云南民族大学学报(自然科学版),2024,33(6):753 - 759　　CN 53 - 1192/N　ISSN 1672 - 8513
doi:10.3969/j.issn.1672 - 8513.2024.06.012　　http://ynmz.cbpt.cnki.net

基于多维度卷积神经网络的 m7G 位点识别

王　煜,李慧敏,唐　轶,胡　梦,陈鹏辉
(云南民族大学 数学与计算机科学学院,云南 昆明, 650500)

摘要: N7 - 甲基鸟苷(N7 - methylguanosine,m7G)修饰在 RNA 修饰中普遍存在,识别 m7G 位点对认识 m7G 功能和深入了解人类疾病具有重要意义.目前关于 m7G 位点的识别方法大多基于传统机器学习,需要手动输入筛选最优特征,存在特征冗余问题.为了解决以上问题,提出一种多维度卷积神经网络,该方法基于卷积神经网络构建,并在卷积的基础上增加空间空洞卷积层,采用空洞空间卷积池化金字塔模块获得多尺度序列信息特征,以扩大模型的感受野,使得提取的特征更加全面.基于相同的 m7G 位点序列数据,将多维度卷积神经网络模型的 m7G 位点预测能力与几种已有算法进行比较,结果表明,多维度卷积神经网络模型的预测性能优于现有算法.

关键词: 多维度卷积神经网络;空洞卷积;m7G 甲基化;深度学习
中图分类号:TP391;Q52　　**文献标志码:**A　　**文章编号:**1672 - 8513(2024)06 - 0753 - 07

RNA 中存在超过 100 种的化学修饰,其中甲基化修饰是 RNA 修饰的主要形式之一,约占 RNA 修饰总量的三分之二[1].当前关于 RNA 甲基化修饰的研究主要集中于 N6 - 甲基腺苷、N1 - 甲基腺苷和 5 - 甲基胞嘧啶等. N7 - 甲基鸟苷 (m7G)是近几年开始研究的一种 RNA 修饰,广泛分布于 tRNA、rRNA 以及真核生物 mRNA 的 5'帽子区[2].研究发现 m7G 在基因表达调控中具有重要的作用,存在于 mRNA 生命周期的多个阶段,如 RNA 剪接[3],mRNA 核输出[4]和翻译[5]等. m7G 还和人类疾病相关,如: m7G 位点可以用来分析患者的基因表达[6],m7G 调控抑癌因子 let - 7e miRNA 和结构性转录因子 HMGA2 轴在结肠癌中发挥抑癌作用[7],m7G 甲基转移酶 METTL1 可以促进缺血后血管生成[8].因此识别 m7G 位点对认识 m7G 功能和深入了解人类疾病具有重要意义.尽管高通量实验方法,如 AlkAniline - Seq[9]、miCLIP - seq[10]等能够精准识别 m7G 位点,但此类方法成本较高且不适用于大数据量的 m7G 位点检测.

随着智能计算技术的发展,近期涌现出大量识别 m7G 位点的方法. Chen 等[11]开发了 iRNA - m7G 模型,该方法融合序列特征,利用支持向量机(support vector machines,SVM)来识别 m7G 位点. Liu 等[12]采用不同的特征提取方法,通过随机森林(random forest,RF)分类器预测 m7G 位点,得到了优于 iRNA - m7G 的 m7G Predictor. Bi 等[13]提出 XG - m7G,采用 6 种特征编码器,选择最优的特征集,应用极端梯度增强(extreme gradient boosting,XGBoost)算法作为分类器来识别 m7G 位点. Yang 等[14]提出了综合特征得到最优特征子集,使用 SVM 作为分类器进行预测. Dai 等[15]通过使用 RNA 序列编码的迭代特征表示算法,并以有监督的迭代方式提高特征表示能力,使用 SVM、XGBoost、RF 和对数几率回归(logistic regression,LR)4 种分类器进行对比,提出能够更好预测 m7G 位点的方法 m7G - IFL.虽然以上方法能有效检测 m7G 位点,但其序列编码都停留在序列基础上,算法的改进都是通过改变特征选取,并对比多种特征选取最优特征来提高准确率.

此外,以上分类器方法都属于传统机器学习范畴,需要手动输入特征,一方面导致特征冗余,另一方面手动提取特征具有复杂性.为了解决以上问题,能够自动提取特征的深度学习算法不断发展且被广泛使用,在分类问题和回归问题都取得了较好的结果.深度学习中的常见网络有卷积神经网络(convolutional neural net-

收稿日期:2022 - 07 - 20.
基金项目:国家自然科学基金(61866040);云南省研究生优质课程建设项目(云学位[2022]8 号);云南民族大学数学与计算机科学学院研究生科研项目(SJXY - 2021 - 015).
作者简介:王煜(1997 -),女,硕士. 主要从事大数据分析研究.
通信作者:李慧敏(1980 -),女,博士,教授,硕士生导师. 主要从事生物信息学与应用统计学研究.

work,CNN)、循环神经网络(recurrent neural network,RNN)和长短时记忆网络(long short term memory,LSTM)等. Ning 等[16]将双向 LSTM 与全连接网络相结合应用于 m7G 位点识别,设计了一种名为 m7G - DLSTM 的方法,但是该方法没能充分考虑数据集位点位置. 数据集中 m7G 位点位于序列中心位置,双向 LSTM 网络和双向 RNN 更关注序列的起点与终点信息. 而 CNN 能有序地识别特征,更适用于对序列等空间数据进行分类,且与 RNN 相比,CNN 可以在所有元素上完全并行计算,从而更好地利用 GPU 硬件[17],所以使用 CNN 来识别位点. 由于普通 CNN 中卷积层和池化层堆叠,容易损失较多数据信息,而空洞空间卷积池化金字塔(atrousspatial pyramid pooling,ASPP)模块能够通过增大感受野(reception field,RF)多维度有效提取信息,减少信息丢失. 因此,提出基于 CNN 和 ASPP 相结合的多维度卷积神经网络(multi - dimensional convolutional neural network,MDCCN),以更全面地提取序列特征,进而提高模型的预测准确率.

1 相关理论

1.1 CNN 网络结构与原理

深度学习中的代表算法 CNN 是一种具有局部连接、权值共享等特点的深层前馈神经网络,通过多个卷积核不断提取特征,进而实现图像分类、自然语言处理等功能. CNN 主要由输入层、卷积层、池化层、压平层、遗忘层和全连接层构成,其结构如图 1 所示.

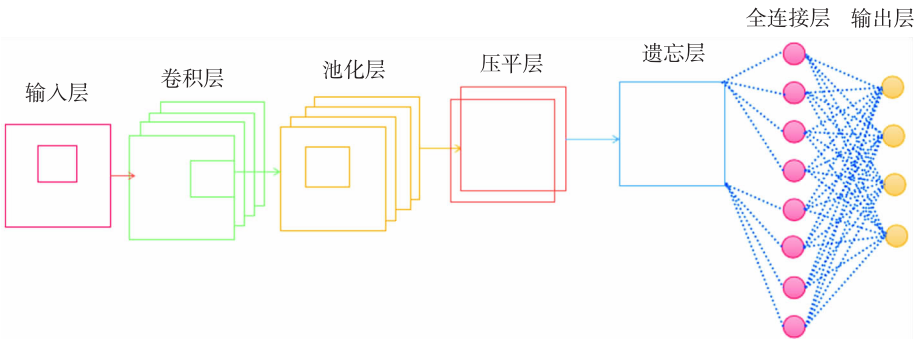


图 1 CNN 结构图

卷积层对数据进行特征的抽取,卷积层中的卷积核在输入数据上进行滑动,与每一个位置上的数据进行点乘运算,得到的输出为特征图. a 表示权重,b 表示偏置,卷积操作如公式(1)所示.

$$C_i = f(a \times X_{i:i+h-1} + b). \tag{1}$$

池化层通过使用区域总体特征代替网络在该区域的输出,从而达到减少网络参数并减小计算量的目的,避免过拟合问题.

压平层实现二维数据的一维化. 遗忘层通过设置参数将一些权重值临时隐藏,缓解过拟合的发生,在一定程度上达到正则化的效果. 全连接层负责完成分类任务,对数据进行输出并得到分类结果,使用 Sigmoid 函数输出分类概率值,以 s 表示模型上一层的输出,Sigmoid 函数如公式(2)所示.

$$g(s) = \frac{1}{1 + e^{-s}}. \tag{2}$$

1.2 ASPP 结构及原理

感受野 RF 是卷积核在特征图上映射到的区域,越大的 RF 包含越多的原始图像特征,特征越具体越有利于模型识别预测. 常用的增大 RF 的方法有卷积时使用较大卷积核或池化时采用较大的步长等,但是卷积核太大会导致计算量大,增大池化步长会损失分辨率. 为了解决以上问题,ASPP[18]被提出. ASPP 通过在标准卷积核中注入空洞,以此来增加模型的 RF.

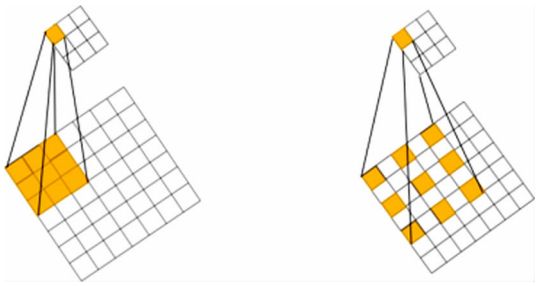
空洞率是在普通卷积的基础上,使相邻权重之间的间隔为空洞率 - 1. 如图 2 所示,图(a)是常见的普通卷积,其空洞率默认为 1,图(b)是空洞率为 2 的卷积,相当于在卷积核中注入 1 个空洞,即在卷积核周围填充权重为 0 的空洞点,将 3 × 3 的卷积核扩张为 5 × 5 的卷积核,增大了 RF 并提高模型分类能力. 设置不同空洞率可以获取不同大小的 RF,从而使得 ASPP 能提取到不同维度的特征. 设空洞率为 d,卷积核大小为 K,空

洞卷积核大小为 K_a 与 K 的关系为:

$$K_a = d \times (K-1)+1. \tag{3}$$

ASPP 将多个卷积层并联在一起,通过设置不同的空洞率,使得卷积层在相同参数下能够得到更大的 RF,相当于使用多个滤波器来探测原始图像,从而在多个尺度上捕获更多上下文信息[18]. ASPP 中的连接层将不同卷积层处理的结果特征图进行融合,形成一个特征更全面的矩阵. 将 ASPP 应用于位点识别中,来实现多维度特征融合. 设一维输入序列为 X_i,卷积核尺寸为 K,滤波器为 w,空洞率为 d,空洞卷积输出 Y_i 定义为:

$$Y_i = \sum_{k=1}^{K} X_{i+d\times k} w. \tag{4}$$

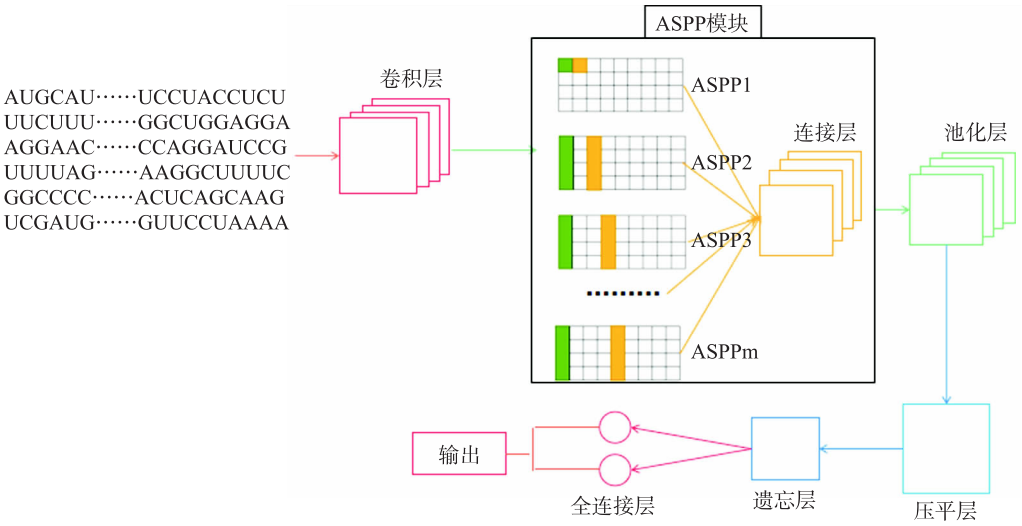


(a)普通卷积　　(b)空洞率 =2 的 ASPP

图 2　ASPP 与 CNN 对比图

1.3 MDCCN 模型

RNA 序列相邻碱基之间具有相关性,进而导致更远碱基之间的关联[19],为了更好地识别序列相邻碱基之间以及碱基与远程碱基之间的关联,将一维 CNN(CNN1D)与 ASPP 相结合,提出了 MDCCN 模型,应用于 m7G 位点识别. MDCCN 与普通 CNN 的主要区别是在卷积层与池化层之间加入 ASPP 模块. 模型结构如图 3 所示.



ASPP 模块中绿色和黄色分别代表卷积核第一次和第二次作用的位置和大小

图 3　多维度模型结构图

MDCNN 模型具体构建过程如下:

(1)将数据中的 RNA 序列编码成网络可以识别的形式输入到卷积层中,并设置卷积层卷积核尺寸. 在卷积层提取信息后,将数据使用批标准化[20](batch normalization,BN)进行归一化,可以加快模型训练时的收敛速度,避免梯度爆炸或消失,并且起到一定的正则化作用. d 表示 BN 层的输入维度,即归一化的上一层卷积层的输出维度,设 $X=(X^{(1)}\cdots X^{(d)})$,$X^{(i)}$ 表示第 i 个维度,对每个维度标准化.

$$\hat{X}^{(K)} = \frac{X^{(K)} - E[X^{(K)}]}{\sqrt{\mathrm{Var}[X^{(K)}]}}. \tag{5}$$

(2)将卷积层输出的特征输入 ASPP 模块中,捕获更长序列间碱基的关联信息,来进行多维度上下文碱基信息提取. 将数据分别输入 m 个不同空洞率的并列卷积层中,通过空洞率来提取不同间隔的碱基信息. 例如,普通卷积核(空洞率为 1)表示提取相邻碱基(间隔 =0)的碱基信息,空洞率为 2 表示提取间隔为 1 的碱基信息,而空洞率为 3 表示提取间隔为 2 的碱基信息. 根据不同空洞率值提取不同尺寸的特征图之后,在连接层中将 m 个特征图融合.

(3)将融合后的特征按照普通 CNN 流程进行池化层、压平层、遗忘层和全连接层的操作,并在全连接层按照 Sigmoid 函数的输出分类. 当输出值大于等于 0.5 时预测为正样本(m7G 甲基化位点),当输出值小于 0.5 时预测为负样本(非 m7G 甲基化位点).

2 实验设计

2.1 数据集

本研究中,人的 m7G 位点数据来源于 Zhang 等[21]的文献,数据从其文献中下载,其中序列长度为41 bp,中间位置是 m7G 位点. 为了避免冗余,减少同源性偏倚[22],采用 CD-HIT 程序去除序列相似性大于 80% 的样本,最终获得 741 条序列作为阳性样本[23]. 阴性数据使用 Chen 等[10]通过实验验证不含 m7G 位点,长度为 41 bp 的 741 条序列. 最终获得 741 条正负样本长度都为 41 bp 的数据集. 为使模型性能更加稳定,采用 10 折交叉验证法将阳性样本和阴性样本随机划分为训练集、验证集和测试集. 具体方法为:将总样本数据集随机划分为 10 等份,其中 2 份作为测试集,2 份作为验证集,其余 6 份作为训练集.

2.2 编码方式

序列数据不能直接被深度学习算法识别,因此在输入网络之前需要对 RNA 序列进行编码转换. 在回归、分类、聚类等机器学习算法中,独热编码都取得了良好的效果,所以使用独热编码来对碱基进行编码. 分别将 A、U、C、G 编码为 $A=[1,0,0,0]^{\mathrm{T}}$,$U=[0,1,0,0]^{\mathrm{T}}$,$C=[0,0,1,0]^{\mathrm{T}}$,$G=[0,0,0,1]^{\mathrm{T}}$. 则序列可以表示为一个维度为 4×41 的矩阵.

2.3 参数设置

网络结构分为 CNN 和 ASPP 模块. CNN 包括一个卷积层、一个池化层、一个压平层、一个遗忘层和一个全连接层. 在卷积层设置 L_2 正则化及偏差量化,池化层采用最大池化. 加入遗忘参数为 0.25 的遗忘层,降低过拟合. ASPP 模块由 4 个具有不同空洞率的卷积层并联和 1 个连接层组成,卷积层卷积核大小和空洞率由常用的参数进行组合,选择最优组合得到,连接层将以上 4 个空洞卷积层得到特征融合.

将碱基做为数据集的最小粒度,设置卷积核的大小与碱基编码的维度一致,将卷积核的大小设为 4×1,将矩阵输入卷积层进行卷积. 模型的学习率设置为 0.01,迭代次数设置为 100,使模型得到最佳的训练效果. 使用非线性函数 ELU(公式(6))作为激活函数,能使模型具有更强的特征表达能力,同时使网络具有稀疏性.

$$\mathrm{ELU}(X)=\begin{cases}x, x>0;\\ \alpha(\mathrm{e}^{x}-1), x\leqslant 0.\end{cases} \tag{6}$$

按照 ELU 函数中 α 取值的常用做法,将公式(6)中取 $\alpha=1$.

最后采用交叉熵函数描述模型预测值与真实值之间的差距.

2.4 评价指标

使用受试者工作特征(receiver operating characteristic,ROC)曲线下与坐标轴围成的面积(area under curve,AUC)、准确率(accuracy,Acc)、特异度(specificity,Sp)、灵敏度(sensitivity,Sn)、马修斯相关系数(matthews correlation coefficient,MCC)和损失值(lossvalue)几个评价指标来比较模型的性能. TP 和 TN 表示分类正确的 2 类:分别表示把正样本和负样本识别正确,FP 和 FN 表示分类错误的 2 类:分别表示把正样本识别为负样本和把负样本识别为正样本. 各个评价指标的公式如(7)~(11)所示.

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}\times100\%. \tag{7}$$

$$\mathrm{Sp}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\times100\%. \tag{8}$$

$$\mathrm{Sn}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\times100\%. \tag{9}$$

$$\mathrm{MCC}=\frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})}}. \tag{10}$$

$$\text{Lossvalue} = -\frac{1}{\text{outputsize}}\sum_{i=1}^{\text{outputsize}} y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i). \tag{11}$$

其中 outputsize 表示输出尺寸. AUC 是判断二分类预测模型优劣的标准,越接近 1,说明模型预测效果越好. Acc 表示所有类别中模型正确分类的比例,越接近 1 说明正确识别概率越高. MCC 是一个描述实际分类与预测分类之间的相关系数,越接近 1 说明相关性越高. Sp、Sn,越接近 1 表示在正样本、负样本集分类正确的概率越高. Lossvalue 越接近 0,说明损失越少.

3　结果分析

3.1　ASPP 模块参数结果对比分析

采用 10 折交叉验证法将阳性样本和阴性样本随机划分为训练集、验证集和测试集. 将以上过程重复 10 次,最后以这 10 次所得 AUC、Acc 的平均值为依据对模型性能及参数进行评估. ASPP 模块常用空洞率范围为 $d \in (1,5)$,文献[17]发现较小的空洞率效果更好,空洞率越大,采样的间隔越大,滤波器中无用的权重就越多. 因此对空洞率 d 在 1 ~5 内进行网格化搜索,当空洞率达到预期目标时(AUC >95%, Acc >95%),停止搜索,取对应的空洞率为 ASPP 模块层参数. 如果达不到预期目标,则取网格化搜索结果中最大 AUC 值对应的空洞率为 ASPP 模块层参数. 研究发现,当空洞率分别为 1、2、3、4 时,AUC 和 Acc 值已达到预期目标,因此取 1、2、3、4 分别作为 ASPP 模块中 4 个空洞卷积层的空洞率. 表 1 同时也列出了几个比较接近预期目标的空洞率取值. 令 K 表示卷积核尺寸,S 表示卷积层步长,RF 的计算公式如下.

$$\text{RF}_i = \text{RF}_{i-1} + (K-1) \times \sum_{k=1}^{i-1} S. \tag{12}$$

由公式(3)和(12)计算可得,在 ASPP 模块中,RF 不断加大,有利于分类器得到更好的效果.

表 1　ASPP 网络层参数及所得结果

ASPP1		ASPP2		ASPP3		ASPP4		AUC/%	Acc/%
卷积核	空洞率	卷积核	空洞率	卷积核	空洞率	卷积核	空洞率		
1 ×1	1	4 ×1	1	4 ×1	2	4 ×1	3	98.79	92.33
1 ×1	1	4 ×1	1	4 ×1	2	4 ×1	4	98.76	92.75
1 ×1	1	4 ×1	1	4 ×1	3	4 ×1	5	97.80	92.75
1 ×1	1	4 ×1	2	4 ×1	3	4 ×1	4	99.40	95.74

3.2　普通 CNN 与 MDCNN 对比分析

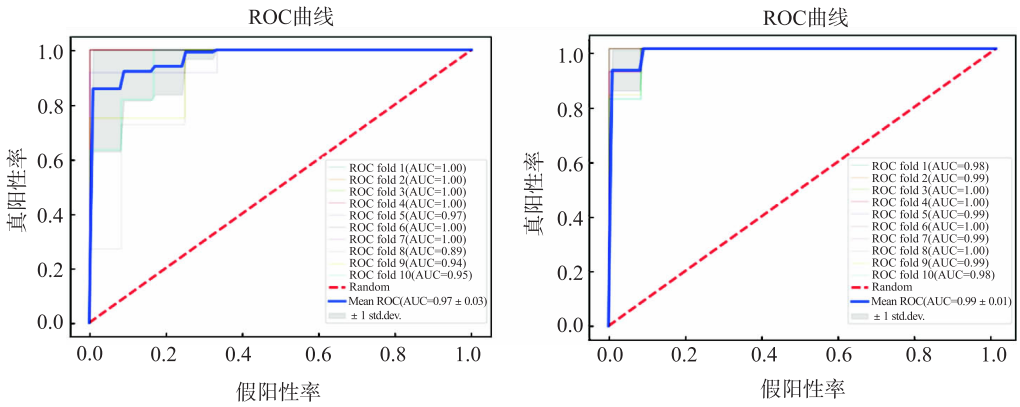
为了对比加入的 ASPP 模块的作用,最后将不同网络模型中得到的值进行对比,对比结果见表 2. 可以看出,MDCNN 中的 Sp、Sn、Acc、MCC 和 AUC 值分别比 CNN 中的对应指标高 1.60%、1.67%、1.69%、2.71% 和 1.73%. ROC 曲线如图 4 所示,图 4(a)为 CNN 的 ROC 曲线图,AUC 的值为(0.97 ±0.03)%,图 4(b)为 MDCNN 的 ROC 曲线图. AUC 的值为(0.99 ±0.01)%,通过 MDCNN 与普通 CNN 上的结果对比,可以发现 MDCNN 使得所有性能指标值都有一定提高,加入的 ASPP 模块有效改进了 CNN,提升了模型预测的准确率.

表 2　模型结果对比　　%

方法	Sn	Sp	Acc	MCC	AUC	Lossvalue
普通 CNN	93.93	94.16	94.05	89.13	97.67	20.68
MDCNN	95.53	95.83	95.74	91.84	99.40	10.39

3.3　MDCNN 与其他方法对比分析

为了验证本研究模型的优劣,将 MDCNN 模型与现有的几种 m7G 位点识别方法进行比较,这几种方法与本研究使用的数据集相同,结果更具有说服力,比较结果见表 3. 从表 3 可以看出,除了 Yang 的模型[14]在 MCC 上,m7G - DLSTM 在 AUC 上未报道,无法比较以外,MDCNN 在各项指标上均为最优. 特别是 MCC 值达到了 91.84%,比目前在该指标上效果最好的 m7G - DLSTM 提高了 4.54%;而 AUC 值则高达 99.4%,比目前 AUC 效果最好的 Yang 的模型提高了 1.20%. 为了更直观地进行模型对比,根据每个模型的性能指标结



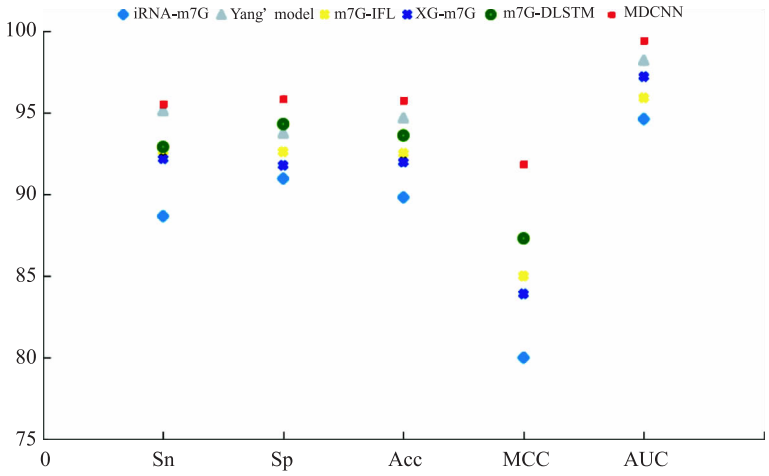
(a)卷积神经网络模型的受试者曲线 (b)多维度卷积神经网络模型的受试者曲线

图 4 普通 CNN 与 MDCNN 的 ROC 曲线图

果做出散点图(图 5). 由图 5 可以看出 MDCNN 的各项性能指标均达到最高,表明 MDCNN 模型优于目前的 m7G 位点识别模型.

表 3 方法对比 %

方法	Sn	Sp	Acc	MCC	AUC
iRNA – m7G[11]	88. 66	90. 96	89. 81	80. 00	94. 60
XG – m7G[13]	92. 17	91. 77	91. 97	83. 90	97. 20
Yang'model[14]	95. 11	93. 74	94. 67	–	98. 20
m7G – IFL[15]	92. 40	92. 60	92. 50	85. 00	95. 90
m7G – DLSTM[16]	92. 90	94. 30	93. 60	87. 30	–
MDCNN	95. 53	95. 83	95. 74	91. 84	99. 40



其中 Yang' model 缺少 MCC 值,m7G – DLSTM 缺少 AUC 值,因此未能在图中显示.

图 5 不同模型性能对比图

4 结语

CNN 是分类预测常用的模型,本文考虑到 RNA 序列相邻碱基之间与更远碱基之间的关联性,为了更好地识别序列相邻碱基之间以及碱基与远程碱基之间的关联,在 CNN 中加入能够多维度识别特征信息的 AS-PP 模块,增大 RF,多维度识别序列特征. 通过在人 m7G 位点数据集上进行实验,发现 MDCNN 网络模型在各项评价指标上不仅优于普通的 CNN 模型,而且优于现有大部分 m7G 位点识别模型,提高了 m7G 位点识别的准确率.

参考文献:

[1] 陈宇晟,杨莹. RNA 修饰类型及调控蛋白[J]. 生命科学,2018,30(4): 391 – 406.

[2] FURUICHI Y. Discovery of m7G – cap in eukaryotic mRNAs[J]. Proceedings of the Japan Academy, Series B,2015,91(8): 394 – 409.

[3] LINDSTROM D L,SQUAZZO S L,MUSTER N,et al. Dual roles for Spt5 in pre – mRNA processing and transcription elongation revealed by identification of Spt5 – associated proteins[J]. Molecular and Cellular Biology,2003,23(4): 1368 – 1378.

[4] LEWIS J D,IZAURFLDE E. The role of the cap structure in RNA processing and nuclear export[J]. European Journal of Biochemistry,1997,247(2): 461 – 469.

[5] MURTHY K G K,PARK P,MANLEY J L. A nuclear micrococcal – sensitive, ATP – dependent exoribonuclease degrades uncapped but not capped RNA substratesx[J]. Nucleic Acids Research,1991,19(10): 2685 – 2692.

[6] KANAMORI – KATAYAMA M,ITOH M,KAWAJI H,et al. Unamplified cap analysis of gene expression on a single – molecule sequencer[J]. Genome Research,2011,21(7): 1150 – 1159.

[7] LIU Y,ZHANG Y,CHI Q,et al. Methyltransferase – like 1(METTL1) served as a tumor suppressor in colon cancer by activating 7 – methyguanosine(m7G) regulated let – 7e miRNA/HMGA2 axis[J]. Life Sciences,2020,249: 117480.

[8] ZHAO Y,KONG L,PEI Z,et al. m7G methyltransferase METTL1 promotes post – ischemic angiogenesis *via* promoting VEGFA mRNA translation[J]. Frontiers in Cell and Developmental Biology,2021,9: 642080.

[9] MARCHAND V,AYADI L,ERNST F G M,et al. AlkAniline – Seq:profiling of m7G and m3C RNA modifications at single nucleotide resolution[J]. Angewandte Chemie International Edition,2018,57(51): 16785 – 16790.

[10] MALBEC L,ZHANG T,CHEN Y S,et al. Dynamic methylome of internal mRNA N7 – methylguanosine and its regulatory role in translation[J]. Cell Research,2019,29(11): 927 – 941.

[11] CHEN W,FENG P,SONG X,et al. iRNA – m7G: identifying N7 – methylguanosine sites by fusing multiple features[J]. Molecular Therapy – Nucleic Acids,2019,18: 269 – 274.

[12] LIU X,LIU Z,MAO X,et al. m7GPredictor: an improved machine learning – based model for predicting internal m7G modifications using sequence properties[J]. Analytical Biochemistry,2020,609: 113905.

[13] BI Y,XIANG D,GE Z,et al. An interpretable prediction model for identifying N7 – methylguanosine sites based on XGBoost and SHAP[J]. Molecular Therapy – Nucleic Acids,2020,22: 362 – 372.

[14] YANG Y H,MA C,WANG J S,et al. Prediction of N7 – methylguanosine sites in human RNA based on optimal sequence features [J]. Genomics,2020,112(6): 4342 – 4347.

[15] DAI C,FENG P,CUI L,et al. Iterative feature representation algorithm to improve the predictive performance of N7 – methylguanosine sites[J]. Briefings in Bioinformatics,2021,22(4): bbaa278.

[16] NING Q,SHENG M. m7G – DLSTM: intergrating directional Double – LSTM and fully connected network for RNA N7 – methlguanosine sites prediction in human[J]. Chemometrics and Intelligent Laboratory Systems,2021,217: 104398.

[17] ZHANG Y,FANG Y,WEI D X. Deep keyphrase generation with a convolutional sequence to sequence model[C]//2017 4th International Conference on Systems and Informatics(ICSAI). IEEE,2017: 1477 – 1485.

[18] CHEN L C,PAPANDREOU G,KOKKINOS I,et al. Deeplab: semantic image segmentation with deep convolutional nets,atrous convolution,and fully connected crfs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,40(4): 834 – 848.

[19] 樊龙江. 生物信息学札记[M]. 杭州:浙江大学出版社,2010.

[20] IOFFE S,SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//International Conference on Machine Learning. PMLR,2015: 448 – 456.

[21] ZHANG L S,LIU C,MA H,et al. Transcriptome – wide mapping of internal N7 – methylguanosine methylome in mammalian mRNA[J]. Molecular Cell,2019,74(6): 1304 – 1316.

[22] ZOU Q,LIN G,JIANG X,et al. Sequence clustering in bioinformatics: an empirical study[J]. Briefings in Bioinformatics,2020,21(1): 1 – 10.

[23] FU L,NIU B,ZHU Z,et al. CD – HIT: accelerated for clustering the next – generation sequencing data[J]. Bioinformatics,2012,28(23): 3150 – 3152.

(下转第 785 页)