基于异质图注意力网络的miRNA与疾病关联预测算法

李政伟[1,2],李佳树[1,2],尤著宏[3],聂 茹[2],赵 欢[2],钟堂波[2]
(1. 中国矿业大学矿山数字化教育部工程研究中心,江苏徐州 221116; 2. 中国矿业大学计算机科学与技术学院,江苏徐州 221116; 3. 西北工业大学计算机学院,陕西西安 710129)

摘 要: 众多实验表明,microRNA(miRNA)的异常表达与人类复杂疾病的产生和演化有关.识别miRNA与疾病间的关联有助于促进临床医学的发展.然而,传统的实验方法往往耗时耗力、效率低下,因此迫切需要高效的计算方法对miRNA与疾病间的潜在关联进行预测.本文提出了一种基于异质图注意力网络的端到端的计算模型来预测miRNA与疾病的关联.该方法通过多头注意力机制捕获异质邻居的结构和属性信息,并将其与中心顶点的属性信息进行融合,从而构建出更具表达能力的miRNA和疾病的特征嵌入,进而通过全连接层对miRNA与疾病间的潜在关联进行预测.5折交叉验证结果显示,该模型分别在HMDD v2.0和HMDD v3.0数据集上取得了93.52%和94.82%的AUC值.此外,关于食管肿瘤的病例研究结果显示,该模型预测的前50个miRNA中有48个得到了证实.上述实验结果表明,该模型可作为一种可靠的工具预测候选疾病的相关miRNA.

关键词: 图注意力网络; miRNA-疾病关联; 异质图; 深度学习
中图分类号: TP399 **文献标识码:** A **文章编号:** 0372-2112(2022)06-1428-08
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20201116

Associations Prediction Algorithm of MiRNAs and Diseases Based on Heterogeneous Graph Attention Network

LI Zheng-wei[1,2], LI Jia-shu[1,2], YOU Zhu-hong[3], NIE Ru[2], ZHAO Huan[2], ZHONG Tang-bo[2]
(1. *Engineering Research Center of Mine Digitalization of Ministry of Education, China University of Mining and Technology, Xuzhou, Jiangsu* 221116, *China*;
2. *School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu* 221116, *China*;
3. *School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi* 710129, *China*)

Abstract: Lots of experiments have shown that the abnormal expression of microRNA(miRNA) is related to the evolution and progression of human complex diseases. Identifying associations between miRNAs and diseases is beneficial to promote the development of clinical medicine. However, traditional experimental methods are often time-consuming and inefficient, so there is an urgent need for efficient computational methods to predict the potential associations between miRNAs and diseases. In this paper, we propose an end-to-end computational model based on heterogeneous graph attention network to predict the associations between miRNAs and diseases. This model captures the structure and attribute information of heterogeneous neighbors via the multi-head attention mechanism, and fuses them with the attribute information of the central vertex to generate more representative feature embeddings of miRNAs and diseases, and then predicts the potential associations between miRNAs and diseases through a fully connected layer. The 5-fold cross-validation results show that our model achieves 93.52% and 94.82% AUC values based on HMDD v2.0 and HMDD v3.0 datasets, respectively. In addition, the case study on esophageal neoplasms shows that 48 of the top 50 miRNAs predicted by our model are confirmed. The above experimental results indicate that our model can be used as a reliable tool to predict candidate disease-related miRNAs.

Key words: graph attention network; microRNA-disease associations; heterogeneous graph; deep learning

收稿日期:2020-10-12;修回日期:2021-04-08;责任编辑:王天慧
基金项目:国家自然科学基金(No.61732012,No.61972399);科技创新2030—"新一代人工智能"重大项目(No.2018AAA0100100)

1 引言

MicroRNA(缩写为miRNA)是一类小的、内源性、非编码单链RNA分子,其长度大约为22个核苷酸,在人类蛋白质编码基因的调控中起到关键作用[1]. 诸多研究分析显示miRNA在众多生物进程中,例如细胞增殖、分化、凋亡、病毒感染等[2],起着至关重要的作用. 同时,miRNA的突变或者异常表达往往会诱导多种人类复杂疾病的产生和演化[3]. 例如,通过单变量Cox回归分析发现,miR-155和miR-150的表达水平对淋巴瘤病人的无进展生存期(Progression-Free-Survival,PFS)有着重要影响[4]. 因此,识别miRNA与疾病间的潜在关联有助于医疗人员从分子角度理解疾病的病理机理,从而促进临床诊断、治疗和预后.

传统的识别miRNA与疾病间潜在关联的生物学湿实验方法主要有Northern杂交[5]、逆转录聚合酶链反应[6]、微阵列分析[7]等. 但是这些方法往往会受到环境影响,且需要大量的资金和时间投入,效率低下. 随着计算机的存储和运算能力的飞速发展,以及大量收集相关miRNA和疾病信息的生物数据库的建立,设计更加高效的计算方法,实现大规模、高置信度地预测miRNA与疾病间的潜在关联,逐渐受到科研人员的广泛关注[8,9].

启发于深度学习理论在生物信息学领域的成功应用[10,11],本文提出一种基于异质图注意力网络的端到端模型即HGATMDA(Heterogeneous Graph Attention Network for MiRNA-Disease Associations Prediction)来预测miRNA与疾病间的潜在关联. 具体而言,首先将集成的miRNA相似性信息、集成的疾病相似性信息以及经实验验证的miRNA-疾病关联整合进miRNA-疾病异质图中,并设计了顶点类型转换矩阵将异质的顶点特征投影至同一向量空间中;其次,采用多头注意力机制聚合异质邻居顶点特征,并将聚合后的特征与中心顶点的属性特征相融合,得到更具有表达能力的miRNA和疾病顶点的特征表示;之后,将miRNA-疾病对特征输入至全连接层(Fully Connected Layer,FCL)中得出预测的概率;最后,根据预测的概率与标签间的损失对整个模型进行端到端的训练. HGATMDA模型的流程图如图1所示.

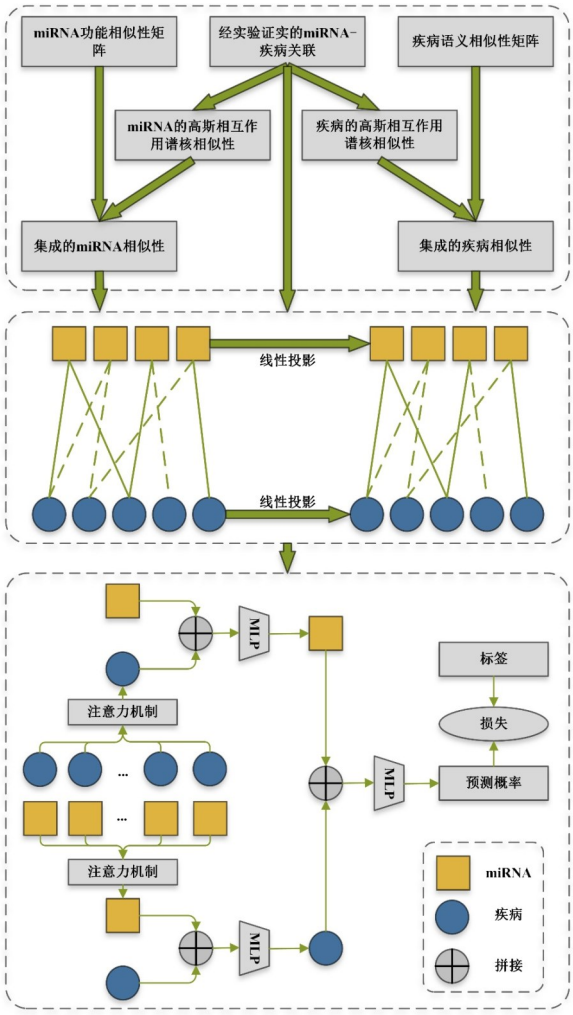


图1　基于异质图注意力网络的miRNA-疾病关联预测模型流程图

2 材料及方法

2.1 人类miRNA-疾病关联

本实验从“https://www.cuilab.cn/hmdd”下载了HMDD v2.0和HMDD v3.0数据集来对模型的预测效果进行验证[12]. 如表1所示,经过数据预处理,HMDD v2.0数据集中包含383种疾病与495种miRNA间5 430条经实验证实的miRNA-疾病关联,HMDD v3.0数据集中包含850种疾病与1 057种miRNA间32 226条经实验证实的miRNA-疾病关联. 为了便于存储,本实验采用二值矩阵$\boldsymbol{A}(\mathrm{nd}\times\mathrm{nm})$来表示miRNA与疾病间的关联,其中nd表示疾病数目,nm表示miRNA数目. 若疾病$d(i)$与miRNA $m(j)$有关联,则二值矩阵$\boldsymbol{A}$对应位置的元素$A(d(i),m(j))$被赋值为1,否则为0.

表1　本文所用miRNA-疾病关联信息

数据集	HMDD v2.0	HMDD v3.0
疾病数目	383	850
miRNA数目	495	1 057
关联数目	5 430	32 226

2.2 MiRNA功能相似性

基于表型相似的疾病可能与功能相似的miRNA发生关联这一基本生物学假设,Wang等人提出一种计算miRNA功能相似性的模型[13]. 本实验从“https://www.cuilab.cn/files/images/cuilab/misim.zip”下载了miRNA功能相似性数据,并构建出长度为nm的方

阵**FSM**来存储miRNA的功能相似性.

2.3 疾病语义相似性

本实验基于美国国家医学图书馆的MeSH(Medical Subject Headings)数据库计算疾病的语义相似性[14]. 疾病间抽象出的数据结构可以用有向无环图(Directed Acyclic Graph,DAG)进行表示. 具体而言,采用$\mathrm{DAG}(d(i))=(d(i),T(d(i)),E(d(i)))$来描述疾病$d(i)$,其中,$T(d(i))$表示包含顶点$d(i)$自身及其祖先顶点的集合,$E(d(i))$表示包含从$d(i)$的祖先顶点到顶点$d(i)$的路径上所有直连的边的集合. 因此,疾病$d(k)$对$d(i)$的语义贡献值计算如下:

$$\boldsymbol{D}1_{d(i)}(d(k))=\begin{cases}1, & d(k)=d(i)\\ \max\left\{\Delta\times\boldsymbol{D}1_{d(i)}(d(k)')\,\middle|\,d(k)'\in C\{d(k)\}\right\}, & d(k)\neq d(i)\end{cases}\tag{1}$$

式(1)中,Δ表示语义贡献衰减因子,设置为0.5;$C\{d(k)\}$表示疾病$d(k)$的孩子顶点集合. 于是,疾病$d(i)$的语义值定义为

$$\mathbf{DS}1(d(i))=\sum_{d(k)\in T(d(i))}\boldsymbol{D}1_{d(i)}(d(k))\tag{2}$$

基于不同疾病间共享的DAG部分越多,就具有更高的语义相似性这一假设(其中共享的DAG部分指不同疾病顶点的祖先顶点的交集),疾病语义相似性矩阵**DSSM**1计算如下:

$$\mathbf{DSSM}1(d(i),d(j))=\frac{\sum_{d(t)\in T(d(i))\cap T(d(j))}\left(\boldsymbol{D}1_{d(i)}(d(t))+\boldsymbol{D}1_{d(j)}(d(t))\right)}{\mathbf{DS}1(d(i))+\mathbf{DS}1(d(j))}\tag{3}$$

由于不同疾病在DAG中出现的次数不尽相同,同一层DAG中的疾病往往也会有不同的疾病语义贡献值,因此,根据疾病在DAG中出现的次数计算另一种疾病$d(k)$对$d(i)$的语义贡献值的计算如下:

$$\boldsymbol{D}2_{d(i)}(d(k))=-\log\left(\frac{\text{包含}d(k)\text{的DAG的数量}}{\mathrm{nd}}\right)\tag{4}$$

相应地,第二种疾病$d(i)$的语义值以及疾病的语义相似性矩阵*DSSM*2计算如下:

$$\mathbf{DS}2(d(i))=\sum_{d(k)\in T(d(i))}\boldsymbol{D}2_{d(i)}(d(k))\tag{5}$$

$$\mathbf{DSSM}2(d(i),d(j))=\frac{\sum_{d(t)\in T(d(i))\cap T(d(j))}\left(\boldsymbol{D}2_{d(i)}(d(t))+\boldsymbol{D}2_{d(j)}(d(t))\right)}{\mathbf{DS}2(d(i))+\mathbf{DS}2(d(j))}\tag{6}$$

整合上述两种疾病语义相似性矩阵,计算最终的疾病语义相似性矩阵**DSSM**如下:

$$\mathbf{DSSM}(d(i),d(j))=\frac{\mathbf{DSSM}1(d(i),d(j))+\mathbf{DSSM}2(d(i),d(j))}{2}\tag{7}$$

2.4 MiRNA与疾病的高斯相互作用谱核相似性

鉴于上述方法得出的miRNA功能相似性矩阵以及疾病语义相似性矩阵具有稀疏性,本实验引入高斯相互作用谱核相似性[15]来进一步完善miRNA和疾病的相似性信息. 根据miRNA $m(i)$是否与每一种疾病发生关联,构建二值向量$\mathbf{IP}(m(i))$表示miRNA的相互作用谱. miRNA的高斯相互作用谱核相似性矩阵**MGSM**为

$$\mathbf{MGSM}(m(i),m(j))=\exp\left(-r_m\|\mathbf{IP}(m(i))-\mathbf{IP}(m(j))\|^2\right)\tag{8}$$

式(8)中,r_m用以调控函数的带宽,可通过规范化参数r_m'计算而得:

$$r_m=\frac{r_m'}{\left(\frac{1}{nm}\sum_{i=1}^{nm}\|\mathbf{IP}(m(i))\|^2\right)}\tag{9}$$

式(9)中,r_m'设置为1. 同样地,疾病的高斯相互作用谱核相似性矩阵**DGSM**可由下式计算:

$$\mathbf{DGSM}(d(i),d(j))=\exp\left(-r_d\|\mathbf{IP}(d(i))-\mathbf{IP}(d(j))\|^2\right)\tag{10}$$

$$r_d=\frac{r_d'}{\left(\frac{1}{\mathrm{nd}}\sum_{i=1}^{nd}\|d(i)\|^2\right)}\tag{11}$$

其中,二值向量$\mathbf{IP}(d(i))$表示疾病$d(i)$是否与每一种miRNA存在关联,r_d'设置为1.

2.5 MiRNA与疾病的集成相似性

本文将miRNA与疾病的高斯相互作用谱核相似性矩阵整合进miRNA的功能相似性矩阵和疾病的语义相似性矩阵中,从而得到集成的miRNA相似性矩阵**IM**与集成的疾病相似性矩阵**ID**.

$$\mathbf{IM}(m(i),m(j))=\begin{cases}\mathbf{MFSM}(m(i),m(j)), & m(i)\text{与}m(j)\text{存在相似性}\\ \mathbf{MGSM}(m(i),m(j)), & \text{其他}\end{cases}\tag{12}$$

$$\mathbf{ID}(d(i),d(j))=\begin{cases}\mathbf{DSSM}(d(i),d(j)), & d(i)\text{与}d(j)\text{存在相似性}\\ \mathbf{DGSM}(d(i),d(j)), & \text{其他}\end{cases}\tag{13}$$

2.6 MiRNA-疾病异质图

本文构建了miRNA-疾病异质图,共包含两类顶点(分别为miRNA顶点与疾病顶点),以及一类边(表示miRNA与疾病间的关联). 其中,miRNA顶点数目为nm,疾病顶点数目为nd,miRNA与疾病间的关联数目为2S. 由于HMDD数据集中经实验证实的miRNA-疾病关联数目远小于miRNA与疾病间的未知关联数目,因此,从所有的未知关联中随机选取S条miRNA-疾病关联作为负样本. 在miRNA和疾病顶点间相应地添加S条正边与S条负边,并将miRNA的集成相似性信息赋给miRNA $m(i)$顶点,作为其属性特征$F_{m(i)}$,即

$$F_{m(i)}=\left(v_1,v_2,v_3,\cdots,v_{nm-1},v_{nm}\right) \tag{14}$$

式(14)中,$F_{m(i)}$表示矩阵**IM**的第i行,v_j表示miRNA $m(i)$与$m(j)$间的集成相似性. 类似地,将疾病的集成相似性信息赋给疾病$d(i)$顶点,作为其属性特征$F_{d(i)}$,即

$$F_{d(i)}=\left(w_1,w_2,w_3,\cdots,w_{nd-1},w_{nd}\right) \tag{15}$$

式(15)中,$F_{d(i)}$表示矩阵**ID**的第i行,w_j表示疾病$d(i)$与$d(j)$间的集成相似性.

2.7 异质图注意力网络

由于miRNA-疾病异质图中的miRNA顶点和疾病顶点分别处于不同的特征空间中,对于每一种类型的顶点(例如类型为Φ_i的顶点),本实验设计了顶点类型转换矩阵$\boldsymbol{W}_{\Phi_i}$将miRNA顶点和疾病顶点投影到同一向量空间中进行计算,即

$$H_i=F_i\boldsymbol{W}_{\Phi_i} \tag{16}$$

式(6)中,F_i和H_i分别表示顶点i的初始属性特征和投影后的属性特征;$\boldsymbol{W}_{\Phi_i}$表示针对类型为Φ_i的顶点的投影矩阵,该矩阵可将不同向量空间的顶点投影至D维的向量空间中. 因此,miRNA顶点和疾病顶点可处在同一个向量空间中进行后续计算. 由于异质邻居顶点对中心顶点存在不同程度的影响,本实验采用多头注意力机制[16,17]聚合异质顶点的邻域信息,并将其与中心顶点的属性信息进行融合,从而得到包含异质图结构与顶点属性信息的miRNA和疾病的有效特征嵌入. 首先计算中心顶点i与其邻居顶点j之间的注意力分数e_{ij}:

$$e_{ij}=\mathrm{LeakyReLU}\left(H_i\left(H_j\right)^{\mathrm{T}}\right) \tag{17}$$

式(17)中,LeakyReLU为非线性激活函数(负输入斜率为0.2). 仅计算顶点$j\in\mathcal{N}_i$的注意力分数e_{ij},其中,$\mathcal{N}_i$表示顶点i的一阶异质邻居顶点集合. 采用softmax函数规范化注意力分数e_{ij},并计算出注意力权重系数α_{ij},即

$$\alpha_{ij}=\mathrm{softmax}_j\left(e_{ij}\right)=\frac{\exp\left(e_{ij}\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(e_{ik}\right)} \tag{18}$$

再根据顶点i的投影特征和注意力权重系数计算出顶点i的异质邻居聚合特征H_i',即

$$H_i'=\sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}H_i\right) \tag{19}$$

式(19)中,$\sigma(\cdot)$表示ELU激活函数. 为了使模型学习到的特征嵌入更加稳定,按照上述公式独立计算K次,并将每次计算的结果拼接起来作为顶点i最终的异质邻居聚合特征H_i',即

$$H_i'=\Big\Vert_{k=1}^{K}\sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^kH_i^k\right) \tag{20}$$

上述过程仅聚合了异质邻居特征,却忽略了中心顶点特征,因此将异质邻居聚合特征H_i'与中心顶点特征F_i拼接,并通过全连接层进行特征融合,表示为

$$Z_i=\sigma\left(g\left(H_i'\oplus F_i\right)\right) \tag{21}$$

式(21)中,$g(\cdot)$表示输出维度为64的全连接层,$\oplus$表示特征拼接操作. 最终分别获得64维度的miRNA嵌入特征Z_m和64维度的疾病嵌入特征Z_d.

2.8 目标优化

为了获得miRNA $m(i)$与疾病$d(j)$间关联的预测概率,将上述得到的miRNA和疾病嵌入特征拼接,并通过全连接层生成预测概率$\hat{y}_{ij}$,即

$$\hat{y}_{ij}=\mathrm{sigmoid}\left(f\left(Z_{m(i)}\oplus Z_{d(j)}\right)\right) \tag{22}$$

式(22)中,$f(\cdot)$表示输入维度为128,输出维度为1的全连接层;sigmoid$(\cdot)$表示非线性激活函数.

本文采用交叉熵损失计算模型的预测值与标签间的损失,表示为

$$L=-\sum_{i,j\in\mathcal{Y}\cup\mathcal{Y}^-}\left(y_{ij}\log\hat{y}_{ij}+\left(1-y_{ij}\right)\log\left(1-\hat{y}_{ij}\right)\right) \tag{23}$$

式(23)中,y_{ij}表示miRNA $m(i)$与疾病$d(j)$间的关联标签;$\mathcal{Y}$和$\mathcal{Y}^-$分别表示正样本和负样本对应的顶点集. 最后,采用反向传播算法对整个模型进行端到端的训练.

3 实验结果与分析

3.1 实现细节

本实验基于深度图库(Deep Graph Library,DGL)[18]实现,后端采用PyTorch框架,并采用Adam作为模型的优化器. 经过网格搜索,设置学习率(Learning Rate)为0.0001,权重衰减(Weight Decay)为5×10^{-3}. 为了防止过拟合,设置丢弃率(Dropout)为0.6. 为了保持较高的计算效率,设置多头注意力头数K为8,投影向量维度D为64. 为了充分训练模型的参数,训练批次(Epochs)设置为1 000.

3.2 评价指标

本文采用准确率(Accuracy)、精确率(Precision)、召

回率(Recall)以及F1值(F1-score)作为模型的评价指标,具体计算公式如下:

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+FP+\text{FN}} \tag{24}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{25}$$

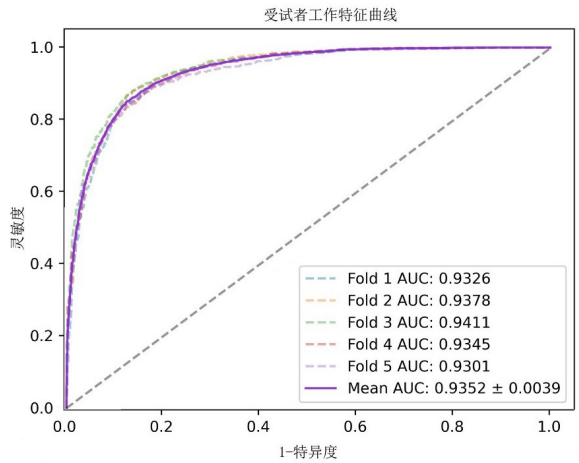
$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{26}$$

$$\text{F1-score}=\frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \tag{27}$$

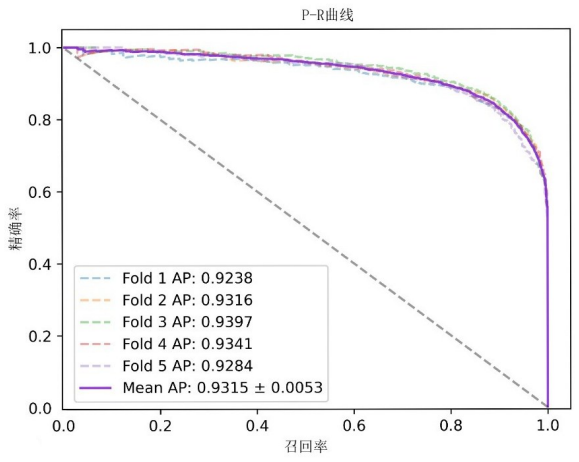
式(24)~(27)中,TP,TN,FP,FN分别表示真正例数、真负例数、假正例数和假负例数.此外,本文还绘制了受试者工作特征(Receiver Operating Characteristic,ROC)曲线以及精确率-召回率(Precision-Recall,P-R)曲线来直观地显示模型的预测能力,并分别计算了ROC曲线下面积(Area Under the Curve,AUC)以及P-R曲线下面积(Average Precision,AP)来综合评估模型的预测能力.

3.3 模型预测能力评估

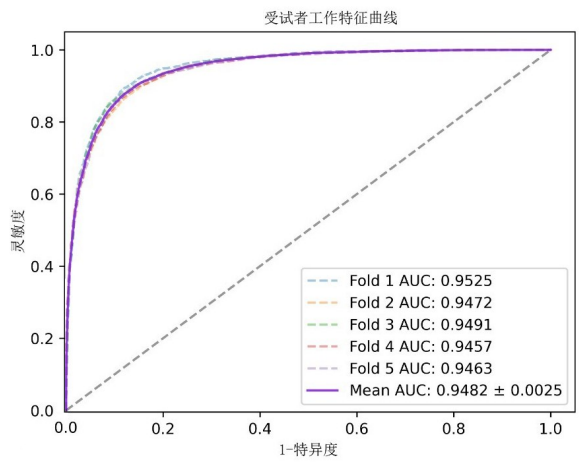
本实验采用5折交叉验证法(5-fold cross-validation)对模型的预测能力进行评估.本文所提模型在HMDD v2.0数据集上的预测结果如表2所示,取得了86.14%的准确率、86.10%的精确率、86.25%的召回率以及86.15%的F1值.所提模型在HMDD v3.0数据集上的预测结果如表3所示,取得了87.85%的准确率、88.02%的精确率、87.64%的召回率以及87.83%的F1值.所提模型的5折交叉验证ROC曲线和P-R曲线如图2所示,该模型在HMDD v2.0数据集上取得了93.52%的AUC值和93.15%的AP值,在HMDD v3.0数据集上取得了94.82%的AUC值和94.66%的AP值.由于HMDD v3.0数据集中包含了更多的样本数量,且深度学习模型在更大的数据集上一般体现出更优的拟合效果,相较于HMDD v2.0数据集,所提模型在HMDD v3.0数据集上关于6项评价指标均表现出更高的值.为方便后续对比实验的展开,接下来的实验均采用



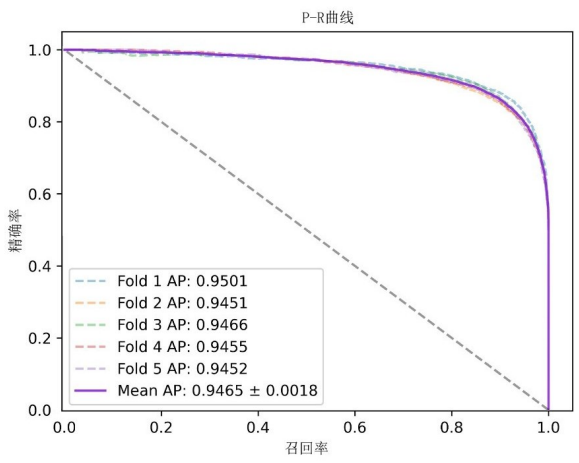
(*a*) 在HMDD v2.0数据集上生成的ROC曲线



(*b*) 在HMDD v2.0数据集上生成的P-R曲线



(*c*) 在HMDD v3.0数据集上生成的ROC曲线



(*d*) 在HMDD v3.0数据集上生成的P-R曲线

图2 所提模型基于5折交叉验证的实验结果图

HMDD v2.0数据集为基准数据集.

表2　所提模型基于5折交叉验证在HMDD v2.0数据集上的实验结果

测试集	准确率/%	精确率/%	召回率/%	F1值/%
1	85.75	84.08	87.94	85.97
2	87.06	85.94	88.34	87.12
3	86.65	88.50	84.27	86.33
4	85.45	85.69	85.84	85.77
5	85.77	86.30	84.86	85.57
均值	86.14	86.10	86.25	86.15

表3　所提模型基于5折交叉验证在HMDD v3.0数据集上的实验结果

测试集	准确率/%	精确率/%	召回率/%	F1值/%
1	88.44	89.08	87.91	88.49
2	87.31	87.11	87.38	87.24
3	88.04	87.83	88.39	88.11
4	87.66	88.41	86.74	87.57
5	87.82	87.68	87.76	87.72
均值	87.85	88.02	87.64	87.83

3.4　特征融合的影响

本实验将中心顶点特征与其邻居聚合特征相融合作为最终的miRNA和疾病的特征. 为了对比这种融合方式对模型预测能力的影响,本实验分别设计了只采用中心顶点特征的模型和只采用异质邻居聚合特征的模型,最终的对比结果如表4所示. 从表中可以看出,本文所提模型在这三个模型中取得了最高的准确率、精确率、F1值、AUC值以及AP值;尽管只采用邻居聚合特征的模型取得了最高的召回率,但其在其他5项指标上均远低于本文所提模型. 本文所提模型以多头注意力机制形式从多个角度探索miRNA-疾病异质图中异质顶点间复杂的交互信息,生成涵盖异质图结构及顶点属性信息的嵌入特征,进一步加强miRNA和疾病特征的表达能力,提高模型的预测能力.

3.5　与其他方法的比较

为了进一步验证本文所提模型的有效性,将其与WBSMDA[19],BNPMDA[20],KBMFMDA[21],WBNPMD[22],M2GMDA[23],KNMBP[24],MCLPMDA[25]等7个模型基于5折交叉验证的平均AUC值进行比较,此外,还对比了不同注意力头数K对所提模型AUC值的影响,详细的对比结果如表5所示. 从表中可以看出,适当增加注意力的头数可以提高模型的预测能力,但过多的注意力头数反而会对模型预测能力起反作用. 最终,本文选择的注意力头数$K=8$,其对应的AUC值为93.52%,在所有8个模型中最高.

3.6　病例研究

为了进一步评估本文所提模型在预测特定疾病潜在的相关miRNA方面的性能,本文针对食管肿瘤疾病开展了病例研究. 首先采用HMDD v2.0数据集对模型进行训练,然后预测与食管肿瘤有潜在相关的前50种miRNA,最后通过dbDEMC[26]和miR2Disease[27]数据库进行验证.

表4　所提模型与未进行特征融合的模型的对比实验结果

模型	准确率/%	精确率/%	召回率/%	F1值/%	AUC值/%	AP值/%
中心顶点特征	84.72	85.41	83.77	84.57	92.47	92.15
邻居聚合特征	78.63	71.62	**95.03**	81.65	91.94	92.01
HGATMDA	**86.14**	**86.10**	86.25	**86.15**	**93.52**	**93.15**

表5　所提模型与其他方法的AUC值的对比结果

模型	AUC值/%
WBSMDA	81.85
BNPMDA	89.80
KBMFMDA	90.08
WBNPMD	91.73
M2GMDA	91.82
KNMBP	93.13
MCLPMDA	93.20
HGATMDA-K1	93.24
HGATMDA-K2	93.34
HGATMDA-K4	93.50
HGATMDA-K8	**93.52**
HGATMDA-K16	93.49

食管肿瘤是一种发生在食管组织中的恶性肿瘤,全球范围内每年大约会有30万人死于食管肿瘤. 本文选择食管肿瘤作为病例研究对象. 实验验证结果如表6所示,通过在dbDEMC和miR2Disease两个数据库中进行核实,模型预测的前25个miRNA中有24个被证实,前50个miRNA中有48个被证实. 因此,本文所提出的模型能有效预测出潜在的疾病相关miRNA,可作为一种便捷的工具指引研究人员开展相关具体的生物实验研究.

4　结论

本文提出了一种基于异质图注意力网络的端到端计算模型(HGATMDA)预测潜在的miRNA与疾病间的关联. 该模型首先将miRNA和疾病间的多重相似性信息建模为异质图,并设计了顶点类型的转换矩阵将异质的顶点特征投影至同一向量空间中;然后采用多头注意力机制聚合中心顶点的异质邻居特征,并将其与中心顶点的特征进行有效融合,得到更具有表达能力的miRNA和疾病特征嵌入;最后,将得到的miRNA和疾病特征嵌入输入至全连接层中对潜在的miRNA与疾病间关联进行预测. 5折交叉验证的结果表明,本文所提模型在多项评价指标上均取得了较为满意的结果.

表6 所提模型预测出的前50个与食管肿瘤有关联的miRNA

miRNA (1-25)	证据	miRNA (26-50)	证据
hsa-mir-17	dbDEMC	hsa-mir-181b	dbDEMC
hsa-mir-221	dbDEMC	hsa-mir-106a	dbDEMC
hsa-mir-29a	dbDEMC	hsa-mir-30a	dbDEMC
hsa-mir-16	dbDEMC	hsa-mir-93	dbDEMC
hsa-mir-222	dbDEMC	hsa-mir-23a	dbDEMC
hsa-mir-18a	dbDEMC	hsa-mir-132	dbDEMC
hsa-mir-19b	dbDEMC	hsa-let-7f	dbDEMC
hsa-mir-125b	dbDEMC	hsa-let-7g	dbDEMC
hsa-mir-1	dbDEMC	hsa-mir-23b	dbDEMC
hsa-mir-29b	dbDEMC	hsa-mir-7	dbDEMC
hsa-mir-122	未证实	hsa-mir-133b	dbDEMC
hsa-mir-206	dbDEMC	hsa-mir-125a	dbDEMC
hsa-mir-106b	dbDEMC	hsa-mir-107	dbDEMC, miR2Disease
hsa-mir-200b	dbDEMC	hsa-mir-429	dbDEMC
hsa-let-7i	dbDEMC	hsa-mir-124	dbDEMC
hsa-mir-181a	dbDEMC	hsa-mir-27b	dbDEMC
hsa-let-7e	dbDEMC	hsa-mir-10b	dbDEMC
hsa-mir-146b	dbDEMC	hsa-mir-199b	dbDEMC
hsa-mir-182	dbDEMC	hsa-mir-30c	dbDEMC
hsa-let-7d	dbDEMC	hsa-mir-96	dbDEMC
hsa-mir-142	dbDEMC	hsa-mir-224	dbDEMC
hsa-mir-15b	dbDEMC	hsa-mir-20b	dbDEMC
hsa-mir-24	dbDEMC	hsa-mir-218	dbDEMC
hsa-mir-195	dbDEMC	hsa-mir-103a	未证实
hsa-mir-9	dbDEMC	hsa-mir-26b	dbDEMC

与未进行特征融合的模型的对比发现,本文所提模型的特征融合策略能够有效提升模型的预测性能. 此外,对食管肿瘤的病例研究结果也显示出所提模型具有良好的预测能力. 上述实验结果均表明,本文提出的计算模型可作为预测miRNA与疾病间潜在关联的可靠工具. 在接下来的研究中,将尝试在模型中嵌入更多的多源信息,如miRNA序列信息、靶基因信息等,以期进一步提升模型的预测性能.

参考文献

[1] AMBROS V. The functions of animal microRNAs[J]. Nature, 2004, 431(7006): 350-355.

[2] XU P Z, GUO M, HAY B A. MicroRNAs and the regulation of cell death[J]. Trends Genet, 2004, 20(12): 617-624.

[3] MELTZER P S. Cancer genomics: Small RNAs with big impacts[J]. Nature, 2005, 435(7043): 745-746.

[4] MONSÁLVEZ V, MONTES-MORENO S, ARTIGA M J, et al. MicroRNAs as prognostic markers in indolent primary cutaneous B-cell lymphoma[J]. Mod Pathol, 2013, 26(2): 171-181.

[5] VÁRALLYAY E, BURGYÁN J, HAVELDA Z. MicroRNA detection by northern blotting using locked nucleic acid probes[J]. Nat Protoc, 2008, 3(2): 190-196.

[6] FREEMAN W M, WALKER S J, VRANA K E. Quantitative RT-PCR: Pitfalls and potential[J]. Biotechniques, 1999, 26(1): 112-122, 124-115.

[7] BASKERVILLE S, BARTEL D P. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes[J]. Rna, 2005, 11(3): 241-247.

[8] SHEN Z, ZHANG Y H, HAN K, et al. MiRNA-disease association prediction with collaborative matrix factorization [J]. Complexity, 2017, 2017: 1-9.

[9] JI B Y, YOU Z H, CHENG L, et al. Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model[J]. Scientific Reports, 2020, 10(1): 6658.

[10] ZHANG Q H, ZHU L, HUANG D S. High-order convolutional neural network architecture for predicting DNA-protein binding sites[J]. IEEE/ACM Trans Comput Biol Bioinform, 2019, 16(4): 1184-1192.

[11] XU W X, ZHU L, HUANG D S. DCDE: An efficient deep convolutional divergence encoding method for human promoter recognition[J]. IEEE Trans Nanobioscience, 2019, 18(2): 136-145.

[12] LI Y, QIU C X, TU J, et al. HMDD v2.0: A database for experimentally supported human microRNA and disease associations[J]. Nucleic Acids Research, 2013, 42(D1): D1070-D1074.

[13] 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.

[14] XUAN P, HAN K, GUO M Z, et al. Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors[J]. PloS one, 2013, 8(8): e70204.

[15] VAN L T, NABUURS S B, MARCHIORI E. Gaussian interaction profile kernels for predicting drug-target interaction[J]. Bioinformatics, 2011, 27(21): 3036-3043.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-06-12) [2020-10-12]. https://arxiv.org/abs/1706.03762.

[17] WANG X, JI H Y, SHI C, et al. Heterogeneous graph attention network [EB/OL]. (2019-03-18) [2020-10-12].

https://arxiv.org/abs/1903.07293.
[18] WANG M J, YU L F, ZHENG D, et al. Deep graph library: Towards efficient and scalable deep learning on graphs[EB/OL]. (2017-09-03) [2020-10-12]. https://arxiv.org/abs/1909.01315v1.
[19] CHEN X, YAN C G, ZHANG X, et al. WBSMDA: Within and between score for miRNA-disease association prediction[J]. Sci Rep, 2016, 6: 21106.
[20] CHEN X, XIE D, WANG L, et al. BNPMDA: Bipartite network projection for miRNA-disease association prediction[J]. Bioinformatics, 2018, 34(18): 3178-3186.
[21] CHEN X, LI S X, YIN J, et al. Potential miRNA-disease association prediction based on kernelized Bayesian matrix factorization[J]. Genomics, 2020, 112(1): 809-819.
[22] XIE G B, FAN Z L, SUN Y P, et al. WBNPMD: Weighted bipartite network projection for microRNA-disease association prediction[J]. Journal of Translational Medicine, 2019, 17(1): 322.
[23] ZHANG L, LIU B L, LI Z W, et al. Predicting miRNA-disease associations by multiple meta-paths fusion graph embedding model[J]. BMC Bioinformatics, 2020, 21(1): 470.
[24] MA Y J, HE T T, GE L X, et al. MiRNA-disease interaction prediction based on kernel neighborhood similarity and multi-network bidirectional propagation[J]. BMC Med Genomics, 2019, 12(10): 185.
[25] YU S P, LIANG C, XIAO Q, et al. MCLPMDA: A novel method for miRNA-disease association prediction based on matrix completion and label propagation[J]. Journal of Cellular and Molecular Medicine, 2019, 23(2): 1427-1438.
[26] YANG Z, REN F, LIU C N, et al. DbDEMC: A database of differentially expressed miRNAs in human cancers[J]. BMC Genomics, 2010, 11 (4): S5-S5.
[27] JIANG Q H, WANG Y D, HAO Y Y, et al. MiR2Disease: A manually curated database for microRNA deregulation in human disease[J]. Nucleic Acids Research, 2009, 37: D98-D104.

作者简介

李政伟　男,1977年生,河南许昌人. 博士. 现为中国矿业大学计算机科学与技术学院副教授、硕士生导师. 主要研究方向为机器学习与模式识别、生物信息学、计算机视觉等.
E-mail: zwli@cumt.edu.cn

李佳树　男,1998年生,江苏徐州人. 现为中国矿业大学计算机科学与技术学院研究生. 主要研究方向为图神经网络、miRNA与疾病的关联预测等.
E-mail: lijiashu7646@163.com

尤著宏　男,1980年生,甘肃兰州人. 博士. 现为西北工业大学计算机学院教授,博士生导师. 主要研究方向为大数据分析、数据挖掘及在生物信息学上的应用等.
E-mail: zhuhongyou@nwpu.edu.cn

聂　茹(通讯作者)　女,1976年生,江苏徐州人. 博士. 现为中国矿业大学计算机科学与技术学院副教授,硕士生导师. 主要研究方向为生物信息学、机器学习和图像处理等.
E-mail: nr@cumt.edu.cn