DOI:10.12113/202302009

基于图自编码器和协同训练预测 miRNA 与疾病的关联

刘立伟[1*],刘晓兰[1],谭者斌[2]
(1.大连交通大学 理学院,辽宁 大连 116028;2.大连交通大学 软件学院,辽宁 大连 116028)

摘 要:近年来,越来越多的生物学实验研究表明,microRNA (miRNA)在人类复杂疾病的发展中发挥着重要作用。因此,预测 miRNA 与疾病之间的关联有助于疾病的准确诊断和有效治疗。由于传统的生物学实验是一种昂贵且耗时的方式,于是许多基于生物学数据的计算模型被提出来预测 miRNA 与疾病的关联。本研究提出了一种端到端的深度学习模型来预测 miRNA-疾病关联关系,称为 MDAGAC。首先,通过整合疾病语义相似性,miRNA 功能相似性和高斯相互作用谱核相似性,构建 miRNA 和疾病的相似性图。然后,通过图自编码器和协同训练来改善标签传播的效果。该模型分别在 miRNA 图和疾病图上建立了两个图自编码器,并对这两个图自编码器进行了协同训练。miRNA 图和疾病图上的图自编码器能够通过初始关联矩阵重构得分矩阵,这相当于在图上传播标签。miRNA-疾病关联的预测概率可以从得分矩阵得到。基于五折交叉验证的实验结果表明,MDAGAC 方法可靠有效,优于现有的几种预测 miRNA-疾病关联的方法。

关键词:microRNA;疾病;关联预测;协同训练;图自编码器;端到端
中图分类号:Q522+.2 **文献标志码**:A **文章编号**:1672-5565(2024)02-116-08

Predicting miRNA-disease associations based on graph autoencoders and collaborative training

LIU Liwei[1*], LIU Xiaolan[1], TAN Zhebin[2]
(1.*School of Science*, *Dalian Jiaotong University*, *Dalian* 116028, *Liaoning*, *China*;
2.*School of Software*, *Dalian Jiaotong University*, *Dalian* 116028, *Liaoning*, *China*)

Abstract: In recent years, increasing biological experiments have shown that microRNA (miRNA) plays an important role in the development of human complex diseases. Therefore, predicting miRNA-disease associations can contribute to accurate diagnosis and effective treatment of diseases. Since traditional biological experiments are expensive and time-consuming, plenty of computational models based on biological data have been proposed to predict MiRNA-disease associations. In this study, we propose an end-to-end deep learning model to predict miRNA-disease associations (MDAGAC). Specifically, we firstly construct the similarity network of miRNA and disease by integrating disease semantic similarity, miRNA functional similarity and Gaussian interaction profile kernel similarity. Then, the effect of label propagation is improved through Graph Autoencoders and Collaborative training. This model implements two graph autoencoders on miRNA graph and disease graph respectively, and trains these two graph autoencoders collaboratively. Graph autoencoders on miRNA graph and disease graph are able to reconstruct score matrix through initial association matrix, which is equivalent to propagate labels on graphs. The prediction probability of MiRNA-disease association can be obtained from the score matrix. The results of the experiment based on 5-fold cross validation show that MDAGAC is reliable and effective and outperforms current MiRNA-disease associations prediction methods.

Keywords: microRNA; Disease; Association prediction; Collaborative training; Graph autoencoder; End-to-end

收稿日期:2023-02-12;**修回日期**:2023-03-21;**网络首发日期**:2023-06-02.
网络首发地址:https://kns.cnki.net/kcms2/detail/23.1513.Q.20230601.1404.008.html
基金项目:海南省计算科学与应用重点实验室开放课题(No. JSKX202102).
*通信作者:刘立伟,男,教授、硕导,研究方向:生物信息学.E-mail:liutree80@163.com.
引用格式:刘立伟,刘晓兰,谭者斌.基于图自编码器和协同训练预测 miRNA 与疾病的关联[J].生物信息学,2024,22(2):116-123.
LIU Liwei, LIU Xiaolan, TAN Zhebin.Predicting miRNA-disease associations based on graph autoencoders and collaborative training[J].Chinese Journal of Bioinformatics,2024,22(2):116-123.

微小核糖核酸(microRNA, 简称 miRNA)是一种长度约为 20-24 个核苷酸的内源性非编码单链 RNA 分子,存在于真核生物中,包括植物和动物,可以在转录后水平调节基因表达[1]。许多研究表明,miRNA 在各种复杂的生物过程中发挥着关键作用,包括细胞生长[2]、细胞分化[3]、细胞增殖[4]、细胞死亡[5]。此外,一系列研究证实, miRNA 与人类疾病的发生和发展密切相关,如食管癌、结肠癌、肝癌和肺癌[6-10]。因此,采用适当的实验或计算方法来探索 miRNA 与疾病之间的联系,可以帮助医务人员从分子角度深入了解各种复杂疾病的病理机制并开发相关的新药[11]。一般来说,传统的生物学实验方法往往效率低下,需要投入大量的时间和金钱。然而,由于实验方法的可靠性,研究人员已经建立了许多权威的生物信息学数据库来存储实验证实的 miRNA-疾病相关性。因此,miRNA-疾病关联预测的计算方法作为传统实验的辅助工具出现。通过对计算模型预测的高概率关联进行实验验证,可以有效地缩短传统实验的时间和成本。

在过去几年中,基于功能相似的 miRNA 往往与相似疾病相关[12],研究人员开发了多种 miRNA-疾病关联预测模型,这些模型可以分为三类[7]。第一种类型的预测模型是基于打分函数的模型,它使用概率分布或统计分析来建立打分函数。例如,Mørk 等[13]提出了 miRNA-蛋白质-疾病关联预测模型 miRPD 来预测潜在的 miRNA-疾病关联。他们基于 miRNA-蛋白质和蛋白质-疾病关联得分定义了 miRNA-疾病关联得分函数。其中,蛋白质被引入作为 miRNA-疾病预测的中介。Chen 等[14]提出了一种 miRNA-疾病关联预测内得分和外得分的计算模型 WBSMDA。他们定义了两种不同类型的函数来计算 miRNA-疾病对的内得分和外得分,并将这两个得分整合以获得最终的关联分数。

第二种类型的预测模型是基于网络算法的模型,它从不同角度利用 miRNA 和疾病的相似性。例如,Xuan 等[15]提出了一种新的预测 miRNA-疾病关联的模型 MIDP。对于有已知关联的疾病,MIDP 在 miRNA 相似性网络中采用随机游走算法来预测与疾病有潜在关联的 miRNA。对于没有任何已知关联的疾病,他们利用 miRNA 相似性网络、疾病相似性网络和已知的 miRNA-疾病关联来构建 miRNA-疾病双层网络。然后,他们在这个双层网络上进行随机游走,因此该模型可以用于没有任何已知关联信息的疾病。此外,Chen 等[16]提出了基于异构图推断的 miRNA-疾病关联预测模型 HGIMDA。他们通过总结 miRNA-疾病异构网络中所有边数为 3 的 miRNA 节点和疾病节点间的路径,定义了未标记的 miRNA-疾病对的关联分数。在此基础上,Chen 等[17]进一步提出了基于矩阵分解和异构图推断的 miRNA-疾病关联预测模型 MDHGI。首先,使用稀疏学习方法重构一个新的 miRNA-疾病关联邻接矩阵。然后,基于重构的邻接矩阵、miRNA 相似矩阵和疾病相似矩阵构建异构图。最后,建立一个迭代方程来预测 miRNA-疾病对的关联概率。此外,You 等[18]提出了基于路径的 miRNA-疾病关联预测模型 PBMDA。首先,在 miRNA-疾病异构网络中搜索 miRNA 节点和疾病节点间所有长度小于等于 3 的路径。然后,基于路径的数量和每个路径的长度来计算所研究的 miRNA 和疾病之间的关联分数。Chen 等[19]进一步提出了一个基于三层异构网络推断的 miRNA-疾病关联预测模型 TLHNMDA。该模型构建了一个包含 miRNA、疾病和 lncRNA 节点的三层异构网络。基于这个三层网络,构建了一个迭代方程来获得 miRNA-疾病对的关联概率。

第三种类型的预测模型是基于机器学习的模型。例如,Chen 等[20]提出了预测 miRNA-疾病关联的正则化最小二乘模型 RLSMDA。他们在正则化最小二乘(RLS)的框架下,分别在 miRNA 和疾病空间构造半监督分类器,然后将两个不同空间的最优分类器组合在一起,得到 miRNA-疾病对的概率。Li 等[21]使用奇异值阈值(SVT)算法建立了 MCMDA 模型。采用矩阵填充算法更新 miRNA-疾病邻接矩阵,得到最终的 miRNA-疾病关联矩阵。此外,Chen 等[22]提出 miRNA-疾病关联预测模型 RKNNMDA。他们利用 KNN 算法获得被研究 miRNA 的 K 个最近邻居,并使用支持向量机对 K 个邻居进行重新排序。然后,通过检查 K 个邻居与候选疾病之间的关联信息,可以计算出被研究的 miRNA 与候选疾病之间的关联分数。同样,作者还从疾病的角度计算了关联分数。最后,他们从两个不同的角度综合了关联得分,用于预测潜在的 miRNA-疾病的关联。此外,Chen 等[23]还提出了基于决策树学习的 miRNA-疾病关联预测模型 EGBMMDA。EGBMMDA 构造了三种不同类型的特征,并将它们连接起来构建特征向量作为输入。在梯度增强的框架下,通过训练回归树获得潜在的 miRNA-疾病关联的概率。

由于深度学习技术可以更好地学习数据的表示形式,并且近年来已经在基因组学和药物研发等许多领域成功应用[24],我们考虑将其应用于 miRNA 与疾病关联的预测。本文提出了一种端到端的深度学习模型 MDAGAC 来预测 miRNA-疾病关联关系。首先从 HMDD v2.0 数据库中收集了人类 miRNA-疾

病关联数据,然后计算了疾病语义相似性、miRNA 功能相似性、疾病和 miRNA 的高斯相互作用谱核相似性并整合疾病和 miRNA 的相似性。然后,通过图自编码器和协同训练来改善标签传播的效果。该模型分别在 miRNA 图和疾病图上建立了两个图自编码器,并对这两个图自编码器进行了协同训练。miRNA 图和疾病图上的图自编码器能够通过初始关联矩阵重构得分矩阵,这相当于在图上传播标签。miRNA-疾病关联的预测概率可以从得分矩阵得到。最后通过五折交叉验证评估了方法的性能,结果显示模型 MDAGAC 在五折交叉验证中获得 0.960 3± 0.003 0 的平均 AUC 和标准差。

1 材料

1.1 数据集

从 HMDD v2.0 数据库中获得人类 miRNA-疾病关联数据[25]。具体来说,有 495 个 miRNA、383 种疾病和 5 430 个经实验验证的 miRNA-疾病关联关系。用 nd 和 nm 分别表示疾病和 miRNA 的数量,利用大小为 nm×nd 的邻接矩阵 A 来表示所有的 miRNA-疾病对。如果 miRNA$m(i)$ 与疾病 $d(j)$ 有关联,则 $A(i,j)$ 等于 1,否则为 0。

1.2 疾病语义相似性 1

使用 MeSH[26] 数据库计算疾病的语义相似性 1。在 MeSH 数据库中,多种疾病间的关联由有向无环图(DAG)[27]表示,其中节点表示疾病,边表示关联。对于疾病 D,我们可以用 $DAG(D)=(D,T(D),E(D))$ 来表示该疾病,其中 $T(D)$ 表示节点集,$E(D)$ 表示边集。定义疾病 D 的语义值 $DV1(D)$ 如下:

$$DV1(D)=\sum_{d\in T(D)} D1_D(d) \tag{1}$$

$$D1_D(d)=\begin{cases}1, \text{if } d=D\\ \max\{\rho * D1_D(d') \mid d'\in \text{childrenofd}\}, \\ \quad \text{if } d\neq D\end{cases} \tag{2}$$

其中,$D1_D(d)$ 表示在 $DAG(D)$ 中的各个节点 d 对疾病 D 的语义值贡献。在 $DAG(D)$ 中,若 $d=D$,则 $D1_D(d)$ 对自身语义值的贡献应该最大,设置为 1;而距离疾病 D 越远的节点对疾病 D 的语义值贡献越小。ρ 为语义值贡献衰减因子,设为 0.5。对于 d_i 和 d_j 两种疾病,在 DAG 中重合部分越多,相似度越大。因此,基于 MeSH,利用如下公式计算疾病 d_i 和 d_j 的语义相似性 $1SS1_d(d_i,d_j)$:

$$SS1_d(d_i,d_j)=\frac{\sum_{t\in T(d_i)\cap T(d_j)}(D1_{d_i}(t)+D1_{d_j}(t))}{DV1(d_i)+DV1(d_j)} \tag{3}$$

1.3 疾病语义相似性 2

在语义相似性 1 中,假设在同一层的不同节点对疾病的语义值贡献相同。但它忽略了 DAG 中疾病数量的因素,对于不常见的疾病 d,对疾病 D 应该有更高的贡献值[28-29]。因此,使用等式(4)来计算 $D2_D(d)$:

$$D2_D(d)=-\log\frac{n_d}{n_{dis}} \tag{4}$$

其中,n_d 表示包含节点 d 的疾病 DAG 的数目,n_{dis} 表示所有疾病 DAG 的数目。

类似地,定义疾病 D 的语义值 $DV2(D)$ 如下:

$$DV2(D)=\sum_{d\in T(D)} D2_D(d) \tag{5}$$

因此,对于疾病 d_i 和疾病 d_j,使用公式(6)来计算语义相似度 $2SS2_d(d_i,d_j)$。

$$SS2_d(d_i,d_j)=\frac{\sum_{t\in T(d_i)\cap T(d_j)}(D2_{d_i}(t)+D2_{d_j}(t))}{DV2(d_i)+DV2(d_j)} \tag{6}$$

1.4 miRNA 功能相似性

一般认为功能相似的 miRNA 往往与语义相似的疾病相关,Wang 等[30]提出了计算 miRNA 功能相似性的方法。从 http://www.cuilab.cn/files/images/cuilab/misim.zip 中可以获得 miRNA 的功能相似性,用矩阵 FS 来表示。

1.5 疾病和 miRNA 的高斯相互作用谱核相似性

由于不能得到所有疾病的 DAG,对于没有 DAG 的疾病,不能通过 DAG 计算疾病语义相似性。所以为了获得更全面的疾病相似性,基于已知 miRNA 与疾病之间的关联,构建了高斯相互作用谱核相似性[31]。疾病 d_i 和疾病 d_j 的高斯相互作用谱核相似性计算如下:

$$GD(d_i,d_j)=\exp(-\gamma_d\|A(*,d_i)-A(*,d_j)\|^2) \tag{7}$$

其中,$A(*,d_i)$ 和 $A(*,d_j)$ 表示对应的疾病和所有 miRNA 的关联信息,分别由 miRNA-疾病的关联矩阵 A 的第 i 列和第 j 列构成的向量。其中 γ_d 控制高斯核的带宽,由下式计算:

$$\gamma_d=\frac{\gamma_d'}{\frac{1}{nd}\sum_{i=1}^{nd}\|A(*,d_i)\|^2} \tag{8}$$

其中,γ_d' 设为 1。类似地,为了能够得到所有 miRNA 的相似性,计算 miRNA$m(i)$ 和 miRNA$m(j)$

的高斯相互作用谱核相似性如下:

$$GM(m_i,m_j)=\exp(-\gamma_m\|A(m_i,*)-A(m_j,*)\|^2) \tag{9}$$

$$\gamma_m=\frac{\gamma_m'}{\frac{1}{nm}\sum_{i=1}^{nm}\|A(m_i,*)\|^2} \tag{10}$$

1.6　MiRNA 和疾病的整合相似性

疾病语义相似性 SS 是一个稀疏矩阵,单独使用这个矩阵很难达到很好的预测效果。此外,高斯相互作用谱核相似性 GD 是通过已知的 miRNA-疾病关联来计算的,这不够准确。因此,有必要将疾病语义相似性 SS 和高斯相互作用谱核相似性 GD 结合起来,以达到良好的预测效果。我们通过一个加权参数将 SS 和 GD 整合为一个疾病相似性矩阵 SD[32]。疾病 d_i 和疾病 d_j 整合后的疾病矩阵如下:

$$SD(d_i,d_j)=\begin{cases}\alpha SS(d_i,d_j)+(1-\alpha)GD(d_i,d_j)\end{cases} \tag{11}$$

$$SS(d_i,d_j)=\frac{SS1_d(d_i,d_j)+SS2_d(d_i,d_j)}{2} \tag{12}$$

其中 α 是权重,范围在 0 和 1 之间。类似地,miRNA$m(i)$ 和 miRNA$m(j)$ 之间的整合相似性矩阵 SM 通过以下公式计算:

$$SM(m_i,m_j)=\beta FS(m_i,m_j)+(1-\beta)GM(m_i,m_j) \tag{13}$$

2　方法

在这项研究中,提出 MDAGAC 模型来预测 miRNA-疾病关联关系。MDAGAC 的流程图如图 1 所示。MDAGAC 的第一步,数据准备,如前节所述,构建了 miRNA-疾病对的关联矩阵 $A(nm\times nd)$、整合 miRNA 相似性矩阵 $SM(nm\times nm)$ 和整合疾病相似性矩阵 $SD(nd\times nd)$;第二步,构造 miRNA 和疾病的相似性图;第三步,分别在 miRNA 图和疾病图上建立两个图自编码器,并对这两个图自编码器进行了协同训练。miRNA 图和疾病图上的图自编码器能够通过初始关联矩阵重构得分矩阵,相当于在图上传播标签。miRNA-疾病关联的预测概率可由评分矩阵得到。

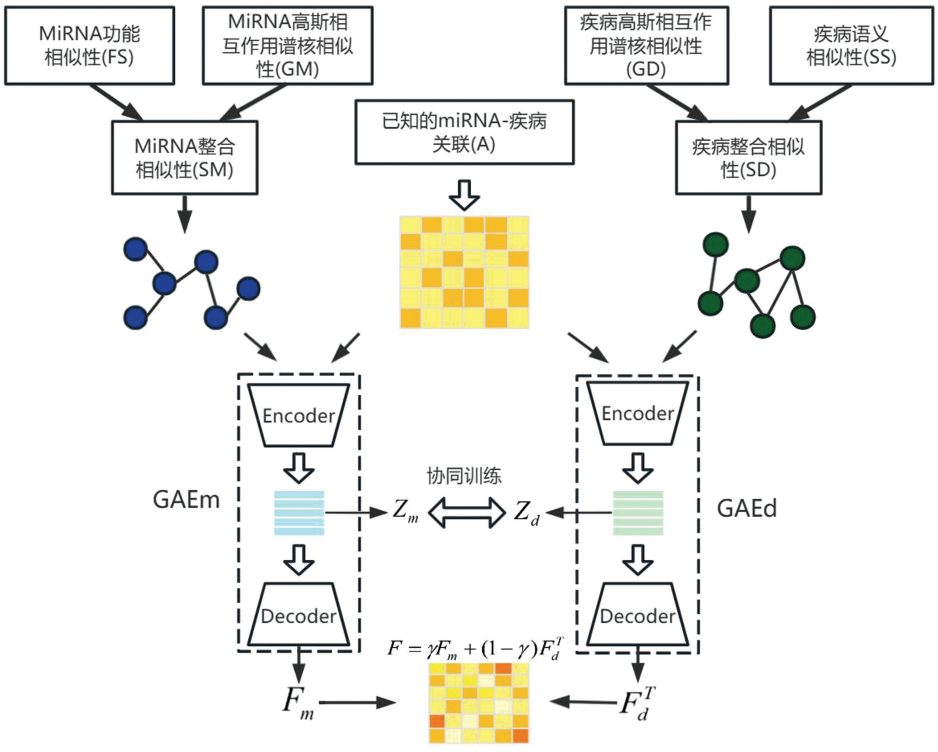

图 1　MDAGAC 算法流程图

Fig.1　Flowchart of MDAGAC

2.1　构造相似性图

相似性矩阵可以看作是 miRNA 图或疾病图的邻接矩阵,因为每个 miRNA 或疾病表示该图的一个节点。根据前人的研究[33],相似性图可以构造如下。以疾病的相似性图构造为例,把 $SD(d_i,d_j)$ 作为两个疾病 d_i 和 d_j 之间的距离。首先,按照其它节

点与节点 i 的距离从小到大进行排序。其次,对于每个疾病节点 i ,选择除其自身之外最近的 κ 个节点。设这 κ 个节点的集合是 $N(i)$ 。第三,如果节点 $j \in N(i)$,则矩阵 C_d 满足 $C_d(i,j)=1$,否则 $C_d(i,j)=0$。所构造的疾病图的自环邻接矩阵 S_d 为

$$S_d = C_d^T \otimes C_d + I \tag{14}$$

其中 $\otimes$ 为 Hadamard 乘积。类似地,也可以得到 miRNA 图的自环邻接矩阵 S_m :

$$S_m = C_m^T \otimes C_m + I \tag{15}$$

2.2 图自编码器

图自编码器[34]是一个具有图卷积层[35]的自编码器。Shi 等[33]证明了以 A 为输入, F 为输出的图自编码器可以模拟标签的传播过程。在 MDAGAC 中,分别在 miRNA 图和疾病图上提出了图自编码器(GAE)[36],用 GAEm 和 GAEd 表示。

首先通过 2 层图卷积编码器学习特征表示 Z_m 和 Z_d :

$$Z_m = \tanh(N_m \cdot \text{Hardswish}(N_m A\, \theta^{(0)})\, \theta^{(1)}) \tag{16}$$

$$Z_d = \tanh(N_d \cdot \text{Hardswish}(N_d A^T \varphi^{(0)})\, \varphi^{(1)}) \tag{17}$$

其中 θ , φ 表示神经网络的权重。N_m 和 N_d 分别表示 miRNA 图和疾病图的归一化邻接矩阵。

$$N_m = D_m^{-1/2} S_m D_m^{-1/2} \tag{18}$$

其中 D_m 是 S_m 的度矩阵。D_m 是通过公式 $D_m(i,i) = \Sigma_j S_m(i,j)$ 计算的对角矩阵。类似的,

$$N_d = D_d^{-1/2} S_d D_d^{-1/2} \tag{19}$$

然后分别对特征表示 Z_m 和 Z_d 进行解码,得到分数矩阵 $F_m \in R^{nm \times nd}$ 和 $F_d \in R^{nd \times nm}$:

$$F_m = \text{sigmoid}(N_m \cdot \text{Hardswish}(N_m Z_m \theta^{(2)})\, \theta^{(3)}) \tag{20}$$

$$F_d = \text{sigmoid}(N_d \cdot \text{Hardswish}(N_d Z_d \varphi^{(2)})\, \varphi^{(3)}) \tag{21}$$

GAEm 的重构误差是预测值和真实标签之间的交叉熵 L_m ,类似的 GAEd 的重构误差是 L_d 。图自编码器可以通过最小化重构误差来训练:

$$L_r = \gamma L_m + (1-\gamma) L_d \tag{22}$$

$$L_m = -\sum_{i,j} A_{ij} \log F_{mij} \tag{23}$$

$$L_d = -\sum_{i,j} A_{ij} \log F_{dij} \tag{24}$$

其中 $\gamma \in (0,1)$ 是平衡从 miRNA 空间和疾病空间捕获信息的权重参数。

2.3 协同训练

最小化式(22)相当于分别在 miRNA 图和疾病图上训练图自编码器。以往的研究表明,整合双方信息的协同训练可以提高预测生物实体关联[33,37]的精度。采用图自编码器学习到的特征表示 Z_m 和 Z_d 来定义协同训练损失[36]:

$$L_c = \frac{1}{2} \| A - Z_m Z_d^T \|_F^2 \tag{25}$$

为了避免过拟合,加入正则化的 Frobenius 范数,总损失可定义为:

$$\min_{\theta,\varphi} L_c + \mu L_r + \lambda \|\theta\|_F^2 + \lambda \|\varphi\|_F^2 \tag{26}$$

其中 μ 是参数, λ 设为 $1\,0^{-7}$ 。最后,通过 F_m 和 F_d 的线性组合得到最优分数矩阵 $F \in R^{nm \times nd}$,

$$F = \gamma F_m + (1-\gamma) F_d^T \tag{27}$$

其中 $F(i,j) \in [0,1]$ 表示 miRNA m_i 与疾病 d_j 之间的预测分数,分数越高表示 miRNA m_i 与疾病 d_j 关联概率越高。MDAGAC 的过程总结为图 1 和表 1。

表 1 MDAGAC 算法

Table 1 MDAGAC algorithm

输入: miRNA 的相似性矩阵 FS 和疾病的相似性矩阵 $SS1$ 和 $SS2$,疾病和 miRNA 的高斯相互作用谱核相似性 GD 和 GM ,初始关联矩阵 A ,参数 α , β , κ , γ , μ

输出:分数矩阵 F

①整合 miRNA 和疾病的相似性矩阵。// 式(11)和(13)

②分别构造 miRNA 相似性图和疾病相似性图。//式(14)和(15)

③分别计算 miRNA 图和疾病图的邻接矩阵 N_m 和 N_d 。//式(18)和(19)

④重复

⑤ Z_m , $F_m \leftarrow$GAEm (N_m , A)//式(16)和(20)

⑥ Z_d , $F_d \leftarrow$GAEp(N_d , A^T)//式(17)和(21)

⑦通过优化式(26)训练 GAEm 和 GAEp。//协作训练

⑧直到收敛

⑨ $F = \gamma F_m + (1-\gamma) F_d^T$ //式(27)

⑩返回 F

3　结　果

3.1　性能评估

首先从 HMDD v2.0[25] 中获得包含 495 个 miRNA 与 383 种疾病之间的5 430 个已知关联的训练数据,然后采用五折交叉验证评估模型 MDAGAC 的准确性。所有已知的 miRNA-疾病关联被随机分为五个大小相等的子集。每个子集依次用作测试集,而其他四个子集用作训练集。如图 2 所示,MDAGAC 的平均 AUC 和标准差为0.960 3±0.003 0,这是五折交叉验证0.961 7,0.955 8,0.959 1,0.965 0和0.959 8的平均值和标准差。

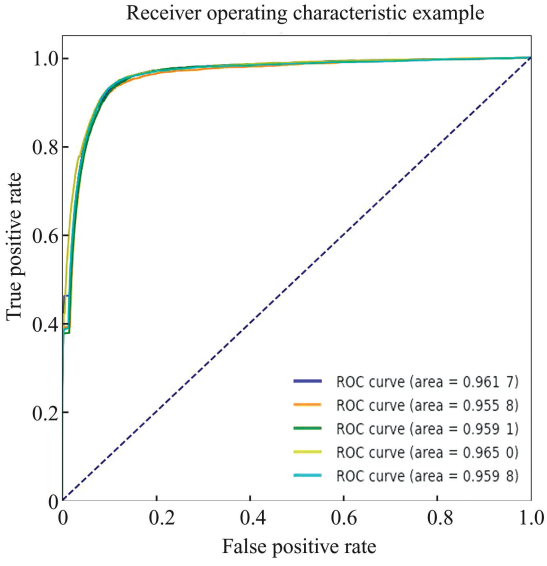

图 2　五折交叉验证中 MDAGAC 的 ROC 曲线

Fig.2　ROC curves of MDAGAC in 5-fold cross-validation

为了进一步证明该模型的优越性能,比较了 MDAGAC 模型与八个最先进模型的预测性能,它们分别是 PBMDA (0.917 2),SAEMDA (0.910 2),EGBMMDA (0.904 8),MDHGI (0.879 4),TLHNMDA (0.879 5),MCMDA (0.876 7),MaxFlow (0.857 9)和 RLSMDA (0.856 9)。为了公平比较,上述模型均基于 HMDD v2.0[25] 进行了五折交叉验证评估。此外,由于上述模型采用了多种不同的评价指标,在此仅利用 AUC 值来综合衡量这些模型的预测性能。比较结果总结在表 2 中。可以看到,我们的模型在这九个模型中实现了最高的 AUC 值。MDAGAC 的优越性能得益于基于图卷积的编码器和端到端的训练方式。

表 2　五折交叉验证中 MDAGAC 与其他模型的性能对比

Table 2　Performance comparison between MDAGAC and other models in 5-fold cross-validation

预测模型	AUC	标准差
MDAGAC	0.960 3	0.003 0
PBMDA	0.917 2	0.000 7
SAEMDA	0.910 2	0.002 9
EGBMMDA	0.904 8	0.001 2
MDHGI	0.879 4	0.002 1
TLHNMDA	0.879 5	0.001 0
MCMDA	0.876 7	0.001 1
MaxFlow	0.857 9	0.001 0
RLSMDA	0.856 9	0.002 0

3.2　参数分析

MDAGAC 中的参数会影响预测性能。在本节中,通过五折交叉验证选择具有最佳平均 AUC 的超参数。为了验证参数 α 和 β 对疾病和 miRNA 的整合相似性矩阵 *SD* 和 *SM* 的有效性,首先在 0 到 1.0 的区间内定义了 11 个等间距的值,并将这些值应用在 α 和 β 上来训练模型。然后通过五折交叉验证来计算每个模型的 AUC 值。如表 3 所示,当 $\alpha = 0.3$ 和 $\beta = 0.3$ 时,模型的预测性能最好。

表 3　不同 α 和 β 值的五折交叉验证结果

Table 3　Results of 5-fold cross-validation for different values of α and β

α	AUC	标准差	β	AUC	标准差
0	0.953 0	0.006 9	0	0.948 6	0.006 2
0.1	0.956 6	0.005 6	0.1	0.959 0	0.004 0
0.2	0.957 8	0.006 9	0.2	0.960 8	0.003 1
0.3	0.960 3	0.003 0	0.3	0.960 3	0.003 0
0.4	0.960 2	0.003 5	0.4	0.958 3	0.003 6
0.5	0.959 9	0.003 1	0.5	0.957 0	0.003 5
0.6	0.959 9	0.003 9	0.6	0.956 4	0.003 4
0.7	0.959 4	0.003 6	0.7	0.957 0	0.003 8
0.8	0.958 4	0.004 5	0.8	0.957 3	0.002 2
0.9	0.956 9	0.003 7	0.9	0.952 8	0.003 3
1.0	0.953 9	0.003 1	1.0	0.909 2	0.003 9

在 MDAGAC 中,通过参数 γ 平衡 miRNA 空间和疾病空间。选择 $\gamma \in \{0.1,0.3,0.5,0.7,0.8,0.9\}$ 来训练模型。结果如表 4 所示,MDAGAC 在 $\gamma = 0.8$ 时具有最佳预测性能。此外,在构造相似性图中用到了参数 k,当 k 设置为 5、8 和 10 时,MDAGAC 的五折交叉验证结果分别为 0.958 1±0.005 0、0.960 3±0.003 0 和 0.958 1±0.003 2。

表 4 不同 γ 值的五折交叉验证结果

Table 4 Results of 5-fold cross-validation for different values of γ

γ	AUC	标准差
0.1	0.945 2	0.002 9
0.3	0.955 2	0.001 9
0.5	0.956 0	0.006 4
0.7	0.959 7	0.002 7
0.8	0.960 3	0.003 0
0.9	0.960 6	0.003 3

采用 pytorch (https://pytorch.org/) 构建 MDAGAC,并应用 Adam 优化器训练模型。然后,将神经网络的随机失活率设置为 0.5,并通过改变学习率 *lr* 来评估 MDAGAC 的预测性能。结果显示在表 5 中,其中最佳学习率值是 0.01。此外,还对不同的隐藏层维度进行模型训练。结果如表 6 所示,我们的模型预测性能随着隐藏层维度的增加而增强。但当维数大于 144 时,AUC 不再有显著提高。因此,将隐藏层的维数设置为 144,以节省模型的计算成本。

表 5 不同 *lr* 值的五折交叉验证结果

Table 5 Results of 5-fold cross-validation for different values of *lr*

lr	AUC	标准差
0.001	0.926 8	0.005 7
0.01	0.960 3	0.003 0
0.05	0.920 6	0.022 4
0.1	0.513 2	0.010 5

表 6 不同隐藏层维度的五折交叉验证结果

Table 6 Results of 5-fold cross-validation for different hidden layer dimensions

维度	AUC	标准差
64	0.948 8	0.004 0
100	0.955 5	0.003 2
125	0.959 3	0.003 8
144	0.960 3	0.003 0
196	0.960 1	0.002 9
225	0.959 5	0.003 5

4 结 论

预测潜在的 miRNA-疾病关联使研究人员能够更好地了解疾病的机制,并促进复杂疾病的诊断、治疗和预防。本研究提出了一种端到端的深度学习模型来预测 miRNA-疾病关联关系,称为 MDAGAC。基于五折交叉验证的实验结果表明:

1) MDAGAC 方法可靠有效,优于现有的几种方法。

2) 与现有的 miRNA-疾病关联预测方法相比,MDAGAC 采用端到端的神经网络模型来协同训练两个 GAE。这种数据驱动的端到端的深度学习模型不仅提高了预测潜在 miRNA-疾病关联的精度,而且为生物信息学的其它领域提供了一种通用的方法。

参考文献(References)

[1] AMBROS V. The functions of animal microRNAs[J]. Nature, 2004, 431:350-355. DOI: 10.1038/nature02871.

[2] BARTEL D P. MicroRNAs: Genomics, biogenesis, mechanism and function[J]. Cell, 2004, 116:281-297. DOI:10.1016/s0092-8674(04)00045-5.

[3] XIAO C, CALADO D P, GALLER G, et al. MiR-150 controls B cell differentiation by targeting the transcription factor c-Myb[J]. Cell, 2016, 165(4):1027. DOI: 10.1016/j.cell.2016.04.056.

[4] JOHNNIDIS J B, HARRIS M H, WHEELER R T, et al. Regulation of progenitor cell proliferation and granulocyte function by microRNA-223[J]. Nature, 2008, 451(7182): 1125-1129. DOI: 10.1038/nature06607.

[5] KIM J H, WOO H R, KIM J, et al. Trifurcate feed-forward regulation of age-dependent cell death involving miR164 in Arabidopsis[J]. Science, 2009, 323(5917):1053-1057. DOI: 10.1126/science.1166386.

[6] AMBROS V. microRNAs: Tiny regulators with great potential [J]. Cell, 2001, 107(7):823-826. DOI: 10.1016/s0092-8674(01)00616-x.

[7] CHEN Xing, XIE Di, ZHAO Qi, et al. MicroRNAs and complex diseases: from experimental results to computational models[J]. Briefings in Bioinformatics, 2019, 20(2):515-539. DOI: 10.1093/bib/bbx130.

[8] 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(D1):D239-D247. DOI:10.1093/nar/gkv1258.

[9] 邹小龙,董雪松,孙学溥.结肠癌中核内 miRNA 的激活调控作用研究[J].生物信息学,2019,17(2):111-115. DOI:10.12113/j.issn.1672-5565.201903009.
ZOU Xiaolong, DONG Xuesong, SUN Xuepu. Activation regulation of nuclear miRNA regulation in colon cancer[J].Chinese Journal of Bioinformatics,2019,17(2):111-115. DOI: 10.12113/j.issn.1672-5565.201903009.

[10] 赵燕伟,王振兴,王熙梓,等.转录因子及 miRNA 调控食管癌耐药机制研究[J].生物信息学, 2022,20(1):56-63. DOI:10.12113/202009005.
ZHAO Yanwei, WANG Zhenxing, WANG Xizi, et al. Role and mechanisms of miRNA and transcription factorsregulating drug resistance of esophageal cancer[J].Chinese Journal of Bioinformatics, 2022, 20(1): 56-63. DOI: 10.12113/202009005.

[11] CHEN Changzheng. microRNAs as oncogenes and tumor

suppressors[J]. New England Journal Medicine, 2005, 353(17):1768-1771. DOI: 10.1056/NEJMp058190.

[12]JIANG Qinghua, HAO Yangyang, WANG Guohua, et al. Prioritization of disease microRNAs through a human phenome-microRNAome network[J]. BMC Systems Biology, 2010, 4:S2. DOI: 10.1186/1752-0509-4-S1-S2.

[13]MØRK S, PLETSCHER-FRANKILD S, PALLEJA CARO A, et al. Protein-driven inference of miRNA-disease associations[J]. Bioinformatics, 2014, 30:392-397. DOI: 10.1093/bioinformatics/btt677.

[14]CHEN Xing, YAN C C, ZHANG Xu, et al. WBSMDA: Within and between score for miRNA-disease association prediction[J]. Scientific Reports, 2016, 6(1):21106. DOI: 10.1038/srep21106.

[15]XUAN Ping, HAN Ke, GUO Yahong, et al. Prediction of potential disease-associated microRNAs based on random walk[J]. Bioinformatics, 2015, 31(11):1805-1815. DOI: 10.1093/bioinformatics/btv039.

[16]CHEN Xing, YAN C C, ZHANG Xu, et al. HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction[J]. Oncotarget, 2016, 7(40):65257-65269. DOI: 10.18632/oncotarget.11251.

[17]CHEN Xing, YIN Jun, QU Jia, et al. MDHGI: Matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction[J]. PLoS Computational Biology,2018,14(8):e1006418. DOI:10.1371/journal.pcbi.1006418.

[18]YOU Zhuhong, HUANG Zhian, ZHU Zexuan, et al. PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction[J]. PLoS Computational Biology, 2017, 13(3):e1005455. DOI: 10.1371/journal.pcbi.1005455.

[19]CHEN Xing, QU Jia, YIN Jun. TLHNMDA: Triple layer heterogeneous network based inference for miRNA-disease association prediction[J]. Frontiers in Genetics, 2018, 9:234. DOI: 10.3389/fgene.2018.00234.

[20]CHEN Xing, YAN Guiying. Semi-supervised learning for potential human microRNA-disease associations inference[J]. Scientific Reports, 2014, 4:5501. DOI: 10.1038/srep05501.

[21]LI Jianqiang, RONG Zhihao, CHEN Xing, et al. MCMDA: Matrix completion for miRNA-disease association prediction[J]. Oncotarget, 2017, 8(13):21187-21199. DOI: 10.18632/oncotarget.15061.

[22]CHEN Xing, WU Qiaofeng, YAN Guiying. RKNNMDA: Ranking-based KNN for miRNA-disease association prediction[J]. RNA Biology, 2017, 14(7):952-962. DOI: 10.1080/15476286.2017.1312226.

[23]CHEN Xing, HUANG Li, XIE Di, et al. EGBMMDA: Extreme gradient boosting machine for miRNA-disease association prediction[J]. Cell Death & Disease, 2018, 9:3. DOI: 10.1038/s41419-017-0003-x.

[24]LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015,521(7553):436-444. DOI:10.1038/nature14539.

[25]LI Yang, QIU Chengxiang, TU Jian, et al. HMDD v2.0: A database for experimentally supported human microRNA and disease associations[J]. Nucleic Acids Research,2014,42(D1): D1070-D1074. DOI:10.1093/nar/gkt1023.

[26]SCHEIBLE R, STRECKER P, YAZIJY S, et al. A multilingual browser platform for medical subject headings[J]. Studies in Health Technology and Informatics, 2022, 289:384-387. DOI: 10.3233/SHTI210939.

[27]LI Yu, KUWAHARA H, YANG Peng, et al. PGCN: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks[J]. bioRxiv, 2019, 532226. DOI: 10.1101/532226.

[28]WANG Lei, YOU Zhuhong, HUANG Y A, et al. An efficient approach based on multi-sources information to predict circRNA-disease associations using deep convolutional neural network[J]. Bioinformatics, 2020,36(13):4038-4046. DOI: 10.1093/bioinformatics/btz825.

[29]WANG Lei, YOU Zhuhong, LI Yangming, et al. GCNCDA: A new method for predicting circRNA-disease associations based on graph convolutional network algorithm[J]. PLOS Computational Biology, 2020, 16(5): e1007568. DOI: 10.1371/journal.pcbi.1007568.

[30]WANG Dong, WANG Jun, LU Ming, et al. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases[J]. Bioinformatics, 2010, 26(13):1644-1650. DOI: 10.1093/bioinformatics/btq241.

[31]VAN LAARHOVEN T, NABUURS S B, MARCHIORI E. Gaussian interaction profile kernels for predicting drug-target interaction[J]. Bioinformatics, 2011, 27(21):3036-3043.DOI: 10.1093/bioinformatics/btr500.

[32]JI Cunmei, GAO Zhen, MA Xu, et al. AEMDA: Inferring miRNA-disease associations based on deep autoencoder[J]. Bioinformatics, 2021, 37(1): 66-72. DOI: 10.1093/bioinformatics/btaa670.

[33]SHI Zhuangwei, ZHANG Han, JIN Chen, et al. A representation learning model based on variational inference and graph autoencoder for predicting lncRNA-disease associations[J]. BMC Bioinformatics, 2021, 22:136. DOI: 10.1186/s12859-021-04073-z.

[34]KIPF T N, WELLING M. Variational graph auto-encoders[J]. arXiv preprint arXiv: 1611.07308, 2016. DOI: 10.48550/arXiv.1611.07308.

[35]KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.DOI: 10.48550/arXiv.1609.02907.

[36]JIN Chen, SHI Zhuangwei, ZHANG Han, et al. Predicting lncRNA-protein interactions based on graph autoencoders and collaborative training[C]// 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2021: 38-43. DOI: 10.1109/BIBM52615.2021.9669316.

[37]HAN Peng, YANG Peng, ZHAO Peilin, et al. GCN-MF: Disease-gene association identification by graph convolutional networks and matrix factorization[C]// Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining,2019: 705-713. DOI: 10.1145/3292500.3330912.