文章编号:1000-2367(2021)02-0031-09 DOI:10.16366/j.cnki.1000-2367.2021.02.005

深度神经网络蛋白质溶解性预测模型设计

王鲜芳[1a,2],刘依锋[1b],杜志勇[1c],朱命冬[1a],李启萌[2]

(1.河南工学院 a.计算机科学与技术学院;b.管理学院;c.自动化学院,
河南 新乡 453003;2.河南师范大学 计算机与信息工程学院,河南 新乡 453007)

摘 要:蛋白质溶解性是生物信息学领域的重要研究课题,通过分析蛋白质溶解性数据,结合特征提取和深度学习技术,设计多种卷积神经网络预测蛋白质溶解性的模型.使用 CD-HIT 对蛋白质原始数据进行降噪,并利用 G-gap 对每个样本进行张量化处理,得到适用于卷积神经网络的特征数据,作为模型其中一路网络的输入;为提高模型预测精度,对每个样本利用 SCRATCH 工具提取 6 维序列特征和 51 维结构特征作为额外特征,作为模型的另一路网络输入.依据数据特点,通过对卷积层的串并联结构调整组合,设计 4 种不同网络模型,实现蛋白质溶解性预测.通过对比试验确定网络结构和参数,结果表明基于深度双路卷积神经网络 DDcCNN(Deep Dual-channel Convolutional Neural Networks)的蛋白质溶解性预测模型整体性能最优,其预测精度、查全率、查准率、MCC(Matthews Correlation Coefficient)等性能指标分别达到 76.31%、65.31%、75.05%、0.55.并通过与基于传统的深度神经网络、支持向量机、随机森林、决策树建立的预测模型以及现有的研究成果进行比较试验,证明了本研究设计的有效性.

关键词:深度双路卷积神经网络;蛋白质溶解性;G-gap 二肽频率;预测模型

中图分类号:TP181 **文献标志码**:A

蛋白质是生命的物质基础,而蛋白质功能是由其分子结构、特征共同决定,其中溶解性是蛋白质关键特征之一[1],因此,研究蛋白质溶解性具有重要的理论和实际意义[2].目前蛋白质可溶性研究方式主要分为两类:试验方法和计算方法.

试验方法是利用大肠杆菌进行特异性表达[3],从而获得蛋白质的可溶性.文献[4]使用大肠杆菌对蛋白质进行特异性表达,但部分蛋白质会形成不溶于水的包涵体[1]进而影响试验进度.针对上述问题,研究者通常使用弱启动子或强变性剂[5]、较低温度[6]或优化其他表达条件[7]等方法来降低包涵体的产生.尽管上述方法有一定的有效性,但这些方法不仅需要昂贵的设备,还会耗费研究者大量的时间和精力.

计算方法是一种替代试验方法的重要方式.通过对蛋白质序列数据进行分析计算[8],利用机器学习算法预测蛋白质溶解性[9].常用的机器学习算法主要为支持向量机[10]、神经网络算法[11]、随机森林[12]等方法.CCSOL[13]是 FEDERICO 等人在 2012 年基于 SVM 建立的预测工具,并首次提出来使用疏水性、β 折叠和 α 螺旋作为主要的特征.PaRSnIP[14]是 REDA 等人在 2017 发布的工具,同时提出高比例暴露残基与蛋白质溶解性成正相关关系、由多个组氨酸组成的三肽和三肽片段与蛋白质溶解性呈负相关关系.SOLpro[15]提取一级序列的 23 组特征用于训练两阶段支持向量机(SVM)架构.PROSO II[16]是 PAWEL 等人使用了带有修改 Cauchy 内核的概率密度窗模型的二级逻辑分类器.但现有研究大多使用 SVM 模型为分类器,对大数据处理能力有限、速度较慢.

深度学习是目前人工智能技术的核心领域[17],相对于 SVM 等“浅层学习”,深度学习模型能够获得更多非线性关系[18].卷积神经网络是深度学习的重要构架之一,在图像检测[19]、人脸识别[20]、音频检索[21]等方面

收稿日期:2019-11-22;**修回日期**:2020-01-07.

基金项目:国家自然科学基金(62072157;61802116);河南省自然科学基金(202300410102);河南工学院博士启动项目(KQ2002).

作者简介(通信作者):王鲜芳(1995—),女,河南洛阳人,河南工学院教授,博士,研究方向为数据挖掘、机器学习及应用,E-mail:2wangfang@163.com.

得到了广泛的应用,并取得较好的效果,但是较少应用在蛋白质溶解性预测研究领域.麻省理工学院 SAMEER KHURANA 等[22]在 2018 年构建了 DeepSol 预测模型,主要采用 onehot 对蛋白质序列进行编码,得到 21 * 1 200 特征矩阵,结合 57 维蛋白质额外特征,构建 25 257 维特征向量作为输入,利用 7 个不同大小的卷积核建立浅层并行卷积神经网络模型,用于预测蛋白质溶解性,尽管取得了一定的效果,但这种方法不但存在输入数据维数过于庞大,耗费了相当大的计算资源,还存在浅层卷积核不能学习更深层次关系的问题.

针对上述问题,本研究在对蛋白质数据进行 CD-HIT[23]降噪处理的基础上,利用 G-gap[24-25]二肽频率计算蛋白质特征矩阵,并结合 57 维额外特征构筑卷积神经网络的输入数据;通过对卷积层的串并联结构调整组合,设计多种蛋白质预测模型;通过对比试验结果,确定模型参数,实现蛋白质溶解性预测.并通过与基于支持向量机、K 近邻、决策树、随机森林、深度神经网络等所建预测模型以及现有的研究成果进行对比试验,得到适合蛋白质数据的溶解性预测模型.

1 数据来源及预处理

本研究采用文献[14]的数据集,为确保训练集内序列异质性,使用 CD-HIT[23]设置冗余值为 0.9 减少数据的序列冗余.最终获得 28 972 条可溶性蛋白质和 40 448 条不溶性蛋白质,共 69 420 条蛋白质溶解数据.但所得数据长度从 19 至 1 696 不等,不能满足所建模型张量化的要求,因此需要进一步处理.

1.1 蛋白质数据离散化处理

本研究采用 $S_n=R_1R_2\cdots R_i\cdots R_L$,$(n\leq K)$ 对蛋白质一级结构进行离散化处理,其中 S_n 为第 n 个蛋白质数据,R_i 为蛋白质序列中第 i 个氨基酸残基,L 为序列长度,K 为数据集中蛋白质序列个数,K 为数据预处理后的样本个数(69 420).

1.2 G-gap 特征提取及重构

为达到蛋白质序列长度一致的目的,本研究使用间隔二肽氨基酸频率(G-gap)对蛋白质序列进行表征,得到 1 * 400 维的蛋白质 G-gap 结构特征数据.G-gap 的定义为 $S_{\text{G-gap}}=(v_1^g,v_2^g,v_3^g,\cdots,v_{400}^g)^T$,其中二肽是由两个氨基酸(20 种)脱水缩合形成的聚合物,故形成二肽的个数为 400,v_i^g 表示 G-gap 二肽的第 $i(i=1,2,3,\cdots,400)$ 个特征的频率.v_i^g 为 $v_i^g=n_u^g/\sum_{u=1}^{400}n_u^g=n_u^g/(L-g-1)$,其中 L 表示蛋白质序列的长度,g 为二肽中两个残基相隔的残基个数,n_u^g 表示蛋白质序列中包含第 u 个二肽特征的个数.

由于一维特征会丢失二肽对应的残基位置信息,本研究利用基坐标为 20 种氨基酸将 G-gap 数据(1 * 400)重构为大小 20 * 20 的矩阵,旨在保留位置信息,并作为一路网络的输入.

1.3 额外特征

为提高预测模型的精度,本研究建立模型时采用文献[14]提供的 57 维额外特征设计另外一路网络的输入,这些额外特征主要分为两类:序列特征和结构特征.序列特征包括:序列长度、分子质量、正电荷数、脂肪系数(AIs)、亲水性平均值(GRAVY)、转角氨基酸频率(Fraction turn-forming residues)共 6 维.结构特征利用 SCRATCH 工具[26]得到,主要有 51 维,其中包含 3 维三态 SS、8 维八态 SS、20 维 FER 值(使用 RSA 截止值范围从 0 到 95%、间隔为 5%得到)以及使用 FER 乘以暴露残基的疏水性指数获得另外 20 维特征.数据预处理的完整流程如图 1 所示.

2 蛋白质溶解性预测模型设计

2.1 模型设计

通过对蛋白质数据计算获得 20 * 20 维 G-gap 特征数据矩阵和 57 维额外特征.针对这些数据,本研究利用卷积层连接高维度向量空间特性[18]挖掘隐藏关系,设计 4 种不同架构的卷积层神经网络模型,实现蛋白质溶解性预测.具体如下:

(1)DCNN(Deep Convolutional Neural Network)深度卷积神经网络模型.使用通道 1 对 G-gap 特征进

行多层卷积运算,通道 1 输出数据与额外特征合并后输入隐藏层进行计算,通过输出层得到蛋白质溶解性.设计的 DCNN 模型结构如图 2 所示.

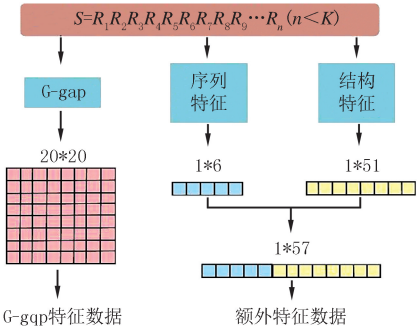

图1　数据预处理过程

Fig.1　The process of data preprocessing

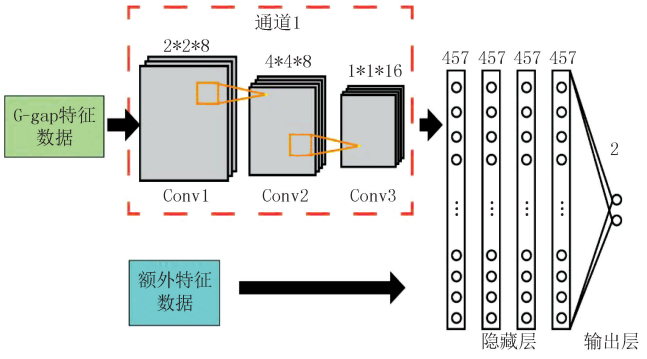

图2　DCNN模型整体结构图

Fig.2　The overall structure of the DCNN model

(2)DDcCNN(Deep Dual-channel Convolutional Neural Networks)深度双路卷积神经网络模型.在 DCNN 模型的基础上,针对额外特征由两种不同特征组成的特点,构建通道 2 进行卷积运算,合并双通道运算结果作为第 1 层隐藏层的输入,经过 4 层隐藏层计算,在输出层得到蛋白质溶解性预测结果.设计的 DDcCNN 模型结构如图 3 所示.

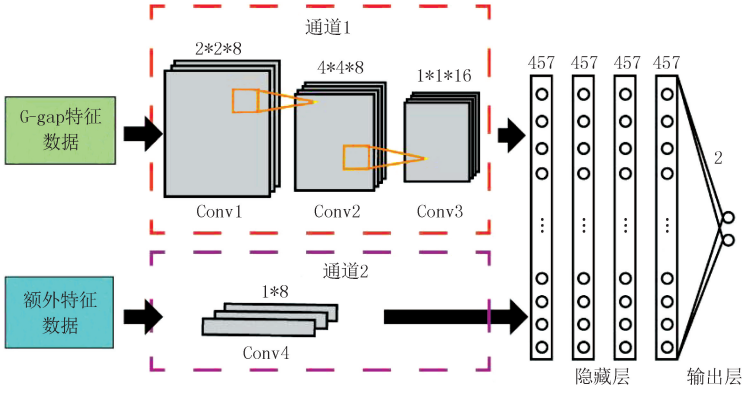

图3　DDcCNN模型整体结构图

Fig.3　The overall structure of the DDcCNN model

(3)DTcCNN(Deep Tri-channel Convolutional Neural Network)深度三路卷积神经网络模型.在 DDcCNN 模型的基础上,针对序列特征和结构特征,设计建立通道 2、通道 3 分别进行 1D 卷积运算,合并三通道运算结果作为第 1 层隐藏层的输入,经过 4 层隐藏层计算,在输出层得到蛋白质的溶解性预测结果.设计的 DTcCNN 模型结构如图 4 所示.

(4)SCNN(Shallow Convolutional Neural Network)浅层并行卷积神经网络模型.在 DDcCNN 模型的基础上保留通道 1,建立 3 个平行 2D 卷积层的浅层卷积神经网络,合并双通道运算结果作为第 1 层隐藏层的输入,经过 4 层隐藏层计算,在输出层得到蛋白质溶解性预测结果.设计的 SCNN 模型结构如图 5 所示.

2.2　模型参数设置

在卷积层中,将卷积运算和 Relu 激励函数定义为一个卷积单元.卷积单元的输入数据进行步长为 1 的卷积计算,并采用 Relu 函数对计算结果进行激励.多个卷积单位以串联方式连接组成卷积通道,双通道结果以并联形式连接.将双通道并联结果输入隐含层进行计算,并对隐含层计算输出数据进行 Softmax 函数激励得到蛋白质溶解性概率.为降低网络训练参数数量和复杂度[27],本研究使用小卷积核设计网络.在模型设计中,分别使用大小为 2 * 2、4 * 4 的卷积核进行隐藏关系计算和大小为 1 * 1 的卷积核进行特征扩展处理.模型中各层参数如表 1 所示.

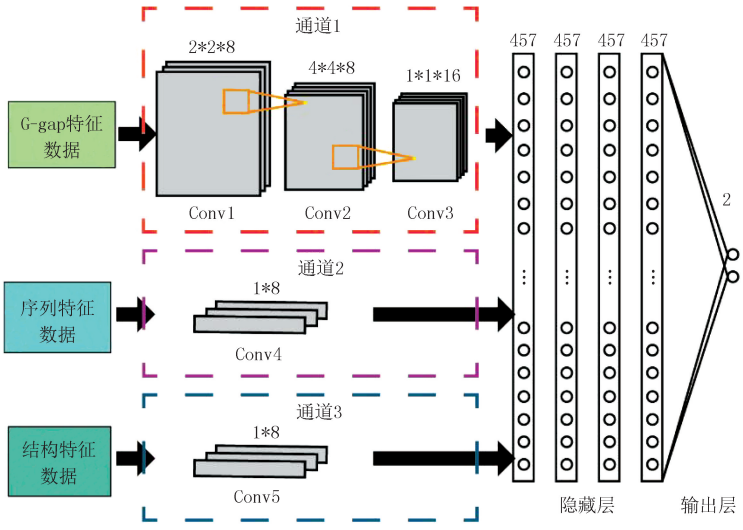

图4 DTcCNN模型整体结构图

Fig. 4 The overall structure of the DTcCNN model

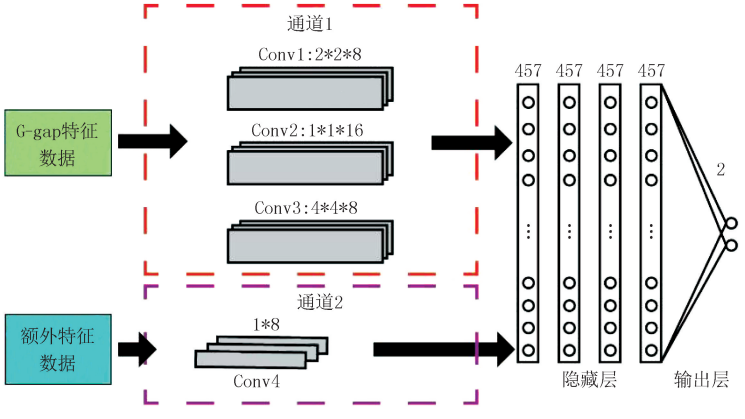

图5 SCNN模型整体结构图

Fig. 5 The overall structure of the SCNN model

表 1 网络模型参数

Tab. 1 The overall parameters of network model

层数	卷积核个数	尺寸	输入	输出	层数	卷积核个数	尺寸	输入	输出
卷积层 1	8	2 * 2	20 * 20 * 1	20 * 20 * 8	隐含层 2	—	457	457	457
卷积层 2	8	4 * 4	20 * 20 * 8	20 * 20 * 8	隐含层 3	—	457	457	457
卷积层 3	16	1 * 1	20 * 20 * 8	20 * 20 * 16	隐含层 4	—	457	457	457
卷积层 4	8	1	57 * 8	57 * 8	输出层	—	2	457	2
隐含层 1	—	457	20 * 20 * 16+57 * 8	457					

2.3 模型训练

在模型训练阶段,利用二元交叉熵作为损失函数,以计算预测值与真实值的差距,其定义为 $CE=-\sum_{n=1}^{N} y^n \ln P(y^n=1 \mid x^n)+(1-y^n)\ln(1-P(y^n=1 \mid x^n))$,其中 x_n 为蛋白质数据集第 n 个数据所提取的特征集合,y_n 为当前蛋白质对应可溶性,N 为训练集合中蛋白质的总数. 设置 dropout 层,随机失活 50% 的神经元防止过拟合[28],并采用 Adam 优化器对计算结果进行训练.在每次迭代动态更新学习率 $L=L_0\times\theta^t$,其中 θ 为学习率衰减因子,t 当前迭代次数.

3　试验过程及结果分析

3.1　评估指标

本研究的评估指标主要包括精确度 $ACC=\frac{TP+TN}{TP+TN+FP+FN}$,查准率 $P=\frac{TP}{TP+FP}$,查全率 $R=\frac{TP}{TP+FN}$,马修斯相关系数(MCC,Matthews Correlation Coefficient)其公式为 $MCC=\frac{(TP\times TN-FP\times FN)}{\sqrt{((TP+FP)(TP+FN)(TN+FP)(TN+FN))}}$,真正例率(敏感度)$TPR=\frac{TP}{TP+FN}$,假正例率 $FPR=\frac{FP}{TN+FP}$,特异度 $TNR=\frac{TN}{FP+TN}$,其中查准率、查全率可绘制 PR 图,真正例率、假正例率可绘制 ROC 图,其中 TP 为真正例、FN 为假反例、FP 为假正例、TN 为真反例.

3.2　DDcCNN 模型试验结果及分析

通过第 2 节算法处理后的数据集个数为 69 420 * (20 * 20,57),运用留出法将数据集 D 划分为 2 个大小不同的互斥子集作为训练集和测试集,且大小比例为 9∶1,如表 2 所示.

表 2　划分后的测试集和训练集

Tab. 2　The number of test set and training set after division

数据集	溶解性	非溶解性	总量	数据集	溶解性	非溶解性	总量
训练集	26 051	36 427	62 478	总量	28 972	40 448	69 420
测试集	2 921	4 021	6 942				

针对不同卷积层、全连接层数进行反复的对比试验以确定最优卷积层数和隐藏层数,具体参数如表 3 所示,试验结果如表 4 所示,其中 DDcCNN_0,…,DDcCNN_6 表示 7 种不同网络模型,C_i 表示卷积层第 i 层卷积核大小,D_i 表示第 i 层全连接层神经元个数. 试验结果如表 4 所示.

表 3　DDcCNN 模型架构参数

Tab. 3　The overall architecture parameters DDcCNN model

参数	模型						
	DDcCNN_0	DDcCNN_1	DDcCNN_2	DDcCNN_3	DDcCNN_4	DDcCNN_5	DDcCNN_6
$C1$	2,2,8	2,2,8	2,2,8	2,2,8	2,2,8	2,2,8	2,2,8
$C2$	—	4,4,8	4,4,8	4,4,8	4,4,8	4,4,8	4,4,8
$C3$	—	—	1,1,16	1,1,16	1,1,16	1,1,16	1,1,26
$D1$	457	457	457	457	457	457	457
$D2$	457	457	—	457	457	457	457
$D3$	457	457	—	—	457	457	457
$D4$	457	457	—	—	—	457	457
$D5$	457	457	—	—	—	—	457

表 4　不同 DDcCNN 模型参数蛋白质溶解性预测结果比较

Tab. 4　The results of contrast experiment of protein solubility prediction by different DDcCNN model parameters

分类器	精确度/%	查全率/%	查准率/%	MCC	分类器	精确度/%	查全率/%	查准率/%	MCC
DDcCNN_0	75.79	64.51	74.23	0.52	DDcCNN_4	76.17	64.48	75.03	0.52
DDcCNN_1	76.20	**66.72**	73.84	0.54	DDcCNN_5	**76.31**	65.31	**75.05**	**0.55**
DDcCNN_2	74.28	58.97	74.24	0.46	DDcCNN_6	76.01	65.00	74.40	0.51
DDcCNN_3	74.94	62.24	73.75	0.48					

从表 4 可以看出,DDcCNN_5 模型相对其他模型总体优势比较明显,其精度为 76.31%、查全率为 65.31%、查准率为 75.05%,MCC 为 0.55,其中精度、查准率、MCC 为最高值,查全率为次高值仅低于 DDcC-

NN_1 模型.为进一步比较不同参数下模型的性能,绘制 ROC 曲线以及 PRC 曲线并计算 AUC 与 ED 值如图 6、图 7、表 5 所示.

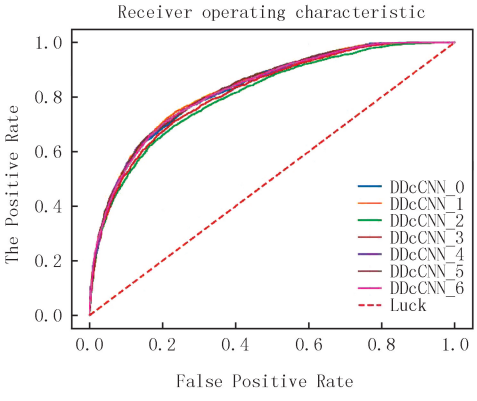

图6 不同网络架构ROC图

Fig.6 ROC curves of different network architectures

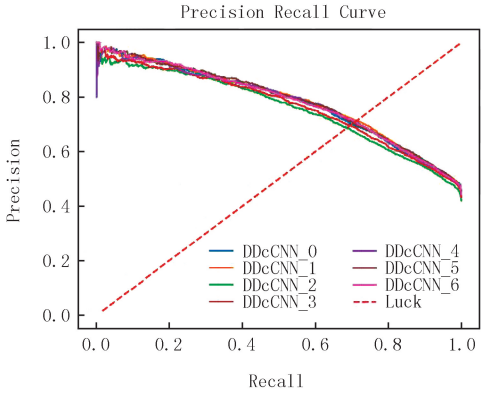

图7 不同网络架构P-R图

Fig.7 P-R curves of different network architectures

表 5 不同网络架构 AUC 与 EB 比较

Tab. 5 Comparison results of AUC and EB of different network architectures

模型	AUC	EB	模型	AUC	EB	模型	AUC	EB	模型	AUC	EB
DDcCNN_0	0.828 0	0.700 3	DDcCNN_2	0.807 7	0.685 8	DDcCNN_4	0.829 2	0.707 9	DDcCNN_6	0.828 6	0.709 3
DDcCNN_1	0.831 7	0.713 1	DDcCNN_3	0.817 7	0.695 5	DDcCNN_5	**0.832 1**	**0.714 4**			

通过比较可知 DDcCNN_5 模型在较大范围内可以包裹其他模型,其中 AUC=0.83,EBP=0.71.而 DDcCNN_1 模型 AUC=0.80,EBP=0.68,可知 DDcCNN_5 的两项系数均高于 DDcCNN_2,故总体性能更优越.

3.3 4 种模型的试验结果比较

为分析本研究所提设计的 4 种不同架构模型 DDcCNN,DCNN,DTcCNN,SCNN 的性能,对其进行了对比试验,主要包括精确度、敏感度、特异性、MCC.结果如表 6 所示.

表 6 不同卷积层模型预测精度

Tab. 6 The results of contrast experiment for prediction accuracy of different convolutional layer models

算法	精确度/%	灵敏度	特异性	MCC	算法	精确度/%	灵敏度	特异性	MCC
DDcCNN	**76.31**	**0.65**	**0.84**	**0.55**	DTcCNN	75.82	0.64	0.83	0.53
DCNN	75.68	0.64	0.83	0.52	SCNN	75.56	0.64	0.83	0.52

由表 6 可知 DCNN,DTcCNN,SCNN 在灵敏度、特异性指标上均为 0.64、0.83,而 DDcCNN 模型相对比高出 0.01,获得 0.65 和 0.84.在精确度、MCC 指标上 DDcCNN 模型获得最高值 76.31、0.55,比次高的 DTcCNN 模型高出 0.49、0.02.由此可知,DDcCNN 模型性能优越于其他架构模型,该模型更适合于蛋白质溶解性预测.

3.4 不同分类器模型的结果比较

为了验证 DDcCNN 模型的优越性,与基于不同分类器模型进行比较试验,主要包括支持向量机 SVM (Support Vector Machines)、决策树 DT(Decision Tree)、随机森林 RF(Random Forest)和深度神经网络 DNN(Deep Neural Network),各种分类器的参数设置如表 7 所示.经过利用相同训练集训练和测试集测试,各个分类器预测结果如表 8 所示.

表 7 分类器参数设置

Tab. 7 The parameter settings for different classifier

分类器	参数	分类器	参数
DNN	Dnn1=457,Dnn2=457,Dnn3=457,Dnn3=457	DT	Critertion="gini",splitter="best"
SVM	C=1.0,kernel='rbf',gamma='auto'	RF	n_estimators=25

表 8　多种分类器蛋白质溶解性预测结果比较

Tab. 8　The results of contrast experiment for protein solubility predictor based on different classifier

分类器	精确度/%	查全率/%	查准率/%	MCC	时间/s	分类器	精确度/%	查全率/%	查准率/%	MCC	时间/s
DDcCNN	**76.31**	65.31	**75.05**	**0.55**	958	RF	68.48	81.14	69.52	0.30	**78**
DNN	75.09	74.27	61.86	0.46	226	DT	67.24	71.47	71.83	0.29	94
SVM	71.47	**94.57**	68.33	0.41	17257						

从表 8 可以看出,DDcCNN 模型在精确度、查准率、MCC 评价指标获得最高值,相对于未使用卷积层的深度神经网络,DDcCNN 模型性能获得提升,其精确度提高 1.22%、查准率提高 13.19%、MCC 提高 0.09,但查全率下降 8.96%,说明卷积层在对蛋白质数据处理上有着正向调节作用.尽管在查全率方面略低于 SVM 模型,但在训练用时比较中,RF 的训练时间最短为 78 s,训练时间最长为 SVM 模型,用时 17 257 s(约 287 min),远远大于卷积神经网络和深度神经网络.但查全率指标为进一步比较各分类器性能,绘制 ROC 曲线,如图 8 所示.

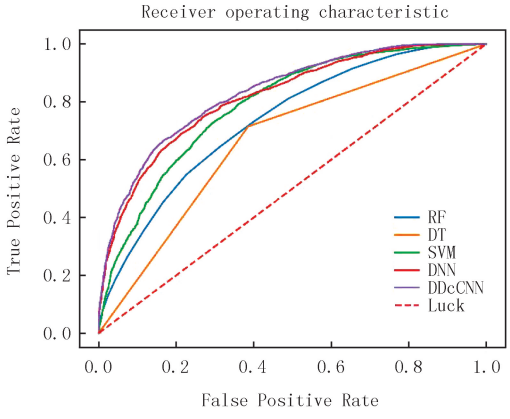

图8　不同分类器的ROC曲线图

Fig. 8　The ROC curve of different classifier

由图 8 可知,DNN 模型与 SVM 模型在 TPR=0.4 时交叉,AUC 分别为 0.81、0.78.在 TPR 小于 0.4 的范围内 DNN 模型包含 SVM 模型,在 TPR 大于 0.4 的范围内 SVM 模型包含 DNN 模型.DDcCNN 模型可以完全包含基于其他分类器的模型,其 AUC 为 0.83 为最高,性能优于基于其他分类器的模型.

3.5　与现有研究方法的结果比较

将本研究最终结果与 CCSOL[13],PaRSnIP[14],SOLpro[15],PROSO II[16],DeepSol[22] 等研究结果进行比较试验,结果如表 9 所示.

表 9　已有研究成果性能比较结果

Tab. 9　The comparison results of our model with existing research results

算法	精确度/%	灵敏度	特异性	MCC	算法	精确度/%	灵敏度	特异性	MCC
DDcCNN	76.31	0.65	**0.84**	**0.55**	PROSO II[16]	75.40	0.73	0.76	0.39
Deepsol[22]	**77.00**	**0.77**	0.78	0.55	CCSOL[13]	54.00	0.54	0.54	0.08
PaRSnIP[14]	74.00	0.74	0.74	0.48	SOLpro[15]	60.00	0.60	0.60	0.20

由表 9 可知,DDcCNN 模型在输入特征向量为 457 维情况下,可以达到精确度 76.31%、灵敏度为 0.65、特异性为 0.84、MCC 为 0.55,特异性、MCC 为最高值,精确度仅低于 Deepsol 模型,说明具有较好的性能.对比于 PaRSnIP,精确度提升 2.31%,特异性、MCC 分别提高 0.10、0.32.与 Deepsol 模型相比虽然精度下降 0.69,但仅使用 457 维输入特征远小于 Deepsol 的 25 257 维特征,说明本研究算法的训练参数更少,可以获得更短的训练时间,同时可以在硬件性能更低的环境下运行.

4　结束语

为实现蛋白质溶解性的预测,本研究通过分析蛋白质序列特性,设计了 4 种不同的深度卷积神经网络模型,分别为双路卷积神经网络模型(DDcCNN)、深度单路卷积神经网络模型(DCNN)、深度 3 路卷积神经网络模型(DTcCNN)以及浅层卷积神经网络模型(SCNN).利用 CD-HIT 对原始数据集进行降噪处理,提取蛋白质序列的 G-gap 特征以及 57 维额外特征,构成输入向量.对 4 种不同架构模型进行对比实验,结果表明利

用 3 层 2D 卷积运算与 1 层 1D 卷积运算共同构建的 DDcCNN 模型整体性能最优,其预测精度、查全率、查准率、MCC 性能指标分别达到了 76.31%、65.31%、75.05%、0.55.对比综合性能较好的 DTcCNN 模型,精确度、灵敏性、特异性、MCC 分别提高 0.49、0.01、0.01、0.03.对比基于 SVM 分类器的预测模型,精确度提升 1.22%、MCC 提升 0.02,同时降低了 94.44%(16 299 s)的训练时间.相较于 PaRSnIP 模型,在使用更少特征的情况下精确度提升 2.31%、MCC 提高 0.32,与使用 25 257 维特征的 DeepSol 模型相比,本研究所提设计的 DDcCNN 模型仅使用 457 维特征,可以获得更快的训练和预测速度,以及更低的硬件要求,具有一定的理论和实际意义.

参 考 文 献

[1] HABIBI N,MOHD H,ALIREZA N,et al.A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli[J].Bmc Bioinformatics,2014.DOI:10.1186/1471-2105-15-134.

[2] SIMON B,ANDY K,ANDERS B,et al.Solubility prediction from first principles:a density of states approach[J].Physical Chemistry Chemical Physics,2018,20(32):20981-20987.

[3] 程姗.5′腺苷单磷酸钠探针同步荧光测定生物样品中蛋白质含量[J].河南师范大学学报(自然科学版),2014,42(4):64-68.
CHENG S.Simultaneous fluorescence determination of protein content in biological samples with 5′adenosine monophosphate probe[J].Journal of Henan Normal University(Natural Science Edition),2014,42(4):64-68.

[4] YANG Z,DAVID R,KIMYEN N,et al.Expression of eukaryotic proteins in soluble form in Escherichia coli[J].Protein Expression and Purification,1998,12(2):159-65.

[5] GREGORY D,CLAUDE E,DENTON N,et al.New fusion protein systems designed to give soluble expression in Escherichia coli[J].Biotechnology and bioengineering,1999,65(4):382-388.

[6] SUSAN T,BALAJI V.Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli[J].Protein Science,2005,14(3):582-592.

[7] HERHUT M,BRANDENBUSCH C,SADOWSKI G.Modeling and prediction of protein solubility using the second osmotic virial coefficient[J].Fluid Phase Equilibria,2016,422:32-42.

[8] ZHANG S,ZHANG T,LIU C.Prediction of apoptosis protein subcellular localization via heterogeneous features and hierarchical extreme learning machine[J].Sar and Qsar in Environmental Research,2019,30(3):209-228.

[9] HAN G,YU Z,VO A,A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC[J].Journal of Theoretical Biology,2014,344:31-39.

[10] 孟光胜,赵志宇.基于两层主动学习策略的 SVM 分类方法[J].河南师范大学学报(自然科学版),2014,42(2):158-162.
MENG G S,ZHAO Z Y.SVM classification method based on two-layer active learning strategy[J].Journal of Henan Normal University(Natural Science Edition),2014,42(2):158-162.

[11] WEI L,DING Y,RAN S,ET AL.Prediction of human protein subcellular localization using deep learning[J].Journal of Parallel and Distributed Computing,2018,117:212-217.

[12] FEDERICO A,DAVIDE C,CARMEN,M,et al.Bioinformatics approaches for improved recombinant protein production in Escherichia coli:protein solubility prediction[J].Briefings in Bioinformatics,2014,15(6):953-62.

[13] FEDERICO A,DAVIDE C,CARMEN,M,et al.ccSOL omics:a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli[J].Bioinformatics,2014,30(20):2975-2977.

[14] REDA R,RAGHVENDRA M,KHALID K,et al.PaRSnIP:sequence-based protein solubility prediction using gradient boosting machine[J].Bioinformatics,2018,34(7):1092-1098.

[15] CHRISTOPHE M,ARLO R,PIERRE B,SOLpro:accurate sequence-based prediction of protein solubility[J].Bioinformatics,2009,25(17):2200-2207.

[16] PAWEL S,GERO D,PHILLIPP T,et al.,PROSO II-a new method for protein solubility prediction[J].Febs Journal,2012,279(12):2192-2200.

[17] 曹周键,王赫,朱建阳.深度学习在引力波数据处理中的应用初探[J].河南师范大学学报(自然科学版),2018,46(2):26-39.
CAO Z J,WANG H,ZHU J Y.Application of Deep Learning in Gravitational Wave Data Processing[J].Journal of Henan Normal University(Natural Science Edition),2018,46(2):26-39.

[18] SAVOJARDO C,BRUCIAFERRI N,TARTARI G,et al.DeepMito:accurate prediction of protein submitochondrial localization using convolutional neural networks[J].Bioinformatics,2019.DOI:10.1093/bioinformatics/btz512.

[19] JANKE,J,CASTELLI M,POPOVIC A.Analysis of the proficiency of fully connected neural networks in the process of classifying digital images.Benchmark of different classification algorithms on high-level image features from convolutional layers[J].Expert Systems with

Applications,2019,135:12-38.

[20] WU W,YIN Y,WANG X,et al.Face Detection With Different Scales Based on Faster R-CNN[J].Ieee Transactions on Cybernetics,2019,49(11):4017-4028.

[21] TUNCER T,DOGAN S,ERTAM F.Automatic voice based disease detection method using one dimensional local binary pattern feature extraction network[J].Applied Acoustics,2019,155:500-506.

[22] SAMEER K,REDA R,KHALID K,et al.DeepSol:a deep learning framework for sequence-based protein solubility prediction[J].Bioinformatics,2018,34(15):2605-2613.

[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.

[24] WANG X,LI H,GAO P,et al.Combining Support Vector Machine with Dual g-gap Dipeptides to Discriminate between Acidic and Alkaline Enzymes[J].Letters in Organic Chemistry,2019,16(4):325-331.

[25] LIN H,CHEN W,DING H.AcalPred:A Sequence-Based Tool for Discriminating between Acidic and Alkaline Enzymes[J].Plos One,2013.DOI:10.1371/journal.pone.0075726.

[26] MAGNAN C,BALDI P.SSpro/ACCpro 5:almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles,machine learning and structural similarity[J].Bioinformatics,2014,30(18):2592-2597.

[27] 李彦冬,郝雷航.卷积神经网络研究综述[J].计算机应用,2016(9):2508-2515.
LI Y D,HE L H.Review of Convolutional Neural Network Research[J].Computer Application,2016(9):2508-2515.

[28] 胡辉.一种结合 Dropblock 和 Dropout 的正则化策略[J].河南师范大学学报(自然科学版).2019,47(6):51-56.
HU H.A regularization strategy combining Dropblock and Dropout[J].Journal of Henan Normal University(Natural Science Edition).2019,47(6):51-56.

Design of protein solubility prediction model based on deep neural network

Wang Xianfang[1a,2], Liu Yifeng[1b], Du Zhiyong[1c], Zhu Mingdong[1a], Li Qimeng[2]

(1.a. School of Computer Science and Technology; b. School of Management;
c. School of Electrical Engineering and Automation, Henan Institute of Technology, Xinxiang 453003, China;
2. School of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China)

Abstract: Protein solubility is an important research in the field of bioinformatics. We designed multiple convolutional neural network models to predict protein solubility based on using combining feature extraction and deep learning technology. CD-HIT was used to denoise the original protein data, and the features of each sample were extracted by G-gap, which were input to a channel of convolutional neural networks. And SCRATCH tool was used to extract 6-dimensional sequence features and 51-dimensional structural features as additional features for each sample to improve the accuracy of the model, which were input as another channel of the model. We analyzed the characteristics of the data, which designed four different network models by adjusting the series-parallel structure of the convolutional layers. The network structure and parameters were determined through comparative experiments. The results showed that the protein solubility prediction model based on Deep Dual-channel Convolutional Neural Networks got the best overall performance. Its prediction accuracy, recall rate, precision rate, MCC (Matthews Correlation Coefficient) indicators reached 76.31%, 65.31%, 75.05%, 0.55, respectively. The verification experiments were established to compare our method to the traditional Deep Neural Networks, Support Vector Machines, Random Forests, Decision Tree, and the models of existing research, the results showed that effectiveness of our method was proved.

Keywords: deep dual-channel convolutional neural networks; protein solubility; G-gap dipeptide frequency; predict model

[责任编校　陈留院　赵晓华]