基于 miRNA 组学的数据增强算法

周丰丰[1]，孙燕杰[2]，范雨思[3*]

(1. 吉林大学计算机科学与技术学院 长春 130012；2. 吉林大学人工智能学院 长春 130012；
3. 吉林大学软件学院 长春 130012)

【摘要】近年来，诸多研究揭示了 miRNA 的表达和疾病之间的关系，特别是其与肿瘤的发生、发展和治疗的密切关联。然而，传统的分子生物学测试方法既耗时又昂贵，患病样本获取困难，不平衡的数据集训练得到的分类器导致患病样本识别准确率低。面对以上挑战，提出了一种新的区分患病样本、健康样本以及挖掘疾病生物标志物的数据增强算法 OCF，使用条件式生成对抗网络进行数据增强，然后用特征选择算法减少特征数量，最后再利用机器学习分类器进行分类识别并筛选出生物标志物进行分析。实验结果表明，该算法具有更好的分类性能，并验证了筛选出的生物标志物的准确性。

关 键 词 计算机应用技术； 深度学习； 特征选择； 生成对抗网络； miRNA
中图分类号 TP399 **文献标志码** A doi:10.12178/1001-0548.2023002

Data Augmentation Algorithm for miRNA Omics-Based Classifications

ZHOU Fengfeng[1], SUN Yanjie[2], and FAN Yusi[3*]

(1. College of Computer Science and Technology, Jilin University Changchun 130012;
2. School of Artificial Intelligence, Jilin University Changchun 130012; 3. College of Software, Jilin University Changchun 130012)

Abstract In recent years, many studies have revealed the relationship between microRNA expression and diseases, especially its close relationship with the occurrence, development and treatment of tumors. However, traditional molecular biology testing methods are time-consuming and expensive, and it is difficult to obtain disease samples. The classifier obtained from imbalanced data set training leads to low accuracy of disease sample recognition. In the face of the above challenges, we propose a new data augmentation algorithm OCF (original data-based conditional generative adversarial network for sample generation) to distinguish health samples from disease samples and mine disease biomarkers, by using conditional generative adversarial networks for data augmentation, followed by feature selection algorithms to reduce the number of features. Finally, the machine learning classifier is used for classification and recognition, and the biomarkers are selected for analysis. The experimental results show that our proposed algorithm has better classification performance, and verify the accuracy of the selected biomarkers.

Key words computer application technology; deep learning; feature selection; generative adversarial networks; miRNA

微小 RNA(microRNA, miRNA) 是一种长度为 20 个左右核苷酸的非编码短 RNA 分子，是人类基因组编码的重要功能元件，在细胞发育和分化等过程中起重要调节作用。miRNA 的发现揭示了一种新的基因调节机制的存在，并被验证在癌症发生、发育和转移等方面都发挥至关重要的作用。此外，miRNA 被认为是识别各种不同癌症类型的潜在生物标志物。文献 [1] 研究表明 miRNA 可在血清中稳定存在，与组织标本 miRNA 表达谱相比，血清 miRNA 具有微创采样、稳定性强、灵敏度高及便于连续监测等优势，成为疾病诊断和预后评估标记物的研究热点。近年来，越来越多的研究揭示了

收稿日期：2023－01－03；修回日期：2023－02－16
基金项目：国家自然科学基金 (62072212，U19A2061)；吉林省中青年科技创新创业卓越人才 (团队) 项目 (创新类)(20210509055RQ)；吉林省大数据智能计算实验室 (20180622002JC)
作者简介：周丰丰 (1977－)，男，博士，教授，主要从事健康大数据方面的研究.
*通信作者：范雨思，E-mail：fan_yusi@163.com

miRNA 的表达和疾病之间存在紧密关联。如 miRNA 与甲状腺乳头状癌的发生、发展和转移关系紧密，在对甲状腺乳头状癌的特异性诊断、治疗及预后评估等方面具有广泛应用前景[2]。另有研究表明 miRNA 在心血管疾病发生早期的诊断和治疗方面也具有重要价值，miRNA 表达的抑制不仅与心脏病有关，与心脏的发育和生长也相关联。由于 miRNA 在特定的细胞通路上具有特异性靶标，因此将其作为诊断的标记物或通过对它的操控以获得治疗作用，都具有较好的生物学机制支持和临床应用前景[3]。miRNA 在血脑屏障和脑神经方面的调节和干预功能也获得了广泛的研究成果，据此构建的 miRNA 调控血管性痴呆病的分子交互网络，为进一步研究该疾病临床诊断方案和靶向治疗药物提供了理论支撑[4]。

但研究仍面临以下挑战：1) 样本获取困难，导致可使用的样本数量少，同时样本的特征数量多，存在部分不相关的特征和冗余特征会降低模型训练的速度、提高计算复杂度，也会影响模型的泛化精度与准确率；2) 数据集存在类别不平衡的问题，健康样本远远多于患病样本，训练得到的分类器模型会更侧重学习多数类即健康样本的特征，从而更准确地识别多数类，而忽略了少数类即患病样本的识别，因此得到的整体结果是不准确的。在现实生活中，准确识别患病样本往往更有意义，所以训练一个分类模型，可以正确识别少数类样本，从而保证最终的分类性能具有多类别的识别平衡性。

1　OCF 算法

针对以上挑战，本文提出了数据增强和特征选择结合的算法 OCF (original data-based conditional generative adversarial network for sample generation)，首先以生成对抗网络[5] 的变种条件式生成对抗网络作为数据增强模型，学习原训练集的分布，生成新的少数类样本，加入原不平衡训练集中，降低其不平衡程度。然后对增强后的训练集分析特征重要性进行特征选择[6]，选择方差最大的 50 个特征来减少样本冗余特征和不相关特征，从而提高分类器模型的训练速度和准确率，最后从选择出的特征中找出疾病相关的生物标志物，算法流程如图 1 所示。

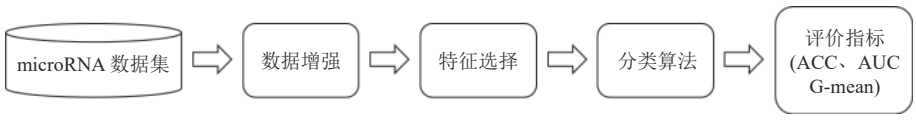


图 1　OCF 算法流程图

算法中的数据增强模块网络结构如图 2 所示，其中生成器和鉴别器均由 3 层全连接神经网络组成，模型参数如表 1 所示。将本文所用数据集按 7∶3 的比例划分为训练集和测试集，分别用每个数据集的训练集去训练鉴别器和生成器并保存训练得到的模型。样本标签 y 作为条件信息参与对生成器和鉴别器的训练，用真实样本和经生成器生成的样本去训练鉴别器。鉴别器的鉴别结果反馈给生成器，让生成器逐渐提升生成的样本质量，直到鉴别器无法区分传入的样本是真实样本还是生成样本时，结束训练。然后把需要的样本标签和数量传入已训练好的生成器，以生成所需的样本。本文提出的 OCF 算法及实验代码可以在如下网址下载：http://www.healthinformaticslab.org/supp/。

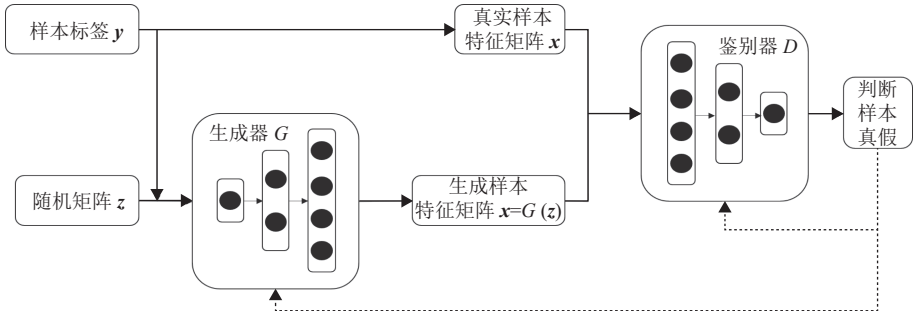


图 2　数据增强模块网络结构

表 1 模型超参数设置

超参数	取值
训练轮数/轮	200
生成器网络层数/层	3
生成器网络节点数/个	641, 1 282, 2 565
鉴别器网络层数/层	3
鉴别器网络节点数/个	2 565, 1 282, 641

2 实验结果与分析

2.1 实验环境和模型参数

本文的实验环境使用了 Python 编程语言 (版本 3.6.13)、PyTorch 框架 (版本 1.7.1)、numpy 库函数 (版本 1.19.2)、pandas 库函数 (版本 1.1.5)、sklearn 库函数 (版本 0.24.2)。计算服务器的 GPU 加速显卡型号为 TITAN RTX(24 GB 显存)，驱动程序版本 455.45.01，CUDA 版本 11.1。

实验数据增强模块的条件式生成对抗网络超参数的取值设置参见表 1。

2.2 数据集和评价指标

本文使用 3 个结构化数据集，具体如表 2 所示，定义一个数据集的不平衡率为多数类样本数量与少数类样本数量的比值。

表 2 数据集信息

数据集	样本数/个	特征数/个	不平衡率
GSE122497	5 531	2 565	8.772
GSE106817	3 079	2 565	8.622
GSE137140	3 744	2 565	1.391

数据集 GSE122497 是食管鳞状细胞癌的大规模血清 miRNA 谱组学数据，共有 5 531 个样本。其中 566 个是食管鳞状细胞癌患病样本，剩下的 4 965 个为非癌症对照样本。数据集 GSE106817 用于卵巢癌筛查的整合细胞外 miRNA 谱分析，包含 3 079 个样本，其中包括 320 个卵巢癌患病样本和 2 759 个非癌症对照样本。数据集 GSE137140 使用血清 miRNA 的血液检测肺癌患者，包含 3 744 个样本，其中包括 1 566 个术前肺癌样本和 2 178 个非癌症对照样本。

表 2 是本文使用的数据集，分别按 7∶3 的比例划分为训练集和测试集。本文所有实验的模型训练、数据增强和特征选择步骤均在训练集上进行。本文所获得的优化模型和特征子集，在没有变动的测试集上进行性能测试。

实验选择 KNN[7] 作为分类器，数据集划分为 70% 训练集和 30% 测试集，对于不平衡数据的分类问题，采用准确率 (ACC)、ROC 曲线下的面积 (AUC) 和 几何平均数 (G-mean) 作为评价指标。

2.3 消融实验结果展示和分析

为了验证本文算法模型 OCF 的有效性，以消融实验来验证各模块的必要性。原数据为未经处理的原始数据的训练集，原数据+特征选择为对未经处理的原始数据的训练集进行特征选择筛选出方差最大的前 50 个特征，原数据+数据增强为对未经处理的原始数据的训练集进行数据增强，使得训练集中多数类样本和少数类样本数量相同，实验结果均在独立测试集上取得。

数据集 GSE122497 和数据集 GSE106817 的消融实验结果如表 3 所示，本文提出的算法 OCF 在数据集 GSE122497 上的 3 个指标比原数据结果分别提升 5.16%、5.21% 和 5.21%；在数据集 GSE106817 上时，3 个指标也均取得了最佳结果，与原数据结果相比分别提升了 6.71%、7.52% 和 7.52%。

表 3 数据集 GSE122497 和 GSE106817 消融实验结果

方法	GSE122497			GSE106817		
	ACC	AUC	G-mean	ACC	AUC	G-mean
原数据	0.942 8	0.943 5	0.943 5	0.915 6	0.911 9	0.911 9
原数据+特征选择	0.930 7	0.940 9	0.940 8	0.923 2	0.936 2	0.936 1
原数据+数据增强	0.942 8	0.943 5	0.943 5	0.915 6	0.911 9	0.911 9
本文方法OCF	**0.994 6**	**0.995 6**	**0.995 6**	**0.982 7**	**0.987 1**	**0.987 1**

数据集 GSE137140 的消融实验结果如表 4 所示，因为数据集 GSE137140 的不平衡率较低，为了验证本文算法的有效性和生成样本与真实世界数据的相似程度，随机选择原数据中 20% 少数类样本和 100% 多数类组成筛选数据，对其数据增强和特征选择后，再与原数据特征选择进行对比，实验结果表明，本文模型生成的新样本组成的数据集表现优于原数据集。

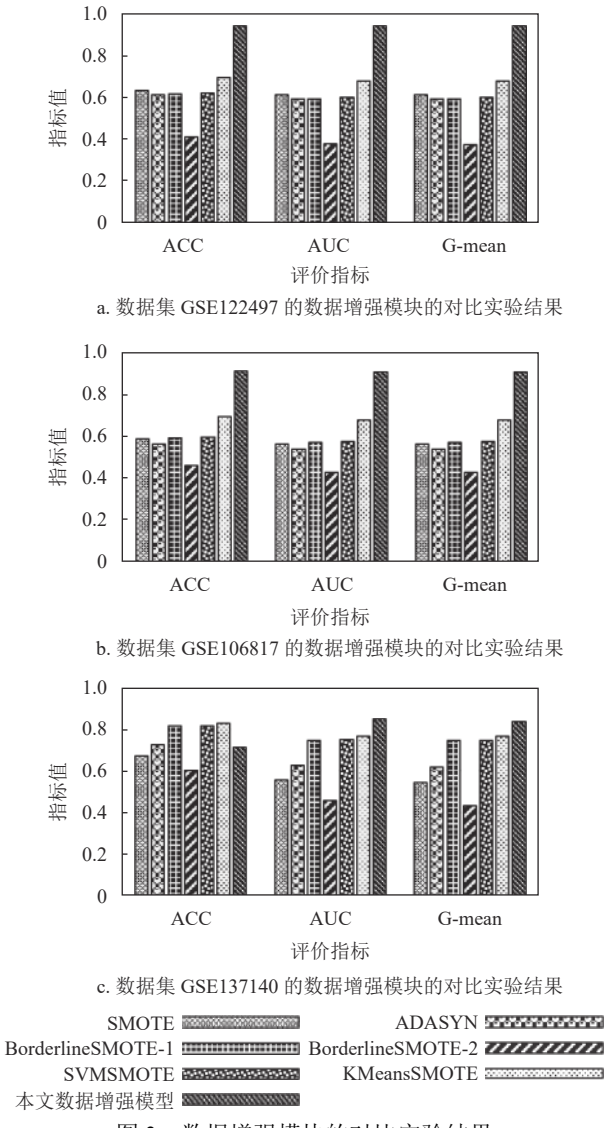
表 4 数据集 GSE137140 消融实验结果

方法	ACC	AUC	G-mean
原数据	0.831 9	0.853 2	0.853 0
原数据+特征选择	0.967 1	0.959 1	0.959 0
筛选数据	0.710 9	0.846 3	0.835 4
筛选数据+特征选择	0.948 4	0.968 8	0.968 6
筛选数据+数据增强	0.710 9	0.846 3	0.835 4
本文方法OCF	**0.997 3**	**0.998 7**	**0.998 7**

以上结果表明，数据增强和特征选择对分类器模型具有较好的性能提升，且二者联合使用会对原始模型的改进程度更大，从而证明了本文算法的有效性，可以更好地优化特征子集，筛选出针对目标问题更有意义的特征子集。

2.4　数据增强模块的对比实验结果

为了证明本文提出的 OCF 算法的有效性，和已有的其他面向不平衡数据分类问题的数据增强算法进行了对比。这些模型包括 SMOTE[8]、ADASYN[9]、Borderline-SMOTE1[10]、Borderline-SMOTE2[10]、SVMSMOTE[11] 及 KMeansSMOTE[12]。实验结果都在独立测试集上取得，验证指标包括准确率 (ACC)、ROC 曲线下的面积 (AUC) 和几何平均数 (G-mean)。图 3 为数据集 GSE122497、数据集 GSE106817和数据集 GSE137140 的数据增强模块的对比实验结果。



a. 数据集 GSE122497 的数据增强模块的对比实验结果

b. 数据集 GSE106817 的数据增强模块的对比实验结果

c. 数据集 GSE137140 的数据增强模块的对比实验结果

图 3　数据增强模块的对比实验结果

与 6 种流行的数据增强算法对比结果表明，除了数据集 GSE137140 的评价指标 ACC 未取得最大值，在其余数据集的各指标上本文算法均取得了最好的结果，提升明显。数据集 GSE122497 和数据集 GSE106817 的指标值都能达到 90% 以上，数据集 GSE137140 随机选择原数据中 20% 少数类样本和 100% 多数类组成筛选数据，本文算法在指标 AUC 和 G-mean 上都取得最大值，也可以证明本文算法训练出了高质量的分类器模型。

2.5　结合特征选择算法的对比实验结果

表 5 为数据集 GSE122497、数据集 GSE106817 和数据集 GSE137140 的数据增强和特征选择结合算法的对比实验结果。6 个数据增强算法结合特征选择算法之后指标值也均有所提升，本文算法在数据集 GSE122497 和数据集 GSE137140 上的 3 个指标的值均达到 99% 以上，在数据集 GSE106817 上的 3 个指标的值均达到 98% 以上，表现出了明显的优势。

以上两组对比实验的结果表明，本文提出的 OCF 算法中的数据增强模块在 3 个指标上都比其他数据增强模型表现好，结合了数据增强和特征选择的 OCF 算法均取得了指标的最大值，这足以证明本文算法的有效性。

2.6　生物标志物分析

用梯度提升决策树算法 XGBoost[13] 选出各个数据集重要性排名前 20 的特征取交集，对候选关键特征进行排序[14]，取排名前 5 名的 miRNA 进行生物标志物分析，miRNA 信息如表 6 所示。

文献 [15] 验证了 hsa-miR-1228-5p 作为生物标志物为肝癌诊断的高准确性，也可用于区分 HCC 患者与健康和肝硬化患者。文献 [16] 验证了 hsa-miR-1228-5p 为具有抗黑素瘤分化相关蛋白 5 抗体阳性亚群的皮肌炎相关间质性肺病的新型生物标志物。文献 [17] 确定了 hsa-miR-4532 为早期和快速识别 COVID-19 患者疾病进展的预测标志物之一。文献 [18] 的结论证明了 hsa-miR-4532 与糖尿病肾病相关，影响 KCNJ11 的表达和磺酰脲刺激的胰岛素分泌。文献 [19] 证明了 hsa-miR-4532 是在卵巢血清癌样本中受到调控的 miRNA 之一，为证明血清 miRNA 谱是卵巢癌的一个有前途的诊断生物标志物提供了重要证据。文献 [20] 表明 hsa-miR-2861 作为检测宫颈癌的新型非侵入性生物标志物。

表 5 数据增强和特征选择结合算法的对比实验结果

算法	GSE122497			GSE106817			GSE137140		
	ACC	AUC	G-mean	ACC	AUC	G-mean	ACC	AUC	G-mean
SMOTE	0.796 4	0.785 4	0.785 4	0.785 7	0.775 1	0.775 0	0.952 8	0.941 3	0.941 2
ADASYN	0.833 1	0.824 6	0.824 6	0.764 1	0.750 4	0.750 3	0.947 5	0.930 9	0.930 8
Borderline-SMOTE1	0.839 2	0.833 7	0.833 7	0.814 9	0.812 1	0.812 1	0.955 5	0.942 6	0.942 5
Borderline-SMOTE2	0.750 0	0.737 1	0.737 0	0.743 5	0.729 8	0.729 7	0.859 4	0.808 9	0.807 3
SVMSMOTE	0.845 8	0.840 7	0.840 7	0.826 8	0.818 6	0.818 6	0.952 8	0.941 3	0.941 2
KMeansSMOTE	0.810 8	0.800 7	0.800 7	0.853 9	0.847 2	0.847 2	0.960 9	0.949 8	0.949 8
本文方法OCF	**0.994 6**	**0.995 6**	**0.995 6**	**0.982 7**	**0.987 1**	**0.987 1**	**0.997 3**	**0.998 7**	**0.998 7**

表 6 miRNA 信息

排名	特征	miRNA名称
1	MIMAT0005582	hsa-miR-1228-5p
2	MIMAT0019071	hsa-miR-4532
3	MIMAT0022946	hsa-miR-1237-5p
4	MIMAT0023712	hsa-miR-6087
5	MIMAT0013802	hsa-miR-2861

3 结 束 语

本文提出了一种基于条件式生成对抗网络的数据增强和特征选择结合的算法，生成现实生活中难获取的少数类样本，降低了数据集的不平衡程度，减少了数据集的特征数量，去掉了冗余特征，筛选出对于目标问题更具备意义的特征子集，训练出优秀的分类器模型。实验结果表明，本文提出的 OCF 算法在各评价指标上都表现最好，可以准确地区分疾病样本和健康样本，并找出疾病相关的生物标志物。期望本文算法可以为早期的疾病诊断、预测和预防提供有价值的信息。

参 考 文 献

[1] 王火强, 王奕然. miRNA 标志物在临床检测的应用[J]. 中国医药导刊, 2022, 24(2): 127-130.
WANG H Q, WANG Y R. Perspective of clinical diagnosis by miRNAs Biomarker[J]. Chinese Journal of Medicinal Guide, 2022, 24(2): 127-130.

[2] 胡仿玲, 张思林, 余杰情, 等. microRNA 作为甲状腺乳头状癌生物标记物的研究进展[J]. 临床耳鼻咽喉头颈外科杂志, 2018, 32(15): 1199-1202.
HU F L, ZHANG S L, YU J Q, et al. Research progress of microRNA as a biomarker of papillary thyroid carcinoma[J]. J Clin Otorhinolaryngol Head Neck Surg(China), 2018, 32(15): 1199-1202.

[3] 齐秀丽, 徐莉, 韩丽丽, 等. microRNA-疾病诊断的潜在生物标记物[J]. 生物化工, 2016, 2(6): 72-74.
QI X L, XU L, HAN L L, et al. microRNAs: Potential biomarkers for disease diagnosis[J]. Biological Chemical Engineering, 2016, 2(6): 72-74.

[4] 孙孟艳, 秦合伟, 牛雨晴, 等. microRNA 与其在血管性痴呆中作用的研究进展[J]. 解放军医学院学报, 2022, 11: 1198-1203.
SUN M Y, QIN H W, NIU Y Q, et al. Research advances in microRNA and its role in vascular dementia[J]. Acad J Chin PLA Med Sch, 2022, 11: 1198-1203.

[5] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//The 2014 Neural Information Processing Systems. Montreal: NIPS Press, 2014: 139-144.

[6] 黄娜, 何泾沙, 吴亚飚. 恶意 PDF 检测中的特征工程研究与改进[J]. 电子科技大学学报, 2022, 51(5): 766-773.
HUANG N, HE J S, WU Y B. Research and improvement of feature engineering for malicious PDF detection[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(5): 766-773.

[7] MAMUN A M, KAWSAR A, MINHTHANG B F, et al. Machine learning-based statistical analysis for early stage detection of cervical cancer[J]. Computers in Biology and Medicine, 2021, 139: 104985.

[8] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[9] HE H B, BAI Y, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]//The 2008 IEEE International Joint Conference on Neural Networks (IJCNN 2008). Hong Kong, China: IEEE Press, 2008: 1322-1328.

[10] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[C]//Advances in Intelligent Computing, ICIC 2005. Berlin, Heidelberg: Springer-Verlag, 2005: 878-887.

[11] NGUYEN H M, COOPER E W, KAMEI K. Borderline over-sampling for imbalanced data classification[J]. Int J Knowl Eng Soft Data Paradigms, 2009, 3: 4-21.

[12] LAST F, DOUZAS G, BACAO F. Oversampling for imbalanced learning based on K-means and SMOTE [EB/OL]. (2017-11-02) https://arXiv.org/abs/1711.00837.

[13] LYU H, ZHANG Y, WANG J S, et al. iRice-MS: An integrated XGBoost model for detecting multitype post-translational modification sites in rice[J]. Briefings in Bioinformatics, 2021, 23(1): bbab486.

[14] 姚旭, 詹秀秀, 刘闯, 等. 基于复杂网络控制理论的肿瘤关键基因预测研究[J]. 电子科技大学学报, 2022, 51(1): 138-147.
YAO X, ZHAN X X, LIU C, et al. Predicting the critical tumor genes based on complex network control theory[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(1): 138-147.

[15] GUAN X Y, TAN Y W, GE G H, et al. A serum microRNA panel as potential biomarkers for hepatocellular carcinoma related with hepatitis B virus[J]. PLoS ONE, 2014, 9(9): e107986.

[16] ZHONG D L, WU C Y, XU D, et al. Plasma-Derived exosomal hsa-miR-4488 and hsa-miR-1228-5p: Novel biomarkers for dermatomyositis-associated interstitial lung disease with anti-melanoma differentiation-associated protein 5 antibody-positive subset[EB/OL]. (2021-07-29). https://www.hindawi.com/journals/bmri/2021/6676107/.

[17] PARRAY A, MIR F A, DOUDIN A, et al. SnoRNAs and miRNAs networks underlying COVID-19 disease severity[J]. Vaccines, 2021, 9(10): 1056.

[18] CHEN Z R, HE F Z, LIU M Z, et al. MIR4532 gene variant rs60432575 influences the expression of KCNJ11 and the sulfonylureas-stimulated insulin secretion[J]. Endocrine, 2019, 63(3): 489-496.

[19] HAMIDI F, GILANI N, BELAGHI R A, et al. Exploration of potential miRNA biomarkers and prediction for ovarian cancer using artificial intelligence[J]. Front Genet, 2021, 12: 724785.

[20] ZHANG Y, ZHANG D, WANG F, et al. Serum miRNAs panel (miR-16-2*, miR-195, miR-2861, miR-497) as novel non-invasive biomarkers for detection of cervical cancer[J]. Sci Rep, 2015, 5: 17942.

编　辑　刘飞阳