。GAN尽管考虑了向量表示的鲁棒性，但是忽视了节点间的关系，没有捕获网络的拓扑结构和语义关系。DeepWalk表现较差的主要原因是针对同质网络设计的网络表示学习算法，忽视了不同节点和链接类型。AspEm在网络分解合并过程中可能会丢失一些重要信息。HeteWalk尽管通过基于元路径的随机游走捕获到网络的异质信息，但是没有学习节点的数据分布，学习到的向量表示鲁棒性不高。在所有对比算法中，AspEm和HeteWalk表现较好，说明考虑网络异质性可以提升预测结果。

本文提出的DisGAN模型超过了所有的对比算法，可以通过对抗学习节点的数据分布，得到更具鲁棒性的表示，能够较好地保留网络结构和异质语义信息。此外，DisGAN模型在基因-疾病关联预测任务上的表现提升更明显，主要是由于异质网络中基因-疾病关联数据相对更多且数据可能更稀疏或存在噪声，因此需要更具鲁棒性的向量表示。

3.4　异质性分析

本节探究每个算法在处理异质性上的能力。实验中采用三折交叉验证，并去除3.1节中部分数据集生成了另外两个只包含两种节点类型的子网络。从图2和图3可以发现，在只包含两种节点类型的子网络上进行关联预测的AUC得分更低，整合3.1节中所有网络数据，构建一个更加复杂的异质网络有明显的益处，尤其是在miRNA-疾病关联预测任务上。这主要是由于miRNA和疾病之间的已知关联数据更稀少，因此单一网络无法保证预测的可靠性。基因相关的数据集可以帮助建立miRNA和疾病之间的间接关联，这些关联很有可能被进行关联预测的算法捕获。整合多方面数据可以加深对复杂疾病的理解，结合间接关系信息，进一步提升预测结果。DisGAN算法能够整合更多来源的异质网络数据。

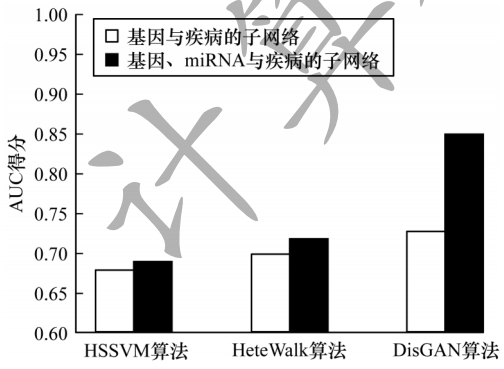

图2　基因-疾病关联预测中不同网络的AUC得分

Fig.2　AUC score on different networks in gene-disease association prediction

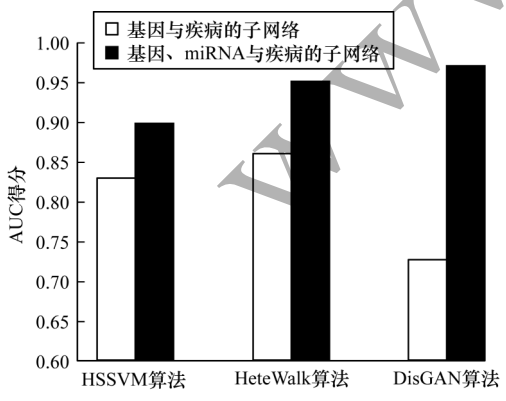

图3　miRNA-疾病关联预测中不同网络的AUC得分

Fig.3　AUC score on different networks in miRNA-disease association prediction

4　结束语

本文提出一种基于GAN的异质网络表示学习算法DisGAN进行疾病关联预测。DisGAN中的判别器和生成器都考虑了网络中的关系捕获异质语义信息，通过对抗学习得到鲁棒的向量表示，并在构建的生物异质网络上进行基因-疾病关联预测和miRNA-疾病关联预测来衡量模型性能表现。实验结果证明了DisGAN算法的有效性和优越性。下一步将整合更多生物数据集来提升DisGAN算法的预测性能。

参考文献

[1] BOTSTEIN D, RISCH N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease[J]. Nature Genetics, 2003, 33(3): 228-237.

[2] XIONG Y, GUO M, RUAN L, et al. Heterogeneous network embedding enabling accurate disease association predictions[J]. BMC Medical Genomics, 2019, 12(10): 186.

[3] SHI Chuan, LI Yitong, ZHANG Jiawei, et al. A survey of heterogeneous information network analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 29(1): 17-37.

[4] DONG Y, CHAWLA N V, SWAMI A. Metapath2vec: scalable representation learning for heterogeneous networks[C]//Proceedings of the 23rd International Conference on Knowledge Discovery and Data Mining. Halifax, Canada: [s. n.], 2017: 158-169.

[5] TANG J, QU M, MEI Q. PTE: predictive text embedding through large-scale heterogeneous text networks[C]// Proceedings of the 21th International Conference on Knowledge Discovery and Data Mining. Sydney, Australia: [s. n.], 2015: 321-332.

[6] SHI Yu, GUI Huan, ZHU Qi, et al. AspEm: embedding learning by aspects in heterogeneous information networks[C]//Proceedings of SIAM International Conference on Data Mining. Washington D. C., USA: IEEE Press, 2018: 144-152.

[7] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of Advances in Neural Information Processing Systems. [S. 1.]: MIT Press, 2014: 635-648.

[8] GOODFELLOW I. NIPS 2016 tutorial: generative adversarial networks[EB/OL]. [2020-02-10]. https://arxiv. org/abs/1701. 00160.

[9] MAAS A L, HANNUN A Y, NG A Y. Rectifier nonlinearities improve neural network acoustic models[C]// Proceedings of International Conference on Machine Learning. Washington D. C., USA: IEEE Press, 2013: 226-238.

[10] KESHAVA PRASAD T S, GOEL R, KANDASAMY K, et al. Human protein reference database—2009 update [J]. Nucleic Acids Research, 2009, 37(1): 767-772.

[11] WANG D, WANG J, LU M, et al. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases[J]. Bioinformatics, 2010, 26(13): 1644-1650.

[12] VAN DRIEL M A, BRUGGEMAN J, VRIEND G, et al. A text-mining analysis of the human phenome[J]. European Journal of Human Genetics, 2006, 14(5): 535-542.

[13] PIÑERO J, BRAVO À, QUERALT-ROSINACH N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants [J]. Nucleic Acids Research, 2017, 45(1): 833-839.

[14] CHOU C H, CHANG N W, SHRESTHA S, et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database [J]. Nucleic Acids Research, 2016, 44(1): 239-247.

[15] CHEN Hailin, ZHANG Zuping. Similarity-based methods for potential human microRNA-disease association prediction[J]. BMC Med Genomics, 2013, 6(1): 12-20.

[16] FAN Y, SIKLENKA K, ARORA S K, et al. miRNet-dissecting miRNA-target interactions and functional associations through network-based visual analysis [J]. Nucleic Acids Research, 2016, 44(1): 135-141.

[17] ZENG Xiangxiang, LIAO Yuanlu, LIU Yuansheng, et al. Prediction and validation of disease genes using HeteSim scores [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, 14(3): 687-695.

[18] SHI Chuan, KONG Xiangnan, HUANG Yue, et al. HeteSim: a general framework for relevance measure in heterogeneous networks [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(10): 2479-2492.

[19] PEROZZI B, AL-RFOU R, SKIENA S. DeepWalk: online learning of social representations[C]//Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2014: 159-168.

[20] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of Advances in Neural Information Processing Systems. Boston, USA: MIT Press, 2013: 635-648.

[21] LOBO J M, JIMENEZ-VALVERDE A, REAL R. AUC: a misleading measure of the performance of predictive distribution models[J]. Global Ecology and Biogeography, 2008, 17(2): 145-151.

编辑 索书志