

基于GAT双聚合运算与归纳式矩阵补全的关联预测

张奕^{1,2}, 郑婧¹, 蔡钢生¹, 王真梅¹

(1. 桂林理工大学 信息科学与工程学院, 广西 桂林 541004; 2. 广西嵌入式技术与智能系统重点实验室, 广西 桂林 541004)

摘要: 可计算模型能够有效替代生物实验进行长链非编码RNA(lncRNA)-疾病的关联预测,但由于存在已知数据稀疏性问题,导致现有模型的预测精度不高。针对这一局限性,提出基于图注意力网络与归纳式矩阵补全技术的双融合机制 lncRNA-疾病关联预测模型(DFMP-LDA)。引入 n 头注意力机制,设计带有双重聚合器的图注意力网络,增强 lncRNA 节点与疾病节点的特征,避免数据稀疏性导致模型预测精度不高的问题。在此基础上,针对传统图注意力网络不能直接应用于潜在 lncRNA-疾病对关联预测的问题,引入归纳式矩阵补全技术,应用增强后的节点特征重建 lncRNA-疾病关联网络,进一步提高模型的预测精度。5 折交叉验证结果表明,DFMP-LDA 预测 lncRNA-疾病关联的 AUC 值为 0.932 2, AUPR 值为 0.770 5,在时间成本上分别较 DMF-LDA、SDLDA、TPGLDA 模型节省 33.89%、32.17%、16.12%,预测性能较优。

关键词: 图注意力网络;归纳式矩阵补全;关联预测;双重聚合器;特征增强

开放科学(资源服务)标志码(OSID):



中文引用格式:张奕,郑婧,蔡钢生,等.基于GAT双聚合运算与归纳式矩阵补全的关联预测[J].计算机工程,2022,48(12):72-78.

英文引用格式:ZHANG Y, ZHENG J, CAI G S, et al. Association prediction based on duplex polymerize operation in GAT and inductive matrix completion[J]. Computer Engineering, 2022, 48(12): 72-78.

Association Prediction Based on Duplex Polymerize Operation in GAT and Inductive Matrix Completion

ZHANG Yi^{1,2}, ZHENG Jing¹, CAI Gangsheng¹, WANG Zhenmei¹

(1. School of Information Science and Engineering, Guilin University of Technology, Guilin, Guangxi 541004, China;

2. Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin, Guangxi 541004, China)

[Abstract] Computational models have been applied in long non-coding RNA(lncRNA)-disease association prediction to effectively replace traditional biological experiments. Due to the sparsity lack of input data, however, the prediction accuracy of existing models remains low. To address this limitations, Dual Fusion Mechanism Prediction model for lncRNA-Disease Association(DFMP-LDA) is proposed based on Graph Attention Network(GAT) and Inductive Martix Completion(IMC). In the first step of DFMP-LDA, a multi-head attention mechanism is introduced to design a GAT with duplex polymerizers, which enhance the features of lncRNA nodes and disease nodes. In the second step, as the traditional GAT cannot be directly applied to the potential lncRNA-disease prediction, IMC technology is introduced to reconstruct the lncRNA-disease association network. The IMC uses the enhanced node features obtained in the first step to improve model accuracy. The results of 5-fold cross-validation show that DFMP-LDA predicts association with an AUC value of 0.932 2 and an AUPR value of 0.770 5, saving 33.89%, 32.17%, 16.12% in time cost compared with DMF-LDA, SDLDA, and TPGLDA, respectively. The experimental results therefore show that DFMP-LDA has better prediction performance than previous prediction frameworks.

[Key words] Graph Attention Network(GAT); Inductive Matrix Completion(IMC); association prediction; duplex polymerizer; feature enhancement

DOI: 10. 19678/j. issn. 1000-3428. 0063069

0 概述

长链非编码 RNA(long non-coding RNA, lncRNA)是非编码 RNA 家族中的一个组成部分,它拥有长度超

过 200 个核苷酸的转录产物^[1]。研究表明 lncRNA 异常表达会导致多种复杂疾病。探寻导致疾病的 lncRNA,有助于理解疾病产生的机理,为疾病治疗及预后提供

基金项目: 国家自然科学基金(62166014);广西自然科学基金面上项目(2020GXNSFAA297255);广西嵌入式技术与智能系统重点实验室项目(2019-01-06)。

作者简介: 张奕(1977—),女,教授、博士,主研方向为生物信息学、机器学习、服务计算;郑婧、蔡钢生、王真梅,硕士研究生。

收稿日期: 2021-10-27 **修回日期:** 2022-01-04 **E-mail:** zjing029@glut.edu.cn

参考^[2]。

由于生物实验费时费力,在现实生活中大多采用可计算模型代替生物实验来实现 lncRNA-疾病的关联预测,为生物实验提供高效的更准确的候选项。目前,常用基于生物网络和基于机器学习这两类计算方法预测 lncRNA-疾病关联。

基于生物网络的方法通常需要构建基因相似性网络,构建完成后,根据 lncRNA-疾病的关联得分大小对候选的 lncRNA 进行排序来预测致病基因。最常用的是标签传播算法,如重启随机游走和 KATZ 算法,它们的主要区别在于不同的传播算法应用的底层网络不同。文献[3]根据 lncRNA 功能相似性网络建立了全局的重启随机游走算法 RWRlncD,从而对潜在的关联信息进行预测。但是该模型不能预测没有任何已知相关 lncRNA 的新疾病或没有任何已知相关疾病的孤立 lncRNA。文献[4]基于“与多种相同 miRNA 有关的 lncRNA 会导致相似疾病”这一生物假设建立了 RWRHLD 模型,从而预测 lncRNA-疾病的关联信息。该模型将 miRNA 信息加入到 lncRNA-lncRNA 网络中,与疾病相似性网络和已知的 lncRNA-疾病关联网络整合成新网络,在这个新网络中实施重启随机游走。但是该模型不适用于预测没有已知 lncRNA-miRNA 相互作用的 lncRNA,模型实用性较差。文献[5]结合已知的 lncRNA-疾病关联、lncRNA 表达谱、lncRNA 功能相似性、疾病语义相似性和高斯相互作用谱核相似性建立基于 KATZ 的 lncRNA-疾病关联预测模型 KATZLDA。虽然该模型可以发现新疾病或孤立 lncRNA,但是模型预测精度不高。

基于机器学习的方法是根据与疾病相关的已知 lncRNA 和没有任何已知关联的 lncRNA-疾病对来训练分类器和建立模型的。文献[6]将已知的疾病-lncRNA 关联和 lncRNA 表达谱信息进行整合,构建了 LRLSLDA 计算模型来预测潜在的 lncRNA-疾病关联。该模型不需要负样本且适用于预测孤立 lncRNA,但是模型最优参数的选取复杂,且模型分别将疾病空间和 lncRNA 空间作为两个分类器,对于同一个 lncRNA-疾病对会产生两个不同的得分,不同分数的选取会得出不同的预测结果。文献[7]基于贝叶斯算法整合已知的与疾病相关的 lncRNA 和多种生物学数据(基因组数据、调控和转录生物数据),预测潜在的 lncRNA-疾病关联。该模型虽然预测性能良好,但是贝叶斯分类器想要提高预测性能必须使用足够多的负样本,而此类研究缺少负样本,随机选择负样本不利于优化贝叶斯分类器的性能。文献[8]提出基于矩阵分解的 lncRNA-疾病关联预测模型 MFLDA。该模型通过矩阵分解将数据转换为低秩矩阵,不同的数据拥有各自的权重,并进一步引入迭代解,同时对权重矩阵和低秩矩阵进行优化。

优化后得到的矩阵用来重建 lncRNA-疾病关联,从而预测出潜在的 lncRNA-疾病关联。MFLDA 具有较好的适用性,很容易集成各种异构数据源来预测不同类型实体之间的关联,但是该模型寻找低秩矩阵最优秩过程复杂,且模型更偏向于选择稀疏的数据矩阵,导致模型预测精度不高。

为弥补上述不足,深度学习技术逐渐成为研究的热点。图作为一种能够抽象出实体与实体之间关系的数据结构得到广泛应用^[9],图结构可以将节点与节点间的关系通过边的权重表现出来。目前,图神经网络主要应用于相邻节点间的信息传递和汇聚。文献[10]将图神经网络中的双向门控循环网络和标签注意力机制结合,提出基于图深度学习的金融文本多标签分类算法,提升了文本分类性能。文献[11]在动态网络异常检测中引入图神经网络,使得结构和属性上的异常可以同时被获知,提升了异常检测的准确度。文献[12]将图神经网络应用到会话序列推荐算法中,引入注意力机制,提出基于复杂结构信息的图神经网络序列推荐算法,提升了会话向量在推荐过程中的准确性。文献[13]将图神经网络用于网络中物理链路路由方案路径建模,实现了对延迟抖动等端到端性能指标的有效预估。

近年来,图注意力网络(Graph Attention Network, GAT)^[14-15]也被应用于一些生物信息学任务中,如文献[16]提出一种新的基于图注意力网络的方法 GATMDA 识别微生物-疾病关联,文献[17]基于图注意力网络提出预测 circRNA-疾病关联的方法 GATCDA,文献[18]基于图注意力网络预测药物 ADMET 分类。但截止到目前,较少有使用图注意力网络进行 lncRNA-疾病关联预测的工作。另一方面,归纳矩阵补全(Inductive Matrix Completion, IMC)技术广泛应用于生物信息领域^[19-21],但也普遍存在预测精度不高的问题。如何更好地将生物信息相似性网络与归纳矩阵补全技术相结合,进一步提升预测性能,是有待研究的一个问题。

本文提出一种基于图注意力网络和归纳矩阵补全技术的双融合机制 lncRNA-疾病关联预测模型(Dual Fusion Mechanism Prediction model for lncRNA-Disease Association, DFMP-LDA)。引入 n 头注意力机制,设计带有双重聚合器的图注意力网络。传统的图注意力网络虽然可以稳定自我注意的过程,但由于节点的独立性,训练后的节点特征不明显,本文通过设计双重聚合器增强 lncRNA 节点与疾病节点特征,避免传统可计算模型中因已知数据稀疏性导致预测精度不高的问题。在此基础上,利用归纳矩阵补全技术恢复 lncRNA-疾病关联矩阵中缺失的元素,应用增强后的节点特征重建 lncRNA-疾病关联,并使用 Adam 优化器^[22]进一步提高预测精度。

1 DFMP-LDA 模型

1.1 模型框架

如图1所示,DFMP-LDA模型框架主要包括3个步骤,具体如下:

1)相似性网络建立。整合疾病集成相似性网络 $D_s^i \in \mathbb{R}^{nd \times nd}$ 和 lncRNA 集成相似性网络 $F_s^i \in \mathbb{R}^{nl \times nl}$, 得到 lncRNA-疾病的特征矩阵 $X \in \mathbb{R}^{(nl+nd) \times (nl+nd)}$ 。其中, nl 和 nd 代表 lncRNA 和疾病的数量。

2)lncRNA 特征、疾病特征增强。使用带有双重

聚合器的 n 头注意力网络训练特征矩阵 X , 先计算矩阵 X 中任意节点 i 与邻居节点集的注意力分数, 再将节点 i 的特征和邻居节点集特征进行“加”和“连接”双重聚合操作, 得到带有注意力分数的特征矩阵 $Z \in \mathbb{R}^{(nl+nd) \times (nl+nd)}$ 。

3) lncRNA-疾病关联重建。将第2)步得到的特征矩阵 Z 进行分解, 得到疾病特征矩阵 $Z^D \in \mathbb{R}^{nd \times (nl+nd)}$ 和 lncRNA 特征矩阵 $Z^L \in \mathbb{R}^{nl \times (nl+nd)}$, 通过归纳矩阵补全技术重建原始 A_{LD} 关联网络, 得到新的补全后的 lncRNA-疾病关联 $Q \in \mathbb{R}^{nl \times nd}$, 再通过 Adam 优化器进行模型优化。

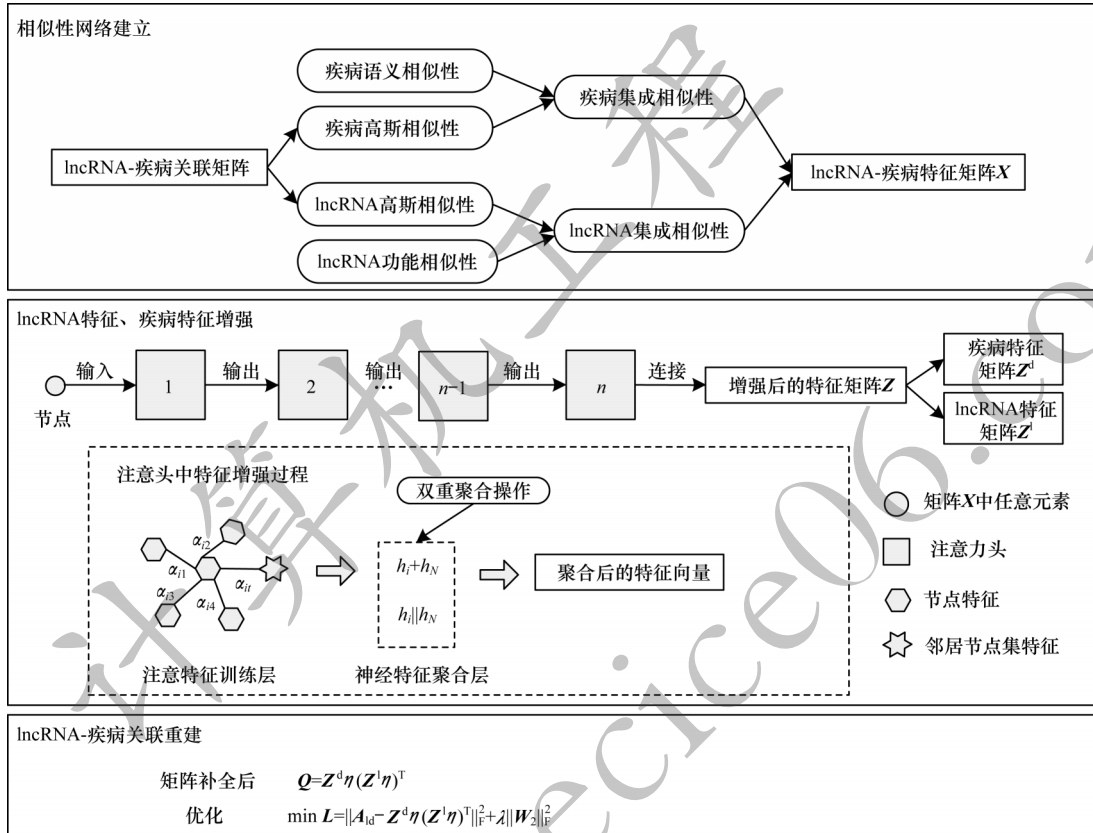


图1 DFMP-LDA模型框架

Fig.1 Framework of DFMP-LDA model

1.2 相似性网络建立

1.2.1 疾病语义相似性网络建立

利用文献[23]提出的有向无环图(Directed Acyclic Graph, DAG)计算疾病之间的语义相似性。任意疾病 d_i 对疾病 d_j 的语义贡献值用 $D_{d_i}(d_j)$ 表示, 计算公式如下:

$$D_{d_i}(d_j) = \begin{cases} 1, & d_i = d_j \\ \max \{ \gamma D_{d_i}(d_{j'}) | d_{j'} \in d_i \text{ 的孩子节点集} \}, & d_i \neq d_j \end{cases} \quad (1)$$

其中: 参数 γ 为语义贡献系数, 参考文献[23]的研究结果, 将 γ 设为其最优值 0.5。

由文献[23]可知, 两种疾病的 DAG 图的重叠部分越多, 两者相似程度越高。矩阵 $D_s \in \mathbb{R}^{nd \times nd}$ 表示疾病语义相似性网络, 矩阵元素 $D_s(d_i, d_j)$ 表示疾病 d_i 和 d_j 的语义相似性, 计算公式如下:

$$D_s(d_i, d_j) = \frac{\sum_{d_m \in (T_{d_i} \cap T_{d_j})} (D_{d_i}(d_m) + D_{d_j}(d_m))}{S(d_i) + S(d_j)} \quad (2)$$

其中: T_{d_i} 表示疾病 d_i 的 DAG 图; $S(d_i)$ 表示疾病 d_i 的语义值。 $S(d_i)$ 计算公式如下:

$$S(d_i) = \sum_{d_j \in T_{d_i}} D_{d_i}(d_j) \quad (3)$$

1.2.2 lncRNA 功能相似性网络建立

由文献[23]可知, 功能相似的 lncRNA 往往与表型相似的疾病有关, 计算两个 lncRNA 的功能相似性首先要理解疾病的语义相似性及其与 lncRNA 之间的关系。用集合 $D = \{d_1, d_2, \dots, d_i, \dots, d_{nd}\}$ 表示疾病集, $\max(d_i, D)$ 表示任意疾病 d_i 在疾病集合 D 中语义相似性最大值, 如式(4)所示:

$$\max(d_i, D) = \max_{1 \leq l \leq nd} (D_s(d_i, d_l)) \quad (4)$$

矩阵 $F_s \in \mathbb{R}^{nl \times nl}$ 表示 lncRNA 功能相似性网络, 矩阵元素 $F_s(l_i, l_j)$ 表示 lncRNA l_i 和 l_j 之间的功能相似性, 计算公式如式(5)所示:

$$F_s(l_i, l_j) = \frac{\sum_{1 \leq i \leq m} \max(d_i, D_1) + \sum_{1 \leq j \leq n} \max(d_j, D_2)}{m+n} \quad (5)$$

其中: 集合 D_1 表示与 lncRNA l_i 有关联的疾病集合; 集合 D_2 表示与 lncRNA l_j 有关联的疾病集合; m, n 分别表示集合 D_1 和集合 D_2 中疾病的数目。

1.2.3 高斯谱核相似性网络建立

如果疾病 d_i 与 lncRNA l_j 存在经实验验证的已知关联, 则定义 $I_p(d_i) = 1$; 如果疾病 d_i 与任何 lncRNA 都不存在经实验验证的已知关联, 则定义 $I_p(d_i) = 0$ 。因为某些疾病不具备语义相似性, 所以为了降低数据稀疏性对模型的影响, 将高斯核函数^[24]应用到生物信息节点之间拓扑结构的关联关系网络中。核函数在机器学习以及诸多生物信息分类中被证实是高效有用的方法, 使用高斯核函数计算出的疾病高斯谱核相似性(以下简称高斯相似性)可以代替疾病语义相似性。矩阵 $G_D \in \mathbb{R}^{nd \times nd}$ 表示疾病的高斯相似性网络, 矩阵元素 $G_D(d_i, d_j)$ 表示疾病 d_i 和疾病 d_j 的高斯相似性, 计算公式如式(6)所示:

$$G_D(d_i, d_j) = \exp(-\lambda_D \|I_p(d_i) - I_p(d_j)\|^2) \quad (6)$$

在式(6)中, λ_D 是标准化的核带宽, 计算公式如式(7)所示:

$$\lambda_D = \frac{1}{\frac{1}{nd} \sum_{i=1}^{nd} \|I_p(d_i)\|^2} \quad (7)$$

同理, 用矩阵 $G_L \in \mathbb{R}^{nl \times nl}$ 表示 lncRNA 的高斯相似性网络, 矩阵元素 $G_L(l_i, l_j)$ 表示 lncRNA l_i 和 l_j 的高斯相似性, 计算公式如式(8)所示:

$$G_L(l_i, l_j) = \exp(-\lambda_L \|I_p(l_i) - I_p(l_j)\|^2) \quad (8)$$

在式(8)中, λ_L 是标准化的核带宽, 计算公式如式(9)所示:

$$\lambda_L = \frac{1}{\frac{1}{nl} \sum_{i=1}^{nl} \|I_p(l_i)\|^2} \quad (9)$$

1.2.4 集成相似性网络建立

由于并非所有疾病都可以找到相关的 lncRNA, 如果给定疾病缺乏相关基因, 将无法得到该疾病与其他疾病的语义相似性。为了提高疾病语义相似性的准确性, 将疾病的高斯相似性和疾病语义相似性进行集成。如果疾病 d_i 与疾病 d_j 之间存在语义相似性, 则将 d_i 与 d_j 之间的语义相似性定义为疾病语义相似性 $D_s(d_i, d_j)$ 和疾病高斯相似性 $G_D(d_i, d_j)$ 的平均值, 否则等于疾病高斯相似性 $G_D(d_i, d_j)$, 由此得到疾病集成相似性网络 D_s^1 , 矩阵元素 $D_s^1(d_i, d_j)$ 表示疾病 d_i 与 d_j 的集成相似性, 计算公式如式(10)所示:

$$D_s^1(d_i, d_j) = \begin{cases} D_s(d_i, d_j) + G_D(d_i, d_j), & D_s(d_i, d_j) \neq 0 \\ G_D(d_i, d_j), & D_s(d_i, d_j) = 0 \end{cases} \quad (10)$$

同理, 用矩阵 F_s^1 表示 lncRNA 集成相似性网络, 矩阵元素 $F_s^1(l_i, l_j)$ 表示 lncRNA l_i 和 l_j 的集成相似性, 计算公式如式(11)所示:

$$F_s^1(l_i, l_j) = \begin{cases} F_s(l_i, l_j), & F_s(l_i, l_j) \neq 0 \\ G_L(l_i, l_j), & F_s(l_i, l_j) = 0 \end{cases} \quad (11)$$

将疾病集成相似性网络 D_s^1 和 lncRNA 集成相似性网络 F_s^1 结合, 定义对角矩阵 X 表示 lncRNA-疾病的特征矩阵, 用于后续模型计算。矩阵 X 如式(12)所示:

$$X = \begin{bmatrix} \mathbf{0} & D_s^1 \\ F_s^1 & \mathbf{0} \end{bmatrix} \quad (12)$$

1.3 lncRNA 特征与疾病特征加强

1.3.1 带有双重聚合器的 n 头图注意力网络构建

原始图注意力网络通过注意力分数在节点更新时自适应聚合邻居节点信息, 通过为不同的邻居节点分配不同的权重来学习图上节点的表示。GAT 利用多头注意力机制稳定自我注意的过程, 每个注意力头采用“连接”的方式聚合特征, 对于特征向量的提取效果还有待改进。为更好地提取 lncRNA 特征向量和疾病特征向量, 根据文献[16]设计带有双重聚合器的 n 头图注意力网络, 在每个注意力头设计中加入双重聚合器, 对节点特征进行“加”和“连接”双重操作, 并将前一个注意力头的输出特征作为下一个注意力头的输入特征, 经过 n 次迭代, 构造出带有双重聚合器的 n 头图注意力网络, 达到强化节点间特征的目的。

1.3.2 注意力中的特征增强过程

在注意力中, 特征增强过程具体如下:

1) 注意特征训练层

在特征矩阵 X 中任选一个元素作为节点 i , 根据图注意力网络的设计思想, 计算节点 i 的邻居节点 j 对节点 i 在第 k 次迭代中的注意力分数 e_{ij}^k , 计算公式如下:

$$e_{ij}^k = f(\mathbf{h}_i^k \mathbf{W}, \mathbf{h}_j^k \mathbf{W}) \quad (13)$$

其中: $f(\cdot)$ 表示单层神经网络; \mathbf{h}_i^k 表示节点 i 在第 k 次 ($1 \leq k \leq n$) 迭代过程中的特征向量; $\mathbf{W} \in \mathbb{R}^{(nl+nd) \times l}$ 表示权重矩阵。

为了使特征矩阵 X 中所有节点的注意力分数值在 $[0, 1]$ 区间, 使用 Softmax 函数进行标准化, 标准化后的注意力分数 e_{ij}^k 用 α_{ij}^k 表示, 计算公式如式(14)所示:

$$\alpha_{ij}^k = \frac{\exp(e_{ij}^k)}{\sum_{t \in N_i} \exp(e_{it}^k)} \quad (14)$$

其中: N_i 表示矩阵 X 中节点 i 的邻居节点集合。

$\mathbf{h}_{N_i}^k$ 表示节点 i 在第 k 次迭代时邻居节点集特征,

计算公式如式(15)所示:

$$\mathbf{h}_{N_i}^k = \sum_{t \in N_i} \alpha_{it}^k \mathbf{h}_t^k \quad (15)$$

2) 神经特征聚合层

在原始图注意力网络中, 神经特征聚合层仅仅是将注意特征训练层的特征进行“连接”操作, 为增

强节点特征,本文在注意特征训练层得到节点 i 在第 k 次迭代时的邻居节点集特征 $\mathbf{h}_{N_i}^k$ 后,根据文献[16]设计双重聚合器,通过“加”和“连接”双重聚合操作,实现对特征 \mathbf{h}_i^k 和 $\mathbf{h}_{N_i}^k$ 的聚合。以 \mathbf{Z}^k 表示第 k 次聚合后的特征向量,计算公式如下:

$$\mathbf{Z}^k = \text{LeakyReLU}((\mathbf{h}_i^k + \mathbf{h}_{N_i}^k)\mathbf{W}_1) + \text{LeakyReLU}((\mathbf{h}_i^k \parallel \mathbf{h}_{N_i}^k)\mathbf{W}_1) \quad (16)$$

其中: $\text{LeakyReLU}(\cdot)$ 表示激活函数;“+”表示加操作;“ \parallel ”表示连接操作; $\mathbf{W}_1 \in \mathbb{R}^{(nl+nd) \times k}$ 表示权重矩阵。

最后,每次聚合后的特征 \mathbf{Z}^k 经过 n 头图注意力网络,得到最终的特征矩阵 \mathbf{Z} :

$$\mathbf{Z} = \parallel_{k=1}^n \mathbf{Z}^k = \begin{bmatrix} \mathbf{Z}^D \\ \mathbf{Z}^L \end{bmatrix} \quad (17)$$

其中: \mathbf{Z}^D 表示疾病特征矩阵; \mathbf{Z}^L 表示lncRNA特征矩阵。

注:特征矩阵 \mathbf{Z} 是原始特征矩阵 \mathbf{X} 经过 n 头图注意力网络得到的,故特征矩阵 \mathbf{Z} 的前 nd 行表示疾病特征矩阵,其维数为 $nd \times (nl+nd)$,其余行表示lncRNA特征矩阵。

1.4 lncRNA-疾病关联重建

在lncRNA-疾病关联预测方面,研究者常采用矩阵补全的方式,用低秩的关联矩阵表示lncRNA-疾病的关联关系,通过较少的已知关联恢复原始矩阵^[25]。但传统的矩阵补全技术依赖于现存的lncRNA-疾病关联进行预测,由于关联矩阵中存在整行、整列数据缺失的情况,会导致冷启动发生,因此不能达到理想的预测效果。DFMP-LDA采用归纳式矩阵补全技术,打破传统矩阵补全的局限,使矩阵补全不只是单纯依赖关联矩阵,而是还加入了样本和未标记信息,实现预测未知样本的功能。

DFMP-LDA模型使用上一步推导得到的疾病特征向量 \mathbf{Z}^D 和lncRNA特征向量 \mathbf{Z}^L 补全已知的关联矩阵 \mathbf{A}_{LD} ,重建lncRNA-疾病关联,得到补全后的关联矩阵 \mathbf{Q} ,计算公式如下:

$$\mathbf{Q} = \mathbf{Z}^D \eta (\mathbf{Z}^L \eta)^T \quad (18)$$

在此基础上,通过最小化损失函数实现参数训练,根据文献[22],选择Adam优化器对矩阵 \mathbf{Q} 进行优化,具体优化过程如下:

$$\min L = \|\mathbf{A}_{LD} - \mathbf{Z}^D \eta (\mathbf{Z}^L \eta)^T\|_F^2 + \lambda \|\mathbf{W}_2\|_F^2 \quad (19)$$

其中: L 表示损失函数; η 表示衰减系数; λ 表示平衡正则项的平衡因子,其值设置为1; \mathbf{W}_2 表示权重矩阵。

2 实验与评价

2.1 实验数据集与实验环境

对原始数据库LncRNA Disease v2.0^[26]进行预处理,收集与人类疾病关系密切的lncRNA及其关联,去除重复疾病和lncRNA,最终得到本文使用的数据集Dataset1。Dataset1中含有352个经实验验证的lncRNA-疾病已知关联对,涉及156种lncRNA和190种疾病。

为了建立模型,用矩阵 \mathbf{A}_{LD} 表示352个已知的lncRNA-疾病关联, nl 和 nd 代表lncRNA和疾病的数量。矩阵元素 $A_{LD}(i,j)=1$,表示lncRNA l_i 与疾病 d_j 之间存在经实验验证的已知关联;矩阵元素 $A_{LD}(i,j)=0$,表示lncRNA l_i 与疾病 d_j 之间不存在经实验验证的已知关联。所有实验均在配置Intel Core i5-10210U,1.60 GHz CPU和64位处理器以及Windows 10操作系统的计算机上完成。

2.2 评价指标

本文采用五折交叉验证法,将已知的lncRNA-疾病关联随机分成5组,实验过程中依次选择1组lncRNA-疾病关联(即正样本)和1组相同大小的未知关联lncRNA-疾病对(即负样本)作为测试样本,剩下的4组lncRNA-疾病关联以及其余未知lncRNA-疾病对用来训练模型。通过设置不同的阈值,获得真阳率(True Positive Rate, TPR)、假阳率(False Positive Rate, FPR)、召回率、精度4个模型评价指标,根据这4个评价指标绘制ROC曲线和PR曲线,模型性能通过ROC曲线下面积(AUC)和PR曲线下面积(AUPR)衡量。为了避免随机分组的影响,每组实验重复进行10次,最后根据10次重复实验的平均值计算AUC值和AUPR值。

2.3 参数选择

本节分析注意头数目 n 和Adam优化器中衰减系数 η 对模型DFMP-LDA预测性能的影响。首先根据文献[16]将注意头数目 n 设置为4,分析衰减系数 η 对DFMP-LDA的影响。将参数值 η 从 $5E-6$ 增加到 $5E-1$ (步长为 $E-1$),对数据集Dataset1执行五折交叉验证,得到的AUC值如图2所示。可以看出,当 η 为 $5E-3$ 时,AUC值为最优值0.952 8;当 η 为 $5E-2$ 时,得到AUC的最小值0.822 8。类似地,将 η 设置为 $5E-3$ 后,改变注意头数目 n ,发现当 n 为5时,得到最优值0.932 2,如图3所示。综合以上两步,通过设置注意头数目 n 为5,衰减系数 η 为 $5E-3$,DFMP-LDA获得最佳AUC值0.932 2。

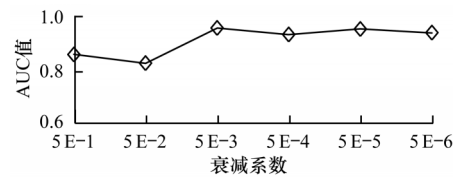


图2 不同衰减系数下的AUC值

Fig.2 AUC values under different delay factors

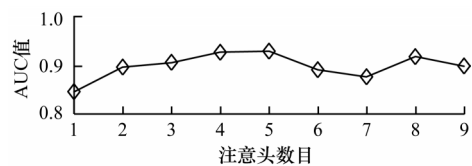


图3 不同数目注意头下的AUC值

Fig.3 AUC values under different number of attention heads

2.4 性能比较

将 DFMP-LDA 与现有的 3 种基于机器学习和基于矩阵分解的计算方法 SDLDA^[27]、DMF-LDA^[28]、TPGLDA^[29]在相同的数据集 Dataset1 上进行比较。SDLDA 使用奇异值分解提取 lncRNA 和疾病的线性特征,使用具有 2 个完全连接层的神经网络学习 lncRNA 和疾病的非线性特征,将线性特征和非线性特征结合成一个向量用于最终预测。DMF-LDA 使用带有一系列非线性隐藏层的神经网络,从 lncRNA-疾病关联矩阵中提取 lncRNA 和疾病的潜在特征,然后将这 2 个特征融合成一个新的向量,用其执行预测任务。TPGLDA 将基因疾病关联与 lncRNA 疾病关联相结合,基于分配算法预测潜在的 lncRNA 疾病关联。五折交叉验证后,得到 DFMP-LDA 与其他 3 种模型的 ROC 曲线、PR 曲线、AUC 值、AUPR 值和预测时间,分别如图 4、图 5 和表 1 所示。

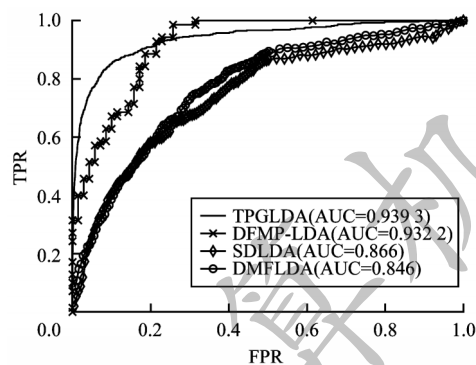


图 4 DFMP-LDA 与其他模型的 ROC 曲线
Fig.4 ROC curves of DFMP-LDA and other models

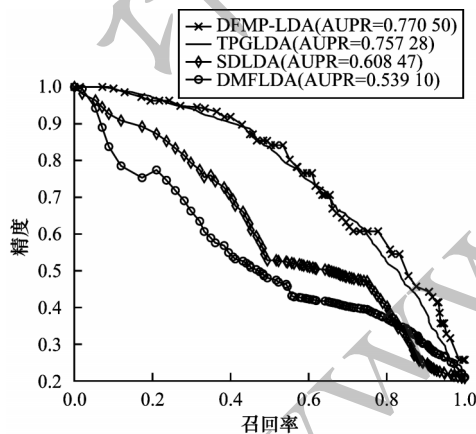


图 5 DFMP-LDA 与其他模型的 PR 曲线
Fig.5 PR curves of DFMP-LDA and other models

表 1 DFMP-LDA 与其他模型的预测性能对比

Table 1 Prediction performance comparison of DFMP-LDA and other models

模型	AUC	AUPR	预测时间/h
DFMP-LDA	0.932 2	0.770 5	0.780 0
TPGLDA	0.939 3	0.752 8	0.930 0
SDLDA	0.866 0	0.608 4	1.150 0
DMF-LDA	0.846 0	0.539 1	1.180 0

由表 1 可知,从 AUC 值和 AUPR 值来看,DFMP-LDA 的预测性能优于 SDLDA 和 DMFLDA,虽然 DFMP-LDA 的 AUC 值比 TPGLDA 低了 0.76%,但是 AUPR 值比 TPGLDA 高 1.75%,而且在预测时间上 DFMP-LDA 较 TPGLDA 节省了 16.12%。从 AUC 值、AUPR 值和预测时间 3 个方面得出,DFMP-LDA 的综合性能最优。

3 结束语

本文建立一种融合图注意力网络和归纳矩阵补全技术的 lncRNA-疾病关联预测模型,该模型利用图注意力网络的 n 头注意力机制对节点及其邻居节点集特征进行加权,并通过注意力中的双重聚合操作进一步增强节点特征。在此基础上,增强后的特征矩阵输入到归纳矩阵补全过程中,补全原始关联矩阵中缺失元素,重建 lncRNA-疾病关联网络。五折交叉验证结果显示,DFMP-LDA 与对比的 3 种计算模型相比 AUPR 值最优,AUC 值分别比 SDLDA 模型和 DMFLDA 模型高 7.64%、10.18%,虽然 AUC 略低于 TPGLDA 模型,但是预测时间节省了 16.12%。以上结果显示,DFMP-LDA 模型是一个可靠的 lncRNA-疾病关联预测模型。

如何整合多种 lncRNA 和疾病的生物信息是未来主要的研究方向。此外,因为无法获得新 lncRNA 和孤立疾病的特征,所以 DFMP-LDA 不能对这些基因和疾病进行预测。后续将考虑结合基因测序等手段收集更多的生物信息,同时对聚合器进行优化,进一步提高预测准确性。

参考文献

[1] 夏天,肖丙秀,郭俊明. 长链非编码 RNA 的作用机制及其研究方法[J]. 遗传,2013,35(3):269-280.
XIA T, XIAO B X, GUO J M. Acting mechanisms and research methods of long noncoding RNAs[J]. Hereditas, 2013,35(3):269-280. (in Chinese)
[2] 王利然,杨丽红,宁文华,等. 长链非编码 RNA 调控血管新生的研究进展[J]. 中国比较医学杂志,2021,31(4):143-149.
WANG L R, YANG L H, NING W H, et al. Research progress of long non-coding RNA involvement in angiogenesis[J]. Chinese Journal of Comparative Medicine, 2021,31(4):143-149. (in Chinese)
[3] SUN J, SHI H B, WANG Z Z, et al. Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network[J]. Molecular BioSystems, 2014,10(8):2074-2081.
[4] ZHOU M, WANG X J, LI J W, et al. Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network[J]. Molecular BioSystems, 2015,11(3):760-769.
[5] CHEN X. KATZLDA: KATZ measure for the lncRNA-disease association prediction[J]. Scientific Reports, 2015,5(1):16840.
[6] CHEN X, YAN G Y. Novel human lncRNA-disease association inference based on lncRNA expression profiles[J]. Bioinformatics, 2013,29(20):2617-2624.

- [7] ZHAO T, XU J, LIU L, et al. Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features[J]. *Molecular BioSystems*, 2015, 11(1): 126-136.
- [8] FU G Y, WANG J, DOMENICONI C, et al. Matrix factorization-based data fusion for the prediction of lncRNA-disease associations[J]. *Bioinformatics*, 2017, 34(9): 1529-1537.
- [9] 王健宗, 孔令炜, 黄章成, 等. 图神经网络综述[J]. *计算机工程*, 2021, 47(4): 1-12.
WANG J Z, KONG L W, HUANG Z C, et al. Survey of graph neural network[J]. *Computer Engineering*, 2021, 47(4): 1-12. (in Chinese)
- [10] 金雨澄, 王清钦, 高剑, 等. 基于图深度学习的金融文本多标签分类算法[J]. *计算机工程*, 2022, 48(4): 16-21.
JIN Y C, WANG Q Q, GAO J, et al. Multi-label financial text classification algorithm based on graph deep learning[J]. *Computer Engineering*, 2022, 48(4): 16-21. (in Chinese)
- [11] 郭嘉琰, 李荣华, 张岩, 等. 基于图神经网络的动态网络异常检测算法[J]. *软件学报*, 2020, 31(3): 748-762.
GUO J Y, LI R H, ZHANG Y, et al. Graph neural network based anomaly detection in dynamic networks[J]. *Journal of Software*, 2020, 31(3): 748-762. (in Chinese)
- [12] 胡承佐, 王庆梅, 李迪超, 等. 基于复杂结构信息的图神经网络序列推荐算法[J]. *计算机工程*, 2022, 48(5): 82-90, 97.
HU C Z, WANG Q M, LI D C, et al. Sequence recommendation algorithm of graph neural networks based on complex structure information[J]. *Computer Engineering*, 2022, 48(5): 82-90, 97. (in Chinese)
- [13] 车向北, 康文倩, 邓彬, 等. 一种基于图神经网络的SDN路由性能预测模型[J]. *电子学报*, 2021, 49(3): 484-491.
CHE X B, KANG W Q, DENG B, et al. A prediction model of SDN routing performance based on graph neural network[J]. *Acta Electronica Sinica*, 2021, 49(3): 484-491. (in Chinese)
- [14] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks [C]//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: Vancouver Convention Center, 2018: 1-12.
- [15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. [2021-10-21]. <https://arxiv.org/abs/1706.03762>.
- [16] LONG Y H, LUO J W, ZHANG Y, et al. Predicting human microbe-disease associations via graph attention networks with inductive matrix completion[J]. *Briefings in Bioinformatics*, 2020, 22(3): 1-13.
- [17] BIAN C, LEI X J, WU F X. GATCDA: predicting circRNA-disease associations based on graph attention network[J]. *Cancers*, 2021, 13(11): 2595.
- [18] 顾耀文, 张博文, 郑思, 等. 基于图注意力网络的药物ADMET分类预测模型构建方法[J]. *数据分析与知识发现*, 2021, 5(8): 76-85.
GU Y W, ZHANG B W, ZHENG S, et al. Predicting drug ADMET properties based on graph attention network[J]. *Data Analysis and Knowledge Discovery*, 2021, 5(8): 76-85. (in Chinese)
- [19] LU C Q, YANG M Y, LUO F, et al. Prediction of lncRNA-disease associations based on inductive matrix completion[J]. *Bioinformatics*, 2018, 34(19): 3357-3364.
- [20] HUANG L, LI X H, GUO P F, et al. Matrix completion with side information and its applications in predicting the antigenicity of influenza viruses[J]. *Bioinformatics*, 2017, 33(20): 3195-3201.
- [21] CHEN X, SUN L G, ZHAO Y. NCMCMDA: miRNA-disease association prediction through neighborhood constraint matrix completion[J]. *Briefings in Bioinformatics*, 2020, 22(1): 485-496.
- [22] KINGMA D P, BA J L. Adam: a method for stochastic optimization [C]//Proceedings of International Conference on Learning Representations. New York, USA: ACM Press, 2015: 1-15.
- [23] 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.
- [24] VAN LAARHOVEN T, NABUURS S B, MARCHIORI E. Gaussian interaction profile kernels for predicting drug-target interaction[J]. *Bioinformatics*, 2011, 27(21): 3036-3043.
- [25] 陈蕾, 陈松灿. 矩阵补全模型及其算法研究综述[J]. *软件学报*, 2017, 28(6): 1547-1564.
CHEN L, CHEN S C. Survey on matrix completion models and algorithms[J]. *Journal of Software*, 2017, 28(6): 1547-1564. (in Chinese)
- [26] CHEN G, WANG Z Y, WANG D Q, et al. lncRNADisease: a database for long-non-coding RNA-associated diseases[J]. *Nucleic Acids Research*, 2012, 41(D1): D983-D986.
- [27] ZENG M, LU C Q, ZHANG F H, et al. SDLDA: lncRNA-disease association prediction based on singular value decomposition and deep learning[J]. *Methods*, 2020, 179: 73-80.
- [28] ZENG M, LU C, FEI Z, et al. DMFLDA: a deep learning framework for predicting lncRNA-disease associations[J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021, 18(6): 2353-2363.
- [29] DING L, WANG M H, SUN D D, et al. TPGLDA: novel prediction of associations between lncRNAs and diseases via lncRNA-disease-gene tripartite graph[J]. *Scientific Reports*, 2018, 8(1): 1065.

编辑 金胡考