文章编号：1672-6987(2018)01-0106-08；DOI：10.16351/j.1672-6987.2018.01.017

基于集成学习的人类 LncRNA 大数据基因预测

于 彬，李 珊，陈 成，陈瑞欣，田保光

（青岛科技大学 数理学院，山东 青岛 266061）

摘 要： 长非编码 RNA（LncRNA）在表观遗传调控、转录后调控和人类疾病中发挥着重要作用，利用机器学习方法从海量的 RNA 数据中识别出 LncRNA 十分必要。本研究提出一种基于集成学习的 LncRNA 大数据基因预测新方法。首先提取序列碱基出现频率的 86 个特征作为原始特征集合，其次，基于 GA-SVM 选取出最优特征，以 SVM 五折交叉验证的准确率作为适应度，最后构建 AdaBoost 算法与 SVM 相结合的基因预测模型（AdaBoost-SVM）。实验结果表明：AdaBoost-SVM 模型对测试集 LncRNA 的预测准确率为 89.26%，优于 RF、SVM 和 DWT-SVM3 种预测模型的结果。

关键词： 长非编码 RNA；基因预测；集成学习；AdaBoost 算法；支持向量机

中图分类号： Q 811.4 **文献标志码：** A

引用格式： 于彬，李珊，陈成，等. 基于集成学习的人类 LncRNA 大数据基因预测[J]. 青岛科技大学学报(自然科学版)，2018，39(1)：106-113.

YU Bin，LI Shan，CHEN Cheng，et al. Prediction of human LncRNA big data genes based on ensemble learning[J]. Journal of Qingdao University of Science and Technology(Natural Science Edition)，2018，39(1)：106-113.

Prediction of Human LncRNA Big Data Genes Based on Ensemble Learning

YU Bin，LI Shan，CHEN Cheng，CHEN Ruixin，TIAN Baoguang

(College of Mathematics and Physics，Qingdao University of Science and Technology，Qingdao 266061，China)

Abstract：Long non-coding RNA（LncRNA）plays an important role in epigenetic regulation，post-transcriptional regulation and human diseases. It is of great necessity to identify LncRNA from vast amounts of RNA data by using machine learning. This paper presents a new method for predicting LncRNA big data genes based on ensemble learning. Firstly，such 86 features as frequency of occurrence of base sequence are extracted as initial characteristic sets. Secondly，the optimal features based on GA-SVM are selected，and 5-fold cross-validation accuracy of SVM is employed as fitness. Lastly，gene prediction model（AdaBoost-SVM）combined by AdaBoost algorithm and SVM is constructed. The experimental results show that the prediction accuracy of test set LncRNA based on AdaBoost-SVM model is 89.26%，which is better than that of the RF，SVM and DWT-SVM models.

Key words：long non-coding RNA；gene prediction；ensemble learning；AdaBoost algorithm；support vector machine

收稿日期： 2017-05-02

基金项目： 国家自然科学基金项目(51572136)；山东省自然科学基金项目(ZR2014FL021)；山东省高等学校科技计划项目(J17KA159).

作者简介： 于 彬(1977—)，男，副教授.

随着高通量测序技术的发展和应用以及生物技术与信息技术的融合,产生了大量基因组学和蛋白质组学数据,利用大数据挖掘方法探索生物学规律显得尤为重要[1-2]。非编码 RNA 是(non-coding RNA,ncRNA)指不编码蛋白质和酶,以 RNA 形式发挥作用的一类分子。在以前 ncRNA 基因被认为是"暗物质"和"垃圾 DNA"[3],海量的生物医学数据表明,ncRNA 在人类的生命活动中扮演着十分重要的角色[4]。长度大于 200 nt 的 RNA 分子被称为长非编码 RNA(LncRNA),研究表明 LncRNA 在肿瘤抑制、细胞凋亡和基因调控发挥着重要作用,对其深入研究可能揭示一个由 RNA 介导的遗传信息表达调控网络,从而为人类疾病的研究和治疗提供新的思路[5-6]。LncRNA 的工作主要是鉴定新的 ncRNA 并分析其功能,但是要从海量的 RNA 数据中识别出 ncRNA 是一件具有挑战的事,存在着测序深度、测序偏好和测序错误等问题。随着计算技术的快速发展,利用生物信息学方法预测新的 LncRNA 成为 RNA 组学的热点[7-8]。

生物分析技术的不断推出和更新,生物医学数据迅速积累,使得利用机器学习方法从海量 RNA 序列中识别出 LncRNA 成为可能。机器学习可以综合序列、结构和表达数据对新的 lncRNA 进行预测,其中支持向量机、随机森林、贝叶斯决策,决策树等监督学习方法[8-14]已经成功用于预测 ncRNA。CHANG 等[11]构建基于支持向量机的细菌小 ncRNA 的预测模型,利用大肠杆菌小非编码 RNA 数据验证了方法的有效性。也有利用非监督学习对 ncRNA 进行分析和预测,运用上下文敏感的隐马尔科夫链模型预测 microRNA 前体[12],该模型根据已知 microRNA 前体的序列,自动整合序列特征对未知 microRNA 前体进行预测。赵英杰等[13]建立基于支持向量数据描述的 ncRNA 基因识别模型,测试集从 NONCODE 数据库中的各种生物体中选取 240 条序列作为目标样本,采用支持向量数据描述的方法对测试集进行五折交叉验证,预测准确率最高为 89.01%。WANG 等[14]构建基于 GA-SVM 的人类长非编码 RNA 基因预测模型,根据 GA-SVM 方法选取的最优特征子集,利用人类 LncRNA 基因数据验证了方法的有效性。于彬等[8]提出基于离散小波变换与支持向量机相结合的基因预测模型(DWT-SVM),利用人类 ncRNA 数据验证了模型的有效性。但是上述方法构建的训练集中可能会有一些难以区分的样本,使得预测结果有所偏差,减弱模型的学习能力。

本研究提出一种基于集成学习的 LncRNA 基因预测新方法。首先从 GENCODE 数据库和 UCSC 数据库中提取人的 LncRNA 和 mRNA 序列数据,并选取碱基出现频率等 86 个特征作为原始材料。其次通过 GA-SVM 选取 49 个最优特征,最优特征包括单核苷酸出现频率、二核苷酸出现频率和三核苷酸出现频率。最后根据最优特征子集,构建基于 AdaBoost-SVM 的 LncRNA 基因预测模型,通过 AdaBoost 算法更新样本权重,利用多个分类器进行最终决策。与 RF、SVM 和 DWT-SVM 等模型的预测结果进行比较,可以发现基于集成学习的 AdaBoost-SVM 模型的预测效果最好,对于研究 LncRNA 的结构和功能具有一定的意义。

1　材料与方法

1.1　数据来源

本研究选取 GENCODE 数据库[15]中的人类 LncRNA 序列作为正例,通过剔除标记异常的、序列异常的 RNA,最后得到了 24 251 条 LncRNA 序列。为测试人类 LncRNA 的预测效果,随机选取 17 000条人类 LncRNA 序列作为分类器的训练集的正例,利用训练集筛选最优特征子集和确定最优参数,剩余的 7 251 条序列作为测试集的正例,LncRNA 序列数据可以从网址 http://www.gencodegenes.org/releases/current.html 下载。

LncRNA 序列和 mRNA 序列之间没有重叠,为区分人类 LncRNA 序列与其它序列的区别并验证模型的预测效果,本研究从 UCSC 数据库[16]中随机选取 24 251 条人类 mRNA 作为反例,以人类的 17 000 条 mRNA 序列作为训练集的反例,选取剩余的 7 251 条序列作为测试集的反例。mRNA 序列数据可以从网址 http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/下载。

1.2　特征提取

构造初始的特征集合对于选择最优子集是非常重要的,本研究将 LncRNA 序列看成特殊的字符串,选取字符出现频率作为训练特征,即选取每条序列的单核苷酸出现频率,二核苷酸对出现频率(AA%,…,CC%,(G-C)%,(A-T)%)和三核苷酸出现频率(AAA%,AAT%,…,CCT%,CCC%)共 86 个序列特征,作为初始的特征集合[6,8],见表 1。

表1 碱基频率特征

Table 1 Frequency characteristics of base

特征描述	特征
单核苷酸出现频率	A%,G%,C%,T%
二核苷酸出现频率	AA%,AG%,AC%,AT%,GA%,GG%,GC%,GT%,CA%,CG%,CC%,CT%,TA%,TG%,TC%,TT%,(G-C)%,(A-T)%
三核苷酸出现频率	AAA%,AAG%,AAC%,AAT%,GGA%,GGG%,GGC%,GGT%,CCA%,CCG%,CCC%,CCT%,TTA%,TTG%,TTC%,TTT%,ATA%,ATG%,ATC%,ATT%,AGA%,AGG%,AGC%,AGT%,ACA%,ACG%,ACC%,ACT%,GAA%,GAG%,GAC%,GAT%,GCA%,GCG%,GCC%,GCT%,GTA%,GTG%,GTC%,GTT%,CAA%,CAG%,CAC%,CAT%,CGA%,CGG%,CGC%,CGT%,CTA%,CTG%,CTC%,CTT%,TAA%,TAG%,TAC%,TAT%,TGA%,TGG%,TGC%,TGT%,TCA%,TCG%,TCC%,TCT%

1.3 遗传算法

遗传算法(genetic algorithm,GA)是由美国的HOLLAND教授首次提出并发展起来的一种随机自适应的全局搜索算法[17]。该算法模拟自然遗传过程中的繁殖、交叉和基因突变。利用遗传算子(选择、交叉和变异)产生新一组候选解,以实现优胜劣汰的过程。遗传算法的操作对象是一组可行解组成的群体,其中每一个个体表示一个编码字符串,使算法具有良好的全局优化性。

个体是遗传算法的基本对象和结构,一定数量的个体构成种群。串是个体的表现形式,对应于生物界的染色体可以是二进制的,也可以是实值型的。

$A=a_1a_2\cdots a_N$。

其中:A 称为个体,a_i 称为基因,表示个体的特征。

适应度函数是个体空间 Ω 到实数空间 $\mathbb{R}$ 的一种映射:

$f:\Omega\rightarrow\mathbb{R}$。

$f(A)\in\mathbb{R}$ 表示适应度,其值越大说明该个体的生存能力越大,有更多的繁殖机会,适应度函数也称优化目标函数。

种群空间到个体空间的映射称为交叉映射,即随机确定一个基因作为交叉点,按照交叉率 P_c 交换两个个体的后半部分。个体空间到个体空间的随机映射称为变异,即按照变异概率 P_m 随机改变一个基因位置的值。

1.4 支持向量机

支持向量机(support vector machine,SVM)是一种基于统计学习理论的机器学习方法[18],它将输入样本集合映射到高维空间,构造最优超平面,使得超平面到两个样本集之间的距离到达最大。针对维数高、样本小、非线性的生物信息数据,SVM表现出了优良的性能[19-20]。

假设 n 个样本的训练集

$$D=\{(x_i,y_i)\mid i=1,2,\cdots,n\},\ x\in\mathbb{R}^n,\ y\in\{-1,1\}。\tag{1}$$

SVM是寻找一个超平面将正例和反例没有错误的分开,且两类样本集间距最大,即

$$\varphi(w,\xi)=\frac{1}{2}(w^{\mathrm{T}}w)+C(\sum_{i=1}^{n}\xi_i),$$

$$s.t.\ |w^{\mathrm{T}}x_i+b|-1+\xi_i\geqslant 0,\quad i=1,2,\cdots,n。\tag{2}$$

其中:$\varphi(w,\xi)$ 表示优化函数,C 是给定的常数,表示惩罚参数,$\xi_i\geqslant 0$,表示松弛变量。

利用拉格朗日函数求解此优化问题,可以转化为如下凸二次规划的对偶问题:

$$\max\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_iy_j\alpha_i\alpha_jK(x_i,x_j),$$

$$s.t.\ \sum_{i=1}^{n}y_i\alpha_i=0,0\leqslant\alpha_i\leqslant C,\ i=1,2,\cdots,n。\tag{3}$$

其中:α_i 表示拉格朗日乘子,$K(x_i,x_j)$ 为核函数。相应的判别函数:

$$f(x)=\mathrm{sgn}\{\sum_{i=1}^{n}\alpha_i^*y_iK(x_i,x)+b^*\}。\tag{4}$$

本研究选用径向基核函数 $K(x_i,x_j)=\exp(-\frac{\|x_i-x_j\|^2}{\sigma^2})$ 作为核函数。

针对线性不可分情况SVM通过核函数将样本空间映射到一个高维特征空间,通过训练数据寻找一个最大间隔,构造最优分类面使输入的样本线性可分,可以有效提高模型的泛化能力。本研究使用的是CHANG和LIN开发的LIBSVM软件[21],可以从网址 http://www.csie.ntu.edu.tw/~cjlin/libsvm/下载。

1.5 集成学习

集成学习(ensemble learning)是一种新的机器学习范式,通常是基于多种分类器集成进行最终的

决策,可以有效提高分类器的泛化能力[22]。常用的集成学习方法有 AdaBoost 算法和 Bagging 算法。1990 年,SCHAPIRE[23] 通过一个构造方法证明多个基本分类器可以集成强分类器,奠定了集成学习的理论基础。1995 年,FREUND 和 SCHAPIRE[24] 提出了 AdaBoost 算法,可以容易的应用于实践中,成为比较流行的集成学习算法。AdaBoost 算法是不需要知道弱学习分类器的误差,且强分类器的精度依赖于所有弱分类器的分类精度,这样可以深入挖掘弱分类器的潜力。

AdaBoost 算法是通过每个样本对应的权重来实现,给定训练样本集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\},y\in\{-1,1\}\}$。初始时每个样本对应的权重是相同的,即 $U_1=\frac{1}{n}$,并按照此样本分布对训练样本集进行训练,训练出一弱分类器 h_1。对于 h_1 错分的样本适当增加该权重,未错分的样本适当降低其权重,以此更新训练集样本的分布。在新的样本分布下,再对基本学习分类器进行训练,得到 h_2。反复迭代 T 次,得到 T 个弱分类器,最终的集成分类器是每个基本分类器的加权投票,本研究选用 SVM 作为基分类器。AdaBoost 算法具体流程如下:

1)给定训练样本集,每一个训练样本集有初始化权重 $D_1(i)=\frac{1}{n},(i=1,2,\cdots,n)$。

2)计算基本分类器的训练偏差 $\varepsilon_t=\sum_{i=1}^{n}D_t(i)\{h_t(x_i)\neq y_i\}$。

3)循环迭代 T 次,并对每个训练样本的权重进行更新

$$D_{t+1}(x)=\frac{D_t(x)}{Z_t}\times\begin{cases}\exp(-\alpha_t), & h_t(x)=y,\\ \exp(\alpha_t), & h_t(x)\neq y。\end{cases}=\frac{D_t(x)\exp(-\alpha_t y h_t(x))}{Z_t}。\tag{5}$$

其中:Z_t 是标准化因子,$\alpha_t=\frac{1}{2}\ln(\frac{1-\varepsilon_t}{\varepsilon_t})$,$h_t$ 是弱分类器。

4)最终的强分类器是 H 通过多个带权重的基本分类器得到,H 可以描述为

$$H(x)=\mathrm{sgn}(\sum_{t=1}^{T}\alpha_t h_t(x))。\tag{6}$$

1.6　评价指标

为评估预测模型的性能,采用敏感度(SE)、特异度(SP)、准确率(ACC)、和 Matthews 相关系数(MCC)对预测结果进行评价[8,14]。定义如下:

$$SE=\frac{TP}{TP+FN},\tag{7}$$

$$SP=\frac{TN}{TN+FP},\tag{8}$$

$$ACC=\frac{TP+TN}{TP+TN+FN+FP},\tag{9}$$

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

上述评价指标中的 TP、TN、FP 和 FN 分别代表真阳性、真阴性、假阳性和假阴性。其中 TP 表示正确预测的正例样本数,TN 表示正确预测的反例样本数,FP 表示反例样本被错误预测为正例样本数,FN 表示正例样本被错误预测为反例样本数。

1.7　基于集成学习的 LncRNA 基因预测方法

本研究提出基于集成学习的人类 LncRNA 基因预测方法称之为 AdaBoost-SVM,计算流程如图 1 所示。仿真实验环境:Windows Server 2012 R2 Intel(R) Xeon(TM) CPU E5-2650@2.30 GHz 32.0 GB 的内存,MATLAB 2014a 编程实现。

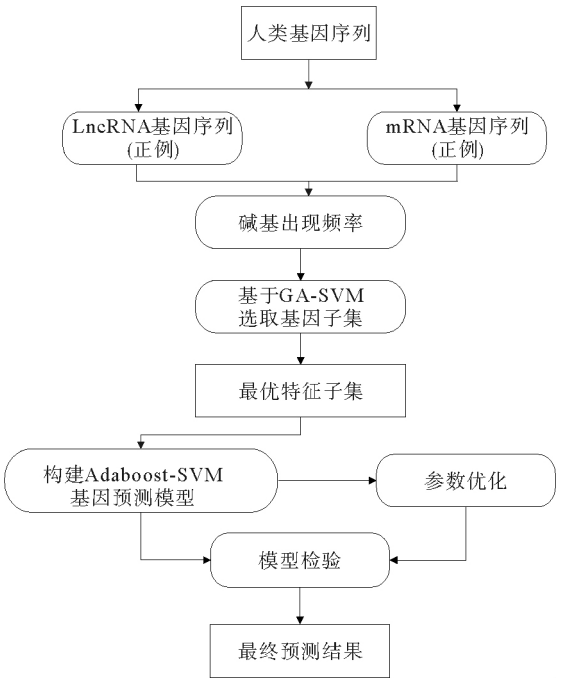


图 1　基于 AdaBoost-SVM 的 LncRNA 基因预测流程图

Fig. 1　Flow chart of LncRNA gene prediction based on AdaBoost-SVM

AdaBoost-SVM 预测方法的步骤描述为

1)输入训练集中的人类 LncRNA(正例)和 mRNA(反例)的基因序列数据,以及所对应的类别

标签。

2)根据表1所示的碱基频率特征,提取LncRNA和mRNA基因序列的86个特征,将碱基序列转化为数值信号,于是训练集可表示为

$D=\{(x_i,y_i)|i=1,2,\cdots,34\ 000\}$, $x\in\mathbb{R}^{86}$, $y\in\{-1,1\}$。

3)选用二进制编码个体,并以SVM五折交叉验证的测试准确率作为适应度,基于GA-SVM方法选择出特征子集,通过交叉、变异选择新的训练集S_1,相应的测试集为S_2。

$S_1=\{(x_i,y_i)|i=1,2,\cdots,34\ 000\}$, $x\in\mathbb{R}^{49}$, $y\in\{-1,1\}$,

$S_2=\{(x_i,y_i)|i=1,2,\cdots,14\ 502\}$, $x\in\mathbb{R}^{49}$, $y\in\{-1,1\}$。

4)构建AdaBoost-SVM预测模型。首先确定训练集S_1中每个训练样本的初始化权重为$\frac{1}{34\ 000}$,将S_1输入到SVM中;其次计算训练偏差$\varepsilon_t=\sum_{i=1}^{n}D_t(i)\{h_t(x_i)\neq y_i\}$,通过式(5)更新权重,错分的样本适当增加该权重,未错分的样本适当降低其权重;再次通过重采样法得到训练样本集对基分类器进行训练,循环迭代T次;最后根据式(6)预测测试集的类别标签。

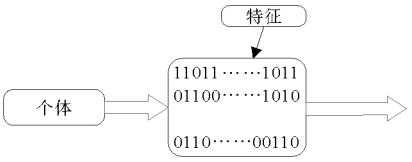
5)根据SE、SP、ACC、MCC的计算结果,与其它预测模型的结果进行比较,评价AdaBoost-SVM预测模型的优劣。

2 结果与讨论

GA-SVM是一种有效的特征特取方法,可以将遗传算法的遍历搜索能力和SVM的学习能力相结合,根据自然选择的生存法则选择较为重要的特征子集。GA-SVM已经被广泛用于生物信息学领域,如在肿瘤基因分类中提取特征基因[25]。本研究用0或1字符串向量作为个体,以SVM五折交叉验证的预测准确率作为适应度,通过交叉、变异和选择生成最优特征子集。初始种群大小为40,迭代次数为4,具体步骤如下:

1)编码。把每一个候选解编码为向量,从n个特征中选择a个特征,可以用n位的0或者1的字符串向量表示,其中1表示选择该特征,0表示未选择该特征,编码操作如图2所示。

随机产生40个个体的种群,每个个体基因γ_i被选择的概率为α,即



1表示选取该特征;0表示不选取该特征。

图2 编码过程

Fig. 2 Process of encoding

$$p(\gamma_i=1)=\alpha,\ \gamma_i:\text{Benoulli}(\alpha),\ (0\leqslant\alpha\leqslant1,i=1,2,\cdots,n) \tag{11}$$

其中$\gamma_i,(i=1,2,\cdots,n)$服从Benoulli分布。

2)计算适应度。根据公式(4)的判别函数预测样本的类别标签,以支持向量机五折交叉验证的分类准确率作为适应度。五折交叉验证的基本思想是:将训练集随机分成5组,选取4组训练,剩余1组进行测试,以5个模型测试准确率的平均值作为验证指标。

3)交叉、变异和选择。

交叉:选用单点交叉对染色体进行操作,交叉可以组合双亲的优良特性,从而生成良好的可行解。单点交叉对每一组配对的个体,随机选择位置交叉点,依照设定的交叉概率在其位置交叉处互换部分染色体,从而生成两个新的个体。

变异:在生物的遗传和进化过程中,可能会因为一些偶然的因素产生突变,于是遗传算法中也引入了变异这个操作,依照设定的变异概率,将染色体编码串中的等位基因用其它染色体上的等位基因替代,从而形成新的个体,这里通过0,1互换来实现。

随着种群的进化,交叉率越来越低,变异率越来越高。于是交叉率P_c和变异率P_m是动态变化的,参数的变化与进化代数相关,P_c和P_m定义如下

$$P_c=1-\frac{g}{G\times4},$$

$$P_m=\frac{g}{G\times16}。$$

其中:G表示总的进化代数,数值是4。g表示当前进化代数,因此P_c值的范围在0.75~0.937 5之间变化,P_m值的范围在0.015 625~0.062 5之间变化。

选择:采用赌轮选择方法生成新一代种群,假设个体i的适应度为E_i,由于本研究设置群体大小为40,则个体被选中是概率P_i为

$$P_i=\frac{E_i}{\sum_{i=1}^{40}E_i}。 \tag{12}$$

由式(12)可知,适应度大的个体被选择的概率

就会越大。

4)选取最优特征子集。以适应能力强的个体包含的特征构造新的特征空间,重复步骤 2)、3)步,直到满足迭代次数。

本研究通过设定个体基因 γ_i 被选择的概率 α 的大小,可以生成不同的特征子集合。随机选取人的 LncRNA 序列作为正例(共 17 000 条),mRNA 作为反例(共 17 000 条),训练集数据共34 000 条序列,以训练集五折交叉验证的预测准确率选取最优特征子集,依次设定 α 为 0.35、0.4、0.45、0.5、1,基于 GA-SVM 产成 5 个特征子集,集合大小依次为 39、49、51、55、86。采用敏感度 SE、特异度 SP、准确率 ACC 和 Matthews 相关系数 MCC 来评价最优特征子集的预测效果,预测结果如表 2 所示,最优特征子集的特征及其描述如表 3 所示。

表 2　基于五折交叉验证的特征子集的预测结果比较

Table 2　Performance comparison of the feature sets using 5-fold cross validation

特征子集	SE/%	SP/%	ACC/%	MCC/%	C	g
39 个特征	86.86	88.49	87.68	75.36	5.2780	0.011 840
49 个特征	87.61	88.81	88.21	76.43	1.7411	0.020 617
51 个特征	87.01	88.76	87.88	75.78	1.7411	0.035 897
55 个特征	85.61	88.47	87.04	74.11	1.7411	0.020 617
86 个特征	86.96	88.72	87.84	75.69	1.7411	0.035 897

由表 2 可知,最优特征子集的大小为 49 个特征,准确率 ACC 最高,为 88.21%,与其它 4 个特征子集得到的结果相差 0.33%～1.17%。敏感度 SE 最高,为 87.61%,其实质是 LncRNA 基因预测的准确率,与其它 4 个特征得到的结果相差 0.6%～2%。Matthews 相关系数 MCC 也最高,为 76.43%,与其它 4 个特征得到的结果相差 0.65%～2.32%。此时 SVM 径向基核函数的最佳参数 $C=1.7411$,$g=0.020\ 617$。

表 3　最优特征子集的 49 个特征及其描述

Table 3　49 features of the optimal feature subset and their description

特征描述	特征
单核苷酸出现频率	A%,G%,C%
二核苷酸出现频率	AT%,AG%,AC%,GT%,GC%,CA%,CG%,(G-C)%,(A-T)%
三核苷酸出现频率	AAA%,AAT%,ATT%,AGT%,AGC%,ACT%,ACC%,TAA%,TAG%,TAC%,TTT%,TTG%,TTC%,TGT%,TGG%,TGC%,TCA%,TCG%,GAA%,GTA%,GTG%,GTC%,GGA%,GGG%,GGC%,GCA%,CAG%,CAC%,CTT%,CTG%,CTC%,CGA%,CGT%,CCA%,CCT%,CCG%,CCC%

由以上分析可知,基于遗传算法与 SVM 相结合的特征提取方法,可以充分利用两者的优点,具有较好的收敛速度和较高的预测准确率。GA-SVM 选取的最优特征子集可以有效减少序列信息的混淆,能够较好的描述 LncRNA 的本质属性,使 LncRNA 和 mRNA 序列之间具有相差较大的序列特征信息,从而为集成学习预测人类 LncRNA 基因提供可靠的特征信息。

由于 AdaBoost 算法可以和其它算法结合使用,预测模型的好坏与所选用的基本分类器密切相关,本研究首先把 SVM 作为基本分类器进行 LncRNA 基因预测。在 AdaBoost-SVM 预测模型中,首先根据 GA-SVM 选取的最优特征子集构造新的训练集和测试集,将样本及其类别标签输入到 SVM 中。其次增加 SVM 错分样本的权重,减少未错分样本的权重,也就是利用式(5)更新权重。最后设定迭代次数为 10,根据式(6)预测测试集的类别标签。实验表明 AdaBoost-SVM 模型对测试集数据的敏感度为 89.26%,特异度为 89.04%,预测准确率为 89.15%,MCC 为 78.29%。

为了便于比较,本研究又提出了基于随机森林的 LncRNA 基因预测模型,称之为 RF。RF 是一种

基于决策树的分类器算法[26]，采用 Bagging 方法从训练集中随机选取训练子集构建基本学习器，同时在构建每棵决策树时，在每个节点处随机选取特征子集进行分裂，该预测模型的基决策树的数目设定为100。

以人类 LncRNA 序列作为训练集的正例(共17 000条)，人类 mRNA 序列作为训练集的反例(共17 000条)，测试集包含14 502条序列。基于 GA-SVM 方法选取的49个最佳特征子集，得到 AdaBoost-SVM 模型、RF 模型、未进行集成学习的 SVM 模型及 DWT-SVM 预测模型[8]的 LncRNA 基因测试集数据主要预测结果，4种方法的预测结果如表4所示。

表4 测试集的人类 LncRNA 基因预测结果

Table 4 Prediction results of human LncRNA gene in test set

预测模型	TP	TN	SE/%	SP/%	ACC/%	MCC/%
AdaBoost-SVM	6 472	6 456	89.26	89.04	89.15	78.29
RF	6 293	6 076	86.79	83.80	85.29	70.61
SVM	6 280	6 452	86.61	88.98	87.79	75.61
DWT-SVM	6 374	6 457	87.91	89.05	88.48	76.96

由表4可知，AdaBoost-SVM 模型的敏感度为89.26%，其实质就是 LncRNA 基因预测的准确率，特异度为89.04%，准确率为89.15%，MCC 为78.29%。AdaBoost-SVM 模型的敏感度比 DWT-SVM 模型高1.35%，比 RF 模型高2.47%，比 SVM 模型高2.65%。AdaBoost-SVM 模型的准确率比 DWT-SVM 模型高0.67%，比 RF 模型高3.86%，比 SVM 模型高1.36%。AdaBoost-SVM 模型的 MCC 也是最高的，比其它3种预测模型高1.33%～7.68%。

从表4中可以看出，AdaBoost-SVM 模型的预测效果最好，DWT-SVM 模型和 SVM 模型次之，RF 模型的预测效果最差。AdaBoost-SVM 模型可以有效预测人类 LncRNA，这是因为：(1)GA-SVM 选取的最优特征子集可以减少序列信息的混淆，挖掘 LncRNA 序列的本质特征，使正例和反例序列之间具有相差较大的序列信息，从而为集成学习预测 LncRNA 提供可靠的特征信息；(2)AdaBoost 算法根据基分类器的表现调整训练样本分布，重采样后使得 SVM 更加关注分类错误的样本。而且使用加权投票规则，使学习能力强的基本分类器赋予了更高的投票权重，算法同时解决了基本分类器的生成和集成的问题；(3)该模型可以结合支持向量机优良的预测性能，深入挖掘 SVM 的训练学习潜力，能够有效减小训练误差，提高分类精度。

AdaBoost-SVM 预测模型可以打破原始样本分布，重新采样后使 SVM 更加关注难学习的样本，不需要先验知识，可获得比单一学习器显著优越的泛化性能。而且 AdaBoost 算法不需要预先知道基分类器错误率上限，当基分类器的数量趋于无穷时，强分类器的错误率趋近于零。利用该模型可以从海量的 RNA 序列中成功识别出 LncRNA，对于研究 LncRNA 的结构和功能具有重要意义，从而为肿瘤抑制、精准医疗提供理论依据。AdaBoost-SVM 模型也可以应用到亚细胞定位预测、蛋白质相互作用预测及药物靶点预测等生物信息学研究热点领域。

3 结 语

本研究提出了一种基于集成学习的人类 LncRNA 基因预测新方法，首先选择单核苷酸、二核苷酸出现频率等86个特征作为原始数据，其次基于 GA-SVM 方法确定最优特征子集，最后构建基于集成学习的 LncRNA 大数据基因预测模型。AdaBoost-SVM 模型的预测准确率为89.26%，优于 RF、SVM 和 DWT-SVM 3种预测方法。实验表明该模型具有较强的泛化能力，可以有效预测 LncRNA，对于研究 LncRNA 的结构和功能具有重要意义。但是 AdaBoost 易受噪声干扰，执行效果依赖于 SVM 的预测性能，并且基分类器的训练时间偏长。目前国际上对于 LncRNA 的预测仍处在起步和探索阶段，随着高通量测序技术的发展，使得基于机器学习的生物信息学方法预测 LncRNA 显得尤为重要。选择性集成学习是一种应用广泛的集成学习方法，利用选择性集成学习进行 LncRNA 基因预测将是下一步的研究方向。

参 考 文 献

[1] 金晨，瞿坤. 生物信息技术在表观遗传调控机制研究中的应用[J]. 中国科学：生命科学，2017，47(1)：116-124.
JIN Chen，QU Kun. Application of bioinformatics techniques in

revealing the mechanisms of epigenetic regulation[J]. Scientia Sinica Vitae, 2017,47(1):116-124.

[2] 宁康, 陈挺. 生物医学大数据的现状与展望[J]. 科学通报, 2015, 60(5/6):534-546.
NING Kang, CHEN Ting. Big data for biomedical research: Current status and prospective[J]. Chinese Science Bulletin, 2015, 60(5/6):534-546.

[3] PONTING C P, BELGARD T G. Transcribed dark matter: Meaning or myth?[J]. Human Molecular Genetics, 2010, 19(2): R162-R168.

[4] QUINN L, FINND S P, CUFFEE S, et al. Non-coding RNA repertoires in malignant pleural mesothelioma[J]. Lung Cancer, 2015, 90(3): 417-426.

[5] CESANA M, CACCHIARELLI D, LEQNINI I, et al. A long noncoding RNA controls muscle differentiation by functioningas a competing endogenous RNA[J]. Cell, 2011, 147(2): 358-369.

[6] CABILI M N, TRAPNELL C, GOFF L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses[J]. Genes & Development, 2011, 25(18): 1915-1927.

[7] SCHRAGA S, BERNSTEIN D A, MUMBACH M R, et al. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA[J]. Cell, 2014, 159(1): 148-162.

[8] 于彬, 陈成, 刘健, 等. 基于支持向量机的人类 ncRNA 基因预测[J]. 青岛科技大学学报(自然科学版), 2017, 38(2): 112-118.
YU Bin, CHEN Cheng, LIU Jian, et al. Prediction of human non-coding RNA genes based on support vector machine[J]. Journal of Qingdao University of Science and Technology (Natural Science Edition), 2017, 38(2): 112-118.

[9] WANG Y Q, CHEN X W, JIANG W, et al. Predicting human microRNA precursors based on an optimized feature subset generated by GA-SVM[J]. Genomics, 2011, 98(2): 73-78.

[10] NG K L, MISHRA S K. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures[J]. Bioinformatics, 2007, 23(11): 1321-1330.

[11] CHANG T H, WU L C, LIN J H, et al. Prediction of small non-coding RNA in bacterial genomes using support vector machines[J]. Expert Systems with Applications, 2010, 37(8): 5549-5557.

[12] AGARWAL S, VAZ C, BHATTACHARGA A, et al. Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM)[J]. BMC Bioinformatics, 2010, 11(1): 1-7.

[13] 赵英杰, 王正志. 基于支持向量机描述的非编码 RNA 基因识别[J]. 生物医学工程学杂志, 2010, 27(4): 779-784.
ZHAO Yingjie, WANG Zhengzhi. Support vector data description for finding non-coding RNA gene[J]. Journal of Biomedical Engineering, 2010, 27(4): 779-784.

[14] WANG Y Q, LI Y, WANG Q, et al. Computational identification of human long intergenic non-coding RNAs using a GA-SVM algorithm[J]. Gene, 2014, 533(1): 94-99.

[15] DERRIEN T, JOHNSON T R, BUSSOTTI G, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression[J]. Genome Research, 2012, 22(9): 1775-1789.

[16] KAROLCHIK D, DERRIEN T, JOHNSON R, et al. The UCSC Genome Browser database: 2014 update[J]. Nucleic Acids Research, 2014, 42: D764-D770.

[17] HOLLAND J H. Genetic algorithms[J]. Scientific American, 1992,267:66-72.

[18] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer-Verlag New York Inc, 1995.

[19] YU B, ZHANG Y. The analysis of colon cancer gene expression profiles and the extraction of informative genes[J]. Journal of Computational and Theoretical Nanoscience, 2013, 10(5): 1097-1103.

[20] YU B, ZHANG Y, ZHAO L K. Cancer classification by a hybrid method using microarray gene expression data[J]. Journal of Computational and Theoretical Nanoscience, 2015, 12(10): 3194-3200.

[21] CHANG C C, LIN C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27.

[22] 江峰, 张有强, 杜军威, 等. 一种基于抽样与约简的集成学习算法[J]. 青岛科技大学学报(自然科学版), 2016, 37(4): 451-456.
JIANG Feng, ZHANG Youqiang, DU Junwei, et al. An ensemble learning algorithm based on sampling and reduction[J]. Journal of Qingdao University of Science and Technology (Natural Science Edition), 2016, 37(4): 451-456.

[23] SCHAPIRE R E. The strength of weak learn ability[J]. Machine Learning, 1990, 5(2): 197-227.

[24] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. European Conference on Computational Learning Theory, 1995, 55(7): 23-37.

[25] YU B, ZHANG Y. The analysis of colon cancer gene expression profiles and the extraction of informative genes[J]. Journal of Computational and Theoretical Nanoscience, 2013, 10(5): 1097-1103.

[26] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

(责任编辑　姜丰辉)