



基于异质网络层次注意力机制的基因功能预测

万美含^{1,2,3}, 熊 贇^{1,2,3}, 朱扬勇^{1,2,3}

(1. 复旦大学 计算机科学技术学院, 上海 200433; 2. 上海市数据科学重点实验室, 上海 200433;
3. 上海先进通信与数据科学研究院, 上海 200433)

摘 要: 基因组测序技术的快速发展使得生物数据库中的基因和基因组序列数据数量迅速增加, 但其中仍有大量基因功能是未知的。为此, 提出基于异质网络层次注意力机制的基因节点表示学习方法 HAGE, 用以预测基因功能。结合多种来源的数据集, 构建一个具有节点属性的基因功能相关异质网络, 在网络中使用层次注意力机制为每一个基因节点学习一个节点嵌入向量, 该向量可用于后续的基因功能预测等任务。实验结果表明, 与 GraphSAGE 和 GAT 等方法相比, HAGE 具有更好的预测性能。

关键词: 基因功能预测; 异质信息网络; 注意力机制; 网络表示学习; 网络嵌入

开放科学(资源服务)标志码(OSID):



中文引用格式: 万美含, 熊贇, 朱扬勇. 基于异质网络层次注意力机制的基因功能预测[J]. 计算机工程, 2020, 46(7): 43-49.

英文引用格式: WAN Meihan, XIONG Yun, ZHU Yangyong. Gene function prediction based on hierarchical attention mechanism in heterogeneous network[J]. Computer Engineering, 2020, 46(7): 43-49.

Gene Function Prediction Based on Hierarchical Attention Mechanism in Heterogeneous Network

WAN Meihan^{1,2,3}, XIONG Yun^{1,2,3}, ZHU Yangyong^{1,2,3}

(1. School of Computer Science and Technology, Fudan University, Shanghai 200433, China;

2. Shanghai Key Laboratory of Data Science, Shanghai 200433, China;

3. Shanghai Institute of Advanced Communications and Data Science, Shanghai 200433, China)

[Abstract] The rapid development of genome sequencing has led to the explosive growth of gene and genomic sequence data in biological databases, in which functions of a large number of genes still remain unknown. Therefore, this paper proposes a gene node representation learning method, HAGE, based on hierarchical attention mechanism in heterogeneous network to predict the function of genes. Firstly, a gene function-related heterogeneous network with node attributes is constructed. Then the hierarchical attention mechanism is used in network to enable each gene node to learn a node embedding vector, which can be used for subsequent tasks such as gene function prediction. Experimental results show that the proposed method has better performance than GraphSAGE, GAT and other methods.

[Key words] gene function prediction; heterogeneous information network; attention mechanism; network representation learning; network embedding

DOI: 10.19678/j.issn.1000-3428.0054805

0 概述

基因是具有功能性的 DNA 片段^[1]。由于可通过功能产物的表达或基因表达调控来影响生物体性状^[2], 因此确定基因的功能是生物学中的核心问题之一, 其对于了解疾病的生化过程、识别和验证新药物

的靶点等都具有重要意义^[3]。

基因组测序的快速发展使得生物数据库中基因和基因组序列的数据规模爆炸式增长, 但其中有大量的基因功能仍是未知的^[4]。因此, 通过已有的基因特性信息对基因的功能进行预测是目前的研究热点。

基金项目: 国家自然科学基金(U1636207, 91546105); 上海市科技发展基金(16JC1400801)。

作者简介: 万美含(1994—), 女, 硕士研究生, 主研方向为数据挖掘; 熊 贇、朱扬勇, 教授、博士生导师。

收稿日期: 2019-05-05 **修回日期:** 2019-07-15 **E-mail:** 16210240040@fudan.edu.cn

目前,已有大量的基因功能预测方法被提出,总体可归为两类:一类是基于 guilt-by-association 原则的方法,即与相似的生物物质(如疾病)相连的基因应共享相同的功能^[5],通过融合不同类型的生物数据,构建一个与基因功能相关的网络来预测基因的功能^[6];另一类是基于基因本体(Gene Ontology, GO)的方法,即基因本体通过结构化的术语以分子功能、生物过程和细胞成分 3 种属性来描述基因,如文献[7-9]利用基因本体计算不同基因之间的相似度,实现对基因功能的准确预测。本文结合上述两类方法,将基因本体数据作为基因节点的属性,使用多种数据源构建一个基因功能相关异质信息网络。

近年来,注意力机制受到学者的关注^[10],且在各个研究领域得到广泛应用。在异质网络表示学习方面,文献[11]构建了 HAN 模型,通过引入层次注意力机制进行异质网络节点表示学习,文献[12]在其基础上使用节点结构特征信息构建了 HANE 模型,但该模型仅适用于无节点属性的异质网络。本文将 HANE 模型扩展到属性异质信息网络(Attributed Heterogeneous Information Network, AHIN)中,构建一个具有节点属性的基因功能相关异质信息网络,并在此基础上提出基于层次注意力机制的基因节点表示学习方法 HAGE。

1 相关定义

本文通过结合多种类型的公开数据集,构建一个具有节点属性的基因功能相关异质信息网络,并在该网络上应用基于层次注意力机制的网络表示学习方法,为每一个基因节点生成一个节点嵌入向量,该向量可用于后续的基因功能预测任务。对上述过程中使用的相关概念进行形式化定义:

定义 1 异质信息网络^[13]是具有多种节点类型或(和)多种边类型的网络,表示为 $G = (V, E, T)$,其中, V 是节点的集合, E 是边的集合。同时, $\varphi: V \rightarrow T_v$ 是节点到节点类型的映射, $\varphi: E \rightarrow T_e$ 是边到边类型的映射, T_v 和 T_e 是预设的节点和边的类型,并满足 $|T_v| + |T_e| > 2$, $T = T_v \cup T_e$ 。

由于本文使用的异质信息网络是基于基因-疾病关系网络、基因-miRNA 关系网络和 miRNA-疾病关系网络生成的,因此其中包含 3 种节点类型(基因、疾病和 miRNA)和 3 种边类型(基因-疾病关系、基因-miRNA 关系和 miRNA-疾病关系)^[14]。

定义 2 网络模式^[15]是定义在节点类型和边类型上的一个有向图,表示为 $S_G = \{T_v, T_e\}$ 。

本文构建的基因功能相关异质信息网络的网络模式如图 1 所示。

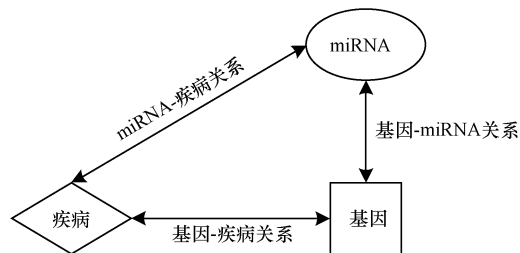


图 1 基因功能相关网络模式

Fig. 1 Gene function correlation network pattern

定义 3 元路径^[16]是定义在网络模式 $S_G = \{T_v, T_e\}$ 上的一条路径 P , 用于描述节点类型 t_{v_1} 到 $t_{v_{l+1}}$ 之间的关系, 表示为 $t_{v_1} \xrightarrow{t_{e_1}} t_{v_2} \cdots \xrightarrow{t_{e_l}} t_{v_{l+1}}$, 本文简称为 $t_{v_1} t_{v_2} \cdots t_{v_{l+1}}$ 。

本文中使用的元路径及其含义如表 1 所示。

表 1 基因功能相关网络中的元路径及其含义
Table 1 Meta-paths and their meanings in gene function correlation network

元路径	缩写	元路径含义
基因-疾病-基因	GDG	与同一疾病相关的基因
基因-miRNA-基因	GMG	受同一 miRNA 调控的基因
基因-miRNA-疾病- miRNA-基因	GMDMG	受同一疾病相关 miRNA 调控的基因

定义 4 基于元路径的邻居给定一个节点 i 和元路径 P , 所有通过元路径 P 与节点 i 连接的节点集合为 N_i^P 。

定义 5 元路径的目标节点为给定一个异质信息网络 $G = (V, E, T)$ 与一个元路径 $P: t_{v_1} t_{v_2} \cdots t_{v_{l+1}}$, G 中所有节点类型为 $t_{v_{l+1}}$ 的节点, 表示为 V_T^P 。

定义 6 异质网络表示学习^[17](异质网络嵌入)给定一个异质信息网络 $G = (V, E, T)$, 对 V 中每一个节点学习一个 d 维向量 $X \in \mathbb{R}^{|V| \times d}$, 其中 $d \ll |V|$ 。

2 HAGE 模型结构

在基因功能相关异质信息网络的基础上, 本文提出基于层次注意力机制的基因节点表示学习方法 HAGE, 为每一个节点学习一个节点嵌入向量。HAGE 模型主要包括 3 个部分, 即节点特征抽取、节点层次的注意力机制和元路径层次的注意力机制。

2.1 节点特征抽取

给定一个异质信息网络 $G = (V, E, T)$ 以及元路径集合 $\{P_1, P_2, \cdots, P_m\}$, 对于 V 中的每一个节点 v_i , 本文从 2 个方面考虑该节点的特征: 节点的属性信息 a_i 以及该节点在网络中的结构特征 f_i 。

在构建的基因功能相关网络中, 基因节点属性 a_i 来自于基因本体数据, 将每个基因对应的本体术语转化为 multi-hot 编码并作为基因节点的属性。

对于节点网络中的结构特征 f_i , 本文使用基于

元路径的连接分布来描述。在异质网络中,不同的元路径具有不同的语义信息,因此,不同元路径下相同节点间的连接分布也是不同的。对于同一对基因节点 A 和 B ,其通过元路径基因-疾病-基因连接的路径与通过元路径基因-miRNA-基因连接的路径完全不同,并且路径的权重和数量也不同,因此,其连接分布也完全不同。

对于节点对 v_i 与 v_j ,定义其基于元路径 P 的连接强度 I_{ij}^P 为 v_i 与 v_j 基于元路径 P 相连的所有路径的权重之和:

$$I_{ij}^P = \sum_{k=1}^{n_{ij}^P} w_k^{P_{ij}} \quad (1)$$

其中, n_{ij}^P 为 v_i 与 v_j 基于元路径 P 相连的所有路径的数量, $w_k^{P_{ij}}$ 为 v_i 与 v_j 基于元路径 P 相连的第 k 条路径的权重。

下面对连接强度矩阵 I^P 进行归一化,作为节点结构特征矩阵 F^P :

$$F_{ij}^P = \frac{I_{ij}^P}{\sum_{k \in V_T^P} X_{ik}^P} \quad (2)$$

最后,将每个节点 v_i 的节点属性与其基于元路径的结构特征进行拼接并作为节点的特征向量:

$$h_i^P = a_i \parallel f_i^P \quad (3)$$

其中, \parallel 表示拼接操作, f_i^P 为节点结构特征矩阵 F^P 的第 i 行。

2.2 节点层次的注意力机制

首先在节点层次上使用注意力机制来学习基于元路径邻居节点的重要性,并通过聚合这些拥有不同权重的邻居节点得到新的特征向量,即如果基因 A 具有功能 f ,其邻居节点中功能与功能 f 相同或更近似的节点应具有更大的权重,通过聚合不同邻居节点的嵌入向量及其权重来更新基因节点 A 的嵌入向量。

给定通过元路径 P 连接的节点对 (v_i, v_j) ,通过自注意力机制^[18]学习节点 v_j 对 v_i 的重要程度 e_{ij}^P ,形式化定义为:

$$e_{ij}^P = a_{\text{node}}(h_i^P, h_j^P; P) \quad (4)$$

其中, a_{node} 是一个深度神经网络,代表节点层次的注意力机制。对于给定的元路径 P ,基于该路径的所有邻居节点共享 a_{node} 。

得到基于元路径 P 的所有节点对 (v_i, v_j) 的重要程度后,对其进行归一化得到权重系数 α_{ij}^P :

$$\alpha_{ij}^P = \text{softmax}(e_{ij}^P) = \frac{\exp(\sigma(a_p^T [Wh_i^P \parallel Wh_j^P]))}{\sum_{k \in N_i^P} \exp(\sigma(a_p^T [Wh_i^P \parallel Wh_k^P]))} \quad (5)$$

其中, σ 是激活函数, W 是权重矩阵, a_p 是基于元路径 P 节点层次的注意力向量。

通过权重系数与基于元路径邻居节点的特征向量可以得到节点 v_i 新的特征向量 z_i^P :

$$z_i^P = \sigma \left(\sum_{j \in N_i^P} \alpha_{ij}^P Wh_j^P \right) \quad (6)$$

为使训练过程更加稳定,将节点层次的注意力机制扩展到多头注意力机制,即基于 K 个独立的节点层次的注意力机制计算 z_i^P ,并对结果进行拼接得到最终的节点向量:

$$z_i^P = \parallel_{k=1}^K \sigma \left(\sum_{j \in N_i^P} \alpha_{ij}^P Wh_j^P \right) \quad (7)$$

给定一系列元路径的集合 $\{P_1, P_2, \dots, P_m\}$,基于节点特征向量并利用节点层次的注意力机制可以得到 m 组新的节点特征向量 $\{Z_{P_1}, Z_{P_2}, \dots, Z_{P_m}\}$ 。

2.3 元路径层次的注意力机制

基于节点层次的注意力机制可以得到不同元路径下新的节点特征向量,为得到最终的节点嵌入向量,需要对不同元路径下的节点特征向量进行融合。

在异质网络中,不同的元路径代表不同的语义信息,因此,需要为不同的元路径分配不同的权重。使用一个元路径层次的注意力机制^[11]来学习不同元路径的重要程度 β_P 。给定元路径的集合 $\{P_1, P_2, \dots, P_m\}$ 以及基于节点层次注意力机制得到的新的节点特征向量 $\{Z_{P_1}, Z_{P_2}, \dots, Z_{P_m}\}$,为每个元路径 P_i 学习一个权重系数 β_{P_i} ,定义为:

$$\beta_{P_i} = a_{\text{meta}}(Z_{P_i}) \quad (8)$$

其中, a_{meta} 是一个深度神经网络,代表元路径层次的注意力机制。不同的元路径将学习到不同的权重,对基因功能预测任务更重要的元路径将具有更大的权重。

为学习不同元路径的重要程度,首先对基于节点层次的注意力机制得到的节点特征向量进行非线性变换,然后将变换后的特征向量与元路径层次的注意力向量 q 的相似度作为元路径的重要程度。因此,对于元路径 P_i ,其重要程度 w_{P_i} 表示为:

$$w_{P_i} = \frac{1}{|V_T^{P_i}|} \sum_{k \in V_T^{P_i}} q^T \cdot \tanh(Wz_k^{P_i} + b) \quad (9)$$

其中, W 是权重矩阵, b 是偏置向量, q 是元路径层次的注意力向量。

得到每条元路径的重要程度 w_i 后,对它们进行归一化处理,得到每条元路径的权重系数 β_i :

$$\beta_{P_i} = \text{softmax}(w_{P_i}) = \frac{\exp(w_{P_i})}{\sum_{k=1}^m \exp(w_{P_k})} \quad (10)$$

对不同元路径下的节点特征向量进行融合,得到最终的节点嵌入矩阵 Z :

$$Z = \sum_{i=1}^m \beta_{P_i} Z_{P_i} \quad (11)$$

为提高模型的精度,本文增加一个全连接层用于分类,并利用部分有标签的节点对模型进行优化,使用交叉熵作为损失函数:

$$\mathcal{L} = - \sum_{l \in V_L} Y_l \ln(CZ_l) \quad (12)$$

其中, V_L 为拥有标签的节点集合, Y_l 为节点的标签, Z_l 为该节点的最终节点嵌入矩阵, C 是分类器的参

数。最后通过反向传播对模型进行优化,学习节点的节点嵌入向量。

2.4 HAGE 算法描述

注意力的计算可以在所有节点和元路径下单独计算,因此,HAGE 模型支持并行运算。给定一个元路径 P ,节点层次的注意力机制时间复杂度为 $O(V_p F_1 F_2 K + E_p F_1 K)$,其中, V_p 是节点的数量, E_p 是基于元路径的节点对的数量, K 是多头注意力机制的数量, F_1 是节点特征的数量, F_2 是输出的节点嵌入向量的维度。总体的时间复杂度与节点数量以及基于元路径的节点对呈线性关系。

HAGE 模型的算法描述如下:

算法 1 HAGE 算法

输入 异质信息网络 $G = (V, E, T)$, 元路径集合 $\{P_1, P_2, \dots, P_m\}$, 节点属性集合 $\{a_i, i \in V\}$, 多头注意力机制数量 K

输出 节点嵌入矩阵 Z

for $i \in V$ do

计算该节点结构特征 f_i^p ;

得到节点初始特征向量 $h_i^p = a_i \parallel f_i^p$;

end

for $P_i \in \{P_1, P_2, \dots, P_m\}$ do

for $k = 1, 2, \dots, K$ do

for $i \in V$ do

得到节点 i 基于元路径的邻居 N_i^p ;

for $j \in N_i^p$ do

计算节点层次权重系数 α_{ij}^p ;

end

计算节点层次的特征向量

$$z_i^p = \sigma \left(\sum_{j \in N_i^p} \alpha_{ij}^p W h_j^p \right);$$

end

拼接得到节点层次的嵌入向量

$$z_i^p = \parallel_{k=1}^K \sigma \left(\sum_{j \in N_i^p} \alpha_{ij}^p W^k h_j^p \right);$$

end

计算元路径层次的权重系数 β_p ;

$$\text{得到最终的节点嵌入矩阵 } Z = \sum_{i=1}^m \beta_p Z_p;$$

end

计算交叉熵 $L = - \sum_{i \in V_i} Y_i \ln (CZ_i)$;

反向传播并更新 HAGE 模型的参数;

return 节点嵌入矩阵 Z

3 实验结果与分析

3.1 实验数据集

本文构建的具有节点属性的基因功能相关异质信息网络使用以下数据集:

1)使用 DisGeNET^[19] 数据集构建基因-疾病关系网络。每条边的权重根据可靠性设为 0~1,选取数据集中权重在 0.3 以上的 3 833 条基因-疾病关系来构建网络。

2)使用 miRTarBase^[20] 数据集构建基因-miRNA 关系网络。miRTarBase 是一个手工收集的经过实验验证的 miRNA 及其靶基因关系的数据集,选取其中 7 150 对经过蛋白质印迹法以及报告基因分析验证的基因-miRNA 关系,并将权重设为 1。

3)使用 2 个数据集构建 miRNA-疾病关系网络。第 1 个数据集来自文献[21]提供的 242 条 miRNA-疾病关系;第 2 个数据集来自 miRNet^[22] 数据集,选取其中疾病名称可以对应到 OMIM 编号的 666 条 miRNA-疾病关系。将 2 个数据集进行融合,去除重复数据后,共有 267 个 miRNA 和 59 个疾病组成的 878 条 miRNA-疾病关系。由于可信度较高,因此将权重设为 1。

4)使用基因本体 GO 数据库^[23-24] 中得到所有基因节点的本体信息,将其作为基因节点的节点属性,共得到 4 402 个基因节点的基因本体信息。

5)使用 MSigDB^[25] 基因集数据库中的基因家族作为节点的标签。MSigDB 将数据库中的基因集按照 PubMed 中文献的定义进行分类,同一家族的基因具有相似的功能性,它们具有同源性或者生物化学活性。结果总共有 1 185 个基因节点获得了所属的基因家族标签。

实验数据集具体描述如表 2 所示。

表 2 实验数据集描述

Table 2 Description of the experimental dataset

节点类型	节点数量	关系	关系数量	基因节点个数				元路径
				训练集	验证集	测试集	类标签	
基因	4 402	基因-疾病	3 833					GDG
疾病	2 869	基因-miRNA	7 150	350	200	635	7	GMG
miRNA	889	miRNA-疾病	878					GMDMG

3.2 对比算法

为评估本文方法的性能,选取以下算法作为对比方法:

1) GraphSAGE^[26]。GraphSAGE 通过聚集局部邻居节点的特征来学习节点的节点嵌入向量。本文使用平均聚合器版本的 GraphSAGE 来证明为不同

邻居节点以及元路径分配不同注意力的重要性。

2) GAT^[27]。GAT 是一个基于注意力机制的同质网络表示学习方法,其注意力系数通过单层前馈神经网络学习。本文在不同元路径上使用 GAT,选择表现最好的作为最终结果。

3) HAGE w/o struc。HAGE w/o struc 是 HAGE

的变种,其仅使用节点属性作为节点初始特征向量,不考虑节点在网络中的结构特征。

4) HAGE w/o node。HAGE w/o node 是 HAGE 的变种,其不使用节点层次的注意力机制,仅为不同的基于元路径的邻居节点分配相同的权重系数。

5) HAGE w/o meta。HAGE w/o meta 是 HAGE 的变种,其不使用元路径层次的注意力机制,仅为不同的元路径分配相同的权重系数。

3.3 实验设置

随机初始化模型参数,并且使用 Adam^[28] 作为模型的优化器。其中,学习率设置为 0.001,正则化参数设置为 0.005,多头注意力机制数量 K 设置为 8,元路径层次的注意力向量 q 的维度为 128,最终的节点嵌入向量维度为 128。实验运行环境为 64 位 Linux 系统,GPU 为 NVIDIA GTX 1080 Ti。

3.4 节点分类

本文使用 Micro-F1、Macro-F1、Average Precision 和 AUC 作为模型评价指标,实验结果如表 3 所示。

表 3 节点分类实验结果

Table 3 Experimental results of node classification

模型	Micro-F1	Macro-F1	Average Precision	AUC
GraphSAGE	0.681 3	0.662 0	0.507 1	0.875 9
GAT	0.697 3	0.699 0	0.568 3	0.895 6
HAGE w/o struc	0.752 8	0.753 8	0.627 4	0.916 7
HAGE w/o node	0.739 6	0.725 4	0.594 9	0.906 9
HAGE w/o meta	0.707 7	0.699 4	0.597 2	0.913 8
HAGE	0.778 6	0.787 9	0.685 8	0.924 7

由表 3 可以看出,在 Micro-F1、Macro-F1、Average Precision 和 AUC 这 4 种不同的指标下,HAGE 模型的分类效果均为最优。相比于 GraphSAGE 和 GAT 2 种同质网络表示学习方法,HAGE 由于考虑异质网络的特点即不同元路径具有不同的语义信息,为不同的元路径分配不同的权重,因此能够取得更好的分类性能。与 HAGE w/o struc、HAGE w/o node 和 HAGE w/o meta 相比,HAGE 的分类效果均有所提升,由此表明同时考虑网络结构特征、节点层次以及元路径层次注意力机制的重要性。

3.5 模型性能分析

为分析本文模型的效率性能,构建不同规模的属性异质信息网络进行实验,结果如表 4 所示。

表 4 本文模型时间效率

Table 4 Time efficiency of the proposed model

节点数量	边数量	运行时间/s
1 000	1 000	171
5 000	5 000	987
10 000	10 000	1 981

3.6 参数敏感性分析

对实验中使用的参数敏感性进行测试,研究不同参数对模型结果的影响。

1) 多头注意力机制数量

为测试多头注意力机制的效果,设置不同 K 值进行测试,当 $K=1$ 时退化为单头注意力机制,实验结果如图 2 所示。可以看出,随着 K 值的增加,AUC 的值也得到提升,当 $K=8$ 时模型的性能最好。

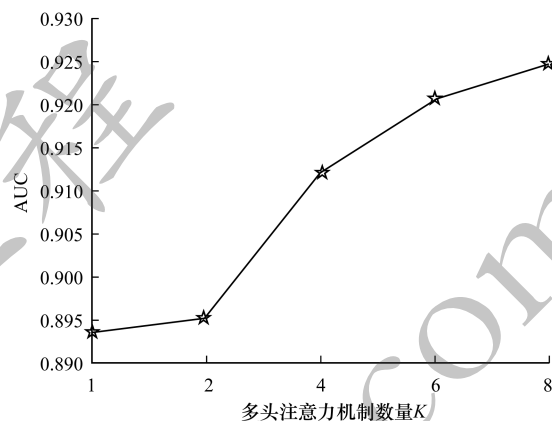


图 2 多头注意力机制数量对 AUC 的影响

Fig. 2 Effect of number of multiple attention mechanism on AUC

2) 元路径层次的注意力向量维度

元路径层次的注意力机制的分类效果受元路径层次的注意力向量 q 的影响,因此,在不同维度的注意力向量 q 下进行测试,实验结果如图 3 所示。可以看出,当注意力向量 q 的维度为 128 时,模型的性能最好。

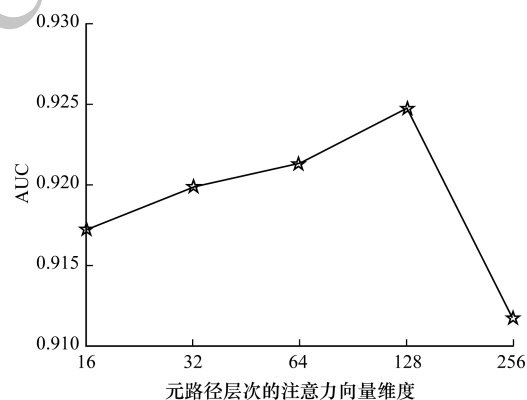


图 3 元路径层次的注意力向量维度对 AUC 的影响

Fig. 3 Effect of attention vector dimensionality in meta-path layer on AUC

3) 节点嵌入向量的维度

模型的分类效果受最终的节点嵌入向量 Z 维度的影响,因此对不同维度的节点嵌入向量 Z 进行测试,实验结果如图 4 所示。可以看出,模型的性能在维度为 128 时效果最好,后续随着维度的继续

增加, AUC 略微降低。

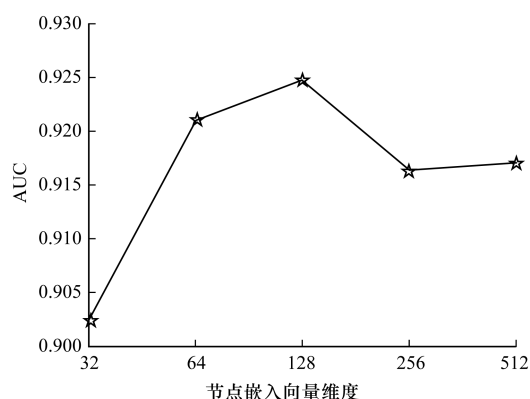


图 4 节点嵌入向量维度对 AUC 的影响

Fig.4 Effect of node embedding vector dimensionality on AUC

3.7 注意力机制性能分析

在学习基因节点的嵌入向量时,本文考虑了不同元路径下的邻居节点以及元路径的重要性,并为它们分配不同的权重系数。为更好地理解权重的意义,分别从节点层次注意力机制以及元路径层次注意力机制方面进行分析。

1) 节点层次注意力机制

本文以基因 CHEK2 为例,其基于元路径基因-疾病-基因(GDG)的邻居如图 5 所示,注意力权重系数如图 6 所示。其中,基因 CHEK2、BRCA2、RB1、BRCA1 和 TP53 同属于家族 tumor suppressors(抑癌基因),RNASEL 属于家族 protein kinases(蛋白激酶),HOXB13 属于家族 homeodomain proteins(同源域蛋白),PIK3CA 属于家族 oncogenes(致癌基因)。

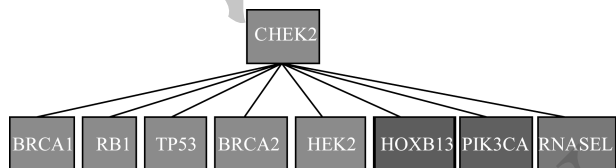


图 5 基因 CHEK2 在元路径 GDG 下的邻居

Fig.5 Neighbors of gene CHEK2 under the meta-path GDG

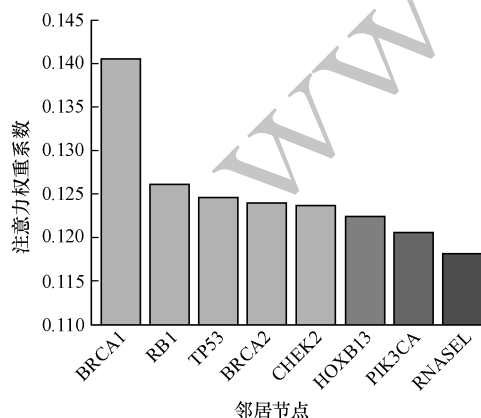


图 6 基因 CHEK2 邻居的权重系数分布

Fig.6 Weight coefficient distribution of neighbors of gene CHEK2

从图 6 可以看出,具有相同或相似功能的同家族的邻居基因节点的权重系数较大,其他家族的邻居基因节点权重系数较小。其中 BRCA1 的权重最高,文献[29]指出 CHEK2 和 BRCA1 参与的 DNA 修复有关,与乳腺癌发生有较密切的关系,因此,它们之间的功能关联更密切。由此可见,本文模型可以较好地学习到基因节点层次的重要性。

2) 元路径层次注意力机制

为分析模型学习到的不同元路径的权重系数是否反映了该元路径对基因功能预测任务的重要性,对比仅使用该元路径进行基因功能预测的结果以及该元路径的注意力权重系数,结果如图 7 所示。

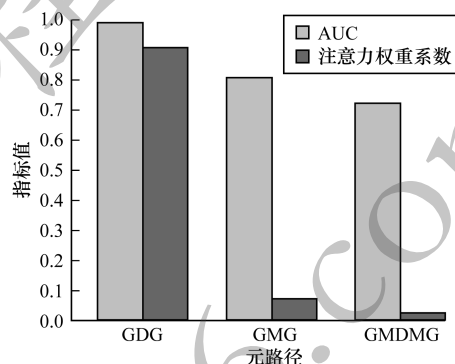


图 7 3 种元路径的 AUC 与注意力权重系数对比

Fig.7 Comparison of AUC and attention weight coefficients of three meta-paths

由图 7 可以看出,单个元路径的预测结果与该元路径的注意力权重系数是成正比的,即该元路径单独进行基因功能预测时得到的 AUC 越高,其注意力权重系数越大。由此可见,本文模型能够较好地学习到不同元路径对基因功能预测任务的重要性。

4 结束语

本文提出基于异质网络层次注意力机制的基因节点表示学习方法 HAGE。结合不同来源的数据集构建一个具有节点属性的基因功能相关网络,使用节点属性以及节点在网络中的结构特征作为节点初始向量,并通过层次注意力机制为每一个基因节点学习一个节点嵌入向量,将其用于后续的基因功能预测任务。实验结果表明,与 GraphSAGE、GAT 等方法相比,本文方法能够取得较好的预测效果。下一步将把本文方法拓展到不同的生物数据集中进行预测,如蛋白质交互网络、miRNA 基因共表达网络和代谢网络等。

参考文献

- [1] WAIN H M, BRUFORD E A, LOVERING R C, et al. Guidelines for human gene nomenclature[J]. Genomics, 2002, 79(4): 464-470.
- [2] PEARSON H. Genetics; what is a gene? [J]. Nature, 2006, 441(7092): 398-401.

- [3] MURALI T M, WU C J, KASIF S. The art of gene function prediction [J]. *Nature Biotechnology*, 2006, 24(12):1474-1475.
- [4] ENRIGHT A J, KUNIN V, OUZOUNIS C A. Protein families and TRIBES in genome sequence space [J]. *Nucleic Acids Research*, 2003, 31(15):4632-4638.
- [5] PAVLIDIS P, GILLIS J. Progress and challenges in the computational prediction of gene function using networks [J]. *FI000 Research*, 2012, 1:1-14.
- [6] LEE I, AMBARU B, THAKKAR P, et al. Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana* [J]. *Nature Biotechnology*, 2010, 28(2):149-156.
- [7] RONG Hejiang, WANG Yadong. Computation method for semantic similarity based on gene ontology [J]. *Intelligent Computer and Applications*, 2019, 9(1):108-113, 118. (in Chinese)
荣河江,王亚东. 基于基因本体的相似度计算方法[J]. *智能计算机与应用*, 2019, 9(1):108-113, 118.
- [8] WEI Wei, XIANG Yang, CHEN Qian. Research on semantic similarity based on gene ontology [J]. *Computer Engineering*, 2010, 36(20):209-210, 219. (in Chinese)
魏韡,向阳,陈千. 基于基因本体的语义相似度研究[J]. *计算机工程*, 2010, 36(20):209-210, 219.
- [9] VAFAGEE F, ROSU D, BROACKES-CARTER F, et al. Novel semantic similarity measure improves an integrative approach to predicting gene functional associations [J]. *BMC Systems Biology*, 2013, 7(1):1-17.
- [10] ZHOU Yujia, DOU Zhicheng, GE Songwei, et al. Dynamic personalized search algorithm based on recursive neural network and attention mechanism [J/OL]. *Chinese Journal of Computers*: 1-16 [2019-04-30]. <http://kns.cnki.net/kcms/detail/11.1826.TP.20190624.1709.002.html>. (in Chinese)
周雨佳,窦志成,葛松玮,等. 基于递归神经网络与注意力机制的动态个性化搜索算法 [J/OL]. *计算机学报*: 1-16 [2019-04-30]. <http://kns.cnki.net/kcms/detail/11.1826.TP.20190624.1709.002.html>.
- [11] WANG Xiao, JI Houye, SHI Chuan, et al. Heterogeneous graph attention network [C]//*Proceedings of the Web Conference 2019*. San Francisco, USA: [s.n.], 2019:1-10.
- [12] ZHOU Sheng, BU Jiajun, WANG Xin, et al. HAHE: hierarchical attentive heterogeneous information network embedding [EB/OL]. [2019-04-30]. <https://arxiv.org/pdf/1902.01475.pdf>.
- [13] SUN Yizhou, HAN Jiawei. Mining heterogeneous information networks: a structural analysis approach [J]. *ACM SIGKDD Explorations Newsletter*, 2013, 14(2):20-28.
- [14] XIONG Yun, RUAN Lu, GUO Mengjie, et al. Predicting disease-related associations by heterogeneous network embedding [C]//*Proceedings of 2018 IEEE International Conference on Bioinformatics and Biomedicine*. Washington D. C., USA: IEEE Press, 2018:548-555.
- [15] SHI Chuan, LI Yitong, ZHANG Jiawei, et al. A survey of heterogeneous information network analysis [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 29(1):17-37.
- [16] SUN Yizhou, HAN Jiawei, YAN Xifeng, et al. PathSim: meta path-based top-k similarity search in heterogeneous information networks [J]. *Proceedings of the VLDB Endowment*, 2011, 4(11):992-1003.
- [17] DONG Y, CHAWLA N V, SWAMI A. Metapath2vec: scalable representation learning for heterogeneous networks [C]//*Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA: ACM Press, 2017:135-144.
- [18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. [2019-03-30]. <https://arxiv.org/pdf/1706.03762.pdf>.
- [19] PINERO 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(D1):833-839.
- [20] 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*, 2015, 44(D1):239-247.
- [21] CHEN Hailin, ZHANG Zuping. Similarity-based methods for potential human microRNA-disease association prediction [J]. *BMC Medical Genomics*, 2013, 6(1):1-5.
- [22] 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(W1):135-141.
- [23] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene ontology: tool for the unification of biology [J]. *Nature Genetics*, 2000, 25(1):25-29.
- [24] Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong [J]. *Nucleic Acids Research*, 2018, 47(D1):330-338.
- [25] SUBRAMANIAN A, TAMAYO P, MOOTHA V K, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles [J]. *Proceedings of the National Academy of Sciences*, 2005, 102(43):15545-15550.
- [26] HAMILTON W L, YING R, LESKOVEC J. Inductive representation learning on large graphs [C]//*proceedings of the 31st Conference on Neural Information Processing System*. Long Beach, USA: [s.n.], 2017:1-11.
- [27] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks [EB/OL]. [2019-03-30]. <https://arxiv.org/pdf/1710.10903.pdf>.
- [28] KINGMA D P, BA J. Adam: a method for stochastic optimization [EB/OL]. [2019-04-10]. <https://arxiv.org/pdf/1412.6980v8.pdf>.
- [29] XIANG Beiting, LU Yunfei, ZHANG Haitian. CHEK2 gene mutation and breast cancer [J]. *Chinese Journal of General Surgery*, 2009, 24(4):331-332. (in Chinese)
向俾庭,陆云飞,张海添. CHEK2 基因突变与乳腺癌 [J]. *中华普通外科杂志*, 2009, 24(4):331-332.