基于集成学习的 N6 甲基化位点预测方法研究

赵媛媛[1],陈进祥[1],李富义[2,3],吴 昊[1],刘全中[1*]
(1. 西北农林科技大学 信息工程学院,陕西 杨凌 712100;
2. 蒙纳士大学数据科学中心,澳大利亚 墨尔本 VIC 3800;
3. 蒙纳士大学生物医学发现研究所和生物化学与分子生物学系,澳大利亚 墨尔本 VIC 3800)

摘 要:N6-甲基腺嘌呤(N6-methyladenine,6mA)是指腺嘌呤第 6 位氮原子的甲基化修饰。6mA 在维持细胞正常的转录活性、DNA 损伤修复、染色质重塑、遗传印记、胚胎发育和肿瘤发生等生物过程中起着非常重要的作用。通过生物实验的方法来鉴定 6mA 位点耗时且昂贵。近年来,研究界提出了一些基于机器学习的 6mA 位点预测方法,但这些预测方法过度依赖一种学习模型,导致模型的泛化能力不足以及预测的准确度不高等问题。集成学习综合多种预测模型的优点,具有较好的泛化能力及预测性能。因此,为了进一步提升 6mA 位点的预测准确性,提出了一种基于 stacking 集成学习的 6mA 位点预测模型 Stack6mAPred。该模型由两层分类器组成,第一层集成了朴素贝叶斯、支持向量机(support vector machine,SVM)和 LightGBM 等三种主流分类器,第二层使用逻辑回归(logistic regression,LR)分类器。Stack6mAPred 利用增强核苷酸组成等 5 种特征对实验已鉴定 6mA 序列和非 6mA 序列进行编码,使用 XGBoost (extreme gradient boosting)算法进行特征选择,去除冗余特征。通过在水稻基准数据集上进行五折交叉验证,与目前性能最优的方法 MM-6mAPred 相比,Stack6mAPred 在敏感性、特异性、准确度、MCC 和 AUC 上均获得了更好的性能,分别提高了 1.7%、1.36%、1.72%、0.06 和 0.031。

关键词:6mA 甲基化;stacking 集成学习;XGBoost;LightGBM;支持向量机
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2021)03-0149-08
doi:10.3969/j.issn.1673-629X.2021.03.026

Research on Prediction Method of N6-methylation Sites Based on Ensemble Learning

ZHAO Yuan-yuan[1],CHEN Jin-xiang[1],LI Fu-yi[2,3],WU Hao[1],LIU Quan-zhong[1*]
(1. School of Information Engineering,Northwest A&F University,Yangling 712100,China;
2. Monash Centre for Data Science,Monash University,Melbourne VIC 3800,Australia;
3. Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology,
Monash University,Melbourne VIC 3800,Australia)

Abstract:N6-methyladenine (6mA) refers to the methylation modification of the nitrogen atom at position 6 of adenine,which plays an important role in maintaining the normal transcriptional activity of cells,DNA damage repair,chromatin remodeling,genetic imprinting,embryonic development and tumorigenesis. However,it is a challenge to detect 6mA sites through experimental methods,which are time-consuming and costly. In recent years,the research community has proposed several machine learning-based approaches for 6mA sites detection. However,the existing 6mA detection approaches heavily rely on a single learning model. A single learning model mainly focuses on some respects to detect 6mA sites,and its accuracy and generalization ability need to be further improved. Ensemble learning can achieve powerful performance by combining multiple models. To address the drawbacks of a single learning model,a stacking ensemble-based 6mA site prediction method called Stack6mAPred is proposed. Stack6mAPred consists of two layers of classifiers. In the first layer,three mainstream classifiers such as Naive Bayes,support vector machine (SVM) and LightGBM are integrated,and in the second layer the logistic regression (LR) classifier is used. Stack6mAPred uses five sequence features to encode the experimentally

收稿日期:2020-04-21 **修回日期**:2020-08-25
基金项目:国家自然科学基金面上项目(61972322);教育部人文社科交叉项目(18YJCZH190);基本科研业务费前沿与交叉科学研究项目(2452019180);中央高校基本科研业务费(2452017342);博士科研启动经费(2452017019)
作者简介:赵媛媛(1995-),女,硕士研究生,研究方向为数据挖掘、计算生物学;通信作者:刘全中(1978-),男,副教授,博士,研究方向为数据挖掘、计算生物学。

identified 6mA sequences and non-6mA sequences into feature vectors, and employs XGBoost (extreme gradient boosting) algorithm to select important features from a high dimension. We conduct a five-fold cross-validation test on the benchmark rice datasets and compare with current best performing method MM-6mAPred. Results show that Stack6mAPred has achieved better performances on five common evaluation metrics, including sensitivity, specificity, accuracy, MCC (Matthews correlation coefficient) and AUC (area under the ROC curve). Performances of these five metrics are increased by 1.7%, 1.36%, 1.72%, 0.06 and 0.032 respectively.

Key words: N6-methyladenine (6mA); stacking ensemble learning; extreme gradient boosting (XGBoost); LightGBM; support vector machine

0 引 言

DNA 甲基化,是指经 DNA 甲基转移酶催化,以 S-腺苷甲硫氨酸(SAM)作为甲基供体,DNA 分子与甲基相连接的过程[1]。在 DNA 的四种碱基中,只有胞嘧啶和腺嘌呤可以被甲基化。近年来,研究者发现了腺嘌呤的第六位氮原子甲基化修饰,即 6-甲基腺嘌呤(N6-methyladenine,6mA)。6mA 甲基化作为一种重要的非永久性但相对长期可遗传的基因修饰,被发现在维持细胞正常的转录活性、DNA 损伤修复能力、染色质重塑、遗传印记、胚胎发育和肿瘤发生中都有着不可替代的作用,成为分子生物学及医学领域的研究热点[2]。

6mA 在 DNA 层面表达丰富度相对较低,在哺乳动物中,平均每百万个腺嘌呤中只有不到 10 个 6mA 位点[3]。目前已经有几种鉴定 6mA 的实验方法,例如甲基化 DNA 免疫沉淀测序(MeDIP-seq)[4],毛细管电泳和激光诱导荧光(CE-LIF)[5] 和单分子实时测序(MRT-seq)[6]。虽然通过实验方法能鉴定 6mA 位点,但实验方法实验周期长、劳动强度大且十分昂贵,很难适合从高通量序列中识别 6mA。基于机器学习的计算方法可以同时处理多条序列中 6mA 位点的鉴定,这种方法省时、省力并且效率高,作为实验方法有效的补充,越来越受到生物界的青睐。

最近,华中农业大学周道绣课题组使用了免疫沉淀测序技术对水稻基因组的 6mA 进行了精确定量和定位,获得了水稻基因组的 6mA 图谱[7]。该数据的获取为构建基于机器学习模型的 6mA 识别方法奠定了数据基础。近年来,研究界提出了一些基于传统的机器学习和深度学习的 6mA 位点预测方法。例如,2019 年 1 月,Chen 等提出了一种基于支持向量机方法(6mA-Pred)鉴定水稻因组中的 6mA 位点[8],模型准确率达到了 83.13%。2019 年 4 月,Tahir 等人提出了一种卷积神经网络(CNN)计算模型(iDNA6mA)[9],从 DNA 输入序列中自动地提取关键特征并训练模型,该模型准确率达到了 86.59%。2019 年 7 月,Pian 基于马尔可夫模型提出了一种新的分类方法(MM-6mAPred)[10],准确率达到 89.72%。2019 年 9 月,Liu 等人[11]提出了基于提升树模型(ExtraTree)对小鼠和水稻基因中的 6mA 位点鉴定方法(csDMA),对于水稻中 6mA 位点的预测达到了 86.1% 的准确度。

在上述方法中,基于深度学习的 iDNA6mA 方法不需要人工设计特征,但其识别性能仍有待提高。基于传统机器学习方法的 6mA 识别方法虽然具有较强的识别能力,但现有的学习模型使用序列要么特征单一,缺乏从多种角度综合考量 6mA 位点;要么特征维度较高且未使用特征选择方法进行特征选择,如 6mA-Pred 等,预测的性能还有很大提升空间。

基于上述的现有研究的不足,为了进一步提升 6mA 位点的预测性能,该研究提出一种基于 stacking 集成学习的 6mA 预测模型 Stack6mAPred。Stack6mAPred 结合了增强核苷酸组成(ENAC)、核苷酸电子-离子相互作用伪电位(EIIP)、核苷酸化学性质(NCP)、Kmer 和核苷酸间隔(diTriKGap)5 种不同类型的特征编码;利用 XGBoost 进行特征选择[12],去除冗余特征;集成了朴素贝叶斯、支持向量机(SVM)、LightGBM 和逻辑回归等 4 种不同的分类器。在真实的水稻基因组数据集上进行了实验,结果表明:提出的 Stack6mAPred 预测模型对 6mA 位点鉴定的准确率达到 91.83%,AUC 达到 0.967。

1 数据集

数据集构建是机器学习模型的基础,基准数据集的质量对构建模型的性能至关重要。该文使用了 Chen 等人[8]提供的水稻 DNA 序列中的 6mA 数据集。该数据集是从美国国家生物技术信息中心(national center for biotechnology information,NCBI)获得,使用 CD-HIT 软件[13]去除同源性超过 60% 的序列。数据集包括 880 个经实验验证的 6mA 位点的序列片段和 880 个非 6mA 位点的序列片段,序列长度均为 41 bp。该数据集已经被多个预测模型使用[8-10]。数据集获取公开站点为 http://lin-group.cn/server/i6mAPred/data。

2 特征提取

不同序列特征对不同问题具有不同的识别能力,最终影响预测模型的性能。为了提取针对 6mA 位点具有较强预测能力的特征,该研究对 iLearn[14] 和 PyFeat[15] 中总结的所有 DNA 序列特征分别进行性能

评估,发现五种对于 6mA 位点具有较强的识别能力的特征:增强核苷酸组成(enhanced nucleic acid composition,ENAC)、核苷酸电子-离子相互作用伪电位(electron - ion interaction pseudopotentials of trinucleotide, EIIP)、核苷酸化学性质(nucleotide chemical property, NCP)、Kmer、核苷酸间隔(ditriKGap)。特征对应的维度和参数设置如表 1 所示。

表 1　实验中使用的特征及参数介绍

特征	参数设置	维度	工具包
ENAC	$N=2$	160	iLearn
EIIP	无参数	41	iLearn
NCP	无参数	123	iLearn
Kmer	$K=5$	1 364	iLearn
ditriKGap	$g=3$	311	PyFeat

2.1　增强核苷酸组成(ENAC)

增强核苷酸组成(ENAC)根据固定长度的序列窗口计算核苷酸组成(nucleic acid composition, NAC)[16],通常可用于编码等长的核苷酸序列。NAC 编码用于计算核苷酸序列中每种核酸类型的频率。序列中四种核苷酸出现频率可以由式(1)计算:

$$f(t)=\frac{N(t)}{N},t\in\{A,C,G,T(U)\} \tag{1}$$

其中,$N(t)$ 是 t 型核苷酸的数目,N 是核苷酸序列的长度。

ENAC 编码的核心是计算固定长度的序列窗口内的 NAC,即该窗口首先从序列的第一位核苷酸开始,依次向后移动,计算窗口内序列的 NAC,直到窗口包含序列的最后一位完成编码过程。实验表明窗口值为 2 时,ENAC 编码的性能达到最优。

2.2　核苷酸电子-离子相互作用伪电位(EIIP)

Nair 等人提出了一种新的特征编码方式[17],通过计算核苷酸中离域电子的能量,将其表示为电子-离子相互作用伪电位(EIIP)进行编码。该编码方式直接使用核苷酸的 EIIP 值取代 DNA 序列中核苷酸 A,G,C 和 T。核苷酸 A,G,C,T 的 EIIP 值分别为 0.126 0、0.134 0、0.080 6 和 0.133 5。EIIP 特征编码维数等于 DNA 序列的长度。

2.3　核苷酸化学性质(NCP)

DNA 中有四种核苷酸,即腺嘌呤(A)、鸟嘌呤(G)、胞嘧啶(C)和胸腺嘧啶(T)。根据化学性质进行分类,四种核苷酸分类结果如表 2 所示。

核苷酸化学性质(NCP)根据每种核苷酸在不同分组内所处类别不同,将每个核苷酸表示为一个 3 维向量,对特征进行编码。每个化学性质分成两类,一个核苷酸在第 1 个类中出现编码为 1,否则编码为 0。因此,根据表 2 中分类可知,A、C、G 和 T 分别被表示为(1,1,1)、(0,1,0)、(1,0,0)和(0,0,1)。

表 2　核苷酸化学性质

化学性质	分类	核苷酸
环状结构	嘌呤	A,G
	嘧啶	C,T
官能团	氨基	A,C
	酮基	G,T
氢键	强氢键	A,T
	弱氢键	G,C

2.4　Kmer

Kmer 特征编码用于计算 DNA 序列中 K 个相邻核苷酸的出现频率,已成功应用于人类基因调控序列预测[18]和增强子识别[19]。Kmer(以 $K=3$ 为例)由式(2)计算:

$$f(t)=\frac{N(t)}{N},t\in\{\mathrm{AAA,AAC,AAG,\cdots,TTT}\} \tag{2}$$

其中,$N(t)$ 是 Kmer 型 t 的次数,N 是核苷酸序列的长度。

文中采用的 Kmer 编码方式,将小于等于 K 的相邻核苷酸频率全部计算,以 $K=3$ 为例,该特征编码将 $K=3,2,1$ 的 Kmer 全部列出。经实验验证,K 取 5 使得 6mA 识别的性能达到最优。

2.5　核苷酸间隔(diTriKGap)

diTriKGap 特征编码通过设置间隔大小 g,统计 DNA 或 RNA 序列内不同间隔序列结构的数目。当设置间隔 $g=1$ 时,序列结构为 XX_XXX;当设置间隔 $g=2$ 时,序列结构为 XX_XXX 和 XX__XXX;当设置间隔 $g=3$ 时,序列结构为 XX_XXX、XX__XXX 和 XX___XXX。例如,当设置间隔 $g=2$ 时,将会统计 DNA 序列中∑AA_AAA,∑AA__AAA,∑AA_AAC,∑AA__AAC,∑AA_AAG,∑AA__AAG,∑AA_AAT,∑AA__AAT…等结构的数量。特征编码的每一列代表序列结构的数目。

经实验验证,当设置 $g=3$ 时,diTriKGap 特征编码性能达到最优。该编码通过工具包 PyFeat 提取。为了去除冗余特征,减少特征维度过多[20]而对模型造成的影响,PyFeat 采用 Adaboost 算法进行特征选择,在保持模型性能的同时,将该特征编码的维度由 3 072 减少到 311。

3　特征选择

特征选择是指从初始特征集中选择相关特征子集的机器学习过程,特征选择能有效地降低特征空间的

维度,去除对分类不重要的和冗余的特征,提高预测模型的预测性能。

梯度提升决策树(gradient boosting decision tree,GBDT)是一种集成模型,基分类器是 CART 树[21],适用于分类和回归问题,同时可用于特征选择。XGBoost(extreme gradient boosting)是陈天奇博士在 2011 年提出的一种基于提升树[21]的集成学习模型。XGBoost 是在 GBDT 的基础上改进,适应范围更广,是对 GBDT 的一种高效实现。XGBoost 中的基分类器是使用 CART 和线性分类器的组合。

XGBoost 根据特征重要性进行排序,以此达到特征选择的目的。如果一个特征在所有决策树中作为划分属性的次数越多,那么该特征就越重要,XGBoost 以此计算每个特征的重要性。

在实验中,使用 XGBoost 进行特征选择,找到最优的特征子集,降低特征维度,流程如图 1 所示。由于 XGBoost 模型参数的选择对特征重要性打分影响较大,在实验中改变模型参数进行了 36 次实验来减少特征选择的误差,最终得到 36 个不同 XGBoost 模型的特征打分表,具体步骤如下所述:

(1)将特征编码输入 XGBoost 模型进行参数打分,得到特征打分表;

(2)重复步骤(1),36 次后得到 36 个互不相同的特征打分表;

(3)将 36 个特征打分表中的特征取交集,计算交集中每个特征的重要性平均值,并进行排序,得到最终的最优特征子集。

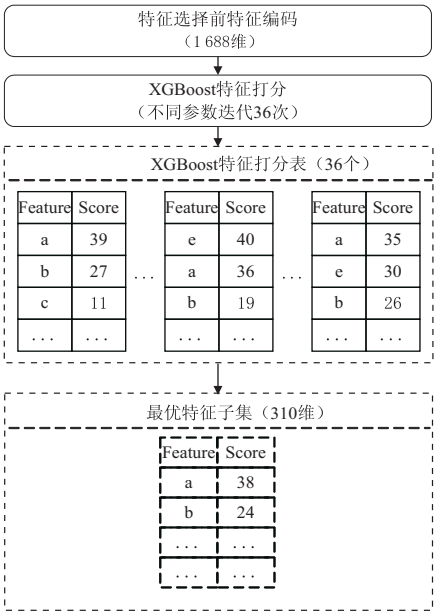

图 1 XGBoost 特征选择流程

4 基于集成学习 6mA 位点识别方法

机器学习中的监督学习目标是训练出一个稳定且各方面性能良好的模型。在许多现实场景中,往往只能得到多个有偏好的模型,也就是弱监督模型。根据组合弱监督模型方法的不同,集成学习分为 bagging[22]、boosting[23]和 stacking[24]三种集成方式。

该研究采用 stacking 集成学习构建 6mA 位点预测模型 Stack6mAPred,该模型集成 4 种主流分类器以及 5 种最优的特征编码构造一个性能更好的 6mA 位点预测模型。

4.1 模型的整体框架

stacking 集成学习模型预测性能的好坏主要取决于基分类器的预测精度和多样性,使用不同参数、不同类型的分类器训练相同的特征,实现不同基学习器之间的强强联合和优势互补。文中提出的 stacking 集成模型框架 Stack6mAPred 如图 2 所示。

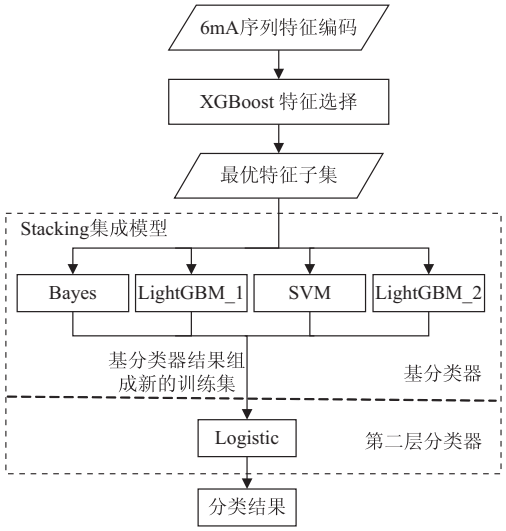

图 2 Stacking 集成模型框架

初始特征编码由 XGBoost 进行特征选择得到最优特征子集后,使用 stacking 集成学习模型训练分类器,得到最终预测结果。本模型中采用 4 个基分类器,分别由朴素贝叶斯(naive Bayes classifiers,NB)、支持向量机(support vector machine,SVM)和 LightGBM 等组成,其中 LightGBM_1 和 LightGBM_2 均使用 LightGBM 算法进行训练,两者仅参数设置不同。第二层分类器使用逻辑回归(LR)。

4.2 朴素贝叶斯

朴素贝叶斯分类器假设特征对于给定类的影响独立于其他特征,是一种较稳定的有监督分类算法,其分类算法基于贝叶斯定理,在处理大规模数据库时有较高的分类准确率。贝叶斯分类器的分类原理是通过某对象的先验概率,利用贝叶斯公式计算出其后验概率,即该对象属于某一类的概率,选择具有最大后验概率的类作为该对象所属的类。

4.3 支持向量机

支持向量机是 Cortes 和 Vapnik 于 1995 年首先提

出的[25],已经广泛应用于生物信息学问题中。其基本思想是将输入数据转换为高维特征空间,然后确定最佳分隔超平面,以此作为决策边界。

SVM 模型有两个非常重要的参数 C 与 gamma。其中 C 是惩罚系数,即对误差的容忍度,C 越高越容易过拟合;gamma 是选择 RBF 作为核函数后,该函数自带的一个参数,决定了数据映射到新的特征空间后的分布。实验中选择 RBF 作为核函数。

为了选择最优的参数使模型达到最佳性能,使用 LibSVM 进行参数寻优。LibSVM 是用来调整 SVM 参数非常有效的手段,应用广泛[26-28],该工具包可在 https://www. csie. ntu. edu. tw/ ~ cjlin/libsvm/免费获取。

LibSVM 采用网格搜索(grid search)进行参数搜索,在 C 和 gamma 组成的二维参数矩阵中,依次实验每一对参数的效果,以此得到全局最优的参数。实验最终确定的参数在 4.6 节进行了详细介绍。

4.4　LightGBM

LightGBM 是个快速的、分布式的、高性能的基于决策树算法的梯度提升(gradient boosting)模型[29],是对 GBDT 的高效实现,可用于排序、分类、回归等机器学习任务中。

该算法在传统 GDBT 算法基础上引入了梯度单边采样(gradient-based one-side sampling,GOSS)和互斥特征合并(exclusive feature bundling,EFB)两种新技术。梯度单边采样(GOSS)算法保留所有的大梯度样本,在小梯度样本中进行随机采样,从而达到提升效率的目的。独立特征合并(EFB)算法通过使用基于直方图(histograms)方法安全地将互斥特征绑定在一起形成一个新的特征,从而减少特征维度。LightGBM 可以在不损失分类器精度的前提下,显著减少模型学习的时间,提升模型的泛化能力。

4.5　集成学习方法

该文采用 stacking 集成学习模型,其中基分类器使用朴素贝叶斯、LightGBM_1、LightGBM_2 和 SVM 等四个模型,第二层分类器使用逻辑回归。该研究中 Stack6mAPred 集成模型的训练步骤如下所述:

(1)输入数据集。假设数据集表示为 $D=\{x_i,y_i\}_{i=1}^{n}$,其中 $n=1\ 760$ 表示样本的个数,x_i 表示第 i 个样本,维度为 621;$y_i=1$ 表示第 i 个样本是正例,否则为负例。集成学习第一步将 D 输入到集成学习模型的第 1 层分类器。

(2)学习出新的数据集 D'。将 D 随机分为 2 个相同大小的子集 $D=D_1\cup D_2$,$D_1\cap D_2=\varnothing$;首先将 D_1 分别输入到第一层的 4 个基分类器,学习出 4 个模型为 $B_{1t}(t=1,2,3,4)$。$\forall s_i=\{x_i,y_i\}\in D_2$,其中 x_i 表示样本 s_i 的特征向量,y_i 表示样本 s_i 的类别,学习出新样本 $s_i'=\{(B_{11}(x_i),B_{12}(x_i),B_{13}(x_i),B_{14}(x_i)),y_i\}$。其中 $B_{1t}(x_i)(t=1,2,3,4)$ 表示模型 B_{1t} 对样本向量 x_i 的预测概率值,4 个概率值作为新的样本 s_i' 的特征值,y_i 作为 s_i' 的类别;用样本子集 D_1 训练第一层的 4 个分类器,预测样本子集 D_2,得到一个新的样本子集 D_2',D_2' 与 D_2 样本数相同,具有 4 个概率特征值。同理,用样本子集 D_2 训练第一层的 4 个分类器,预测样本子集 D_1,得到一个新的样本子集 D_1',D_1' 与 D_1 样本数相同,具有 4 个概率特征值。经过第 1 层 4 个分类模型后,数据集 D 被改造为新的数据集 $\boldsymbol{D}'=D_1'\cup D_2'$,$\boldsymbol{D}'$ 为 1 760×4 的矩阵。

(3)将 $\boldsymbol{D}'$ 作为第二层分类器逻辑回归(LR)的训练集,训练得到最终的集成分类模型 Stack6mAPred。

4.6　模型参数设置

集成模型中各基分类器中参数的选择对模型分类性能影响极大。在该研究中,为了使集成模型分类性能最佳,先人工确定参数最优的大致范围,然后利用网格搜索(grid search)寻找集成模型的全局最优参数。各基分类器用到的参数如表 3 所示。

表 3　基分类器参数设置

分类器	参数设置
朴素贝叶斯	-
LightGBM_1	Learning_rate:0.5 Max_depth:10 Num_leaves:14
支持向量机	C:32768 Gamma:0.0004
LightGBM_2	Learning_rate:0.25 Max_depth:6 Num_leaves:14

4.7　评价指标

为了评估 Stack6mAPred 模型的预测性能,分别使用了 AUC(area under the ROC curve)值、准确度(accuracy,Acc)、特异性(specificity,Sp)、敏感性(sensitivity,Sn)和马修斯相关系数(Matthews correlation coefficient,MCC)共 5 个常用的评价指标。其中 AUC 是 ROC(area under curve)曲线与坐标轴围成的面积。ROC 曲线是根据一系列不同的二分类方式,以真阳性率为纵坐标,假阳性率为横坐标绘制的曲线。

准确度、特异性、敏感性和马修斯相关系数指标定义如式(3)~式(6)所示:

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}\times 100\% \tag{3}$$

$$\mathrm{Sp}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\times 100\% \tag{4}$$

$$Sn=\frac{TP}{TP+TN}\times 100\% \tag{5}$$

$$MCC=\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TN+FN)\times(TP+FP)\times(TN+FP)\times(TP+FN)}} \tag{6}$$

其中,TP 是正确识别的真实 6mA 序列的数量,FN 是错误分类的 6mA 序列的数量,TN 是正确识别的非 6mA 序列的数量,FP 是错误分类的非 6mA 序列的数量。

5 实验结果

5.1 6mA 位点序列分析

为从生物序列方面解释特征编码合理性,本实验使用软件 Two Sample Logos[30] 绘制 6mA 位点的序列标识图,该软件对正例样本序列和负例样本序列进行对比,结果如图 3 所示。

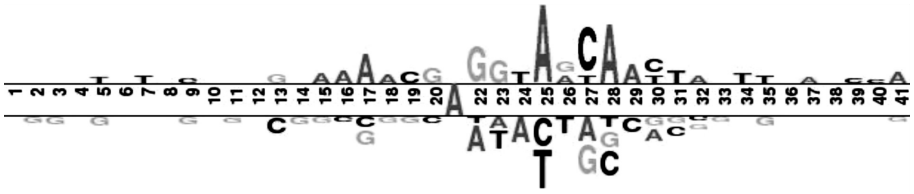

图 3 6mA 序列对比

从序列标识图可以直观地看出序列共有 41 个位点,第 21 号中心点代表 6mA 位点。每个位点上符号的高度代表对应的核苷酸在该位置的出现频率。

从图 3 中可以看出,对于 6mA 位点,在中心位点下游 15 ~ 18、上游 25 位置腺嘌呤(A)出现的概率较高,非 6mA 位点上游 22 ~ 24 位置腺嘌呤(A)出现概率较高,对应文中采用的特征编码 ditriKGap 中 XXX 的结构;在中心位点下游 20、上游 22 ~ 23 位置鸟嘌呤(G)出现概率较高,非 6mA 位点下游 18 ~ 19、14 ~ 15 等位置鸟嘌呤(G)出现概率较高,对应 ditriKGap 中 XX 的结构。由此可以看出,6mA 序列具有一定的规律性,反映出文中采用的特征编码中 Kmer、核苷酸间隔(ditriKGap)对于提升模型准确度的有效性。

5.2 不同编码方式性能比较

为了找到特征子集的最佳组合,用随机森林分类器(100 棵决策树)评估单个特征的性能。表 4 列举了单个特征以及特征组合预测性能。单个特征中 ENAC 达到了最佳性能,其次是 NCP,而 Kmer 的性能表现最差。5 种特征组合可以显著提高模型的准确度,比单个特征中性能最优的 ENAC 在敏感性、特异性、准确率、MCC 以及 AUC 值五个性能指标方面分别提高了 1.1%、3.2%、2.1%、0.043 以及 0.02。证明了综合多个特征从多种角度区分 6mA 位点能提高预测性能。该文整合 ENAC、NCP、EIIP、Kmer 和 diTriKGap 5 种特征对样本序列进行编码。

表 4 不同编码方式性能比较

Encoding schemes	Sn/%	Sp/%	Acc/%	MCC	AUC
1	83.9	85.3	84.6	0.696	0.921
2	82.2	82.5	82.3	0.651	0.913
3	81.8	85.0	83.4	0.672	0.912
4	67.2	67.8	67.5	0.352	0.744

续表 4

Encoding schemes	Sn/%	Sp/%	Acc/%	MCC	AUC
5	82.0	66.1	74.8	0.489	0.823
{1,2,3,4,5}	85.0	88.5	86.7	0.739	0.941

表 4 中,编号 1、2、3、4 和 5 分别代表五个特征编码 ENAC、EIIP、NCP、Kmer 和 diTriKGap,{1,2,3,4,5} 代表五种特征组合。

5.3 特征选择性能评价

为了避免特征冗余以及分类器过拟合,使用了 XGBoost 进行特征选择,选出最优特征子集。在实验中使用的五种特征编码中,diTriKGap 在使用 PyFeat 进行特征提取时,已经使用 AdaBoost 进行特征选择。因此,实验仅对 ENAC、NCP、EIIP 和 Kmer 的特征组合进行特征选择,将特征选择后的结果与 diTriKGap 特征编码合并,构成最优特征子集。实验用随机森林分类器(100 棵决策树)评估特征选择前后的预测性能。

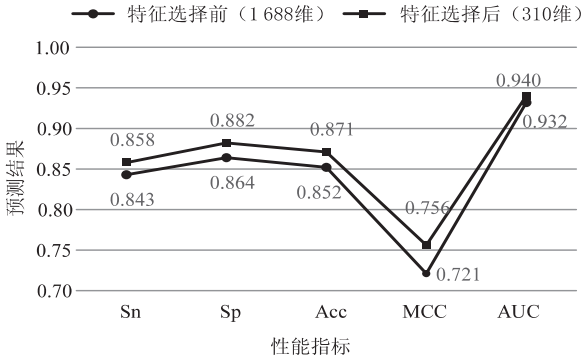

图 4 特征选择前后性能对比

实验证明,该文使用的特征选择方法,有效减少了冗余特征,将原来 1 688 维的特征编码降低为 310 维。从图 4 中的五个性能指标来看,该文使用的 XGBoost 特征选择方法在降低特征编码维度的同时,使模型的性能在所有的评价指标上都有较好的提高。

5.4　不同分类器性能

该文比较了 4 个不同模型和 Stack6mAPred 集成模型的分类性能结果。其敏感性、特异性、准确度、MCC、AUC 值等 5 个性能指标对比如图 5 所示。

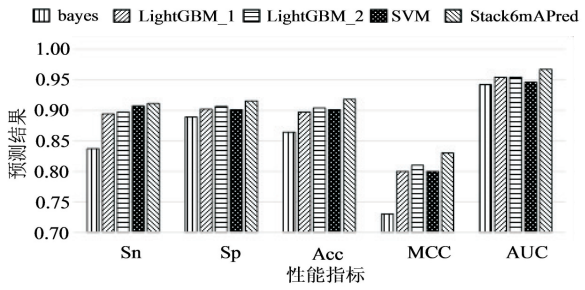

图 5　不同分类器性能对比

以准确率和 AUC 值来说,单个模型均达到了较好的性能,准确率都达到了 86% 以上,AUC 值都在 0.942 以上。其中 LightGBM 是单个分类器中性能最好的分类器,不同参数的两个 LightGBM 分类器的准确率分别达到了 89.7% 和 90.4%,AUC 值均为 0.954;其次支持向量机模型的准确率为 90.1%,AUC 值为 0.946;朴素贝叶斯模型效果最差。与 4 个单一模型相比,Stack6mAPred 集成模型效果最好,最终预测准确率为 91.8%,AUC 值为 0.967,并且敏感性、特异性、马修斯相关系数这些性能指标都有了明显的提升,说明 Stack6mAPred 集成模型将单一模型优势互补,使预测性能有了较好的提升。

5.5　与现有方法性能比较

为了验证 Stack6mAPred 模型的性能,将 Stack6mAPred 与现有模型进行对比。根据调查得知,目前共有 4 种预测 6mA 位点的工具:i6mA-Pred[8]、iDNA6mA[9]、MM-6mAPred[10] 和 csDMA[11]。该研究将 Stack6mAPred 与以上 4 种工具从敏感性、特异性、准确度、MCC 和 AUC 这五个指标进行对比分析,比较结果如表 5 所示。

表 5　与现有方法性能比较

Method	Sn/%	Sp/%	Acc/%	MCC	AUC
i6mA-Pred	82.95	88.86	83.30	0.66	0.886
iDNA6mA	86.70	90.00	86.59	0.73	0.931
csDMA	84.20	88.00	86.10	0.72	0.923
MM-6mAPred	89.32	90.12	90.11	0.77	0.936
Stack6mAPred	91.02	91.48	91.83	0.83	0.967

从结果中可知,Stack6mAPred 预测模型在五个指标上均高于现有工具中最好的 MM-6mAPred,其中 MCC 值提升最为显著,提升了 0.06,敏感性、特异性、准确度和 AUC 分别提升了 1.7%、1.36%、1.72% 和 0.031,这说明了提出的 stacking 集成方法对于 6mA 位点预测的有效性。

6　结束语

该文提出了一种基于集成学习的水稻基因组中的 6mA 位点识别方法 Stack6mAPred。该方法组合了 5 种类型的特征,并且通过 XGBoost 进行特征选择,去除冗余特征,避免了模型过拟合;集成了朴素贝叶斯、支持向量机、LightGBM 和逻辑回归等异构分类器;实现了基分类器之间的强强联合和优势互补,最终构建出一个性能更强的集成预测模型。

该研究以水稻基因组数据为研究对象,构造模型的方法可以迁移到预测其他物种的 N6 甲基化位点识别中。该研究仅仅集成两层模型,根据需要可以集成更多层的模型,需要指出的是随着模型集成的层次增多,训练的时间会有所增加。

参考文献:

[1] 宋　乔,高世超,王培昌. DNA 甲基化的分子机制及其研究进展[J]. 基因组学与应用生物学,2019,38(7):3317-3322.

[2] XIAO C L,ZHU S,HE M,et al. N6-methyladenine DNA modification in the human genome[J]. Molecular Cell,2018,71(2):306-318.

[3] LUO G Z,HE C. DNA N 6-methyladenine in metazoans: functional epigenetic mark or bystander? [J]. Nature Structural & Molecular Biology,2017,24(6):503-506.

[4] POMRANING K R,SMITH K M,FREITAG M. Genome-wide high throughput analysis of DNA methylation in eukaryotes[J]. Methods,2009,47(3):142-150.

[5] FLUSBERG B A,WEBSTER D R,LEE J H,et al. Direct detection of DNA methylation during single-molecule,real-time sequencing[J]. Nat Methods,2010,7(6):461-465.

[6] KRAIS A M, CORNELIUS M G, SCHMEISER H H. Genomic N(6)-methyladenine determination by MEKC with LIF[J]. Electrophoresis,2010,31(21):3548-3551.

[7] CHAO Z,WANG C,LIU H,et al. Identification and analysis of adenine N6-methylation sites in the rice genome[J]. Nature Plants,2018,4(8):554-563.

[8] CHEN W,LV H,NIE F,et al. i6mA-Pred:identifying DNA N6-methyladenine sites in the rice genome[J]. Bioinformatics,2019,35(16):2796-2800.

[9] TAHIR M,TAYARA H,CHONG K T. iDNA6mA (5-step rule):identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou´s 5-step rule[J]. Chemometrics and Intelligent Laboratory Systems,2019,189:96-101.

[10] PIAN C,ZHANG G,LI F,et al. MM-6mAPred:identifying DNA N6-methyladenine sites based on Markov model[J]. Bioinformatics,2020,36(2):388-392.

[11] LIU Z,DONG W,JIANG W,et al. csDMA:an improved bioinformatics tool for identifying DNA 6 mA modifications

via Chou′s 5-step rule[J]. Scientific Report, 2019, 9(1): 13109.

[12] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//ACM SIGKDD international conference on knowledge discovery & data mining. New York: Association for Computing Machinery, 2016: 785-794.

[13] 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.

[14] CHEN Z, ZHAO P, LI F, et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data[J]. Briefings in Bioinformatics, 2020, 21(3): 1047-1057.

[15] MUHAMMOD R, AHMED S, MD FARID D, et al. PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences[J]. Bioinformatics, 2019, 35(19): 3831-3833.

[16] CHEN Z, ZHAO P, LI F, et al. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences[J]. Bioinformatics, 2018, 34(14): 2499-2502.

[17] NAIR A S, SREENADHAN S P. A coding measure scheme empoying electron-ion interaction pseudo potential (EIIP)[J]. Bioinformation, 2006, 1(6): 197-202.

[18] NOBLE W S, KUEHN S, THURMAN R, et al. Predicting the in vivo signature of human gene regulatory sequences[J]. Bioinformatics, 2005, 21(Suppl 1): i338-i343.

[19] LEE D. Discriminative prediction of mammalian enhancers from DNA sequence[J]. Genome Res., 2011, 21(12): 2167-2180.

[20] KEOGH E, MUEEN A. Curse of dimensionality[J]. Indengchem, 2009, 29(1): 48-53.

[21] 李 航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 137-152.

[22] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 73-92.

[23] SCHAPIRE R E, FREUND Y. Boosting: foundations and algorithms[M]. Cambridge, MA: The MIT Press, 2012: 93-128.

[24] WOLPERT D H. Stacked generalization[J]. Neural Networks, 1992, 5(2): 241-259.

[25] CORTES C, VAPNIK V N. Support vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[26] XING P, SU R, GUO F, et al. Identifying N(6)-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine[J]. Scientific Reports, 2017, 7: 46757.

[27] 王慧勤, 雷 刚. 基于 LIBSVM 的风速预测方法研究[J]. 科学技术与工程, 2011, 11(22): 5440-5442.

[28] KNEBEL T, HOCHREITER S, OBERMAYER K. An SMO algorithm for the potential support vector machine[J]. Neural Computation, 2008, 20(1): 271-287.

[29] KE G, MENG Q, FINELY T, et al. LightGBM: a highly efficient gradient boosting decision tree[C]//Neural information processing systems. New York: Curran Associates Inc, 2017: 3148-3156.

[30] VACIC V, IAKOUCHEVA L M, RADIVOJAC P. Two sample logo: a graphical representation of the differences between two sets of sequence alignments[J]. Bioinformatics, 2006, 22(12): 1536-1537.