基于深度学习的 miRNA 与疾病相关性预测算法

王 磊[1],徐 涛[1],宋传东[1],王海峰[1],尤著宏[2],宋克俭[3],闫 欣[4]
(1. 枣庄学院信息科学与工程学院,山东枣庄 277100;2. 中国科学院新疆理化技术研究所,新疆乌鲁木齐 830011;
3. 江西理工大学信息工程学院,江西赣州 341000;4. 枣庄学院外国语学院,山东枣庄 277100)

摘 要: 大量研究表明,microRNA(miRNA)在人类复杂疾病研究中发挥着重要作用.识别 miRNA 与疾病之间的关系对于提高复杂疾病的治疗水平具有重要意义.然而,传统实验方式常受限于小规模和高成本,因此迫切需要计算模拟的方式快速有效地预测 miRNA-疾病间的潜在关系.本文通过结合深度学习的堆叠自动编码器算法与旋转森林分类器对 miRNA-疾病间关系进行预测.该方法能够有效抽取出融合了疾病语义相似性、miRNA 功能相似性和 miRNA 序列信息的高级特征并对其进行准确分类.在交叉验证实验中,该方法在 HMDD v3.0 数据集上取得 90.30% 的预测准确率.此外,我们还在人类复杂疾病乳腺肿瘤上做了案例研究.结果,模型预测得分最高的前 30 个疾病关联 miRNA 中 28 个得到了证实.这些优异的结果表明,该算法是一种有效预测 miRNA-疾病关系的工具,能够为生物实验提供高可靠的疾病关联 miRNA 候选物.

关键词: 深度学习;miRNA-疾病关系;堆叠自动编码器;旋转森林
中图分类号: TP399 **文献标识码:** A **文章编号:** 0372-2112 (2020)05-0870-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.05.006

Prediction Algorithm of Association Between miRNAs and Diseases Based on Deep Learning

WANG Lei[1], XU Tao[1], SONG Chuan-dong[1], WANG Hai-feng[1], YOU Zhu-hong[2], SONG Ke-jian[3], YAN Xin[4]
(1. *College of Information Science and Engineering, Zaozhuang University, Zaozhuang, Shandong* 277100, *China*;
2. *Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Sciences, Urumqi, Xinjiang* 830011, *China*;
3. *School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi* 341000, *China*;
4. *School of Foreign Languages, Zaozhuang University, Zaozhuang, Shandong* 277100, *China*)

Abstract: Numerous studies have shown that microRNA (miRNA) plays important role in the study of human complex diseases. Identifying the association between miRNAs and diseases is important for improving the therapeutic level of complex diseases. However, traditional experimental is often limited to small-scale and high-cost, so computational simulation is urgently needed to quickly and effectively predict the potential miRNAs-disease associations. In this study, a new method is proposed to predict the miRNA-disease association by combining deep learning stacked automatic encoder algorithm with rotation forest classifier. This method can effectively extract high-level features that combine disease semantic similarity, miNRA functional similarity and miRNA sequence information, and accurately classify them. In the cross-validation experiment, this method achieved 90.30% prediction accuracy on the HMDD v3.0 dataset. Furthermore, we have also done case studies on Breast Neoplasms. As a result, 28 of the top 30 miRNA-disease associations were confirmed. These excellent results indicate that this method is an effective tool for predicting miRNA-disease associations, and can provide highly reliable candidate miRNAs for biological experiments.

Key words: deep learning; miRNA-disease association; stacked automatic encoder; rotation forest

收稿日期:2019-07-11;修回日期:2019-10-09;责任编辑:马兰英
基金项目:国家自然科学基金(No.61702444);中国博士后科学基金(No.2019M653804);中科院西部之光(No.2018-XBQNXZ-B-008)

1　引言

MicroRNAs(miRNAs)是一类微小的内源性非编码RNA,它可通过诱导信使 RNA 降解、翻译抑制或其他形态调节机制来抑制靶基因的表达[1]. 研究表明,miRNA在许多生物进程中发挥着重要作用,它们的突变或异常表达会导致多种人类复杂疾病[2]. 因此,miRNA 与疾病间关系的识别有助于人们深入了解复杂疾病的致病机制,为疾病的预防和诊疗提供有效帮助.

随着高通量测序技术的发展,越来越多的 miRNA-疾病关系被发现. 例如,Bang 等人[3]发现,在心肌缺血和视网膜血管发育过程中 mir-23/27/24 参与血管生成并在心血管生成中发挥重要作用. 然而,传统生物实验方法常受到环境的制约,并且存在成本高、周期长等弊端. 因此,迫切需要更有效的计算方法,来实现大规模、高可信地预测 miRNA-疾病间的潜在关系.

在本文中,我们提出了一种基于机器学习的方法来预测 miRNA-疾病关系. 该模型融合了 miRNA 序列、miRNA 功能相似性和疾病语义相似性的多种信源信息. 具体来说,该模型首先根据 miRNA 功能和疾病语义相似性分别计算 miRNA 与疾病的相似性矩阵,并将它们与高斯交互谱核相似性矩阵相结合. 其次,使用自然语言处理技术提取出 miRNA 序列的特征,并与 miRNA-疾病相似性描述符相融合. 然后使用深度学习算法对融合信息的特征进行客观提取,从而挖掘出隐藏的高级抽象特征. 最后,使用旋转森林分类器准确预测出潜在的 miRNA-疾病间关系. 所提模型的流程图如图 1 所示.

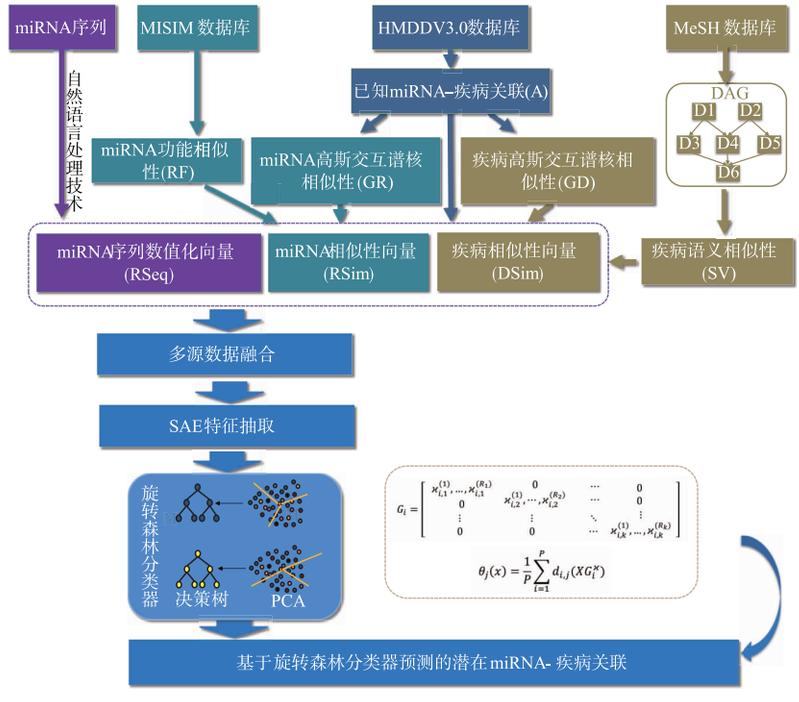

图1　miRNA-疾病关系预测模型流程图

2　材料及方法

2.1　miRNA – 疾病关联数据集

在实验中,我们使用 Li 等人[4]提供的 HMDD v3.0 数据集来对模型进行验证. 经过数据预处理,我们保留了包含 1057 个 miRNA 和 850 个疾病的 32226 条 miRNA-疾病关联数据. 由于 HMDD 数据集中没有提供无关联的 miRNA-疾病数据,因此我们从去除了正样本的所有可能的 miRNA-疾病关联数据中随机选择 32226 对作为负样本集.

基于上述讨论,我们构建了 HMDD 数据集的邻接矩阵 $\boldsymbol{AD}$. 当疾病 $\boldsymbol{d}(i)$ 和 miRNA $\boldsymbol{m}(j)$ 具有关系时,邻接矩阵 $\boldsymbol{AD}$ 对应的元素 $\boldsymbol{AD}(\boldsymbol{d}(i),\boldsymbol{m}(j))$ 被赋值为 1,否则赋值为 0.

2.2　疾病语义相似性模型的构建

本文使用的疾病语义相似性信息来自于美国国家医学图书馆的 MeSH 数据库[5]. 在 MeSH 中,疾病间的关系被描述为有向无环图(DAG),其中节点表示疾病,边表示疾病间的关系. 在 DAG 中,疾病 $\boldsymbol{d}(i)$ 被表示为 $\mathrm{DAG}(\boldsymbol{d}(i))=(\boldsymbol{d}(i),\boldsymbol{N}(\boldsymbol{d}(i)),\boldsymbol{E}(\boldsymbol{d}(i)))$,其中 $\boldsymbol{N}(\boldsymbol{d}(i))$ 是包括疾病 $\boldsymbol{d}(i)$ 在内的,疾病 $\boldsymbol{d}(i)$ 的祖先节点集,$\boldsymbol{E}(\boldsymbol{d}(i))$ 是连接这些疾病的边集. 因此,$\boldsymbol{DAG}(\boldsymbol{d}$

$(i))$中的某一疾病 $\boldsymbol{s}$ 对疾病 $\boldsymbol{d}(i)$ 的语义贡献值可以计算如下：

$$\begin{cases} D_{d(i)}(s)=1, & \text{if } s=d(i) \\ D_{d(i)}(s)=\max\{\varepsilon \cdot D_{d(i)}(s') \mid s' \in children\ of\ s\}, & \text{if } s\neq d(i) \end{cases} \tag{1}$$

此处 ε 为语义贡献因子. 由此，我们可以得到疾病 $\boldsymbol{d}(i)$ 的语义值 $\boldsymbol{DV}$：

$$DV(d(i))=\sum_{s\in N_{d(i)}} D_{d(i)}(s) \tag{2}$$

在这里，我们根据不同疾病间 DAG 共享的部分越多，就具有更高语义相似性的假设构建了语义相似度模型 $\boldsymbol{SV}_1$，它可以计算如下：

$$SV_1(d(i),d(j))=\frac{\sum_{s\in N_{d(i)}\cap N_{d(j)}}(D_{d(i)}(s)+D_{d(j)}(s))}{DV(d(i))+DV(d(j))} \tag{3}$$

在 $\boldsymbol{SV}_1$ 模型中，我们考虑了不同疾病在 DAG 中层次的关系. 然而，不同疾病在 DAG 中出现的次数是不同的. 因此，我们根据疾病的次数计算另一疾病贡献值：

$$D'_{d(i)}(s)=-\log\left(\frac{num(DAGs(s))}{num(diseases)}\right) \tag{4}$$

此处 num(DAGs(s))表示包含疾病 s 的 DAG 的数量，num(diseases)表示疾病的数量. 由此，我们构建了第二种疾病语义相似度模型 $\boldsymbol{SV}_2$：

$$SV_2(d(i),d(j))=\frac{\sum_{s\in N_{d(i)}\cap N_{d(j)}}(D'_{d(i)}(s)+D'_{d(j)}(s))}{DV(d(i))+DV(d(j))} \tag{5}$$

2.3 miRNA 功能相似性模型的构建

基于功能相似的 miRNA 更可能与表型相似的疾病关联的假设，Wang 等人[6]提出了一种计算不同 miRNA 分子间功能相似性的模型. 在实验中，我们使用它来作为 miRNA 的功能相似性.

2.4 高斯交互谱核相似性模型的构建

我们引入高斯交互谱核(GIP)相似性[7]模型来计算 HMDD 中不具有语义相似性的 miRNA－疾病信息. 通过疾病 $\boldsymbol{d}(i)$ 是否与 miRNA 相关，我们定义了向量 $\boldsymbol{V}(\boldsymbol{d}(i))$表示疾病的相互作用谱. GIP 相似性 $\boldsymbol{GD}$ 可计算如下：

$$\begin{aligned} &\boldsymbol{GD}(\boldsymbol{d}(i),\boldsymbol{d}(j)) \\ &=\exp(-\theta_d \parallel V(d(i))-V(d(j)) \parallel^2) \end{aligned} \tag{6}$$

其中 $\boldsymbol{\theta}_d$是函数的宽度参数，它通过规范化参数来计算：

$$\theta_d=\frac{1}{m}\sum_{i=1}^{m} \parallel V(d(i)) \parallel^2 \tag{7}$$

同样地，miRNA 间的 GIP 相似性 $\boldsymbol{GR}$ 可以计算如下：

$$\boldsymbol{GR}(r(i),r(j))=\exp(-\theta_r \parallel V(r(i))-V(r(j)) \parallel^2) \tag{8}$$

$$\theta_r=\frac{1}{n}\sum_{i=1}^{n} \parallel V(r(i)) \parallel^2 \tag{9}$$

2.5 miRNA 序列特征的抽取

我们将 miRNA 序列与其相似向量结合，从而形成最终的 miRNA 描述符. 我们使用自然语言处理技术来提取序列特征，它能够将原始的高维数据转换成低维连续实值向量，并且可以在无监督的情况下从 miRNA 序列中抽取到有效的特征.

在本文中，我们使用 skip-gram 算法来学习 miRNA 的分布表示. 给出一个单词序列，skip-gram 能够使用上下文窗口中单词的共现信息来学习单词表示，并寻找出参数集 $\boldsymbol{\theta}$ 以最大化以下条件概率的乘积：

$$\arg\max_{\theta}\prod_{w\in T}\left[\prod_{c\in C(w)} p(c\mid w;\boldsymbol{\theta})\right] \tag{10}$$

此处 T 是文本集，w 是单词，c 为单词的上下文，$C(w)$ 是文本集 T 中出现单词 w 的上下文中包含的一组单词，p 为条件概率，其定义如下：

$$p(c\mid w;\boldsymbol{\theta})=\frac{\exp(v_c \cdot v_w)}{\sum_{c\in C}\exp(v_c \cdot v_w)} \tag{11}$$

此处 $\boldsymbol{v}_c$和 $\boldsymbol{v}_w$分别是 c 和 w 的列向量，参数 $\boldsymbol{\theta}$ 是以 $\boldsymbol{v}_c$和 $\boldsymbol{v}_w$ 表示的每个维度的特定值. 在实验中，我们使用 6-mers 来转换 miRNA 序列，并通过从公共数据库 miRbase 中下载的全部 miRNA 序列作为训练集来对 skip-gram 算法进行训练.

2.6 多源信息的融合

在本文中，我们最终使用的描述符包含了疾病相似性、miRNA 相似性和 miRNA 序列三种来源的数据.

对疾病来说，我们构建了语义相似性模型 $\boldsymbol{SV}_1$ 和 $\boldsymbol{SV}_2$以及 GIP 相似性模型 GD. 由此得到疾病相似性矩阵 DSim：

$$\begin{aligned} &\boldsymbol{DSim}(\boldsymbol{d}(i),\boldsymbol{d}(j)) \\ &=\begin{cases} \dfrac{\boldsymbol{SV}_1(\boldsymbol{d}(i),\boldsymbol{d}(j))+\boldsymbol{SV}_2(\boldsymbol{d}(i),\boldsymbol{d}(j))}{2}, \\ \qquad \text{如果 } \boldsymbol{d}(i) \text{ 和 } \boldsymbol{d}(j) \text{ 存在语义相似性} \\ \boldsymbol{GD}(\boldsymbol{d}(i),\boldsymbol{d}(j)), \\ \qquad \text{否则} \end{cases} \end{aligned} \tag{12}$$

对 miRNA 来说，我们将功能相似性 RF 与 GIP 模型 GR 结合起来形成 miRNA 相似矩阵 $\boldsymbol{RSim}$，其表示如下：

$$\begin{aligned} &\boldsymbol{RSim}(r(i),r(j)) \\ &=\begin{cases} RF(r(i),r(j)), \\ \qquad \text{如果 } d(i) \text{ 和 } d(j) \text{ 存在功能相似性} \\ GR(r(i),r(j)), \\ \qquad \text{否则} \end{cases} \end{aligned} \tag{13}$$

对于最终的描述符 $\boldsymbol{FV}$，我们还需要整合 miRNA 的

序列信息 RSeq. 最终的描述符 $\boldsymbol{FV}$ 可表示如下:

$$\begin{aligned}&\boldsymbol{FV}(\boldsymbol{d}(i),r(j))\\&=[\boldsymbol{DSim}(d(i)),\boldsymbol{RSim}(r(j)),\boldsymbol{RSeq}(r(j))]\end{aligned}\tag{14}$$

2.7　堆叠自动编码器

堆叠式自动编码器(SAE)[8]是由多个串联的自动编码器(AE)叠加而成的深度学习框架. 典型的 SAE 结构如图 2 所示.

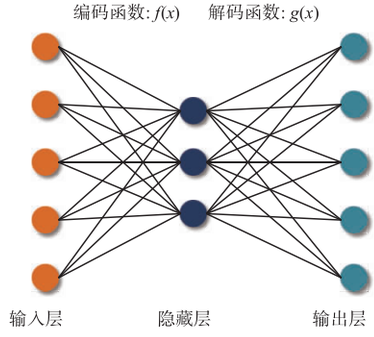

图2　自动编码器结构

通过堆叠多个 AE 来形成 SAE. 图 3 展示了具有三级自动编码器的堆叠自动编码器的结构. 具体流程如下:

(1)在给定初始输入的情况下,对第一层 AE 进行无监督训练,将重构误差减小到设定值.

(2)使用第一个 AE 隐藏层的输出作为第二个 AE 的输入,并按照步骤(1)的方式对第二个 AE 进行训练.

(3)重复第二步,直到所有 AE 完成初始化.

(4)最后一个堆叠 AE 隐藏层的输出作为分类器的输入,然后使用监督方法调整分类器的参数.

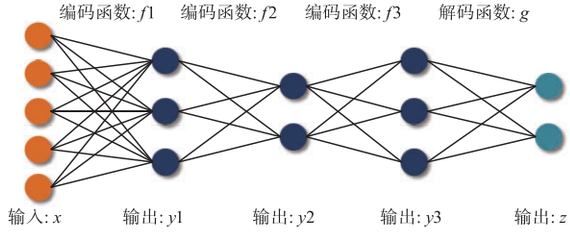

图3　堆叠自动编码器结构

2.8　旋转森林分类器

旋转森林[9](RF)是一种使用线性分析理论和决策树的分类算法. 设 M 为样本集(X,Y),其中 X 表示样本的属性,Y 表示样本的标签. 假设 RF 中决策树用 R 表示. 对于决策树 R_i 的训练过程如下:

(1)将样本集 M 随机分为 k 个子集,每个子集样本数量为 n/k.

(2)在子集 $\boldsymbol{M}(i,j)$ 中选择特征集 X 对应的列形成新的矩阵 $\boldsymbol{X}(i,j)$. 使用 bootstrap 算法从 $\boldsymbol{X}(i,j)$ 中随机选择 75% 作为新的特征集,循环该过程 k 次.

(3)使用主成分分析算法对新的特征集做特征变换,生成系数矩阵 $\boldsymbol{S}(i,j)$.

(4)构造稀疏矩阵 $\boldsymbol{G}_i$,其参数由矩阵 $\boldsymbol{S}(i,j)$ 构成,可表示如下:

$$\boldsymbol{G}_i=\begin{bmatrix}\mathcal{H}_{i,1}^{(1)},\cdots,\mathcal{H}_{i,1}^{(R_1)} & 0 & \cdots & 0\\0 & \mathcal{H}_{i,2}^{(1)},\cdots,\mathcal{H}_{i,2}^{(R_2)} & \cdots & 0\\\vdots & \vdots & \ddots & \vdots\\0 & 0 & \cdots & \mathcal{H}_{i,k}^{(1)},\cdots,\mathcal{H}_{i,k}^{(R_k)}\end{bmatrix}\tag{15}$$

在分类时,根据决策树 R_i 对测试样本 x 产生的结果来判断 x 属于类别 y_i. 然后使用平均组合方法计算出所有决策树的结果,其公式如下:

$$\theta_j(x)=\frac{1}{p}\sum_{i=1}^{p}d_{i,j}(XG_i^{-H})\tag{16}$$

最终,测试样本 x 被旋转森林判定给得分最高的类.

3　实验结果与分析

3.1　评估标准

为了对所提模型公平地评估,我们遵循通用的评估标准[10],包括准确率(Accu.)、敏感率(Sen.)、精确率(Prec.)、马修斯相关系数(MCC)和虚警率(Fa.),它们的定义如下:

$$\text{Accu.}=\frac{TP+TN}{TP+TN+FP+FN}\tag{17}$$

$$\text{Sen.}=\frac{TP}{TP+FN}\tag{18}$$

$$\text{Prec.}=\frac{TP}{TP+FP}\tag{19}$$

$$\text{MCC}=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}\tag{20}$$

$$\text{Fa.}=\frac{FP}{TP+FP}\tag{21}$$

其中 TP、TN、FP 和 FN 分别表示真阳性、真阴性、假阳性和假阴性. 此外,我们还绘制了 ROC 曲线[11]并计算了 ROC 曲线下面积(AUC).

3.2　模型能力评估

在实验中我们使用了五折交叉验证法(5-fold CV)来评估模型能力. 表 1 汇总了模型在 HMDD v3.0 数据集上取得的结果. 从表中我们看到,所提模型取得了 90.30% 的准确率、89.80% 的敏感率、90.71% 的精确率和 80.61% 的马修斯相关系数. 图 4 绘制了所提模型在 HMDD 上生成的 ROC 曲线. 从图中可以看到,模型生成的 AUC 值达到了 90.26%. 这些结果表明,所提模型具有较好的性能,能够有效预测出 miRNA – 疾病关系.

表 1 所提模型在 HMDD 数据集上的实验结果

Test set	Accu.(%)	Sen.(%)	Prec.(%)	MCC(%)	Fa.(%)
1	90.48	89.84	90.98	80.97	9.02
2	90.26	89.42	90.82	80.53	9.18
3	90.17	89.79	90.53	80.34	9.47
4	90.20	90.23	90.10	80.40	9.90
5	90.39	89.73	91.12	80.79	8.88
Average	**90.30 ±0.13**	**89.80 ±0.29**	**90.71 ±0.41**	**80.61 ±0.26**	**9.29 ±0.41**

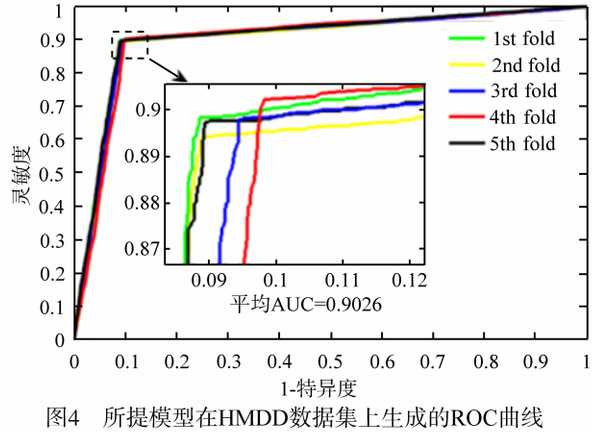

图4 所提模型在HMDD数据集上生成的ROC曲线

3.3 不同分类器模型间的比较

为了验证不同分类器对模型性能的影响,我们对不同分类器模型进行了比较. 在这里我们选择了具有较高分类能力的支持向量机(SVM)分类器[12]和 K 最近邻(KNN)分类器. 在实验中,我们使用相同的数据融合和特征提取方法,仅对分类器进行替换,从而得出哪种分类器更适用于我们的模型.

表 2 列举了 SVM 模型和 KNN 模型生成的 5 – fold CV 结果. 从表 2 中我们看到,SVM 模型在准确率、敏感率、马修斯相关系数和虚警率上分别比所提模型低了 1.67%、4.66%、3.17% 和 0.81%,但是在精确率上高了 0.81%;KNN 模型在上述参数中均低于所提模型. 图 5 和图 6 分别给出了 SVM 模型和 KNN 模型生成的 ROC 曲线,其生成的 AUC 也分别低于所提模型.

表 2 所提模型与 SVM 模型和 KNN 模型实验结果的比较

Model	Accu.(%)	Sen.(%)	Prec.(%)	MCC(%)	Fa.(%)
SVM	88.63 ±0.17	85.14 ±0.25	**91.52 ±0.29**	77.44 ±0.35	**8.48 ±0.29**
KNN	87.53 ±0.20	88.04 ±0.32	87.15 ±0.29	75.06 ±0.41	12.85 ±0.29
Our	**90.30 ±0.13**	**89.80 ±0.29**	90.71 ±0.41	**80.61 ±0.26**	9.29 ±0.41

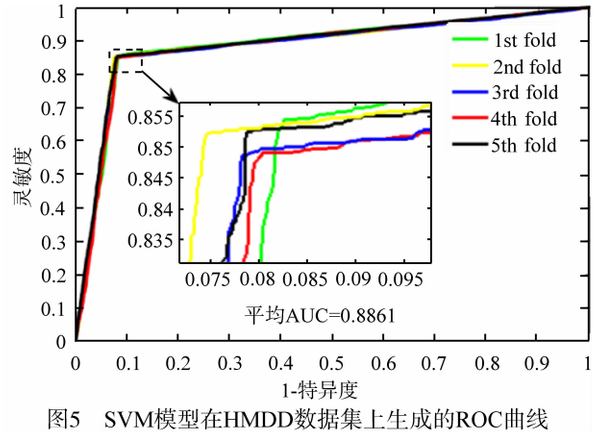

图5 SVM模型在HMDD数据集上生成的ROC曲线

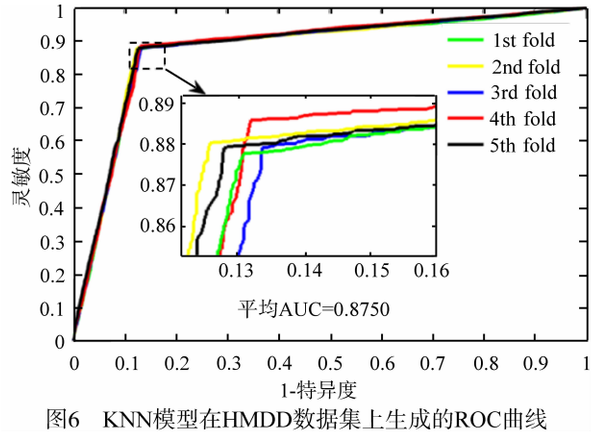

图6 KNN模型在HMDD数据集上生成的ROC曲线

3.4 不同特征提取算法之间的比较

为了检验不同特征提取算法对模型性能的影响,我们对不同的特征提取算法进行了比较. 在这里我们选择了快速傅里叶变换(FFT)和自协方差(AC)算法模型. 表 3 给出了 FFT 模型和 AC 模型生成的 5 – fold CV 结果. 从表中可以看出,FFT 模型在准确率、敏感率、马修斯相关系数和虚警率上分别比我们模型低了 1.60%、5.90%、2.85% 和 2.09%,但在精确率上高了 2.09%;AC 模型在上述参数中分别比我们模型取得的结果低了 2.09%、7.37%、3.67% 和 2.49%,但在精确率上高了 2.49%. 图 7 和图 8 分别给出了 FFT 模型和 AC 模型生成的 ROC 曲线.

3.5 与其他方法的比较

为了评估方法的有效性,我们与 BNPMDA[13]、MaxFlow[14]、MDHGI[15]、MCMDA[16] GOFS[17] 模型生成的 AUC 值进行了比较. 这些模型都使用了五折交叉验证法,并在 HMDD 数据集中实施的实验. 比较结果汇总于表 4 中,从中我们可以看到,我们的方法取得了最高的 AUC 值.

表 3　所提模型与 FFT 模型和 AC 模型实验结果的比较

Model	Accu.(%)	Sen.(%)	Prec.(%)	MCC(%)	Fa.(%)
FFT	88.70 ±0.32	83.90 ±0.58	92.80 ±0.15	77.76 ±0.59	7.20 ±0.15
AV	88.21 ±0.05	82.43 ±0.14	**93.20 ±0.06**	76.94 ±0.10	**6.80 ±0.06**
Our	**90.30 ±0.13**	**89.80 ±0.29**	90.71 ±0.41	80.**61 ±0.26**	9.29 ±0.41

表 4　所提模型与其他方法间 AUC 结果的比较

Method	Our Model	BNPMDA	MaxFlow	MDHGI	MCMDA	miRGOFS
AUC(%)	**90.26**	89.80	86.93	87.94	87.67	87.70

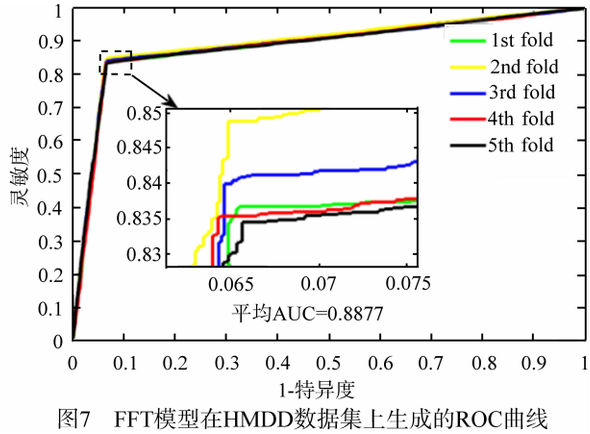

图7　FFT模型在HMDD数据集上生成的ROC曲线

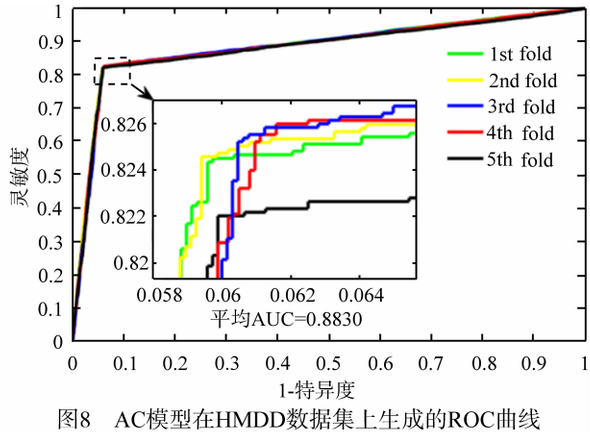

图8　AC模型在HMDD数据集上生成的ROC曲线

3.6　案例研究

为了进一步评估所提模型预测 miNRA－疾病关联的真实性能,我们针对乳腺肿瘤进行了案例研究. 在实验中我们将 HMDD v3.0 数据集中提供的所有已知的 miRNA－疾病关联作为训练集,对所有可能的 miRNA 与乳腺肿瘤之间的关系进行预测. 然后选择预测得分最高的前 30 个 miRNA 在 dbDEMC 和 miR2Disease 数据库[18]中进行验证.

乳腺肿瘤是发生在乳腺组织的肿瘤,约占乳腺疾病的三分之二. 大量实验表明许多 miRNA 与乳腺肿瘤有关,因此我们选择乳腺肿瘤作为病例研究来检验所提模型的预测疾病相关 miRNA 的能力. 结果如表 5 所示,在 dbDEMC 和 miR2Disease 数据库提供的实验数据中,前 30 个预测的疾病关联 miRNA 中的 28 个得到了验证.

研究表明,miRNA 的异常表达能够引起疾病的产生,因此基因的异常表达有可能是疾病产生的根源. 通过所提模型,我们短时间预测出了潜在的关联关系,为生物实验提供了可能性最大的候选物,大大减少了生物实验的工作量. 与生物实验相比,基于计算的模型在实验设备、试验周期及成本上都有一定的优势,从整体上更适用于现阶段的研究,并能为人类对疾病分子机制的认知提供更广阔的视野.

表 5　所提模型预测的与乳腺肿瘤相关的前 30 个 miRNA

miRNA(prediction score 1－15)	Evidence	miRNA(prediction score 16－30)	Evidence
hsa-mir-526b	dbDEMC	hsa-mir-200c	dbDEMC miR2Disease
hsa-mir-520g	dbDEMC	hsa-mir-181b-1	unconfirmed
hsa-mir-520f	dbDEMC	hsa-mir-122	dbDEMC miR2Disease
hsa-mir-520e	dbDEMC	hsa-mir-506	dbDEMC
hsa-mir-325	dbDEMC	hsa-mir-370	dbDEMC
hsa-mir-302f	dbDEMC	hsa-mir-30a	dbDEMC
hsa-mir-616	dbDEMC	hsa-mir-181a	dbDEMC miR2Disease
hsa-mir-492	dbDEMC	hsa-mir-103a-2	dbDEMC
hsa-mir-520c	dbDEMC	hsa-mir-103a-1	unconfirmed
hsa-mir-520b	dbDEMC	hsa-mir-1	dbDEMC
hsa-mir-498	dbDEMC	hsa-mir-205	dbDEMC miR2Disease
hsa-mir-340	dbDEMC	hsa-mir-10b	dbDEMC miR2Disease
hsa-mir-30c-2	dbDEMC	hsa-mir-10a	dbDEMC
hsa-mir-30c-1	dbDEMC	hsa-let-7f-1	dbDEMC miR2Disease
hsa-mir-224	dbDEMC	hsa-mir-637	dbDEMC

4 总结

在本文中,我们提出了一种基于深度学习的预测miRNA-疾病关系的模型. 该模型能够充分利用堆叠自动编码器的性能客观自动地从融合了疾病语义、miRNA 功能和 miRNA 序列的多源信息中提取其高级抽象特征,并利用旋转森林分类器有效预测出 miRNA 与疾病之间的关系. 在 HMDD v3.0 数据集上交叉验证的结果表明,该模型整体性能优异. 在与不同分类器与特征抽取算法的比较中,该模型也取得了最佳的结果. 此外,我们利用该模型还对乳腺肿瘤疾病做了案例分析,其预测结果得到了相关实验的支持. 以上结果表明,我们提出的 miRNA 与疾病关系预测模型是一个可靠模型,能够为生物实验提供高可信的疾病相关 miRNA 候选物. 在将来的研究中,我们将继续对深度学习算法进行优化,以期待能够确定更好的预测结果.

参考文献

[1] Bartel D P. MicroRNAs: genomics, biogenesis, mechanism, and function[J]. Cell, 2004, 116(2): 281 – 297.

[2] Gunter M, Thomas T. Mechanisms of gene silencing by double-stranded RNA[J]. Nature, 2004, 431(7006): 343 – 349.

[3] Bang C, Fiedler J, Thum T. Cardiovascular importance of the microRNA - 23/27/24 family[J]. Microcirculation, 2012, 19(3): 208 – 214.

[4] Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q. HMDD v2.0: a database for experimentally supported human microRNA and disease associations[J]. Nucleic Acids Research, 2013, 42(D1): D1070 – D1074.

[5] Wang L, You Z-H, Chen X, Li Y-M, Dong Y-N, Li L-P, Zheng K. LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities[J]. PLoS Computational Biology, 2019, 15(3): e1006865.

[6] Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases[J]. Bioinformatics, 2010, 26(13): 1644 – 1650.

[7] van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction[J]. Bioinformatics, 2011, 27(21): 3036 – 3043.

[8] Wang L, You Z-H, Xia S-X, Liu F, Chen X, Yan X, Zhou Y. Advancing the prediction accuracy of protein-protein interactions by utilizing evolutionary information from position-specific scoring matrix and ensemble classifier[J]. Journal of Theoretical Biology, 2017, 418: 105 – 110.

[9] Wang L, Wang H-F, Liu S-R, Yan X, Song K-J. Predicting Protein-Protein Interactions from Matrix-Based Protein Sequence Using Convolution Neural Network and Feature-Selective Rotation Forest[J]. Scientific Reports, 2019, 9(1): 9848.

[10] Wang L, Yan X, Liu M-L, Song K-J, Sun X-F, Pan W-W. Prediction of RNA-protein interactions by combining deep convolutional neural network with feature selection ensemble method[J]. Journal of Theoretical Biology, 2019, 461: 230 – 238.

[11] Wang L, You Z-H, Xia S-X, Chen X, Yan X, Zhou Y, Liu F. An improved efficient rotation forest algorithm to predict the interactions among proteins[J]. Soft Computing, 2018, 22(10): 3373 – 3381.

[12] Wang L, You Z-H, Yan X, Xia S-X, Liu F, Li L-P, Zhang W, Zhou Y. Using two-dimensional principal component analysis and rotation forest for prediction of protein-protein interactions[J]. Scientific Reports, 2018, 8(1): 12874.

[13] Chen X, Xie D, Wang L, Zhao Q, Liu H. BNPMDA: Bipartite network projection for MiRNA-Disease association prediction[J]. Bioinformatics, 2018, 34(18): 3178 – 3186.

[14] Yu H, Chen X, Lu L. Large-scale prediction of microRNA-disease associations by combinatorial prioritization algorithm[J]. Scientific Reports, 2017, 7: 43792.

[15] Chen X, Yin J, Qu J, Huang L. MDHGI: Matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction[J]. PLoS Computational Biology, 2018, 14(8): e1006418.

[16] Li J-Q, Rong Z-H, Chen X, Yan G-Y, You Z-H. MCMDA: Matrix completion for MiRNA-disease association prediction[J]. Oncotarget, 2017, 8(13): 21187.

[17] Yang Y, Fu X, Qu W, Xiao Y, Shen H-B. MiRGOFS: a GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA-disease association[J]. Bioinformatics, 2018, 34(20): 3547 – 3556.

[18] Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. miR2Disease: a manually curated database for microRNA deregulation in human disease[J]. Nucleic Acids Research, 2008, 37(suppl_1): D98 – D104.

作者简介

王 磊 男,1982 年 1 月出生,山东枣庄人. 2018 年在中国矿业大学获博士学位,现为中科院新疆理化技术研究所博士后. 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: leiwang@ms.xjb.ac.cn

徐　涛　男,1978 年 11 月出生,山东枣庄人,2009 年在华东师范大学获硕士学位,现为枣庄学院副教授. 主要研究方向计算机网络、分布式计算等.
E-mail:xutao@ uzz. edu. cn

尤著宏(通讯作者)　男,1980 年 8 月出生,甘肃兰州人,2010 年在中国科学技术大学获博士学位,现为中科院新疆理化技术研究所研究员. 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@ ms. xjb. ac. cn