面向不平衡多组学癌症数据的特征表征算法

周丰丰[1,2]，郭 喆[1]，范雨思[2]

（1. 吉林大学 人工智能学院，长春 130012；2. 吉林大学 计算机科学与技术学院，长春 130012）

摘 要：针对癌症疾病数据结构复杂、预测困难、数据不平衡和患者隐私保护等一系列的问题，提出了一种特征表征算法ImFeatures，解决了癌症数据的不平衡问题，丰富了样本结构。联合癌症转录组和甲基化2种组学数据作为真实样本，通过逻辑回归（LR）和随机森林（RF）2种特征选择后，得到的负样本被随机划分并结合等量的正样本，输入本文提出的特征表征模型，生成学习到关键特征信息的表征样本，以提高模型预测能力。实验结果表明，在经过特征表征后的11种常见癌症数据集上，本文提出的结合特征筛选和特征表征的算法的准确率（Acc）均超过了80.00%，其中有5种癌症的预测准确率超过了95.00%，可以有效提升癌症疾病的预测准确率。

关键词：计算机应用技术；特征表征；生物信息学；多组学数据；特征筛选；机器学习

中图分类号：TP399 **文献标志码**：A **文章编号**：1671-5497（2025）06-2089-08

DOI：10.13229/j.cnki.jdxbgxb.20231196

Feature representation algorithm for imbalanced classification of multi-omics cancer data

ZHOU Feng-feng[1,2]，GUO Zhe[1]，FAN Yu-si[2]

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

Abstract：Aiming at a series of problems such as complex data structure，difficult prediction，data imbalance，and patient privacy protection in cancer diseases，a feature representation algorithm ImFeatures was proposed to solve the problem of imbalanced cancer data and enrich the sample structure. By combining two types of cancer transcriptome and methylation data as real samples，negative samples obtained after feature selection by logistic regression and random forest were randomly divided and combined with equal numbers of positive samples. The feature representation model proposed was used to generate sample representations that learn key feature information，thereby improving the predictive ability of the model. The experimental results show that on 11 common cancer datasets after feature characterization，the accuracy（Acc）of the algorithm combining feature selection and feature representation proposed in this paper exceeds 80.00% in all cases，and five cancer types even receive accuracies over 95.00%，which can effectively improve the prediction accuracies of cancer diseases.

收稿日期：2023-11-03.

基金项目：吉林省中青年科技创新创业卓越人才（团队）项目（创新类）（20210509055RQ）；中国自然科学基金项目（62072212，U19A2061）；吉林省大数据智能计算实验室项目（20180622002JC）.

作者简介：周丰丰（1977-），男，教授，博士. 研究方向：健康大数据. E-mail：fengfengzhou@gmail.com

Key words: computer application technology; feature representation; bioinformatics; multi-omics data; feature selection; machine learning

0 引 言

癌症[1]是导致全球人类死亡的主要原因之一,对人类健康造成了巨大负担。癌症的发病机制复杂,涉及多个基因、环境因素和遗传因素的交互作用。癌症的早期诊断和治疗是降低癌症死亡率的有效路径之一。基因序列分析是一种有效的手段,研究人员对比分析癌症患者与普通患者的甲基化和转录组等基因的序列数据,可以更好地推进癌症早期筛查。

DNA甲基化[2]是一种表观遗传修饰,它在不改变DNA序列的情况下,调控基因的表达。甲基化数据可以揭示癌症相关甲基化位点的变化,如整体低甲基化和局部高甲基化等。这些变化可能影响基因表达和染色质结构,进而影响癌症的发生、发展和治疗响应。转录组数据关注基因的表达水平[3]。通过研究转录组数据,可以了解基因在癌症中的表达模式和调控关系,有助于判断肿瘤的不同阶段和预测预后生物标志物的识别,有助于患者的进一步诊断和治疗。细胞层面上,由于单一组学数据可能无法提供肿瘤细胞的全面特征,研究人员通过整合多组学数据(例如基因组学、转录组学、蛋白质组学、代谢组学等)以更深入地理解癌症的发生和发展机制,为精准医疗提供依据[4]。

近年来,随着人工智能技术的发展,通过机器学习和深度学习等技术,可以更精确识别癌症细胞,提高癌症诊断准确性[5]。虽然机器学习在癌症预测方面具有巨大潜力,但在实际应用中仍有许多需要克服的问题:①数据分布不平衡。癌症预测模型通常依赖于大量的患者数据,然而这些数据在大自然中分布和获取都是严重不平衡的[6],且这些数据可能存在缺失值、异常值或噪声。②数据价值密度低。高通量测序技术[7]在临床医学上的不断发展,可以快速、准确地获取大量基因组信息。获取的数据特征可能包含大量的信息,但并非所有信息都对研究具有价值,造成特征冗余。③隐私泄露问题。在处理患者数据时,需要遵循严格的伦理和隐私规定[8]。这可能导致数据收集和共享的困难,从而影响机器学习模型的准确性和稳定性。

针对这些问题,本文提出了一种基于不平衡多组学癌症数据的特征表征方法,通过对真实的癌症数据的特征进行表征[9],同时生成虚拟的癌症患者数据,并且将其加入训练集中,既平衡了数据集又保护了患者的隐私,揭示癌症数据内在规律和本质,增加模型的预测泛化能力,为癌症预测领域和科学研究提供有力支持。

1 实验材料及方法

1.1 ImFeatures算法

本文提出了一种基于不平衡多组学癌症数据的特征表征算法ImFeatures,其处理流程包括数据预处理、特征筛选、特征表征、分类4个部分,如图1所示。

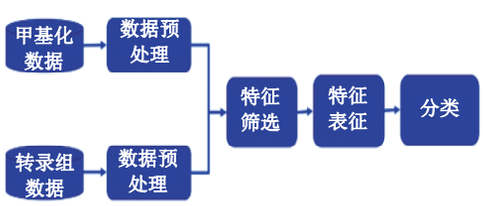

图1 ImFeatures算法流程图

Fig. 1 ImFeatures algorithm procedure

ImFeatures算法首先使用高通量测序技术得到转录组数据及甲基化数据,通过数据预处理的缺失值处理和方差筛选2个部分得到关键特征。缺失值处理将特征值有缺失的样本删除,方差筛选计算每组特征的方差的平均值作为方差筛选阈值[10]。将预处理后的转录组数据和甲基化数据连接起来组成包含甲基化信息和转录组信息的多组学数据。然后将其特征筛选使用了LR-RF方法[11],首先使用逻辑回归(Logistic regression,LR)特征筛选方法对结合后的多组学特征进行初筛,选择了重要性排序前10 000个特征,然后使用随机森林(Random forest,RF)特征选择方法对初筛后的特征重要性排序,选择了重要性排序前200个特征。

本文使用的特征表征模块是基于生成对抗网络的一种改进算法,特征表征模块网络结构如图2所示,首先将不平衡的癌症数据集根据多数类样本与少数类样本不平衡的比率Lr随机划分为Lr的整数份,每份包含全部少数类样本和1/Lr

份随机选择的多数类样本，得到Lr份平衡的癌症数据集。本文使用的所有数据集的$Lr>1$，因此在本文中多数类样本对应负样本，少数类样本对应正样本。将这Lr份数据集分批次输入到数据增强模型中，得到Lr份根据真实样本生成的包含真实特征信息的虚拟样本。在特征表征模块网络中，编码器、生成器和鉴别器共同协作，通过特征表征和相互竞争实现生成任务。

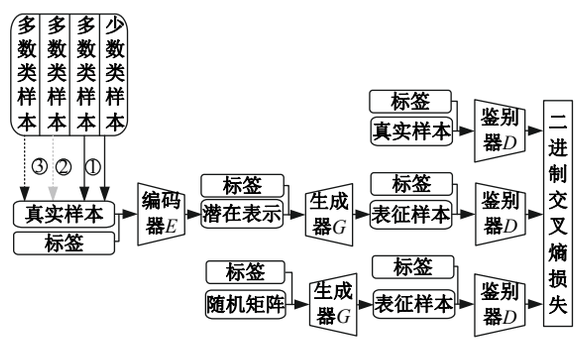

图 2　特征表征模块网络结构

Fig. 2　Feature representation module network structure

首先，将癌症数据及其标签输入一个编码器中，从输入数据中提取特征信息，输出真实数据的分布的均值和方差，得到数据的潜在表示。在这个过程中，编码器将输入数据编码成一组连续的变量，这些变量可以捕捉到数据的内在结构和不稳定性。编码器采用全连接神经网络（Multi-layer perceptron，MLP）结构，通过层层处理将输入数据转化为特征表征。接着，将编码器得到的数据的潜在表示和随机矩阵分别与标签结合输入生成器，根据从编码器提取的特征信息生成新的数据。生成器采用全连接神经网络结构，通过逆向层层处理将特征信息还原为原始数据的结构。生成器的学习过程是通过不断优化生成器参数，使其生成的数据更加接近真实数据分布。然后，将真实数据和生成器的输出样本与对应标签结合，输入鉴别器，判断输入数据是真实数据还是生成数据。鉴别器通过训练学习输入数据的特征，不断地调整内部参数，以提高对真实数据和生成数据的识别能力，从而能够准确地区分真实数据和生成数据。

在特征表征模块训练过程中，编码器、生成器和鉴别器相互竞争，生成器试图学习到能够欺骗鉴别器的特征，而鉴别器则努力提高对生成数据的识别能力，以最小化鉴别器的输出误差。在这种竞争关系中，生成器和鉴别器不断地优化各自的表现，最终使生成器能够在潜在空间中捕捉到足够逼真的特征，从而生成高质量的样本。这种特征表征方法使得模型在无监督学习中能够有效地捕捉数据内部的潜在结构，并生成具有较高质量的样本。

1.2　实验环境和模型参数

本文使用了 Python 编程语言（版本 3.7.0）、PyTorch 框架（版本 1.7.1）、numpy 库函数（版本 1.18.5）、pandas 库函数（版本 1.1.5）、sklearn 库函数（版本 1.0.2）搭建特征表征模型框架。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 10.1。

特征表征的编码器、生成器和鉴别器模型具体参数如表 1 所示。

表 1　特征表征模型超参数设置

Table 1　Feature representation model hyperparameter settings

参数	取值
编码器层数/层	3
编码器网络节点数/个	512，256，128
编码器每层激活函数	ReLu，ReLu，ReLu
生成器层数/层	4
生成器网络节点数/个	128，256，512，1 024
生成器激活函数	ReLu，ReLu，ReLu，Tanh
鉴别器层数/层	4
鉴别器网络节点数	512，256，128，1
鉴别器激活函数	Leaky ReLu，Leaky ReLu，Leaky ReLu，Sigmoid

1.3　数据集和评价指标

本文选择了 11 个不同的常见癌症数据集，这些癌症数据集来源于癌症基因组图谱（The cancer genome atlas，TCGA）项目，它们分别代表了不同类型的癌症。下面是对这些数据集的分类和简单描述：肾上腺皮质癌（Adrenocortical carcinoma，ACC）是发生于肾上腺皮质的一种罕见的高度侵袭性的恶性内分泌肿瘤。膀胱癌（Bladder cancer，BLCA）起源于膀胱黏膜上皮细胞，是一种常见的泌尿系统恶性肿瘤[12]。乳腺癌（Breast cancer，BRCA）是一种发生在乳腺上皮或导管上皮的恶性肿瘤，是女性最常见的癌症类型之一[13]。胆管癌（Cholangiocarcinoma，CHOL）起源于胆管上皮细胞，是一种较为罕见的肝癌类型[14]。结直肠癌（Colon adenocarcinoma，COAD）是一种发生在结肠和直肠的恶性肿瘤，起源于结

直肠上皮细胞。肾嫌色细胞癌(Kidney chromophobe,KICH),肾透明细胞癌(Kidney renal clear cell carcinoma,KIRC)。肾脏乳头状癌(Kidney papillary carcinoma,KIRP)是发生在肾脏的恶性肿瘤,起源于肾脏小管上皮细胞[15]。肺腺癌(Lung adenocarcinoma,LUAD)起源于支气管黏膜上皮,少数起源于大支气管的黏液腺,是肺癌的一种,属于非小细胞癌。肺鳞状细胞癌(Lung squamous cell carcinoma,LUSC)是一种发生在肺部的恶性肿瘤,起源于肺鳞状上皮细胞[16]。间皮瘤(Mesothelioma,MESO)是一种原发于腹膜间皮细胞的肿瘤。这些数据集被广泛用于研究癌症的分子机制、预测癌症的预后和寻找新的治疗靶点。且每个数据集都包含转录组与甲基化2个组学的数据。其中,将癌症Ⅰ期与Ⅱ期的患者规定为负样本,标签设置为0,对应的将癌症Ⅲ期与Ⅳ期的患者规定为正样本,标签设置为1。Lr表示样本不平衡的比率,为负样本数量与正样本数量的比值。数据集基本信息如表2所示。当$Lr>1$时,负样本对应图2中多数类样本,正样本对应少数类样本;当$Lr<1$时,则相反。

表2 数据集信息

Table 2 Aataset information

数据集	甲基化特征数	转录组特征数	样本数	正样本数	负样本数	不平衡率$Lr(N/P)$
ACC	394 014	60 483	76	8	68	8.50
BLCA	382 024	60 483	444	166	278	1.67
BRCA	363 736	60 483	1 112	298	814	2.73
CHOL	378 735	60 483	53	8	45	5.63
COAD	374 805	60 483	459	165	294	1.78
KICH	391 298	60 483	58	8	50	6.25
KIRC	382 720	60 483	274	17	257	15.12
KIRP	383 105	60 483	123	47	76	1.62
LUAD	369 182	60 483	352	7	345	4.24
LUSC	370 619	60 483	639	122	517	5.37
MESO	381 223	60 483	465	73	392	1.13

2 实验结果及分析

2.1 特征表征模型实验结果

由于使用平衡数据集训练模型性能优于使用不平衡数据集训练得到的模型,因此本文随机选择平衡的正负样本作为训练和测试样本,训练集和测试集的划分采用五折交叉验证方式进行,在训练集中对比是否加入经过特征表征得到的虚拟样本。

为了验证本文提出的基于不平衡多组学癌症数据的特征表征方法的有效性,本文使用了K最邻近(K-nearest neighbor,KNN)、支持向量机(Support vector machine,SVM)、逻辑回归(Logistic regression,LR)分类器、朴素贝叶斯(Naive Bayes,NB)分类器这4种常用的机器学习方法对真实数据和真实数据加特征表征数据分别进行训练,使用相同的测试集进行测试。为了多方面评价模型的效果,本文使用了常用的机器学习评价指标,即准确率(Acc)、ROC曲线下方面积(Area under curve,AUC)。实验结果如表3所示。

实验结果表明,在11个癌症数据集上表现最好的结果均发生在真实样本经过本文提出的特征表征模型后的训练集上,且对于同一种分类方法,每个经过特征表征模型后的效果均优于特征表征前。因此可以得出结论,本文提出的特征表征模型在不平衡多组学癌症数据的分类预测上是有效的。在4种分类方法中,KNN在特征表征前后的训练集上的结果有7次表现最好,因此认为KNN是最适合癌症数据集分类的机器学习模型,下述实验使用KNN进行验证。

2.2 数据增强模块的对比实验结果

多组学癌症数据是指通过多种组学技术(如基因组学、转录组学、蛋白质组学、代谢组学等)对癌症相关样本进行深入研究所产生的数据。多组学癌症数据在癌症研究中具有广泛的应用价值,有助于深入揭示癌症的发病机制、生物学特性和治疗策略。为了验证多组学癌症数据在分类预测方面的有效性,本文对比了仅使用转录组数据和仅使用甲基化数据2种单组学数据训练模型与使用转录组数据和甲基化数据结合的多组学数据2种方法的实验结果。

实验在3种常见数据集BRCA、LUAD、LUSC上进行,使用KNN分类器进行训练预测。实验结果如表4所示。

由表4可以看出,在Acc、马修斯相关系数(Matthews correlation coefficient,MCC)、AUC这3种综合评价指标中,使用多组学数据进行预测的表现均优于单组学数据的表现,在其余的评价指标中,大部分结果也都是优于单组学结果的,而BRCA的Sen预测结果和LUAD的Spe结

表 3　特征表征前后模型预测结果

Table 3　Model prediction results before and after feature representation

数据集	特征表征前				ImFeatures			
	KNN	SVM	LR	NB	KNN	SVM	LR	NB
ACC	0.816 7	0.683 3	0.816 7	0.633 3	0.883 3	0.883 3	**0.950 0**	0.933 3
BLCA	0.705 1	0.704 9	0.680 5	0.714 1	0.852 6	0.774 3	0.773 8	0.717 2
BRCA	0.632 9	0.718 0	0.590 4	0.620 5	0.819 1	0.739 3	0.709 1	0.636 6
CHOL	0.616 7	0.666 7	0.800 0	0.933 3	1.000 0	0.933 3	1.000 0	1.000 0
COAD	0.632 9	0.718 0	0.590 4	0.620 5	0.819 1	0.739 3	0.709 1	0.636 6
KICH	1.000 0	1.000 0	1.000 0	0.933 3	1.000 0	1.000 0	1.000 0	1.000 0
KIRC	0.820 0	0.820 0	0.746 7	0.860 0	0.926 7	0.926 7	0.966 7	0.933 3
KIRP	0.871 9	0.787 7	0.882 5	0.787 7	0.977 8	0.893 0	0.935 7	0.829 8
LUAD	0.708 8	0.709 4	0.680 6	0.774 7	0.869 0	0.799 3	0.803 4	0.778 8
LUSC	0.664 8	0.664 8	0.603 2	0.746 4	0.828 5	0.842 3	0.856 3	0.808 3
MESO	0.750 0	0.722 5	0.763 3	0.854 2	0.815 8	0.802 5	0.815 8	0.855 0
数据集	特征表征前				ImFeatures			
	KNN	SVM	LR	NB	KNN	SVM	LR	NB
ACC	0.850 0	0.750 0	0.850 0	0.650 0	0.900 0	0.900 0	0.950 0	0.900 0
BLCA	0.705 1	0.704 6	0.680 7	0.714 3	0.852 4	0.774 1	0.773 7	0.717 3
BRCA	0.633 1	0.718 2	0.590 4	0.620 7	0.819 2	0.739 5	0.709 3	0.636 7
CHOL	0.600 0	0.650 0	0.850 0	0.900 0	1.000 0	0.900 0	1.000 0	1.000 0
COAD	0.633 1	0.718 2	0.590 4	0.620 7	0.819 2	0.739 5	0.709 3	0.636 7
KICH	1.000 0	1.000 0	1.000 0	0.950 0	1.000 0	1.000 0	1.000 0	1.000 0
KIRC	0.816 7	0.816 7	0.750 0	0.866 7	0.933 3	0.916 7	0.966 7	0.933 3
KIRP	0.868 9	0.782 2	0.878 9	0.785 6	0.977 8	0.891 1	0.935 6	0.827 8
LUAD	0.708 7	0.710 7	0.682 2	0.776 2	0.870 0	0.800 7	0.804 3	0.780 3
LUSC	0.667 1	0.665 7	0.603 8	0.748 6	0.828 1	0.843 8	0.856 2	0.810 0
MESO	0.755 4	0.728 6	0.764 3	0.858 9	0.819 6	0.807 1	0.816 1	0.857 1

表 4 单组学与多组学数据预测结果

Table 4　Prediction results of single-omics and multi-omics data

数据名称	准确率	敏感性	特异性	精确率	召回率	马修斯相关系数	ROC 曲线下面积
BRCA-Transcriptomics	0.604 0	0.406 3	0.802 2	0.667 1	0.406 3	0.225 1	0.604 2
BRCA-Methylation	0.596 6	0.489 3	0.704 1	0.623 5	0.489 3	0.198 1	0.596 7
BRCA-Multi-omics	0.632 9	0.464 8	0.801 3	0.701 3	0.464 8	0.283 1	0.633 1
LUAD-Transcriptomics	0.612 0	0.464 0	0.760 0	0.657 2	0.464 0	0.235 0	0.612 0
LUAD-Methylation	0.645 4	0.456 2	0.835 2	0.737 9	0.456 2	0.316 5	0.645 7
LUAD-Multi-omics	0.708 8	0.639 0	0.778 3	0.745 3	0.639 0	0.423 4	0.708 7
LUSC-Transcriptomics	0.593 0	0.258 4	0.927 9	0.788 9	0.258 4	0.251 9	0.593 2
LUSC-Methylation	0.576 1	0.318 7	0.836 3	0.557 0	0.318 7	0.147 6	0.577 5
LUSC-Multi-omics	0.664 8	0.375 2	0.959 0	0.920 0	0.375 2	0.415 5	0.667 1

果,由于单一组学数据分布不平衡导致训练模型过拟合。

2.3　特征表征方法对比其他表征模型

针对本文提出的特征表征方法,本文在3种常用癌症数据集(BRCA、LUAD、LUSC)上对比了其他常用的表征模型(如GAN[17]、VAE[18]、WGAN[19]),表征后的数据加入原始数据集,训练预测模型采用KNN分类器,使用Acc和AUC 2种评价指标,结果如图3所示。

由图3可以看出,本文方法在4种特征表征方法中表现最好。本文方法在BRCA数据集上的准确率(Acc)超出表现次优的GAN模型11.7%,在LUAD上超出表现次优的WGAN模型15.59%,在LUSC上超出表现次优的VAE模型22.55%。在AUC方面,本文方法也优于其他模型的表现结果。验证了本文提出的模型不仅有优秀的特征表征能力,同时具有较高的稳定性,得到的特征表征数据可以很好地学习到模型的重要特征信息,明显提高了模型的预测准确率。

2.4　特征表征方法在联邦学习上的应用

联邦学习(Federated learning)是一种分布式、隐私保护的机器学习技术,它可以在不共享原

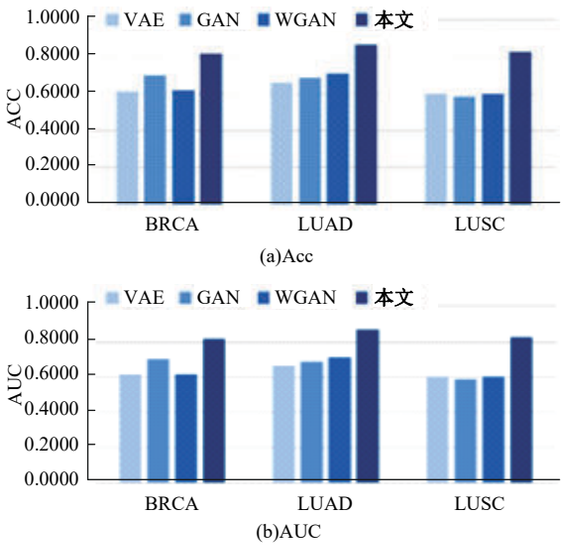

图3 本文模型对比其他特征表征方法结果

Fig. 3 The proposed model comparison with other feature representation methods

始数据的情况下,实现多个客户端之间的协同学习,满足现代应用对数据安全、隐私保护等需求。为了保护患者隐私,本文对比了在3种常见癌症数据集(BRCA、LUAD、LUSC)上训练模型仅使用特征表征数据和使用真实数据的2种结果。同时对比了常用癌症数据不平衡问题的解决方法SMOTE[20]的结果与结合SMOTE和ImFeatures 2种方法的结果。结果如表5所示。

结果表明,本文提出的特征表征方法在3种癌症数据集上都有最好的结果,其中在BRCA数据集上,准确率(Acc)优于使用真实原始样本的结果18.09%,优于使用SMOTE方法14.71%;在LUAD数据集上,准确率(Acc)优于使用真实原始样本的结果14.77%,优于使用SMOTE方法16.83%;在LUSC数据集上,准确率(Acc)优于使用真实原始样本的结果14.30%,优于使用SMOTE方法20.50%。在BRCA和LUSC数据集上,结合2种方法的结果优于仅使用一种方法,在LUAD数据集上,仅使用本文方法表现更好。因此可以看出在预测阶段,不使用真实数据样本而替代以虚拟样本,就避免了患者隐私数据的泄漏。并且仅使用特征表征数据进行预测,用虚拟样本代替真实样本,同样可以达到很好的效果。综上所述,本文提出的特征表征方法不仅可以很好地保护患者隐私,防止数据泄漏,在提升模型预测效果方面也有很好的表现。

2.5 对比其他同类研究方法的结果

本节对比了本文方法与文献[21]提出的方法(OCF),3个数据集分别为食管鳞状细胞癌、卵巢癌和肺癌患者的miRNA谱组学数据集,为了验证本文方法的有效性,数据集、分类器、训练集、测试集划分等全部指标均与文献[21]的相同。结果如表6所示,其中,几何均值(Geometric mean,G-means)是综合衡量分类模型对正类(少数类)和负类(多数类)的识别能力,尤其适用于类别不平衡数据。

实验结果表明,本文方法的准确率(Acc)均大于等于文献[21]的结果,其中,表现最好的是在GSE106817数据集上,超出了OCF的结果1.3%。对于AUC和G-means指标,本文方法在GSE122497和GSE106817数据集上的结果均优于OCF得到的结果,尽管在GSE137140数据集上AUC的结果没有超出OCF得到的结果,可能是由于本文方法更适用于明显不平衡数据。同时,本文方法流程首先进行特征选择然后进行特征表征,在特征选择时样本数量少于OCF方法,

表5 特征表征在联邦学习上的表现结果

Table 5 Performance of feature representation in federated learning

数据集	算法	准确率	敏感性	特异性	精确率	召回率
BRCA	Base	0.632 9	0.464 8	0.801 3	0.701 3	0.464 8
	SMOTE	0.666 7	0.584 7	0.748 3	0.695 6	0.584 7
	ImFeatures	0.813 8	0.815 2	0.812 2	0.813 5	0.815 2
	SMOTE+ImFeatures	0.831 5	0.818 7	0.844 2	0.839 9	0.818 7
LUAD	Base	0.700 6	0.623 0	0.779 3	0.744 7	0.623 0
	SMOTE	0.680 0	0.623 0	0.737 0	0.715 9	0.623 0
	ImFeatures	0.848 3	0.819 0	0.877 0	0.867 6	0.819 0
	SMOTE+ImFeatures	0.828 2	0.802 3	0.853 3	0.853 6	0.802 3
LUSC	Base	0.664 8	0.383 8	0.945 7	0.894 4	0.383 8
	SMOTE	0.602 8	0.316 2	0.889 5	0.770 0	0.316 2
	ImFeatures	0.807 8	0.632 4	0.985 7	0.977 8	0.632 4
	SMOTE+ImFeatures	0.814 7	0.659 0	0.971 4	0.957 8	0.659 0

表 6　与其他同类研究方法对比的结果

Table 6　Comparison results with other similar research methods

数据集	OCF 方法			本文方法		
	Acc	AUC	G-means	Acc	AUC	G-means
GSE122497	0.994 6	0.995 6	0.995 6	0.997 6	0.998 7	0.998 6
GSE106817	0.982 7	0.987 1	0.987 1	0.995 7	0.997 6	0.997 6
GSE137140	0.997 3	0.998 7	0.998 7	0.997 3	0.997 1	0.997 1

并且经过计算 OCF 特征表征方法运行一次的 FLOPS 为 O(8229160)，而本文方法运行一次的 FLOPS 为 O(1018497)。因此，本文提出的 Im-Features 方法可以明显缩短运行时间。这表明，本文方法不仅在模型准确性上有很好的表现，同时在整个模型的运行时间上也更少。

3　结束语

本文提出了一种基于不平衡多组学癌症数据的特征表征方法，在保护患者隐私的同时用于对癌症数据的预测分析。转录组和甲基化 2 种组学数据的结合包含了更多的癌症信息，特征筛选过程选择了重要性排名高的关键信息特征。特征表征方法包含 3 个神经网络，将真实数据和标签数据同时输入编码器、生成器和鉴别器，通过多个隐含层得到表征信息并加入训练集，最终明显提高模型预测的准确性。经实验验证，本文提出的特征表征方法对于癌症的整体预测具有很好的性能，在 11 个常见的癌症数据集上使用 4 种常见的机器学习分类器预测，均超出了特征表征前的结果，与其他常用方法对比均有很好的表现。综上所述，本文提出的特征表征方法具有较强的提取关键特征并学习的能力，并且在联邦学习等研究上也有不错的表现。

参考文献：

[1] 安云鹤，李宝明，李越，等. 癌症基因组测序方案制定的研究进展[J]. 中国生物工程杂志，2014，34(11)：9-17.
An Yun-he，Li Bao-ming，Li Yue，et al. Progress in cancer genome-sequencing study design[J]. China Biotechnology，2014，34(11)：9-17.

[2] 周丰丰，张亦弛. 基于稀疏自编码器的无监督特征工程算法 BioSAE[J]. 吉林大学学报：工学版，2022，52(7)：1645-1656.
Zhou Feng-feng，Zhang Yi-chi. Unsupervised feature engineering algorithm BioSAE based on sparse autoencoder[J]. Journal of Jilin University (Engineering and Technology Edition)，2022，52(7)：1645-1656.

[3] Chen X Y，Yu Y Z W，Zheng H Y，et al. Single-cell transcriptome analysis reveals dynamic changes of the preclinical A549 cancer models，and the mechanism of dacomitinib[J]. European Journal of Pharmacology，2023，960：No. 176046.

[4] 白天，周春光，王喆，等. 代谢组学中机器学习研究进展[J]. 吉林大学学报：信息科学版，2008，26(2)：163-168.
Bai Tian，Zhou Chun-guang，Wang Zhe，et al. Advances of machine learning in metabonomics[J]. Journal of Jilin University (Information Science Edition)，2008，26(2)：163-168.

[5] 高美虹，尚学群. 利用人工智能预测癌症的易感性、复发性和生存期[J]. 生物化学与生物物理进展，2022，49(9)：1687-1702.
Gao Mei-hong，Shang Xue-qun. Artificial intelligence-based prediction for cancer susceptibility，recurrence and survival[J]. Progress in Biochemistry and Biophysics，2022，49(9)：1687-1702.

[6] 刘富，梁艺馨，侯涛，等. 模糊 c-harmonic 均值算法在不平衡数据上改进[J]. 吉林大学学报：工学版，2021，51(4)：1447-1453.
Liu Fu，Liang Yi-xin，Hou Tao，et al. Improvement of fuzzy c-harmonic mean algorithm on unbalanced data[J]. Journal of Jilin University (Engineering and Technology Edition)，2021，51(4)：1447-1453.

[7] 章莓，潘飞燕，章卫国，等. 高通量测序在无创产前遗传学诊断中的应用价值[J]. 中国卫生检验杂志，2022，32(10)：1249-1253.
Zhang Yang，Pan Fei-yan，Zhang Wei-guo，et al. Application value of high -throughput sequencing non-invasive prenatal testing in prenatal genetic diagnosis[J]. Chin J Health Lab Tec，2022，32(10)：1249-1253.

[8] 方朝剑，胡新荣. 基于模糊近似度的隐私敏感数据过滤算法[J]. 吉林大学学报：工学版，2023，53(4)：1174-1180.
Fang Chao-jian，Hu Xin-rong. Privacy-sensitive data filtering algorithm based on fuzzy approximation[J]. Journal of Jilin University (Engineering and Technology Edition)，2023，53(4)：1174-1180.

[9] 张浩，李海鹏，彭国琴，等. 多层次特征融合表征的图像情感识别[J]. 计算机辅助设计与图形学学报，2023，35：1-11.
Zhang Hao，Li Hai-peng，Peng Guo-qin，et al. Im-

age emotion recognition via fusion multi-level representations[J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35: 1-11.

[10] Alain L U, Melissa S, Péter O, et al. Transcriptome-based identification of novel endotypes in adult atopic dermatitis[J]. Allergy, 2022, 77(5): 1486-1498.

[11] Sun M, Ding T, Tang X Q, et al. An efficient mixed-model for screening differentially expressed genes of breast cancer based on LR-RF[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, 16(1): 124-130.

[12] Ball M W, Gorin M A, Drake C G, et al. The landscape of whole-genome alterations and pathologic features in genitourinary malignancies: An analysis of the cancer genome atlas[J]. European Urology Focus, 2017, 3(6): 584-589.

[13] Bourgade R, Rabilloud N, Perennec T, et al. Deep learning for detecting BRCA mutations in high-grade ovarian cancer based on an innovative tumor segmentation method from whole slide images[J]. Modern Pathology, 2023, 36(11): No. 100304.

[14] Yildirimtepe C F, Ercan C. RGS10 suppression by DNA methylation is associated with low survival rates in colorectal carcinoma[J]. Pathology - Research and Practice, 2022, 236: No. 154007.

[15] Aysegul C, Cenk A A, Yalcin A K. Novel molecular signatures and potential therapeutics in renal cell carcinomas: Insights from a comparative analysis of subtypes[J]. Genomics, 2020, 112(5): 3166-3178.

[16] Turab N A A, Murtaza R S A, Imtaiyaz H M. Pan-cancer analysis of Chromobox (CBX) genes for prognostic significance and cancer classification[J]. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 2023, 1869(1): No. 166561.

[17] Goodfellow I J, Jean P A, Mehdi M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014: 2672-2680.

[18] Kingma D P. Welling M. Auto-encoding variational Bayes[J/OL]. [2023-10-22]. https://arxiv.org/pdf/1312.6114v5

[19] Precup D, Teh Y W. Wasserstein Generative Adversarial Networks[C]//Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017, 70: 214-223.

[20] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique [J]. J. Artif. Int. Res.,2002, 16(1): 321-357.

[21] 周丰丰, 孙燕杰, 范雨思. 基于miRNA组学的数据增强算法[J]. 电子科技大学学报, 2023, 52(2): 182-187.
Zhou Feng-feng, Sun Yan-jie, Fan Yu-si. Data augmentation algorithm for miRNA omics-based classifications[J]. Journal of University of Electronic Science and Technology of China, 2023, 52(2): 182-187.