面向环状RNA-疾病关联预测的物态分析优化算法

王 政[1],王 磊[1],尤著宏[2],王 磊[3*],赵博伟[4]

(1. 西安理工大学计算机科学与工程学院,陕西西安 710048;2. 西北工业大学计算机学院,陕西西安 710072;
3. 中国矿业大学计算机科学与技术学院,江苏徐州 221116;4. 浙江大学药学院,浙江杭州 310058)

摘 要: 大量研究表明,环状RNA(circRiboNucleic Acid)作为一种内源性非编码RNA,在多种人类复杂疾病的发生和发展中扮演着关键角色. 它们通过充当分子海绵、调节基因转录或与蛋白质相互作用等多种机制参与疾病相关信号通路的调控. 解析环状RNA与疾病间的关联关系,对于深入理解疾病发生机制、发现新型生物标志物以及推动精准医疗的发展具有至关重要的科学价值. 然而,传统实验方法成本高、周期长、通量有限,严重制约了环状RNA与疾病间关联关系的大规模解析. 因此,发展高效、低成本的计算方法,对推动环状RNA与疾病关联的解析研究至关重要. 本文据此提出了一种基于演化计算的预测模型ES-NMGCDA. 该模型首先构建了多种环状RNA与疾病的多源相似性网络,随后加入物态分析优化算法(State Analysis Optimization Algorithm,SAOA)对多源相似性网络进行融合与优化,最终利用因果森林分类器实现环状RNA-疾病关联关系的精准预测. ES-NMGCDA通过将物态分析优化算法的强大搜索优势与因果森林的卓越推理能力相结合,实现了对环状RNA与疾病间潜在关联的高精度、高稳健性预测. 为全面评估ES-NMGCDA模型的性能,我们在广泛使用的公共基准数据集CircR2Disease上进行了严格的5折交叉验证. 实验结果表明,本模型在测试中达到了93.80%的预测准确率,同时在精确率、敏感率等多项指标上均表现优异,显著优于多种现有基线方法. 此外,为进一步验证模型在真实生物医学场景下的实用价值,我们还开展了两项案例研究:在环状RNA与疾病间关联性的案例研究中,模型预测得分最高的前20个环状RNA-疾病关联对中,有18个获得了最新文献的支持;而在针对乳腺癌的案例研究中,模型预测出的前50个环状RNA中有43个已被证实与乳腺癌密切相关. 这些结果一致表明,ES-NMGCDA模型不仅能够为后续分子生物学实验提供高可信度的候选环状RNA分子清单,显著缩短研究周期并降低实验成本,也为深入理解环状RNA在复杂疾病中的作用机制提供了新的数据支持和理论依据.

关键词: 多源相似性网络;环状RNA-疾病;演化计算;物态分析优化算法;因果森林;潜在关联

基金项目: 国家自然科学基金(No.62172355,No.62176146,No.62325308)

中图分类号: TP399 **文献标识码:** A **文章编号:** 0372-2112(2025)09-3103-14

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250436

Optimization Algorithm for State Analysis of CircRNA-Disease Association Prediction

WANG Zheng[1], WANG Lei[1], YOU Zhu-hong[2], WANG Lei[3*], ZHAO Bo-wei[4]

(1. *School of Computer Science and Engineering*, *Xi'an University of Technology*, *Xi'an*, *Shaanxi* 710048, *China*;
2. *School of Computer Science*, *Northwestern Polytechnical University*, *Xi'an*, *Shaanxi* 710072, *China*;
3. *School of Computer Science and Technology*, *China University of Mining and Technology*, *Xuzhou*, *Jiangsu* 221116, *China*;
4. *College of Pharmaceutical Sciences*, *Zhejiang University*, *Hangzhou*, *Zhejiang* 310058, *China*)

Abstract: Extensive studies have shown that circular RNA (circRiboNucleic Acid), as a type of endogenous non-coding RNA, plays a key role in the occurrence and development of various complex human diseases. Through mechanisms such as acting as molecular sponges, regulating gene transcription, or interacting with proteins, circRNAs participate in the regulation of disease-related signaling pathways. Analyzing the associations between circRNAs and diseases is of crucial scientific value for deepening the understanding of disease mechanisms, discovering novel biomarkers, and advancing precision medicine. However, traditional experimental methods are constrained by high costs, long cycles, and limited throughput, which severely restrict large-scale analysis of circRNA-disease associations. Thus, developing efficient and low-cost

收稿日期:2025-05-29;录用日期:2025-09-24;责任编辑:覃怀银
*通讯作者:王磊

computational methods is essential for promoting research in this field. In response, this paper proposes a prediction model named ES-NMGCDA based on evolutionary computation. The model first constructs multi-source similarity networks of circRNAs and diseases, then incorporates the state analysis optimization algorithm (SAOA) to integrate and optimize these multi-source similarity networks, and finally employs a causal forest classifier to achieve accurate prediction of circRNA-disease associations. By integrating the powerful search advantage of SAOA with the superior inference capability of causal forests, ES-NMGCDA enables highly accurate and robust prediction of potential circRNA-disease associations. To comprehensively evaluate the performance of the ES-NMGCDA model, we conducted rigorous 5-fold cross-validation on the widely used public benchmark dataset CircR2Disease. Experimental results demonstrate that the model achieved a prediction accuracy of 93.80%, while also excelling in multiple metrics such as precision and sensitivity, significantly outperforming several existing baseline methods. Furthermore, to validate the model's practical utility in real biomedical scenarios, we carried out two case studies. In the case study on circRNA-disease associations, 18 out of the top 20 circRNA-disease pairs with the highest prediction scores were supported by recent literature. In the case study focused on breast cancer, 43 out of the top 50 predicted circRNAs were confirmed to be closely associated with the disease. These results consistently indicate that the ES-NMGCDA model not only provides highly reliable candidate circRNA molecules for subsequent molecular biology experiments, significantly shortening research cycles and reducing experimental costs, but also offers new data support and theoretical foundations for understanding the role of circRNAs in complex diseases.

Key words: multi-source similarity network; circRNA-disease; evolutionary computation; state analysis optimization algorithm; causal forest; potential association

Foundation Item(s): National Natural Science Foundation of China (No.62172355, No.62176146), No.62325308)

1 引言

近年来,随着非编码RNA研究的深入,环状RNA(circRiboNucleic Acid)因其独特的闭合环状结构和高度稳定性,逐渐成为生命科学领域的研究热点. 大量研究表明,环状RNA通过调控基因表达、充当分子海绵或与蛋白质相互作用等方式,广泛参与癌症、神经退行性疾病、心血管疾病等复杂人类疾病的发生与发展[1~3]. 解析环状RNA与疾病的关联机制不仅有助于揭示疾病分子机理,更可为精准医疗提供潜在的诊断标志物和治疗靶点,具有重要的科学价值与应用前景[4,5].

然而,使用传统湿实验的环状RNA-疾病关联发现方法面临显著的挑战:一方面,实验验证周期长、成本高昂,难以满足大规模关联筛选的需求[6];另一方面,疾病相关环状RNA的功能具有高度组织特异性和动态性,这使得仅凭单一实验数据难以全面解析其复杂的调控网络[7,8]. 尽管现有研究提出了多种计算模型,例如基于知识图注意力网络[9]、矩阵分解[10]等方法,试图从生物信息学角度预测环状RNA-疾病关联,然而这些方法通常受到数据稀疏性以及多源异构信息融合不足等问题的限制,从而制约了其预测精度与泛化能力[11]. 在此背景下,如何利用智能计算技术高效整合多维度生物数据,并从中挖掘出高可信的环状RNA-疾病关联,已成为当前生物信息学领域亟待解决的关键问题. 针对上述挑战,本文提出一种基于演化计算的物态分析优化算法结合因果森林分类器的预测模型,并将该模型命名为ES-NMGCDA. ES-NMGCDA模型预测的实现流程如图1所示.

首先,我们利用环状RNA-疾病关联数据集分别构建了环状RNA与疾病的多源相似性网络. 具体来说,环状RNA的多源相似性网络包括了环状RNA的功能相似性网络、环状RNA的高斯相互作用谱核相似性网络、环状RNA的非负矩阵分解相似性网络、环状RNA的图嵌入相似性网络以及环状RNA的多任务深度分解相似性网络. 同样,疾病的多源相似性网络则包括了疾病的语义相似性网络、疾病的高斯相互作用谱核相似性网络、疾病的非负矩阵分解相似性网络、疾病的图嵌入相似性网络以及疾病的多任务深度分解相似性网络. 然后,通过物态分析优化算法创新性地融合环状RNA与疾病的多源相似性网络特征,并动态优化多源相似性网络中的融合结构,以增强多源相似性网络之间关联信息的表达. 最后,引入因果森林分类器,依托因果推理原理对环状RNA与疾病间的非线性因果关系进行建模,从而有效识别复杂生物系统中潜在的交互效应. 本文所提方法的完整实现代码及相关数据已公开在GitHub平台:https://github.com/look0012/ES-NMGCDA,以便重现实验结果和后续深入研究.

2 材料及方法

2.1 环状RNA-疾病关联数据集

本研究在Fan等人[12]构建的CircR2Disease数据集的基础上开展模型验证. 基于"序列相似的环状RNA在功能上可能具有一定相似性"这一基本假设,我们首先对该数据集进行了系统清洗. 原始数据集共包含661条环状RNA与100种疾病之间的关联记录. 数据清洗

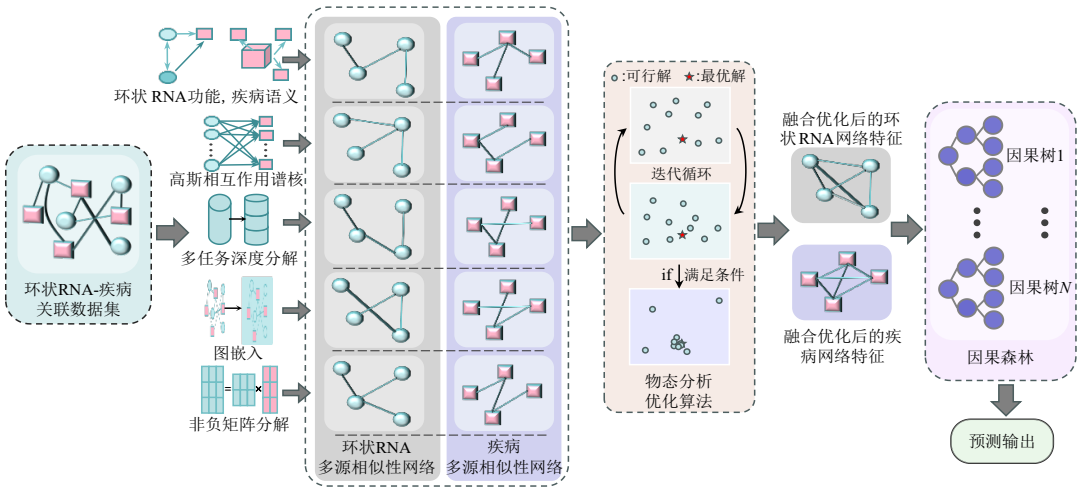


图1　环状RNA-疾病关联预测模型流程图

的具体判定标准如下：保留所有具有可靠序列信息的环状RNA-疾病关联记录，剔除缺乏对应序列信息的环状RNA条目. 清洗后，最终保留了561条环状RNA与100种疾病之间的有效关联，去噪比例为15.13%（即清洗后数据规模为原始数据集的84.87%）. 这一处理有效提升了数据集的可靠性，为后续建模提供了高质量的基础. 为确保数据的可靠性，我们还选取了已经过实验证实的607条环状RNA-疾病关联数据作为正样本集. 鉴于原始CircR2Disease数据集中未提供明确的负样本集，因此我们排除了所有已知的正样本后在剩余的561×100–607=55 493条环状RNA-疾病潜在关联数据中通过随机抽样的方法选取了607条作为负样本集，从而保证正负样本的数量平衡. 虽然这种方法有可能会将未被证实但实际存在关联的环状RNA-疾病潜在关联样本作为负样本，但该方法选出的负样本集只占所有环状RNA-疾病潜在关联样本的607÷(561×100–607)≈1.09%，所以从机器学习和概率的角度来看，这样小的假阴性误差对模型性能的影响可以忽略不计.

基于上述处理，我们构建了CircR2Disease数据集的环状RNA-疾病邻接矩阵$\boldsymbol{A}$. 该矩阵的定义如下：对于矩阵元素$\boldsymbol{A}\left(c(i),d(j)\right)$，当环状RNA$c(i)$与疾病$d(j)$存在已知关联时，$\boldsymbol{A}\left(c(i),d(j)\right)$赋值为1，反之则赋值为0. 这种邻接矩阵表示方法既完整保留了原始数据的关联信息，又为后续机器学习模型提供了规范化的数据输入.

2.2　疾病语义相似性网络构建

本研究基于美国国家医学图书馆提供的医学主题词（Medical Subject Headings，MeSH）数据库[13]构建疾病语义相似性网络. MeSH作为权威的生物医学主题分类系统，其层次化的疾病分类体系为计算疾病间的语义相似度提供了可靠依据. 参考已有研究[14]，我们利用MeSH数据库构建有向无环图（Directed Acyclic Graph，DAG）来表征疾病之间的语义关系. 对于给定疾病d，其对应的DAG定义为$\mathrm{DAG}_d=(d,N_d,E_d)$，其中$N_d$表示疾病$d$及其所有祖先疾病构成的节点集合，$E_d$表示疾病间的有向边集合. 因此，在$\mathrm{DAG}_d$中的任意疾病$s$对目标疾病$d$的语义贡献值$D_d(s)$计算如下：

$$\begin{cases} D_d(s)=1, & \text{if } s=d \\ D_d(s)=\max\left\{\vartheta\cdot D_d(s')\,|\,s'\in\text{children of }s\right\}, & \text{if } s\neq d \end{cases} \tag{1}$$

其中，ϑ为语义贡献衰减因子. 基于此，我们累加节点集合N_d中所有疾病节点的语义贡献值，即可得出疾病d的语义值$\mathrm{DV}(d)$：

$$\mathrm{DV}(d)=\sum_{s\in N_d}D_d(s) \tag{2}$$

在这里，我们基于疾病间DAG拓扑结构共享程度和疾病间语义相似性呈正相关的理论假设，给定两个疾病$d(i)$和$d(j)$，则它们之间的疾病语义相似度网络模型$D_{\mathrm{MSH}}(i,j)$构建方法如下：

$$D_{\mathrm{MSH}}\left(d(i),d(j)\right)=\frac{\sum_{s\in N_{d(i)},N_{d(j)}}\left(D_{d(i)}(s)+D_{d(j)}(s)\right)}{\mathrm{DV}\left(d(i)\right)+\mathrm{DV}\left(d(j)\right)} \tag{3}$$

2.3　环状RNA功能相似性网络构建

基于共享语义相似性疾病组的环状RNA在功能上也更具有相似性的这一假设[15]，因此，我们通过计算环状RNA关联疾病组间的语义相似性来度量其功能相似性. 具体而言，给定两个环状RNA$c(i)$和$c(j)$，它们之间的功能相似度C_{FS}可以计算如下：

$$C_{\mathrm{FS}}\left(c(i),c(j)\right)=\frac{\sum_{d\in D(i)}S\left(d(i),D(j)\right)+\sum_{d\in D(j)}S\left(d(j),D(i)\right)}{\left|D(i)\right|+\left|D(j)\right|} \tag{4}$$

$$S\left(d,D(i)\right)=\max_{1\leqslant k\leqslant D(i)}\left(D_{\mathrm{MSH}}\left(d,d(k)\right)\right) \tag{5}$$

其中，$d(i)$和$d(j)$分别表示疾病集合$D(i)$和$D(j)$中的元

素，$|D(i)|$和$|D(j)|$分别表示$D(i)$和$D(j)$的疾病数量.

2.4 高斯相互作用谱核的相似性网络构建

高斯相互作用谱核(Gaussian Interaction Profile Kernel，GIPK)是基于径向基函数(Radial Basis Function，RBF)构建的标量核函数，其核心特征是通过非线性映射捕捉数据间的潜在关联模式. 该核函数通过捕捉数据样本间的非线性关联特征，可以有效提升相似性度量的准确性[16]. Li等人[17]将GIPK核函数引入到环状RNA-疾病关联预测领域，创新性地构建了环状RNA与疾病的GIPK相似性度量模型. 本研究基于该模型，分别构建了环状RNA的GIPK相似性网络C_{GIPK}和疾病的GIPK相似性网络D_{GIPK}.

2.5 非负矩阵分解的相似性网络构建

本研究采用非负矩阵分解(Non-negative Matrix Factorization，NMF)技术[18]对环状RNA-疾病邻接矩阵$\boldsymbol{A}$进行特征提取，其数学表达式为

$$\boldsymbol{A}\approx\boldsymbol{W}\cdot\boldsymbol{H} \tag{6}$$

其中，$\boldsymbol{W}\in\mathbb{R}^{m\times k}$为环状RNA的潜在特征矩阵(行向量表示单个环状RNA的隐式特征)，$\boldsymbol{H}\in\mathbb{R}^{k\times n}$为疾病的潜在特征矩阵(列向量表示单个疾病的隐式特征)，k为潜在因子维度. 此分解通过低维子空间捕捉环状RNA-疾病关联的潜在模式，同时非负性约束$(\boldsymbol{W}\geqslant 0,\boldsymbol{H}\geqslant 0)$保障了特征的生物学可解释性.

为优化分解效果，设计目标函数以最小化原始矩阵与重构矩阵的Frobenius范数：

$$\min_{W\geqslant 0,H\geqslant 0}\left\|\boldsymbol{A}-\boldsymbol{W}\boldsymbol{H}\right\|_{\mathrm{F}}^{2} \tag{7}$$

基于迭代更新后的矩阵$\boldsymbol{W}$和$\boldsymbol{H}$，环状RNA的NMF相似性网络C_{NMF}和疾病的NMF相似性网络D_{NMF}可以分别计算如下：

$$C_{\mathrm{NMF}}\left(c(i),c(j)\right)=\frac{w_i\cdot w_j}{\left\|w_i\right\|\cdot\left\|w_j\right\|} \tag{8}$$

$$D_{\mathrm{NMF}}\left(d(i),d(j)\right)=\frac{h_i\cdot h_j}{\left\|h_i\right\|\cdot\left\|h_j\right\|} \tag{9}$$

其中，$w_i,w_j\in\mathbb{R}^k$和$h_i,h_j\in\mathbb{R}^k$为矩阵$\boldsymbol{W}$和$\boldsymbol{H}$的第i行和第j行向量.

2.6 图嵌入的相似性网络构建

本研究基于图嵌入[19]方法构建环状RNA和疾病的图嵌入相似性(Graph Embedding Similarity，GES)网络模型. 该方法通过拉普拉斯特征映射将环状RNA-疾病关联网络映射到低维空间，捕捉其拓扑结构中的潜在关联模式. 该方法具体实现流程如下：

首先，构建异构网络的联合邻接矩阵. 给定环状RNA-疾病邻接矩阵$\boldsymbol{A}$，再结合零矩阵$\boldsymbol{0}$扩展构造出包含环状RNA和疾病节点的联合邻接矩阵$\boldsymbol{A}$：

$$\boldsymbol{A}=\begin{bmatrix}0 & \boldsymbol{A}\\ \boldsymbol{A}^{\mathrm{T}} & 0\end{bmatrix} \tag{10}$$

然后，给定嵌入维度d且定义度矩阵$\boldsymbol{D}=\mathrm{diag}\left(\sum_j\boldsymbol{A}_{ij}\right)$，根据下式进而构造归一化拉普拉斯矩阵$\boldsymbol{L}$：

$$\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2} \tag{11}$$

最后，对矩阵$\boldsymbol{L}$进行特征分解. 需要特别说明的是，在此过程中所得到的第1个特征向量为常数向量，其对应的特征值为零. 由于该向量在嵌入过程中不提供任何有效的判别信息，因此予以舍弃. 随后，选取第2至第$d+1$个最小非零特征值所对应的特征向量$\boldsymbol{V}$，构成低维嵌入表示. 将该低维嵌入矩阵$\boldsymbol{E}$划分为环状RNA子矩阵$\boldsymbol{E}_c$和疾病子矩阵$\boldsymbol{E}_d$，并分别用于计算环状RNA和疾病的图嵌入相似度C_{GES}和D_{GES}，具体计算方式如下：

$$C_{\mathrm{GES}}\left(c(i),c(j)\right)=\frac{\boldsymbol{E}_{c(i)}\cdot\boldsymbol{E}_{c(j)}}{\left\|\boldsymbol{E}_{c(i)}\right\|\cdot\left\|\boldsymbol{E}_{c(j)}\right\|} \tag{12}$$

$$D_{\mathrm{GES}}\left(d(i),d(j)\right)=\frac{\boldsymbol{E}_{d(i)}\cdot\boldsymbol{E}_{d(j)}}{\left\|\boldsymbol{E}_{d(i)}\right\|\cdot\left\|\boldsymbol{E}_{d(j)}\right\|} \tag{13}$$

2.7 多任务深度矩阵分解的相似性网络构建

多任务深度矩阵分解(Multi-task Deep Matrix Factorization，MDMF)模型可以通过联合学习环状RNA和疾病的潜在特征空间，实现异构数据的高效融合. 本研究使用MDMF模型将环状RNA功能相似性$\boldsymbol{C}_{\mathrm{FS}}$和疾病语义相似性信息与环状RNA-疾病邻接矩阵$\boldsymbol{A}$信息相融合，从而构建出环状RNA和疾病的多任务深度矩阵分解相似性网络模型. MDMF模型的核心数学模型如下：

假设给定环状RNA-疾病邻接矩阵$\boldsymbol{A}$、环状RNA功能相似性矩阵$\boldsymbol{C}_{\mathrm{FS}}$及疾病语义相似性矩阵$\boldsymbol{D}_{\mathrm{MSH}}$，MDMF的目标是通过联合优化以下分解过程：

$$\boldsymbol{A}\approx\boldsymbol{U}\cdot\boldsymbol{V}^{\mathrm{T}} \tag{14}$$

其中，$\boldsymbol{U}\in\mathbb{R}^{m\times k}$为环状RNA的潜在因子矩阵，$\boldsymbol{V}^{\mathrm{T}}\in\mathbb{R}^{k\times n}$为疾病的潜在因子矩阵，$k$为潜在因子维度. 同时，引入环状RNA特征映射矩阵$\boldsymbol{W}_c{}^{\mathrm{T}}$和疾病特征映射矩阵$\boldsymbol{W}_d{}^{\mathrm{T}}$，将先验特征投影至潜在空间：

$$\boldsymbol{C}_{\mathrm{FS}}\approx\boldsymbol{U}\cdot\boldsymbol{W}_c{}^{\mathrm{T}} \tag{15}$$

$$\boldsymbol{D}_{\mathrm{MSH}}\approx\boldsymbol{V}\cdot\boldsymbol{W}_d{}^{T} \tag{16}$$

模型通过最小化以下多任务损失函数实现参数学习：

$$\min_{\boldsymbol{U}>0,\boldsymbol{V}>0}\left(\left\|\boldsymbol{A}-\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\right\|^2+\lambda\left(\left\|\boldsymbol{U}-\boldsymbol{C}_{\mathrm{FS}}\boldsymbol{W}_c\right\|^2+\left\|\boldsymbol{V}-\boldsymbol{D}_{\mathrm{MSH}}\boldsymbol{W}_d\right\|^2\right)+\gamma\left(\left\|\boldsymbol{W}_c\right\|^2+\left\|\boldsymbol{W}_d\right\|^2\right)\right) \tag{17}$$

其中，λ控制特征对齐强度，γ为L2正则化系数，用于防止过拟合. 经优化后，环状RNA和疾病的相似性矩阵

$\boldsymbol{C}_{\mathrm{MDMF}}$和$\boldsymbol{D}_{\mathrm{MDMF}}$分别计算如下：

$$\boldsymbol{C}_{\mathrm{MDMF}}=\boldsymbol{U}\cdot\boldsymbol{U}^{\mathrm{T}} \tag{18}$$

$$\boldsymbol{D}_{\mathrm{MDMF}}=\boldsymbol{V}\cdot\boldsymbol{V}^{\mathrm{T}} \tag{19}$$

2.8　多源信息网络融合

本研究通过整合环状RNA功能相似性、疾病语义相似性、环状RNA与疾病间的GIPK、NMF、GES及MDMF等多种异质相似性数据，创新性地提出了一种基于演化计算的物态分析优化算法（State Analysis Optimization Algorithm，SAOA）. 该算法以多源相似性网络作为输入，通过模拟物体在固态、液态和气态之间的相变行为，经迭代优化获取多源相似性网络的最优融合方案. 该方法有效克服了传统线性加权融合方法在特征交互建模方面的不足，能够自适应地搜索最优融合策略，从而显著提升融合效果. 最终构建的融合模型为揭示环状RNA在疾病进程中的调控机制提供了可解释的计算框架. 物态分析优化算法的具体流程如算法1所示.

算法1　物态分析优化算法

输入：种群大小, 最大迭代次数, 下界, 上界, 融合维度, 多源相似性网络数据

输出：全局最优适应度, 全局最优解(最优融合方案)

步骤1: 初始化;

步骤1.1: 生成多源相似性网络信息相关的初始种群;

步骤1.2: 计算相关种群初始适应度;

步骤1.3: 初始化参数;

步骤2: 迭代优化;

步骤2.1: 精英选择;

步骤2.2: 精英个体反向学习生成动态种群;

步骤2.3: 计算相关种群多样性;

步骤2.4: 动态调整状态概率;

步骤2.5: 根据状态概率随机选择状态(固态/液态/气态)并执行对应搜索策略;

步骤2.6: 多样性维持(每30次迭代或多样性低于阈值时,随机重置15%个体);

步骤2.7: 更新全局最优解;

步骤3: 返回结果.

SAOA是一种基于物态转变机制的元启发式优化算法，其核心思想是通过模拟固态、液态、气态三种物理状态的动态转换，实现全局探索与局部开发的平衡. 以下是其核心内容及关键数学公式：

在本研究中的固态算子部分，我们采用精英引导的高斯扰动策略进行局部开发，其算子对应的位置更新公式为

$$X_{\mathrm{new}}=X_{\mathrm{gbest}}+\sigma\cdot\mathrm{N}(0,1) \tag{20}$$

其中，σ表示扰动幅度的大小随迭代次数的增加而衰减，N(0, 1)是标准的高斯分布. 对于液态算子我们采用自适应差分进化实现平衡探索与开发，所以液态对应的位置更新公式为

$$\begin{aligned}\boldsymbol{V}&=X_{r1}+F\cdot(X_{r2}-X_{r3}),\\ X_{\mathrm{new}}&=\mathrm{crossover}(X_{\mathrm{old}},\boldsymbol{V},c)\end{aligned} \tag{21}$$

其中，$\boldsymbol{V}$表示差分向量，F表示缩放因子，其大小为$0.5+0.5\cos(\pi\cdot t/T)$，$t$和$T$则分别表示当前迭代次数和最大迭代次数，crossover表示交叉操作，c表示交叉概率. 最后气态算子采用莱昂（Levy）飞行驱动的随机探索扩大全局搜索能力，所以气态对应的位置更新公式为

$$X_{\mathrm{new}}=X_{\mathrm{old}}+s\cdot L(\beta)\cdot\left(\mathrm{rand}()-0.5\right) \tag{22}$$

其中，s表示步长缩放因子，其大小为$0.2\mathrm{e}^{-t/T}$，$L(\beta)$是Levy飞行步长.

在多源信息网络融合任务中，我们将SAOA的目标函数设计为最大化融合矩阵的信息熵：

$$H=-\sum_{i,j}P_{ij}\log_2\left(P_{ij}+\varepsilon\right) \tag{23}$$

其中，P_{ij}表示融合矩阵的元素，ε是一个极小值，它是为了避免$\log_2(0)$的情况出现，该算法通过上述数学机制，在多源信息网络融合中实现了自适应权重分配与高维特征互补，为生物医学数据分析提供了高效的优化框架.

2.9　因果森林分类器

因果森林（Causal Forest，CF）[20]是一种基于随机森林框架的机器学习算法，旨在估计异质性处理效应. 设M为样本集(X,T,Y)，其中X表示样本协变量，T表示样本处理变量，Y表示样本标签. 假设CF中因果树用C表示. 对于因果树C_i的训练过程如下：

（1）从样本集M中有放回地随机抽取样本，生成与原始样本集同大小的子集M_i.

（2）在子集M_i中，随机选取$\left\lfloor\sqrt{\dim(X)}\right\rfloor$个特征，其中$\lfloor\cdot\rfloor$表示向下取整，$\dim(X)$表示子集$M_i$的总特征数.

（3）在节点分裂时，选择特征和切分点使得左右子节点的处理效应差异最大化，目标函数为

$$\Delta=(\tilde{T}_{\mathrm{NL}}-\tilde{T}_{\mathrm{NR}}) \tag{24}$$

其中，$\tilde{T}_{\mathrm{NL}}$和$\tilde{T}_{\mathrm{NR}}$分别为左右子节点的处理效应估计值. 继续重复该步骤构建多个因果树直至满足停止条件.

（4）在叶节点中，通过线性回归计算出第i个因果树的效应估计$\tilde{T}_i(x)$. 在推断时，根据所有因果树C_i对测试样本x的处理效应估计进行聚合，其公式如下：

$$\tilde{T}(x)=\frac{1}{B}\sum_{i=1}^{B}\tilde{T}_i(x) \tag{25}$$

其中，B为决策树总数. 最终，测试样本x的处理效应由聚合结果判定，并通过得分高低来给定标签类别.

3　实验结果与分析

3.1　评估标准

为公正评估ES-NMGCDA模型的性能，我们依据通

用的评估标准[21]，以准确率（Acc）、敏感率（Sen）、精确率（Pre）、F1值（F1）和马修斯相关系数（Matthews Correlation Coefficient，MCC）五项指标对模型进行综合评估．各指标具体定义如下：

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{26}$$

$$\mathrm{Sen}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{27}$$

$$\mathrm{Pre}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{28}$$

$$\mathrm{F1}=\frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{29}$$

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{30}$$

在评估过程中，TP、TN、FP和FN分别表示分类结果中的四种情况，其分别对应为真阳性、真阴性、假阳性和假阴性样本．基于这些基本统计量，我们进一步计算了以下指标：准确率（Acc）表示所有样本中被正确预测的比例；敏感率（Sen）表示在所有真实阳性样本中正确预测为阳性的比例；精确率（Pre）表示在所有预测为阳性的样本中真正为阳性的比例；F1值为敏感度与精确率的调和平均数，能够综合反映模型在这两方面的性能；马修斯相关系数（MCC）是一种综合性指标，尤其在正负样本不平衡的场景中具有更好的平衡性与可解释性，可为模型性能提供更稳健的评估．此外，我们还绘制了受试者工作特征（Receiver Operating Characteristic，ROC）曲线并计算了ROC曲线下面积（Area Under the Curve，AUC）以进一步评估模型在不同分类阈值下的综合预测能力．

3.2 模型性能评估

本实验采用五折交叉验证对模型性能进行评估．如表1所示，ES-NMGCDA模型在CircR2Disease数据集上表现出良好的综合性能，平均准确率为93.80%，敏感率为95.04%，精确率为92.79%，F1值为93.89%，马修斯相关系数为88.39%．此外，图2绘制了模型在该数据集上的ROC曲线，其模型的AUC值达到0.867 2．上述多项评估结果一致表明，ES-NMGCDA模型在环状RNA-疾病关联预测任务中具有良好的判别能力和稳定性．

表1 ES-NMGCDA模型在CircR2Disease数据集上的实验结果

测试集	Acc/%	Sen/%	Pre/%	F1值/%	MCC/%	AUC
1	95.45	95.83	95.04	95.44	91.32	0.884 4
2	92.56	93.55	92.06	92.80	86.21	0.779 8
3	94.63	96.61	92.68	94.61	89.83	0.894 0
4	94.63	94.07	94.87	94.47	89.82	0.912 7
5	91.74	95.12	89.31	92.13	84.78	0.865 1
均值	93.80	95.04	92.79	93.89	88.39	0.867 2

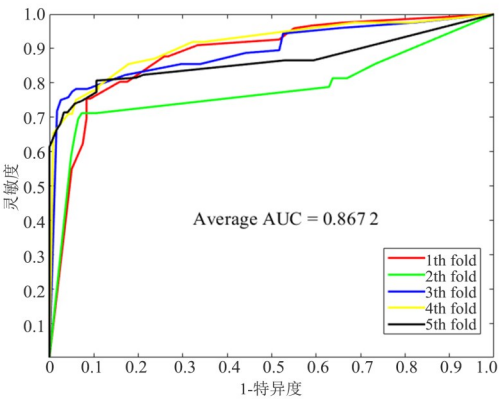


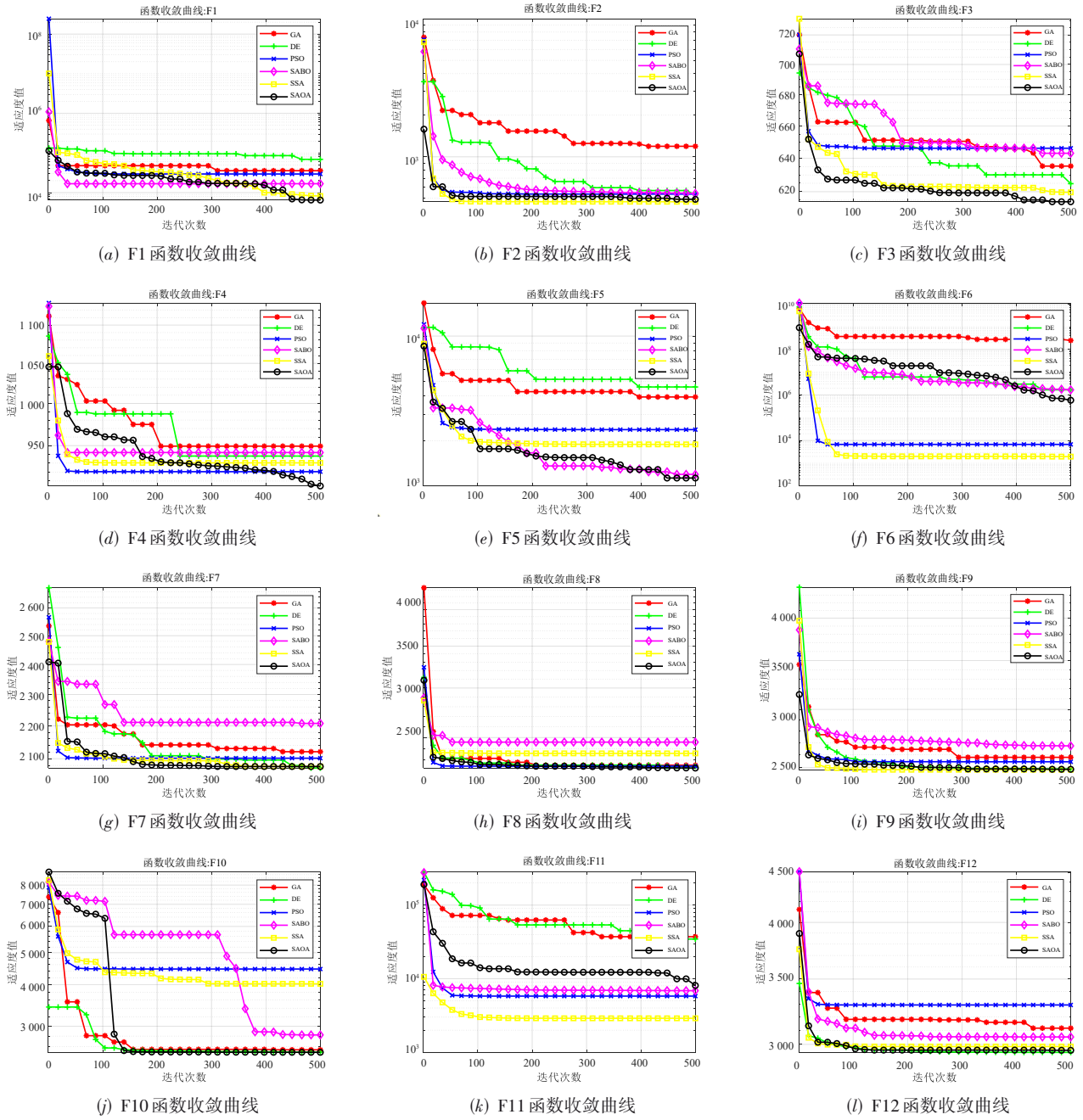
图2 模型在CircR2Disease上生成的ROC曲线

3.3 不同优化算法之间的比较

为了更进一步验证SAOA算法的优越性，本研究选取了五种具有突出代表性的优化算法作为对比对象，包括遗传算法（Genetic Algorithm，GA）[22]、差分进化算法（Differential Evolution，DE）[23]、粒子群优化算法（Particle Swarm Optimization，PSO）[24]、减法平均优化算法（Subtraction-Average-Based Optimizer，SABO）[25]、麻雀搜索算法（Sparrow Search Algorithm，SSA）[26]．在CEC2022基准测试函数上将SAOA算法与上述算法进行综合性能比较，实验结果如图3所示．由图3可知，在种群规模、维度大小和迭代次数相同的条件下，SAOA寻优性能表现较为出色，SSA次之，SABO稍逊．对于单峰函数F1，SAOA能寻到更好的最优值，且不易陷入局部最优．对于多峰函数F2~F4，SAOA寻优结果依然能保持较高收敛精度，表明其具有较强的局部开发能力．对于混合函数F5~F8，SAOA也表现出较为优异的性能，对于组合函数F9~F12，SAOA具有更强的稳定性，表明SAOA在全局搜索方面也具备一定优势．综上所述，SAOA算法与多种代表性优化算法相比，仍表现出优异的竞争力和综合性能．

3.4 不同组合预测模型之间的比较

为探究不同融合策略与不同分类器组合对模型性能的影响，我们系统地比较了多种融合策略与多种分类器组合搭配的实验效果，具体结果如表2所示．本实验选取了多种具有代表性的机器学习分类器进行对比，包括支持向量机（Support Vector Machine，SVM）[27]、*K*最近邻（*K*-Nearest Neighbors，KNN）[28]、反向传播（Back Propagation，BP）神经网络[29]和随机森林（Random Forest，RF）[30]．在表2中，演化策略（Evolution Strategy，ES）代表采用基于演化算法的优化融合策略，而集成策略（Integration Strategy，IS）表示仅使用传统集成融合策略．从表中ES与不同分类器的组合实验结果可以看出，因果森林（CF）分类器的预测性能显著优于其他分类器．为比较不同融合方案的效果，我们进一步将基于



(*a*) F1 函数收敛曲线　(*b*) F2 函数收敛曲线　(*c*) F3 函数收敛曲线

(*d*) F4 函数收敛曲线　(*e*) F5 函数收敛曲线　(*f*) F6 函数收敛曲线

(*g*) F7 函数收敛曲线　(*h*) F8 函数收敛曲线　(*i*) F9 函数收敛曲线

(*j*) F10 函数收敛曲线　(*k*) F11 函数收敛曲线　(*l*) F12 函数收敛曲线

图3　六种优化算法在不同测试函数上的收敛曲线对比

演化计算的ES策略与环状RNA-疾病关联预测中常用的IS策略进行了对比. ES+CF与IS+CF的对比实验结果表明:在ES策略所构建的9种非同源相似性融合方案中,有8种方案(N-M-G-G、N-M-Ges、N-M-Gipk、N-G-G、N-M、NMF、MDMF和GES)的多项性能评价指标均显著优于IS策略下的对应方案. 这一结果充分表明,基于演化算法的融合策略ES在多源相似性网络融合方面明显优于传统融合策略IS. 另外,一个值得关注的发现是:在融合策略IS中,其最高准确率89.17%是由NMF单组合融合方案获得,而该方案在融合策略ES中的准确率进一步提升至90.74%. 同时,融合策略ES能达到的全局最高准确率93.80%也显著优于IS策略的峰值表现,分析IS策略下的具体表现可以发现,NMF、GIPK和GES等单组合方案的准确率普遍优于其他复杂组合. 这表明,传统的IS策略只适用于单组合方式的融合方案,而ES策略不仅兼容单组合方式的融合方案,更能有效挖掘多种复杂组合方式的潜力,展现出更优的适应性与性能上限. 最后,为探究各相似性分量在ES策略中发挥的

表2 不同组合模型在CircR2Disease数据集上进行5折交叉验证后的平均实验结果

组合模型	评价指标	N-M-G-G	N-M-Ges(所用组合)	N-M-Gipk	N-G-G	N-M	NMF	MDMF	GIPK	GES
ES+CF模型(所提模型)	Acc/%	76.86	**93.80**	77.69	75.29	91.32	90.74	80.33	77.19	92.64
	Sen/%	92.54	95.04	82.43	85.13	98.15	**98.16**	98.03	92.81	97.72
	Pre/%	70.39	**92.79**	75.38	71.09	86.30	85.46	72.44	70.69	88.58
	F1值/%	79.93	**93.89**	78.63	77.43	91.83	91.36	83.29	80.13	92.92
	MCC/%	62.66	**88.39**	65.18	61.82	84.02	83.00	66.40	63.03	86.25
	AUC	0.822 4	0.867 2	0.821 0	0.796 1	**0.868 7**	0.861 8	0.829 4	0.801 5	0.858 8
IS+CF模型	Acc/%	69.34	66.03	65.62	67.85	66.45	89.17	67.11	83.39	81.16
	Sen/%	60.88	62.74	59.22	63.44	62.68	97.20	57.33	91.42	92.12
	Pre/%	73.31	67.56	67.86	69.55	68.48	83.73	71.52	78.89	75.51
	F1值/%	66.37	64.59	63.17	66.30	64.87	89.95	63.62	84.62	82.96
	MCC/%	56.78	54.78	54.50	56.10	54.96	80.46	55.12	71.97	68.60
	AUC	0.626 3	0.600 5	0.644 6	0.636 8	0.621 5	0.853 1	0.617 8	0.857 5	0.855 3
ES+BP模型	Acc/%	78.35	88.26	72.73	74.13	80.25	75.95	75.70	78.18	78.76
	Sen/%	82.58	95.19	57.40	69.05	76.63	64.34	71.18	73.47	79.49
	Pre/%	76.14	83.62	82.69	77.97	83.11	85.95	81.92	81.10	79.19
	F1值/%	79.18	89.00	62.82	71.88	75.94	69.37	70.82	76.47	78.59
	MCC/%	66.94	79.10	55.45	61.12	68.24	62.43	61.64	65.51	67.07
	AUC	0.782 8	0.882 8	0.719 3	0.741 4	0.807 4	0.769 8	0.764 8	0.780 1	0.788 0
ES+KNN模型	Acc/%	59.01	69.42	60.83	58.76	69.17	68.93	69.01	60.33	67.60
	Sen/%	50.56	73.30	55.09	55.22	74.34	71.39	83.23	55.74	72.73
	Pre/%	60.99	67.89	62.28	59.67	67.44	68.11	65.18	61.56	66.41
	F1值/%	55.24	70.43	58.28	56.93	70.46	69.57	72.69	58.46	69.13
	MCC/%	51.06	57.30	51.81	50.93	56.95	57.13	55.29	52.11	55.93
	AUC	0.675 6	0.767 7	0.687 8	0.647 4	0.777 7	0.786 2	0.771 1	0.685 7	0.781 9
ES+RF模型	Acc/%	85.12	93.06	86.12	86.86	92.73	92.15	89.42	85.21	78.76
	Sen/%	92.83	95.91	92.58	93.19	91.87	94.51	92.96	92.32	79.49
	Pre/%	80.40	90.76	82.13	82.76	93.46	90.27	86.73	80.76	79.19
	F1值/%	86.10	93.25	86.95	87.63	92.64	92.32	89.72	86.09	78.59
	MCC/%	74.39	87.05	75.82	76.95	86.52	85.53	81.03	74.48	67.07
	AUC	0.937 4	0.983 0	0.938 2	0.933 6	0.974 2	0.976 6	0.960 3	0.946 0	0.788 0
ES+SVM模型	Acc/%	92.40	92.98	91.98	92.31	92.64	92.48	92.31	92.15	91.98
	Sen/%	84.81	86.00	83.96	84.42	85.26	85.12	85.30	84.31	83.99
	Pre/%	100.00	100.00	100.00	100.00	100.00	99.80	99.20	100.00	100.00
	F1值/%	91.74	92.44	91.27	91.50	92.01	91.84	91.70	91.45	91.28
	MCC/%	85.82	86.83	85.06	85.62	86.24	85.92	85.70	85.37	85.09
	AUC	0.975 8	0.984 0	0.968 1	0.974 1	0.985 1	0.983 9	0.972 4	0.972 3	0.982 0

注：N-M-G-G 、N-M-Ges 、N-M-Gipk、N-G-G、N-M分别表示NMF-MDMF-GIPK-GES、NMF-MDMF-GES、NMF-MDMF-GIPK、NMF-GIPK-GES、NMF-MDMF的相似性组合.加粗数字表示在同一行(即同一评价指标)中的最高值.

作用，我们还做了相关的多源相似性网络组合消融实验. 实验结果表明，在ES融合策略框架下，采用N-M-Ges的组合方式对不同分类器的预测效果都有较大提升，分析其原因是ES融合策略下的N-M-Ges组合方式能够最大程度提升特征互补并减少特征冗余. 因此，本文中ES-NMGCDA模型就是采用ES策略下的N-M-Ges组合方式并结合CF分类器实现环状RNA-疾病关联预测. 此外，我们还对该模型的运行效率进行了系统评估. 实验在以下环境中进行：Windows 10操作系统，32 GB内存，Intel® Core™ i7-10750H @ 2.60 GHz处理器，MATLAB R2022b仿真软件. 评估结果表明，该模型完成全部预测任务的总耗时为58.47 s，内存占用为13.69 MB.

3.5　模型泛化能力评估

尽管所提出的模型在CircR2Disease数据集上已展现出优异的预测性能,为进一步验证其泛化能力与普适性,本研究选取了三个独立数据集:CircAtlas v2.0[31]、Circ2Disease[32]和CircRNADisease[33]进行系统性验证.本次实验严格采用相同的5折交叉验证流程及评价指标对模型的预测结果进行验证.如表3所示,模型在三个独立数据集上的平均准确率分别为89.81%、87.52%和86.45%,从实验结果的整体来看,该模型能够表现出稳定的分类性能.由此可见,ES-NMGCDA模型不仅能够有效拓展至未知环状RNA-疾病关联的大规模挖掘,还可为新型疾病相关环状RNA的识别提供高置信度的候选分子清单,从而显著降低后续湿实验验证的成本与工作量.

表3　ES-NMGCDA模型在CircAtlas v2.0、Circ2Disease和CircRNADisease独立数据集上的实验结果

数据集名称	测试集	准确率/%	敏感率/%	精确率/%	F1值/%	MCC/%	AUC
CircAtlas v2.0	1	92.58	100.00	87.29	93.22	86.07	0.861 1
	2	89.03	95.83	85.64	90.45	79.96	0.911 1
	3	87.10	99.31	78.69	87.80	77.11	0.823 6
	4	89.03	98.10	83.33	90.12	80.08	0.878 3
	5	91.29	98.63	85.21	91.43	84.01	0.846 8
	均值	89.81	98.38	84.03	90.60	81.45	0.864 2
Circ2Disease	1	89.11	89.47	91.07	90.27	80.31	0.787 3
	2	86.14	86.27	86.27	86.27	76.12	0.783 7
	3	86.14	88.68	85.45	87.04	76.00	0.843 9
	4	90.10	100.00	82.46	90.38	81.97	0.863 7
	5	86.14	100.00	75.86	86.27	75.65	0.768 3
	均值	87.52	92.89	84.22	88.05	78.01	0.809 4
CircRNADisease	1	85.95	89.23	85.29	87.22	75.59	0.853 8
	2	89.26	90.91	89.55	90.23	80.63	0.814 0
	3	86.78	86.27	83.02	84.62	76.58	0.810 8
	4	84.30	89.83	80.30	84.80	73.41	0.873 0
	5	85.95	91.67	82.09	86.61	75.71	0.842 6
	均值	86.45	89.58	84.05	86.69	76.38	0.838 9

3.6　与其他方法的比较

为全面评估ES-NMGCDA模型在环状RNA-疾病关联预测任务中的性能表现,本研究选取了9种具有代表性的主流方法进行对比,包括MSMCDA[34]、MGRCDA[35]、NMFCDA[10]、KFDAE[36]、iCDA-CGR[37]、MNMDCDA[17]、GNNCDA[38]、iGRLCDA[39]、GGCDA[40].这些方法均发表于权威期刊,并分别基于多视角共享单元结合多通道注意机制、元推荐理论、非负矩阵分解、核融合结合深度自动编码器、混沌博弈表示、高阶图卷积网络结合深度神经网络、图神经网络、图表示学习与图形注意网络结合图卷积网络等不同理论构建,覆盖了当前关联预测中的主要技术路线,具有良好的领域代表性和对比全面性.实验在CircR2Disease和Circ2Disease两个数据集上进行.CircR2Disease作为广泛采用的基准数据集,有助于实现与已有研究的公平对比;而Circ2Disease数据集规模适中、标注规范,常被用于验证模型在跨数据场景下的泛化能力,其引入进一步增强了性能评估的稳健性与说服力.评价方面,我们采用准确率(Acc)、敏感率(Sen)、精确率(Pre)、F1值(F1)、MCC和AUC值共六项指标,以确保从分类性能、类别平衡性和排名能力等多个维度进行综合评估.表4列出了各对比方法在5折交叉验证下的平均结果.从实验结果来看,在CircR2Disease数据集上,ES-NMGCDA模型在准确率(93.80%)、F1值(93.89%)和MCC(88.39%)三项关键指标上均优于所有对比方法.在Circ2Disease数据集上,本模型在精确率与F1值上排名第一,在其余指标上也均处于前列.值得注意的是,因Circ2Disease数据集样本量有限,所有方法在该数据集上均存在一定程度的欠拟合,导致性能普遍低于CircR2Disease数据集上的实验结果.另外,从AUC指标来看,ES-NMGCDA模型的AUC值虽普遍略低于其他方法但仍处于可比区间,我们分析这可能与AUC指标对样本量和数据分布高度敏感的特性有关——在有限样本中,ES-NMGCDA模型对正负样本的排序能力更容易受到统计波动的影响.尽管如此,ES-NMGCDA模型在多项指标中依然表现出优异的稳定性和综合竞争力.

表4 ES-NMGCDA模型与其他方法在CircR2Disease和Circ2Disease数据集上的对比结果

数据集名称	模型名称	准确率/%	敏感率/%	精确率/%	F1值/%	MCC/%	AUC
CircR2Disease	**ES-NMGCDA**	**93.80**	95.04	92.79	**93.89**	**88.39**	0.867 2
	MSMCDA	92.30	NA	89.70	92.50	79.96	**0.976 0**
	MGRCDA	92.49	91.69	NA	92.38	84.94	0.929 8
	NMFCDA	92.56	**96.65**	89.41	92.83	85.51	0.927 8
	KFDAE	91.23	NA	**93.48**	91.23	NA	0.973 8
	iCDA-CGR	81.95	88.08	78.46	82.97	NA	0.804 9
	MNMDCDA	88.69	94.07	85.00	89.28	77.87	0.951 6
	GNNCDA	87.79	92.00	84.87	88.28	75.89	0.937 5
	GGCDA	90.63	NA	85.11	91.34	NA	0.972 6
Circ2Disease	**ES-NMGCDA**	89.11	89.47	**91.07**	**90.27**	80.31	0.787 3
	MSMCDA	88.30	NA	81.80	87.50	79.96	0.933 0
	MGRCDA	**90.66**	81.30	NA	89.66	**82.75**	0.911 6
	NMFCDA	89.03	**98.10**	83.33	90.12	80.08	0.878 3
	iCDA-CGR	71.30	66.30	73.80	69.74	42.92	0.780 9
	MNMDCDA	81.30	85.19	79.24	81.95	63.07	0.898 0
	GNNCDA	87.04	90.37	84.82	87.43	74.38	**0.946 0**
	iGRLCDA	71.30	66.30	73.80	69.74	42.92	0.780 9

注:加粗数字表示在数据集CircR2Disease和Circ2Disease内处于同一列(即同一评价指标)中的最高值.

3.7 案例研究

3.7.1 环状RNA与疾病间关联性的案例研究

为进一步评估ES-NMGCDA模型对未知环状RNA-疾病关联的预测能力,本研究开展了环状RNA-疾病关联的案例分析. 首先,基于CircR2Disease数据集对模型进行训练,随后模型输出关联置信度最高的前20个环状RNA-疾病对. 通过系统检索近期发表文献及相关数据库后对这些预测结果进行验证,其实验结果如表5所示. 该实验结果表明,在ES-NMGCDA模型预测的前20个环状RNA-疾病对中有18个均得到已有文献或数据

表5 ES-NMGCDA模型在CircR2Disease数据集上预测的前20个环状RNA–疾病关联

No.	环状RNA名称	疾病名称	PMID
1	circRNA_002581	Nonalcoholic steatohepatitis	27677588
2	hsa_circ_0001212	Cervical carcinoma	28080204
3	hsa_circRNA_001379	Papillary thyroid carcinoma	28288173
4	hsa_circRNA_103410	Active pulmonary tuberculosis	未证实
5	hsa_circRNA_103454	Papillary thyroid carcinoma	28288173
6	hsa_circRNA_100777	Papillary thyroid carcinoma	28288173
7	hsa_circ_0057093	Diabetes retinopathy	29288268
8	hsa_circ_0061893	Radiation-induced liver fibrosis	28774651
9	circRNA7535	Esophageal squamous cell carcinoma	29218114
10	hsa_circ_0005870	Hypertension	28534714
11	circRNA120	Esophageal squamous cell carcinoma	29218114
12	hsa_circ_0018168	Acne	29573483
13	circRNA2918	Esophageal squamous cell carcinoma	29218114
14	circRNA-000284	Cervical cancer	29511454
15	hsa_circ_101222	Rheumatoid arthritis	未证实
16	hsa_circ_0000673	Ovarian endometriosis	29334789
17	hsa_circ_0013255	Radiation-induced liver fibrosis	28774651
18	hsa_circ_0000615	Hypertension	28824721
19	hsa_circ_0018289	Cervical cancer	29156822
20	hsa_circ_0026372	Diabetes retinopathy	29288268

库的实验支持,确认其关联性. 该结果证明,ES-NMGCDA模型在识别新型环状RNA-疾病关联方面具有较高的准确性和可靠性,具备指导进一步临床前研究的应用价值.

3.7.2　环状RNA-乳腺癌疾病的关系案例研究

疾病预测是生物医学领域的核心挑战,而利用计算模型从海量数据中挖掘生物标志物是关键途径[41]. 基于此,本文还选取乳腺癌作为案例,对ES-NMGCDA模型的预测性能进行专项评估,以验证其在发现环状RNA标志物方面的能力. 我们首先基于CircR2Disease数据集训练模型,随后使用该模型进行预测,并最终筛选出与乳腺癌潜在关联性最高的前50个环状RNA. 经文献与数据库检索验证,预测结果中排名前25的环状RNA已全部获得文献或数据库支持,前50个环状RNA中有43个被证实与乳腺癌存在关联(详见表6). 该结果表明,ES-NMGCDA模型能够高效、准确地识别潜在的疾病相关环状RNA,可作为指导后续生物学实验研究的可靠工具.

表6　ES-NMGCDA模型预测出的前50个与乳腺癌有关联的环状RNA

环状RNA名称(1~25)	PMID	环状RNA名称(26~50)	PMID
hsa_circ_0001667	28803498	hsa_circ_0001721	28744405
hsa_circ_104821	28484086	hsa_circ_0086241	28803498
hsa_circ_104689	28484086	hsa_circ_0002113	28803498
circRNA-001283	29431182	circ-Foxo3	27886165
circBCL11B	29221160	hsa_circ_0003838	28803498
circMED13	29221160	hsa_circ_0085495	28803498
circDENND4C	28739726	hsa_circ_0000098	28744405
hsa_circ_0004214	28622299	circVRK1	29221160
hsa_circ_0093859	29593432	hsa_circRNA_101308	未证实
circRNA-000911	29431182	hsa_circ_0091822	未证实
circBRIP	29221160	hsa_circ_0000732	28744405
circOLA	29221160	hsa_circ_0006054	28484086
hsa_circ_0000981	28744405	hsa_circ_0000911	28744405
circRNA-001175	29431182	hsa_circ_0068033	29045858
hsa_circ_0008945	28744405	hsa_circ_0001283	28744405
hsa_circ_0008717	28744405	hsa_circ_0011946	29593432
hsa_circ_0001821	27928058	hsa_circ_103110	28484086
hsa_circ_0000893	28744405	circETFA	29221160
hsa_circ_0001982	28933584	hsa_circ_0007534	29593432
hsa_circ_0108942	29045858	hsa_circ_0092276	28803498
hsa_circ_0018293	28744405	hsa_circ_0092509	未证实
hsa_circ_100219	28484086	circRNA9671	未证实
hsa_circ_0001785	29045858	hsa_circ_0000172	未证实
hsa_circ_0002874	28803498	hsa_circ_0098964	未证实
hsa_circ_0006528	28803498	mmu_circRNA_010567	未证实

4　结论

本研究提出了一种基于演化计算的环状RNA-疾病关联预测模型ES-NMGCDA. 该模型通过物态分析优化算法,实现了多源相似性网络的深度融合与优化,整合的特征包括环状RNA功能相似性、疾病语义相似性,以及环状RNA与疾病之间的GIPK、NMF、GES和MDMF等多种相似性度量. 基于优化后的融合特征,进一步采用因果森林分类器实现对环状RNA-疾病关联的有效预测.

在CircR2Disease基准数据集上进行的5折交叉验证表明,ES-NMGCDA模型在准确率与F1值等核心指标上均显著优于现有方法. 优化算法对比实验及不同预测模型的比较实验进一步验证了本文所提出算法在寻优效果和整体预测性能上的显著优势. 此外,在泛化能力验证中,ES-NMGCDA模型同样表现出最优的预测性能. 在案例研究方面,ES-NMGCDA模型在未知环状RNA-疾病关联预测中展现出良好的应用潜力. 结果显示,在模型预测排名前20的环状RNA-疾病对中,有18对已得到已有实验数据的支持;而在乳腺癌专项病例研究中,预测前25位的环状RNA均经验证与乳腺癌存在关联,前50个环状RNA候选物中有43个环状RNA也被证实的确与乳腺癌存在关联. 这些实证结果一致表明,ES-NMGCDA模型不仅具有出色的预测准确性,还能够为环状RNA相关的生物医学研究提供高可信度的候选分子库.

尽管该模型展现出良好的预测性能,但仍存在一定的局限性. 多源信息融合过程中对高维相似性矩阵的整合与优化显著增加了计算与存储开销,导致模型收敛效率降低、训练时间延长,这在一定程度上限制了其在大规模应用中的实用性与可扩展性. 在后续研究中,我们将重点优化模型结构与训练机制,通过引入稀疏表征学习、迭代剪枝或增量融合等策略,旨在有效降低多源融合模块的计算复杂度和运行开销,并在保证预测精度的同时显著提升计算效率,从而增强该模型在真实生物医学研究场景中的适用性.

参考文献

[1] 范吉林, 朱婷婷, 田晓玲, 等. 非编码RNA调节心肌缺血再灌注损伤中自噬的作用及机制[J]. 中国组织工程研究, 2022, 26(35): 5716-5723.
FAN J L, ZHU T T, TIAN X L, et al. Effect and mechanism of non-coding RNA regulating autophagy in myocardial ischemia-reperfusion injury[J]. Chinese Journal of Tissue Engineering Research, 2022, 26(35): 5716-5723. (in Chinese)

[2] JIANG C, WANG L, YU C Q, et al. MuGNet-CMI: Multi-head hybrid graph neural network for predicting circRNA-miRNA

interactions with global high-order and local low-order information[J/OL]. IEEE Transactions on Big Data. (2025-08-29) [2025-09-24]. https://ieeexplore.ieee.org/document/11145166.

[3] WEI M M, WANG L, ZHAO B W, et al. Integrating transformer and graph attention network for circRNA-miRNA interaction prediction[J]. IEEE Journal of Biomedical and Health Informatics, 2025, 29(8): 6105-6113.

[4] 高轶昳, 刘菲, 黄卫, 等. 环状RNA与神经、肌肉功能及相关疾病的研究进展[J]. 第二军医大学学报, 2021, 42(3): 301-307.
GAO Y Y, LIU F, HUANG W, et al. Circular RNA and nerve, muscle functions and related diseases: Research progress[J]. Academic Journal of Second Military Medical University, 2021, 42(3): 301-307. (in Chinese)

[5] 雷秀娟, 张文祥, 刘恋. 基于多数据融合的circRNA-疾病关联关系预测[J]. 中国科学: 信息科学, 2021, 51(6): 927-939.
LEI X J, ZHANG W X, LIU L. Prediction of circRNA- disease association based on multi-data fusion[J]. Scientia Sinica (Informationis), 2021, 51(6): 927-939. (in Chinese)

[6] WEI M M, WANG L, SU X R, et al. Multi-hop graph structural modeling for cancer-related circRNA-miRNA interaction prediction[J]. Pattern Recognition, 2026, 170: 112078.

[7] 黄钧鸿, 黄巧娟, 李斌, 等. 基于高通量测序的RNA信息解析技术[J]. 生命科学, 2021, 33(3): 267-280.
HUANG J H, HUANG Q J, LI B, et al. Bioinformatic methods for analyzing noncoding RNAs from high-throughput sequencing data[J]. Chinese Bulletin of Life Sciences, 2021, 33(3): 267-280. (in Chinese)

[8] YU C Q, JIANG C, WANG L, et al. iHofman: A predictive model integrating high-order and low-order features with weighted attention mechanisms for circRNA-miRNA interactions[J]. BMC Biology, 2025, 23(1): 162.

[9] LAN W, DONG Y, CHEN Q F, et al. KGANCDA: Predicting circRNA-disease associations based on knowledge graph attention network[J]. Briefings in Bioinformatics, 2022, 23(1): bbab494.

[10] WANG L, YOU Z H, ZHOU X, et al. NMFCDA: Combining randomization-based neural network with non-negative matrix factorization for predicting CircRNA-disease association[J]. Applied Soft Computing, 2021, 110: 107629.

[11] WANG L, LI S L, SU X R, et al. Prediction of Budd-Chiari syndrome based on attention mechanisms of high-risk factors in multi-hop graph learning[J]. Science China Information Sciences, 2025, 68(7): 179102.

[12] FAN C Y, LEI X J, FANG Z Q, et al. CircR2Disease: A manually curated database for experimentally supported circular RNAs associated with various diseases[J]. Database, 2018, 2018: bay044.

[13] WANG L, YOU Z H, CHEN X, et al. 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.

[14] 王磊, 徐涛, 宋传东, 等. 基于深度学习的miRNA与疾病相关性预测算法[J]. 电子学报, 2020, 48(5): 870-877.
WANG L, XU T, SONG C D, et al. Prediction algorithm of association between miRNAs and diseases based on deep learning[J]. Acta Electronica Sinica, 2020, 48(5): 870-877. (in Chinese)

[15] 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.

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

[17] LI Y, HU X G, WANG L, et al. MNMDCDA: Prediction of circRNA-disease associations by learning mixed neighborhood information from multiple distances[J]. Briefings in Bioinformatics, 2022, 23(6): bbac479.

[18] 苗启广, 王宝树. 图像融合的非负矩阵分解算法[J]. 计算机辅助设计与图形学学报, 2005, 17(9): 2029-2032.
MIAO Q G, WANG B S. A novel algorithm of multi-sensor image fusion using non-negative matrix factorization[J]. Journal of Computer Aided Design & Computer Graphics, 2005, 17(9): 2029-2032. (in Chinese)

[19] 陈劲松, 孟祥武, 纪威宇, 等. 基于多维上下文感知图嵌入模型的兴趣点推荐[J]. 软件学报, 2020, 31(12): 3700-3715.
CHEN J S, MENG X W, JI W Y, et al. POI recommendation based on multidimensional context-aware graph embedding model[J]. Journal of Software, 2020, 31(12): 3700-3715. (in Chinese)

[20] ATHEY S, WAGER S. Estimating treatment effects with causal forests: An application[J]. Observational Studies, 2019, 5(2): 37-51.

[21] LI Y, LIU X Z, YOU Z H, et al. A computational approach for predicting drug-target interactions from protein sequence and drug substructure fingerprint information[J]. International Journal of Intelligent Systems, 2021, 36(1): 593-609.

[22] REEVES C R, ROWE J E. No free lunch for GAs[M]// Genetic Algorithms—Principles and Perspectives. Boston: Springer , 2002: 95-109.

[23] CUEVAS E, ZALDIVAR D, PÉREZ-CISNEROS M. A novel multi-threshold segmentation approach based on

differential evolution optimization[J]. Expert Systems with Applications, 2010, 37(7): 5265-5271.
[24] MARINI F, WALCZAK B. Particle swarm optimization (PSO). A tutorial[J]. Chemometrics and Intelligent Laboratory Systems, 2015, 149: 153-165.
[25] TROJOVSKÝ P, DEHGHANI M. Subtraction-average-based optimizer: A new swarm-inspired metaheuristic algorithm for solving optimization problems[J]. Biomimetics, 2023, 8(2): 149.
[26] XUE J K, SHEN B. A novel swarm intelligence optimization approach: Sparrow search algorithm[J]. Systems Science & Control Engineering, 2020, 8(1): 22-34.
[27] CHANDRA M A, BEDI S S. Survey on SVM and their application in imageclassification[J]. International Journal of Information Technology, 2021, 13(5): 1-11.
[28] GUO G D, WANG H, BELL D, et al. KNN model-based approach in classification[C]//On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Berlin: Springer, 2003: 986-996.
[29] YU S W, ZHU K J, DIAO F Q. A dynamic all parameters adaptive BP neural networks model and its application on oil reservoir prediction[J]. Applied Mathematics and Computation, 2008, 195(1): 66-75.
[30] RESENDE P A A, DRUMMOND A C. A survey of random forest based methods for intrusion detection systems[J]. ACM Computing Surveys, 2018, 51(3): 1-36.
[31] WU W Y, JI P F, ZHAO F Q. CircAtlas: An integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes[J]. Genome Biology, 2020, 21(1): 101.
[32] YAO D X, ZHANG L, ZHENG M Y, et al. Circ2Disease: A manually curated database of experimentally validated circRNAs in human disease[J]. Scientific Reports, 2018, 8(1): 11018.
[33] ZHAO Z, WANG K Y, WU F, et al. circRNA disease: A manually curated database of experimentally supported circRNA-disease associations[J]. Cell Death & Disease, 2018, 9(5): 475.
[34] ZHANG X, ZOU Q, NIU M T, et al. Predicting circRNA-disease associations with shared units and multi-channel attention mechanisms[J]. Bioinformatics, 2025, 41(3): btaf088.
[35] WANG L, YOU Z H, HUANG D S, et al. MGRCDA: Metagraph recommendation method for predicting CircRNA-disease association[J]. IEEE Transactions on Cybernetics, 2023, 53(1): 67-75.
[36] KANG W Y, GAO Y L, WANG Y, et al. KFDAE: CircRNA-disease associations prediction based on kernel fusion and deep auto-encoder[J]. IEEE Journal of Biomedical and Health Informatics, 2024, 28(5): 3178-3185.
[37] ZHENG K, YOU Z H, LI J Q, et al. iCDA-CGR: Identification of circRNA-disease associations based on Chaos Game Representation[J]. PLoS Computational Biology, 2020, 16(5): e1007872.
[38] 李扬, 胡学钢, 王磊, 等. 基于图神经网络的环状RNA生物标志物筛选预测算法[J]. 中国科学: 信息科学, 2023, 53(11): 2214-2229.
LI Y, HU X G, WANG L, et al. Prediction algorithm for screening circRNA biomarker based on graph neural network[J]. Scientia Sinica (Informationis), 2023, 53(11): 2214-2229. (in Chinese)
[39] ZHANG H Y, WANG L, YOU Z H, et al. iGRLCDA: Identifying circRNA-disease association based on graph representation learning[J]. Briefings in Bioinformatics, 2022, 23(3): bbac083.
[40] CAO R F, HE C, WEI P J, et al. Prediction of circRNA-disease associations based on the combination of multi-head graph attention network and graph convolutional network[J]. Biomolecules, 2022, 12(7): 932.
[41] 漆华妹, 胡宇轩, 袁正一. 一种基于降噪自动编码器和宽度学习的增量式疾病预测模型[J]. 电子学报, 2023, 51(6): 1474-1485.
QI H M, HU Y X, YUAN Z Y. An incremental disease prediction model based on denoising autoencoder with broad learning system[J]. Acta Electronica Sinica, 2023, 51(6): 1474-1485. (in Chinese)

作者简介

王　政　男,1991年1月出生于陕西省临潼市. 现为西安理工大学计算机科学与工程学院博士研究生. 主要研究方向为演化计算、数据挖掘和机器学习.
E-mail: xywangzheng0971@163. com

王　磊　男,1972年5月出生于陕西省白水市. 现为西安理工大学计算机科学与工程学院教授、博士生导师. 主要研究方向为进化算法、神经网络与数据挖掘. 中国电子学会会员编号: E1900027585.
E-mail: leiwang@xaut.edu.cn

尤著宏 男,1980年8月出生于甘肃省兰州市. 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、数据挖掘及其在生物信息学上的应用等. 中国电子学会会员编号:E190188767M.

E-mail: zhuhongyou@nwpu.edu.cn

王 磊 男,1982年1月出生于山东省枣庄市. 现为中国矿业大学计算机科学与技术学院、人工智能学院教授,博士生导师. 获国家自然科学基金、省自然科学基金等项目7项. 在国内外发表学术论文150余篇. 中国电子学会会员编号:E1900027585.

E-mail: leiwang@cumt.edu.cn

赵博伟 男,1996年12月出生于陕西省宝鸡市. 现为浙江大学药学院特别资助博士后、助理研究员. 主要研究方向为机器学习、知识图谱及其在生物信息学中的应用.

E-mail: zhaobowei@zju.edu.cn