计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900169

基于堆叠自动编码器的 miRNA-疾病关联预测方法

刘 丹 赵 森 颜志良 赵 静 王会青
太原理工大学信息与计算机学院 太原 030606
(869095763@qq.com)

摘 要 作为一类小的非编码 RNA,miRNA 的异常调控与人类疾病的发生和发展密切相关,研究 miRNA 与疾病的关联对于了解人类疾病致病机制具有重要意义。机器学习方法被广泛应用于 miRNA-疾病关联预测,然而现有方法仅仅考虑了 miRNA 与疾病相似性网络信息,忽略了相似性网络的拓扑结构。因此,文中提出基于堆叠自动编码器的 miRNA-疾病关联预测模型 SAEMDA,该模型采用重启随机游走获取 miRNA 与疾病相似性网络的拓扑结构特征,用堆叠自动编码器提取 miRNA 与疾病的抽象低维特征,将得到的低维特征输入深度神经网络进行 miRNA-疾病关联预测。SAEMDA 模型在 5 折交叉验证中取得了较好的结果,并在结肠癌和肺癌两个案例中进行了验证。在结肠癌的案例中,此模型预测的前 50 个 miRNA-疾病关联中的 45 个 miRNA 在数据库中得到了验证;在肺癌的案例中,排名前 50 的 miRNA 均在数据库中得到了验证。

关键词:miRNA-疾病关联;相似性网络;拓扑结构;重启随机游走;堆叠自动编码器
中图法分类号 TP391

miRNA-disease Association Prediction Model Based on Stacked Autoencoder

LIU Dan,ZHAO Sen,YAN Zhi-liang,ZHAO Jing and WANG Hui-qing
College of Information and Computer,Taiyuan University of Technology,Taiyuan 030606,China

Abstract As a group of small non-coding RNA,the abnormal regulation of miRNA is closely related to the occurrence and development of human diseases. The study on the associations between miRNA and disease is important for understanding the pathogenic mechanism of human diseases. Machine learning methods are widely used to predict miRNA-disease associations. However, existing methods only consider the information of miRNA and disease similarity networks,ignoring the topology structure of the similarity networks. Therefore,SAEMDA model based on stacked autoencoder is proposed in this paper,it gets the topological structure features of miRNA and disease similarity networks by restart random walk,obtains the abstract low dimensional features of miRNA and disease by stacked autoencoder,and the low dimensional features are input into deep neural network for miRNA-disease associations prediction. SAEMDA model has achieved great results in 5-fold cross-validation,and it has been validated in cases of colon cancer and lung cancer additionally. As for colon cancer,45 of the top 50 miRNA-disease associations predicted by this model are verified in the database;and in the cases of lung cancer,all the top 50 miRNAs are verified in the database.

Keywords miRNA-disease associations,Similarity networks,Topological structure,Random walk,Stacked autoencoder

1 引言

MicroRNA(miRNA)是一类小的非编码 RNA,由约 20～25 个核苷酸组成,miRNA 的异常调控会导致许多人类疾病的发生[1-2]。因此,研究 miRNA 与疾病的关联对于了解人类疾病致病机制具有重要意义。

目前的 miRNA-疾病关联预测方法主要包括生物方法、网络方法、传统的机器学习方法和深度学习方法。通过生物手段预测 miRNA-疾病关联成本高且耗时长,多种基于网络的计算方法被提出,用于 miRNA-疾病关联预测。You 等[3]提出了一种基于网络路径的计算方法,但其结合 miRNA 与疾病高斯交互谱核相似性构建异构图,使预测有一定的偏差。RWRMDA[4]以与疾病相关的已知 miRNA 为种子mi-RNA,通过在 miRNA 的功能相似性网络上实施重启随机游走来预测疾病相关的 miRNA,但 RWRMDA 无法预测潜在的 miRNA-疾病关联。近年来,传统的机器学习方法因其自身的高效性和预测结果的可靠性而备受关注,基于传统机器学习的 miRNA-疾病关联预测方法相比基于网络的方法具有更好的预测性能。Chen 等[5]将 miRNA 相似性网络数据与疾病相似性网络数据作为特征,手工选取前 100 种重要特征输入随机

到稿日期:2020-09-23 返修日期:2021-01-23 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目 :山西省重点研发计划项目(201903D121151);山西省研究生教育改革课题(2019JG020153)
This work was supported by the Key Research and Development Plan of Shanxi Province(201903D121151) and Graduate Education Reform of Shanxi Province(2019JG020153).
通信作者:王会青(1013208257@qq.com)

森林模型进行预测。Yao 等[6]结合 miRNA 与疾病相似性网络数据作为 miRNA-疾病对特征，利用随机森林特征评分机制筛选出 100 种特征进行 miRNA-疾病关联预测。这些方法利用传统机器学习模型提取了 miRNA 与疾病的低维特征，但无法获取 miRNA 与疾病相似性网络数据中的高阶抽象信息。

深度学习方法可以通过非线性函数结合低层特征形成抽象的高层特征，从而发现数据的有效特征表示。自动编码器作为一种深度学习方法，因其能够提取输入数据的抽象特征，在 miRNA 与疾病关联预测问题中得到了广泛应用。Zhang 等[7]利用变分自动编码器提取 miRNA 与疾病相似性网络特征，并将结果输入回归模型进行预测。Peng 等[8]结合 miRNA 之间的相似性、疾病之间的相似性和蛋白质之间的相互作用，利用自动编码器提取低维抽象特征，采用卷积神经网络实现了 miRNA-疾病关联预测。Chen 等[9]结合 miRNA 与疾病相似性网络数据，利用堆叠自动编码器与支持向量机进行 miRNA-疾病关联预测。Wang 等[10]利用堆叠自动编码器与随机森林进行 miRNA-疾病关联预测。

上述基于自动编码器的 miRNA-疾病关联预测方法将 miRNA 与疾病相似性网络数据直接作为输入，而 miRNA 与疾病相似性网络仅仅描述了两个直接连接的 miRNA(疾病)节点之间的相似性，没有考虑节点间的间接关系和网络的拓扑连接模式，忽略了相似性网络中的拓扑结构信息。研究表明，网络分析中的关键步骤是研究网络中点、线之间的结构关系。捕获相似性网络的拓扑结构信息有助于提高模型的预测性能，在基因优先级排序等生物信息学领域得到了成功应用[11-12]。

因此，本文引入重启随机游走方法和堆叠自动编码器构建了基于堆叠自动编码器的 miRNA-疾病关联预测模型 SAEMDA，用于捕获相似性网络的拓扑结构信息[13]，学习 miRNA(疾病)相似性网络的高阶抽象信息，从而实现 miRNA-疾病关联预测。实验结果表明，本文模型 SAEMDA 具有预测未知 miRNA-疾病关联的潜能。

2　SAEMDA 模型

基于相似性较高的 miRNA 趋向于与相似的疾病相关联这一假设[14]，本文提出 miRNA-疾病关联预测模型 SAEMDA。

本文模型 SAEMDA 首先对 miRNA(疾病)相似性网络进行重启随机游走，捕获 miRNA(疾病)相似性网络的拓扑结构信息，得到了带有拓扑结构信息的 miRNA(疾病)相似性矩阵；然后将带有拓扑结构信息的 miRNA(疾病)相似性矩阵输入堆叠自动编码器，以获取 miRNA(疾病)的低维特征表示；最后将 miRNA 和疾病的特征进行拼接，输入深度神经网络 DNN，分析潜在的 miRNA-疾病关联。本文模型 SAEMDA 的框架图如图 1 所示。

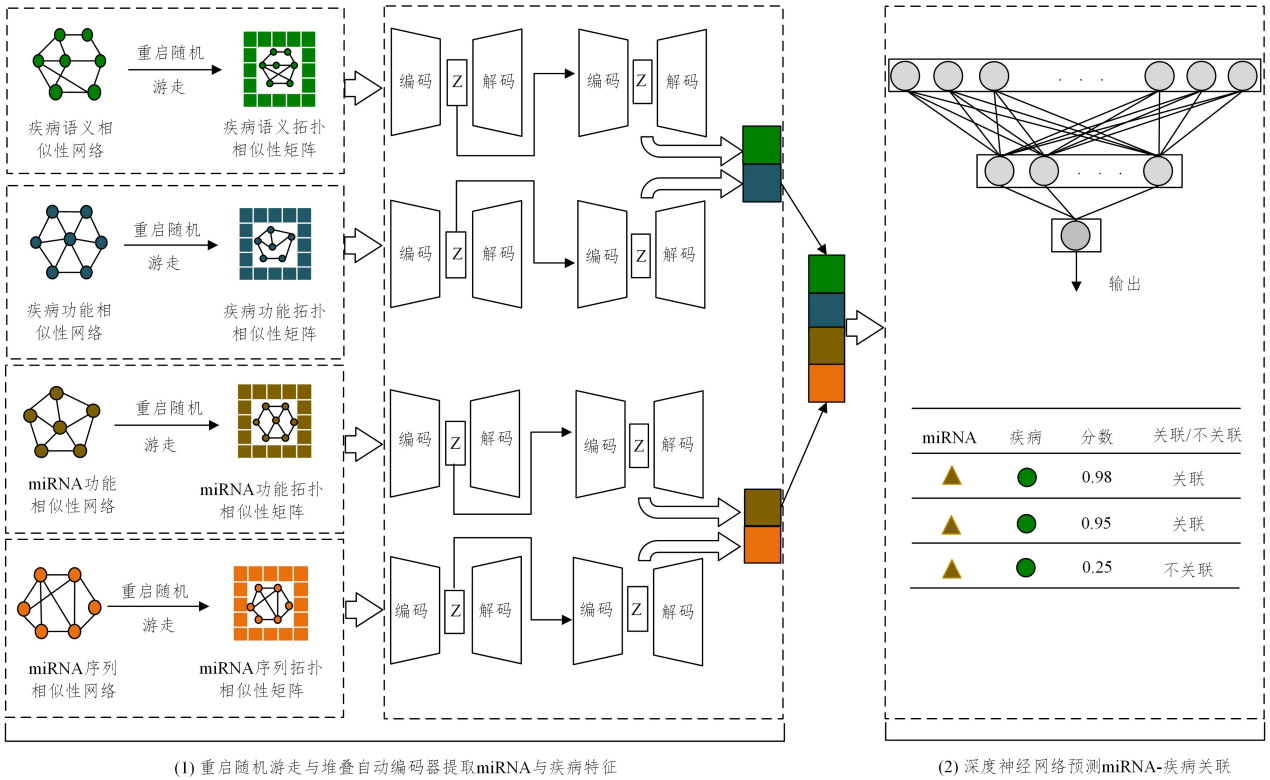

图 1　本文模型 SAEMDA 的框架图

Fig. 1　Flowchart of proposed SAEMDA model

2.1　材料

2.1.1　miRNA-疾病关联数据

miRNA-疾病关联数据库的出现为开发 miRNA-疾病关联预测方法提供了可能，这些数据库包括人类 miRNA-疾病关联数据库(HMDD)[15]与记录人类癌症中差异表达的 miRNAs 数据库(dbDEMC)[16]。

本文的 miRNA-疾病关联数据来自 HMDDv2.0 数据库，HMDDv2.0 为 miRNA-疾病关联数据提供了实验支持，被广泛应用于 miRNA-疾病关联预测方法中。本文从 HMDDv2.0

中获得了包括 495 种 miRNA 和 383 种疾病之间的共 5 430 个 miRNA-疾病关联数据，并定义 miRNA-疾病关联邻接矩阵 $\boldsymbol{A}\in R^{n_m\times n_d}$，如果 miRNA m_i 和疾病 d_j 之间存在关联，则 $\boldsymbol{A}(m_i,d_j)=1$，反之，$\boldsymbol{A}(m_i,d_j)=0$，其中 n_m 表示 miRNA 的数量，n_d 表示疾病的数量，$0\leqslant m_i<n_m$，$0\leqslant d_j<n_d$。

2.1.2 miRNA 与疾病相似性数据

本文从 MISIM 数据库[17]中获得了 miRNA 功能相似性网络数据 MFS，$MFS(m_p,m_q)$ 表示 miRNA 中的 m_p 和 m_q 之间的功能相似性值。

根据 Wang 等[17]提出的疾病语义相似性计算方法，本文得到疾病的语义相似性网络数据 DSS，$DSS(d_i,d_j)$ 表示疾病 d_i 和疾病 d_j 之间的语义相似性值。

本文从 Jiang 等[14]的工作中获得了 miRNA 序列相似性网络数据 MSS 与疾病功能相似性网络数据 DFS，$MSS(m_p,m_q)$ 表示 miRNA 中的 m_p 和 m_q 之间的序列相似性值，$DFS(d_i,d_j)$ 表示疾病 d_i 和疾病 d_j 之间的功能相似性值。

2.2 重启随机游走获取拓扑结构信息

miRNA 与疾病相似性网络中的节点表示 miRNA 与疾病实体，边表示 miRNA 与疾病间的相互作用关系[18]。重启随机游走方法从网络中的某一节点开始，每一步可选择移动到相邻节点或者开始节点，直至获取到每一个节点的向量表示。经过重启随机游走后可以得到网络的拓扑结构信息，同时可以捕获网络中任意两个节点之间的多种关系。

本文在 miRNA 功能相似性网络与序列相似性网络、疾病的语义相似性网络与功能相似性网络上采用重启随机游走方法来获取拓扑结构信息。以 miRNA 功能相似性网络为例，具体过程如下：将 miRNA 功能相似性网络 MFS 作为输入，对 MFS 执行重启随机游走，获得带有拓扑结构信息的 miRNA 功能相似性矩阵。重启随机游走的过程如式(1)所示：

$$\boldsymbol{p}_i^{(t)}=(1-\alpha)\boldsymbol{p}_i^{(t-1)}\boldsymbol{A}'+\alpha\boldsymbol{p}_i^{(0)} \tag{1}$$

其中，$\boldsymbol{p}_i^{(t)}$ 是一个行向量，表示从 miRNA 节点 m_i 开始在第 t 步到达网络中各个 miRNA 节点的概率。$\boldsymbol{p}_i^{(0)}$ 是初始的 one-hot 向量，$\boldsymbol{A}'$ 是 MFS 经过行规范化后的转移概率矩阵，α 为重启概率，$1-\alpha$ 表示移动到相邻 miRNA 节点的概率。

miRNA 功能相似性网络 MFS 中每一个 miRNA 节点经过重启随机游走后的特征向量 $\boldsymbol{r}_i$ 的计算过程如式(2)所示：

$$\boldsymbol{r}_i=\sum_{t=1}^{T}\boldsymbol{p}_i^{(t)} \tag{2}$$

其中，T 表示随机游走的总步数。本文为 miRNA 功能相似网络构建 $\boldsymbol{R}^{n_m\times n_m}$ 矩阵，表示 n_m 个 miRNA 重启随机游走后的特征向量。同理，本文对其他 3 种相似性网络进行重启随机游走，获取其带有拓扑结构特征的相似性矩阵。

2.3 堆叠自动编码器学习 miRNA 与疾病特征

堆叠自动编码器中的每一层都是以低一层的输入为基础来提取特征，逐层学习原始数据的多种表达，以获取更抽象的特征，适合完成复杂的分类任务[19]，在 miRNA 靶点等生物信息学领域中被广泛应用[20]。

堆叠自动编码器通过叠加多个自动编码器来实现低维特征的提取。自动编码器模型运行分为编码与解码两个阶段，编码阶段将原始输入数据映射到低维非线性空间，而解码阶段则通过编码阶段获得的低维非线性特征重构输入数据。基本的堆叠自动编码器结构如图 2 所示，其中 n 表示 miRNA 的个数。

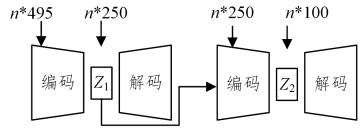

图 2 堆叠自动编码器

Fig. 2 Stacked autoencoder

本文将重启随机游走后带有拓扑结构信息的 miRNA 功能相似性矩阵和序列相似性矩阵、疾病的语义相似性矩阵和功能相似性矩阵分别输入堆叠自动编码器，以学习网络中的低维抽象信息，获取 miRNA 与疾病节点的低维特征表示。自动编码器的定义如式(3)和式(4)所示：

$$\boldsymbol{h}_i=f(\boldsymbol{W}_1\boldsymbol{r}_i+\boldsymbol{b}_1) \tag{3}$$

$$\boldsymbol{o}_i=g(\boldsymbol{W}_2\boldsymbol{h}_i+\boldsymbol{b}_2) \tag{4}$$

其中，f 为编码器编码层的编码函数，$\boldsymbol{r}_i$ 为重启随机游走后 miRNA 相似性矩阵的第 i 个 miRNA 的向量，$\boldsymbol{h}_i$ 表示编码后的低维 miRNA 向量；g 为解码函数，$\boldsymbol{o}_i$ 为解码层重构的 miRNA 特征向量；本文将 f 和 g 均设置为 sigmoid 函数，且设置 $\Theta=\{\theta_1,\theta_2\}=\{\boldsymbol{w}_1,\boldsymbol{b}_1,\boldsymbol{W}_2,\boldsymbol{b}_2\}$ 为需要学习的参数。

本文采用均方误差来计算 miRNA 拓扑相似性特征与重构阶段得到的 miRNA 特征之间的误差，并定义自动编码器的损失函数如式(5)所示：

$$L=\min_{\theta\in\Theta}\sum_{i=1}^{n_m}\|\boldsymbol{o}_i-\boldsymbol{r}_i\| \tag{5}$$

其中，$\boldsymbol{o}_i$ 为重构出的 miRNA 特征，$\boldsymbol{r}_i$ 为 miRNA 输入堆叠自动编码器前的 miRNA 拓扑相似性特征，n_m 为 miRNA 的个数。

本文采用两层堆叠编码器来提取 miRNA 与疾病的低维特征，堆叠自动编码器的具体设置如下：提取 miRNA 特征的堆叠自动编码器的各层节点数依次为[495,250,100]；提取疾病特征的堆叠自动编码器的各层节点数依次为[383,300,150]；堆叠自动编码器的学习率为 0.01；第一层与第二层编码器的迭代次数分别为 1000 和 2500；batch_size 值均为 128。

2.4 深度神经网络预测 miRNA-疾病关联

miRNA-疾病关联预测可以抽象为二分类问题。本文将通过堆叠自动编码器获取的 miRNA 与疾病节点的低维特征进行拼接得到对应的 miRNA-疾病对特征，将 miRNA-疾病对特征输入深度神经网络 DNN 进行 miRNA-疾病关联预测。如果 DNN 的分类结果为 1，则相应 miRNA-疾病对存在潜在关联；若分类结果为 0，则不存在关联。DNN 变换函数如式(6)所示：

$$\boldsymbol{o}_j=f(\boldsymbol{w}_i\boldsymbol{o}_i+\boldsymbol{b}_i) \tag{6}$$

其中，$\boldsymbol{w}_i$ 表示 DNN 中第 i 个隐藏层和下一隐藏层之间需要学习的权重矩阵，$\boldsymbol{o}_i$ 表示第 i 层隐藏层的输入，$\boldsymbol{b}_i$ 表示偏置向量，$\boldsymbol{o}_j$ 表示第 i 层的输出向量。

本文将 DNN 的中间层 f 设置为非线性激活函数 relu；将最终输出层函数设置为 sigmoid，输出值表示 DNN 预测的当

前 miRNA-疾病对之间相互关联的概率。

本文基于 tensorflow 与 keras 进行 DNN 的搭建和训练，利用最小化交叉熵损失函数优化 DNN 模型，交叉熵损失函数如式(7)所示：

$$C=-\frac{1}{N}\sum_{i=1}^{N}[y^{(i)}\log\hat{y}^{(i)}+(1-y^{(i)})\log(1-\hat{y}^{(i)})] \quad (7)$$

其中，C 表示交叉熵损失函数的输出值，N 表示训练样本的数量，i 表示不同样本的下标，$y^{(i)}$ 表示训练样本 i 的真实标签，$\hat{y}^{(i)}$ 表示预测输出。

在模型训练优化过程中，本文选择 Adam 优化器对交叉熵损失函数进行优化，以获取 DNN 的最优参数；采用 dropout 和 early stopping 技术防止过拟合，并设置 dropout 率为 0.5，early stopping 参数为 10。本文设置 DNN 的各层节点数为[500，450，225，1]，设置学习率为 0.000 1。

本文的实验环境为 Windows 10，16 GB 内存，Intel(R) Core(TM) i7-8750 CPU@ 2.20 GHz，显卡版本为 Nvidia GeForce GTX 1050 Ti。

3　实验结果与分析

本文从不同方面分析了所提模型 SAEMDA 的性能。首先，讨论了重启随机游走方法与堆叠自动编码器在本文模型 SAEMDA 中所起到的作用，分析了采用两种 miRNA 与两种疾病相似性网络作为模型输入数据的原因；然后，将本文模型 SAEMDA 与 3 种 miRNA-疾病关联预测模型进行对比；最后，本文提供了两种案例来验证所提模型的可靠性。

3.1　评估指标

miRNA-疾病关联问题中，已知的 miRNA-疾病关联数量远小于未知的关联数量，本文按照如下方式进行阴阳数据集的构建：将 miRNA-疾病关联数据中的 5430 个已知关联作为阳性数据集，从所有未经验证的 184 155 个 miRNA-疾病对中随机挑选 5430 个 miRNA-疾病对作为阴性数据集。本文采用 5 折交叉验证来评估本文模型 SAEMDA 的性能，在每折交叉验证中将 5430 个阳性样本和 5430 个阴性样本分为 5 个子集，选取其中 1 个子集作为测试集，剩余 4 个子集作为训练集，用于训练 DNN 模型。

为了准确评估模型的预测性能，本文选取 10 次 5 折交叉验证的平均 AUC、AUPR、precision、recall 与 F1 作为评估指标，其中 AUC 为以 FPR 为横坐标、TPR 为纵坐标的 ROC 曲线下面积，AUPR 为以 recall 为横坐标、precision 为纵坐标的 P-R 曲线下面积，各指标计算过程如式(8)－式(12)所示：

$$TPR=TP/(TP+FN) \quad (8)$$

$$FPR=FP/(FP+TN) \quad (9)$$

$$recall=TP/(TP+FN) \quad (10)$$

$$precision=TP/(TP+FP) \quad (11)$$

$$F1=2*precision*recall/(precision+recall) \quad (12)$$

3.2　消融性测试

为了验证重启随机游走过程以及堆叠编码器组件在本文模型 SAEMDA 中所起到的作用，本文设置了两组对比实验。第一组实验 DNN，表示将 miRNA 与疾病相似性网络数据直接输入 DNN 进行 miRNA-疾病关联预测；第二组实验 RWR_DNN，首先对 miRNA 与疾病相似性网络实施重启随机游走，再将重启随机游走后的 miRNA 与疾病相似性网络数据输入 DNN 进行 miRNA-疾病关联预测。表 1 列出了 DNN、RWR_DNN 和本文模型 SAEMDA 在 10 次 5 折交叉验证下的平均 AUC、AUPR、precision、recall 和 F1 值。从表 1 中可以发现，RWR_DNN 相比 DNN 获得了更高的 recall 和 F1 指标，而较高的 recall 值表明：采用重启随机游走获取 miRNA 与疾病相似性网络的拓扑结构信息，有利于识别更多真正的 miRNA-疾病关联对[21]。而本文模型 SAEMDA 各指标都高于 RWR_DNN，表明堆叠自动编码器可以学习到 miRNA 与疾病相似性网络中重要的特征，进一步提升了模型的预测性能。

表 1　DNN、RWR_DNN 与 SAEMDA 的对比结果

Table 1　Comparison results of DNN，RWR_DNN and SAEMDA

	AUC	AUPR	precision	recall	F1
DNN	0.9322	0.9279	0.8688	0.8101	0.8324
RWR_DNN	0.9199	0.9203	0.8380	0.8507	0.8442
SAEMDA	**0.9329**	**0.9315**	0.8483	**0.8697**	**0.8587**

表 2 列出了不同相似性网络组合作为所提模型 SAEMDA 输入数据时的 10 次 5 折交叉验证的平均 AUC、AUPR、precision、recall 和 F1 值。从表 2 中可以观察到，分别采用两种 miRNA 和疾病相似性网络作为输入较分别采用单个相似性网络取得了更好的结果。因此，本文使用两种 miRNA 相似性网络与两种疾病相似性网络数据作为所提模型 SAEMDA 的输入。

表 2　不同相似性网络组合下 SAEMDA 模型的 5 折交叉验证结果

Table 2　5-fold cross validation results of SAEMDA model with different similarity network combinations

相似性网络组合	AUC	AUPR	precision	recall	F1
miRNA 序列相似性网络＋疾病功能相似性网络	0.8850	0.8860	0.8143	0.7998	0.8069
miRNA 序列相似性网络＋疾病语义相似性网络	0.8975	0.9002	0.8213	0.8161	0.8181
miRNA 功能相似性网络＋疾病功能相似性网络	0.9309	0.9304	0.8435	0.8685	0.8557
miRNA 功能相似性网络＋疾病语义相似性网络	0.9324	0.9297	0.8460	0.8672	0.8564
miRNA 功能相似性网络＋miRNA 序列相似性网络＋疾病功能相似性网络＋疾病语义相似性网络	**0.9329**	**0.9315**	**0.8483**	**0.8697**	**0.8587**

3.3　与现有方法的对比

为了评估本文模型 SAEMDA 的有效性，本文将其与 RFMDA[5]、IRFMDA[6] 和 MDA-CNN[9] 模型进行了对比，5 折交叉验证的对比结果如图 3 所示。本文模型 SAEMDA 的 AUC、AUPR、precision、recall 和 F1 指标分别为 0.932 9，0.931 5，0.848 3，0.869 7 和 0.858 7，其各项指标均优于对比方法。

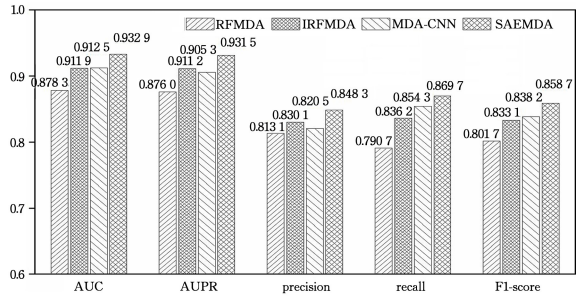

图 3 本文模型 SAEMDA、RFMDA、IRFMDA 与 MDA-CNN 的 5 折交叉验证指标对比

Fig. 3 Comparison of 5-fold cross validation index between proposed SAEMDA, RFMDA, IRFMDA and MDA-CNN

从图 3 可以看出，RFMDA 模型在各项指标上始终低于其他 3 种模型，而 IRFMDA 则采用随机森林自带的特征评分机制选取特征，取得了比 RFMDA 更好的效果，表明手工选取特征会出现人为偏差，而选取更好的特征有利于 miRNA-疾病关联预测。MDA-CNN 方法通过编码器提取 miRNA 与疾病相似性网络特征，采用卷积神经网络 CNN 进行 miRNA-疾病关联预测，相比 RFMDA 与 IRFMDA 这些传统机器学习方法取得了更高的 AUC、recall 和 F1 值，表明提取相似性网络中的非线性特征有利于 miRNA-疾病关联预测。本文模型 SAEMDA 采用堆叠自动编码器逐层提取 miRNA 与疾病相似性网络中的高阶抽象信息，同时考虑了 miRNA 与疾病相似性网络中的拓扑结构特征，取得了比 MDA-CNN 更好的结果。图 4 给出了 SAEMDA，RFMDA，IRFMDA 和 MDA-CNN 的 ROC 曲线和 P-R 曲线，可以看出本文模型 SAEMDA 的 AUC 和 AUPR 值最大，性能更优。

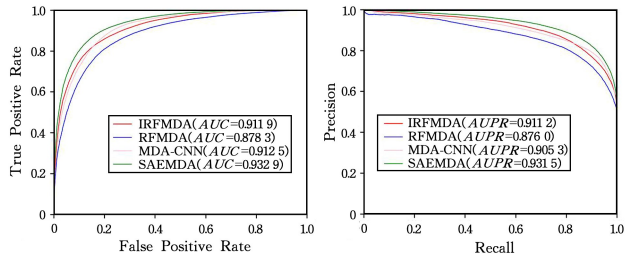

图 4 本文模型 SAEMDA、RFMDA、IRFMDA 与 MDA-CNN 的 5 折交叉验证的 AUC 与 AUPR

Fig. 4 AUC and AUPR of the proposed SAEMDA, RFMDA, IRFMDA and MDA-CNN in 5-fold cross validation

3.4 案例分析

为了进一步体现 SAEMDA 模型在真实案例中的可靠性，本文在结肠癌和肺癌两种案例上进行了验证。首先检测了结肠癌案例的预测情况，各种复杂的原因导致治疗结肠癌仍存在困难。Yu 等[22]发现 miR-21 的下调能够使化疗耐药性结肠癌细胞更容易接受治疗方案，这表明发现并利用 miRNA在结肠癌发病机制中的作用能有效促进对结肠癌的治疗。

本文将模型 SAEMDA 用于预测与结肠癌相关的 miRNA，并将排名在前 50 的 miRNA 在数据库 dbDEMC2.0 中进行验证，如果数据库存在预测出的 miRNA-结肠癌关联，则代表预测成功。表 3 列出了与结肠癌关联的前 50 个 miRNA 的验证结果，第 1 列、第 3 列和第 5 列表示预测结果中与结肠癌相关联的 miRNA，第 2 列、第 4 列和第 6 列表示在 dbDEMC2.0 数据库中的验证情况，可以看到：与结肠癌相关的前 50 个 miRNA 中，有 45 个 miRNA 在 dbDEMC2.0 中得到了验证。

表 3 SAEMDA 预测的结肠癌相关 miRNA 验证结果

Table 3 Validation results of colon cancer-related miRNA predicted by SAEMDA

miRNA	Evidence	miRNA	Evidence	miRNA	Evidence
hsa-mir-21	dbDEMC	hsa-mir-31	dbDEMC	hsa-mir-23a	dbDEMC
hsa-mir-155	dbDEMC	hsa-mir-19b	dbDEMC	hsa-mir-34c	unconfirmed
hsa-mir-92a	dbDEMC	hsa-mir-18a	dbDEMC	hsa-mir-451a	dbDEMC
hsa-mir-29a	dbDEMC	hsa-mir-16	unconfirmed	hsa-mir-19a	dbDEMC
hsa-mir-221	dbDEMC	hsa-mir-107	dbDEMC	hsa-mir-24	dbDEMC
hsa-mir-200c	unconfirmed	hsa-mir-148a	dbDEMC	hsa-mir-20a	dbDEMC
hsa-mir-30a	dbDEMC	hsa-mir-196a	dbDEMC	hsa-mir-137	dbDEMC
hsa-mir-29b	dbDEMC	hsa-mir-223	dbDEMC	hsa-mir-100	dbDEMC
hsa-mir-142	unconfirmed	hsa-mir-133b	dbDEMC	hsa-let-7f	dbDEMC
hsa-mir-203	dbDEMC	hsa-mir-210	dbDEMC	hsa-mir-28	dbDEMC
hsa-mir-143	dbDEMC	hsa-mir-140	dbDEMC	hsa-mir-192	dbDEMC
hsa-mir-34a	dbDEMC	hsa-mir-183	dbDEMC	hsa-let-7i	dbDEMC
hsa-mir-199a	unconfirmed	hsa-mir-375	dbDEMC	hsa-mir-574	dbDEMC
hsa-mir-99a	dbDEMC	hsa-let-7a	dbDEMC	hsa-mir-214	dbDEMC
hsa-mir-125b	dbDEMC	hsa-mir-15a	dbDEMC	hsa-mir-191	dbDEMC
hsa-mir-1	dbDEMC	hsa-mir-182	dbDEMC	hsa-mir-205	dbDEMC
hsa-mir-146a	dbDEMC	hsa-mir-200b	dbDEMC		

肺癌是人群中最常被诊断出的疾病[23]，在第二个案例中，将本文模型 SAEMDA 在肺癌数据库上进行验证。首先，在已知的 miRNA-疾病关联数据中去除与肺癌相关的 miRNA，重新训练本文模型 SAEMDA；然后，利用训练好的 SAEMDA 模型预测每种 miRNA 与肺癌的关联分数，并对这些关联预测分数进行排序。

表 4 列出了与肺癌相关的排名前 50 的 miRNA，我们将表 4 中的 miRNA 在数据库 dbDEMC2.0 与 HMDDv3.2 中进行验证，表 4 中的第 2 列、第 4 列和第 6 列表示与肺癌相关联的 miRNA 在数据库中的验证情况，H&D 表示

在 dbDEMC2.0 与 HMDDv3.2 数据库中都得到了验证。可以发现：排名前 50 的 miRNA-肺癌的关联在数据库中均得到了验证，这表明本文模型 SAEMDA 可以用于发现未知的 miRNA-疾病关联。

表 4　SAEMDA 预测的肺癌相关 miRNA 验证结果

Table 4　Validation results of lung cancer-related miRNA predicted by SAEMDA

miRNA	数据库	miRNA	数据库	miRNA	数据库
hsa-mir-21	H&D	hsa-mir-133a	H&D	hsa-mir-195	H &D
hsa-mir-125b	H&D	hsa-mir-223	H&D	hsa-mir-146b	H &D
hsa-mir-155	H&D	hsa-mir-222	H&D	hsa-mir-210	H &D
hsa-mir-221	H&D	hsa-mir-106b	H&D	hsa-mir-126	H&D
hsa-mir-34c	H&D	hsa-mir-199a	H &D	hsa-mir-24	H&D
hsa-mir-16	H&D	hsa-mir-145	H&D	hsa-mir-499a	HMDDv3.2
hsa-mir-7	H&D	hsa-mir-200b	H&D	hsa-mir-100	H&D
hsa-mir-30a	H&D	hsa-mir-92a	H&D	hsa-mir-375	H&D
hsa-mir-34a	H&D	hsa-mir-22	H&D	hsa-mir-30c	H&D
hsa-mir-19b	H&D	hsa-mir-148a	H&D	hsa-mir-193b	dbDEMC2.0
hsa-mir-17	H&D	hsa-mir-335	H&D	hsa-let-7d	H&D
hsa-mir-20a	H&D	hsa-mir-29a	H&D	hsa-mir-128	dbDEMC2.0
hsa-mir-15b	dbDEMC2.0	hsa-mir-146a	dbDEMC2.0	hsa-mir-101	H&D
hsa-mir-34b	H&D	hsa-let-7b	H&D	hsa-mir-137	H&D
hsa-mir-200a	H&D	hsa-mir-31	H&D	hsa-mir-18a	H&D
hsa-mir-99a	H&D	hsa-mir-451a	H&D	hsa-mir-181a	H&D
hsa-mir-1	H&D	hsa-mir-196a	H&D		

结束语　miRNA 是一种具有重要生物学功能的非编码 RNA，miRNA 的非正常调控与许多人类疾病有关，因此 miRNA-疾病关联预测问题引起了人们广泛的关注。本文提出的 miRNA-疾病关联预测模型 SAEMDA，在 miRNA 与疾病相似性网络上进行重启随机游走，捕获网络的拓扑结构信息，并将获取拓扑结构信息后的特征输入堆叠自动编码器提，取低维抽象特征，然后将结果输入到深度神经网络中进行预测。实验结果表明，SAEMDA 模型在 miRNA-疾病关联预测问题中表现出了较好的性能。但是，SAEMDA 模型也存在一定限制，本文仅考虑了两种相似性网络，融合更多的相似性网络信息可能会得到更好的结果，我们将在以后的工作中分析更多的相似性网络信息，来提高模型的表现能力。

参 考 文 献

[1] ZHANG J X, SONG W, CHEN Z H, et al. Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis[J]. The Lancet Oncology, 2013, 14(13): 1295-1306.

[2] SU Y, DENG M F, XIONG W, et al. MicroRNA-26a/death-associated protein kinase 1 signaling induces synucleinopathy and dopaminergic neuron degeneration in Parkinson's disease [J]. Biological Psychiatry, 2019, 85(9): 769-781.

[3] YOU Z, HUANG Z A, ZHU Z X, et al. PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction [J]. PLoS Computational Biology, 2017, 13(3): e1005455.

[4] CHEN X, LIU M X, YANG Y. RWRMDA: predicting novel human microRNA-disease associations [J]. Molecular Biosystems, 2012, 8(10): 2792-2798.

[5] CHEN X, WANG C C, YIN J, et al. Novel human miRNA-disease association inference based on random forest[J]. Molecular Therapy-Nucleic Acids, 2018, 13: 568-579.

[6] YAO D, ZHAN X, KWOH C K. An improved random forest-based computational model for predicting novel miRNA-disease associations[J]. BMC Bioinformatics, 2019, 20(1): 1-14.

[7] ZHANG L, CHEN X, YIN J. Prediction of Potential miRNA-Disease Associations Through a Novel Unsupervised Deep Learning Framework with Variational Autoencoder[J]. Cells, 2019, 8(9): 1040.

[8] PENG J, HUI W, LI Q, et al. A learning-based framework for miRNA-disease association identification using neural networks [J]. Bioinformatics, 2019, 35(21): 4364-4371.

[9] CHEN X, GONG Y, ZHANG D H, et al. DRMDA: deep representations-based miRNA-disease association prediction[J]. Journal of Cellular and Molecular Medicine, 2017, 22(1): 472-485.

[10] WANG L, XU T, SONG C D. Prediction algorithm of miRNA and disease correlation based on deep learning[J]. Acta Electronica Sinica, 2020, 447(5): 40-47.

[11] KÖHLER S, BAUER S, HORN D, et al. Walking the interactome for prioritization of candidate disease genes[J]. The American Journal of Human Genetics, 2008, 82(4): 949-958.

[12] TANG J Q, WU J L, LIAO Y X, et al. Protein function prediction based on double weighted voting[J]. Computer Science, 2019, 46(4): 222-227.

[13] WANG H, LE Z C, GONG X, et al. Summary of link prediction methods based on feature classification[J]. Computer Science, 2020, 47(8): 302-312.

[14] JIANG L, DING Y, TANG J, et al. MDA-SKF: similarity kernel fusion for accurately discovering miRNA-disease association[J]. Frontiers in Genetics, 2018, 9: 618.

[15] LI Y, QIU C, TU J, et al. HMDD v2.0: a database for experimentally supported human microRNA and disease associations [J]. Nucleic Acids Research, 2014, 42(D1): D1070-D1074.

[16] YANG Z, REN F, LIU C, et al. dbDEMC: a database of differentially expressed miRNAs in human cancers [J]. BMC Genomics, 2010, 11(4): 1-8.

[17] 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.

[18] RUAN L, XIONG Y. Research on Functional Similarity of miRNA Based on Network Representation Learning[J]. Computer Engineering, 2019, 45(2): 154-159.

[19] ZHU Y X, FENG W, GUO X H. Application progress of deep learning method in brain image of Alzheimer's disease[J]. Medical Review, 2019, 25(18): 3562-3566.

[20] LI Y N, HU Y J, GAN W, et al. Survey on Target Site Prediction of Human miRNA Based on Deep Learning[J]. Computer Science, 2021, 48(1): 209-216.

[21] XUAN P, DONG Y, GUO Y, et al. Dual convolutional neural network based method for predicting disease-related miRNAs [J]. International Journal of Molecular Sciences, 2018, 19(12): 3732.

[22] YU Y, NANGIA-MAKKER P, FARHANA L, et al. miR-21 and miR-145 cooperation in regulation of colon cancer stem cells[J]. Molecular Cancer, 2015, 14(1): 1-11.

[23] BRAY F, FERLAY J, SOERJOMATARAM I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: a Cancer Journal for Clinicians, 2018, 68(6): 394-424.

LIU Dan, born in 1995, is a member of China Computer Federation. Her main research interests include intelligent information processing and bioinformatics.

WANG Hui-qing, born in 1978, Ph. D, associate professor, is a member of China Computer Federation. Her main research interests include intelligent information processing and bioinformatics.