计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201100195

核小体定位预测的集成学习方法

陈 伟 李 杭 李维华
云南大学信息学院 昆明 650500
(2810925735@qq.com)

摘 要 核小体定位指DNA双螺旋相对于组蛋白的位置,并在DNA的转录阶段起着重要的调节作用。依靠生物实验的手段测得核小体定位会消耗大量的时间和资源,因此基于计算方法利用DNA序列进行核小体定位预测成为了一个重要的研究方向。针对核小体定位预测中单一模型和单一编码在DNA序列特征表示和学习方面的不足,文中提出了一种端到端的集成深度学习模型FuseENup,利用3种编码方式从多个维度表示DNA数据,利用不同的模型从不同维度提取数据中隐含的关键特征,构造了一种全新的DNA序列表征模型。在4种数据集上进行20倍交叉验证,相比当前针对核小体定位预测问题综合性能最优的模型CORENup,FuseENup的准确度(Accuracy)和精度(Precision)在HS数据集上提高了3%和9%,在DM数据集上提高了2%和6%,在E数据集上提高了1%和4%,相比其他的机器学习和深度学习基准模型,FuseENup具有更好的性能。实验结果表明,FuseENup能提高核小体定位的预测准确度,说明了该方法的有效性和科学性。

关键词:核小体定位;深度学习;集成学习方法;DNA序列编码;交叉验证
中图法分类号 TP183

Ensemble Learning Method for Nucleosome Localization Prediction

CHEN Wei,LI Hang and LI Wei-hua
School of Information Science and Engineering,Yunnan University,Kunming 650500,China

Abstract Nucleosome localization refers to the position of DNA double helix relative to histone,and plays an important regulatory role in DNA transcription. It takes a lot of time and resources to detect nucleosome localization by biological experiments. Therefore,it is an important research direction to predict nucleosome localization by using DNA sequences based on computational methods. Aiming at the shortcomings of single model and single code in DNA sequence feature representation and learning in nucleosome location prediction,this paper proposes an end-to-end ensemble deep learning model FuseENup,which uses three coding methods to represent DNA data from multiple dimensions. Different models extract the key features hidden in the data from different dimensions,and construct a new DNA sequence representation model. Performing 20-fold cross-validation on the four data sets,compared to the current model CORENup with the best comprehensive performance for the nucleosome localization prediction problem,the accuracy and precision of FuseENup are improved by 3% and 9% on the HS data set,increases 2% and 6% on the DM data set,1% and 4% on the E data set. Compared with other machine learning and deep learning benchmark models,FuseENup has better performance. Experiments show that FuseENup can improve the prediction accuracy of nucleosomes localization,which shows the effectiveness and scientificity of the method.

Keywords Nucleosome localization,Deep learning,Ensemble learning method,DNA sequence coding,Cross-validation

1 引言

核小体也称作核体或核仁小体,是由DNA和5种组蛋白构成的复合物,是除精子外的真核生物染色质的基本结构单元和基本功能单元[1]。5种组蛋白分别是H1,H3,H4,H2A和H2B,其中H3,H4,H2A和H2B含有两个大分子,它们之间相互作用形成一个组蛋白八聚体[2]。大约有200pd长的DNA分子缠绕这个组蛋白八聚体1.75圈组成核小体的核心颗粒,这一部分DNA序列称作核小体DNA[1,3]。H1组蛋白结合DNA构成连接区,将核小体核心颗粒相连,这一部分DNA序列称为连接区DNA[2,4]。

研究表明,核小体的定位受多种因素共同作用,如组蛋白乙酰化[5]、重构蛋白[6]、驱动细胞染色质结构[7]等,但是受DNA序列的影响较明显,不同的DNA序列和组蛋白八聚体

到稿日期:2020-11-26 返修日期:2021-04-19 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:云南省教育厅科学研究基金(2019J0006);云南省创新团队项目(2018HC019)
This work was supported by the Scientific Research Fundation of the Education Department of Yunnan Province China(2019J0006) and Innovative Research Team of Yunnan Province,China(2018HC019).
通信作者:李维华(lywey@163.com)

的亲和力不同[8]。研究表明[9-11]，许多与核小体相关的序列含有近似周期特征的核苷酸分布，通过对体外和体内重建的核小体序列图进行比较得出，每个位置的相对占据率是不相同的，这在一定程度上表明，核小体在体外的间距不像在体内的核小体那样有规律。这些观察结果证实了基因组 DNA 编码核小体位置的结论，说明了 DNA 对核小体定位的重要性，为研究利用 DNA 预测核小体定位提供了可行性。

核小体的定位在很多生物新陈代谢过程中发挥着重要作用，如 DNA 的复制、修复、重组和基因的表达调控，以及染色体和拮抗转录因子的形成，核小体的定位还与生物体的疾病有着重大关系，如肾小球肾炎、皮疹、脱发、红斑狼疮、自身免疫型肝炎等疾病[12-15]。因此，了解核小体的结构和功能在生物学中具有重要的意义。

核小体定位预测问题的研究由来已久，不同于高通量测序技术[16]会消耗大量资源，使用该计算方法用于核小体定位预测成为了目前的研究热点。Xing 等[17]结合 DNA 数据信息熵和支持向量机，充分利用统计学方法表示 DNA 数据，在核小体定位方面取得了较好的效果。Segal 等[18]结合多条 DNA 组合的频率、分布信息和隐马尔可夫链对核小体定位进行预测。Lieleg 等[19]利用碱基对偏转角度来表示核小体 DNA 序列特征，成功地对核小体定位进行了预测。Meher 等[20]利用 K-mer 编码将 DNA 序列向量化与随机森林相结合，对核小体定位进行了预测。Chen 等[21]根据核小体 DNA 序列和连接 DNA 序列之间的物理化学性质的差异性来预测核小体定位。这些方法为核小体的预测提供了基础，但是他们往往需要较多的先验知识和人工特征。

为了弥补基于传统机器学习预测的不足，以及充分发挥深度学习在特征提取和模型预测中的优势，深度学习也被应用到核小体定位预测中。例如，Bosco 等[22]将 DNA 序列进行 One-hot 编码，利用卷积神经网络提取特征向量的局部特征来进行核小体定位预测。Gangi 等[23]利用递归神经网络自动提取 DNA 序列远程特征，使核小体定位预测取得了较好的效果。Di 等[24]提出了堆叠卷积层和长短期记忆层的神经网络模型，自动从短期和长期依赖序列中提取特征，从而进行核小体定位预测。Zhang 等[25]提出了一个新的卷积神经网络，将门控神经网络与多种模式的反应和 DNA 序列长期关联相结合，用于预测核小体的定位。这些基于深度学习的预测方法虽然在一定程度上提高了模型预测的准确度，但是模型的输入局限于 DNA 序列的单一编码，限制了深度学习在特征提取方面的优势以及预测模型的性能。

为了充分表示 DNA 序列并挖掘序列中蕴含的关键特征，本文首次提出了一种端到端的集成深度学习模型 FuseE-Nup，并设计了 3 种编码方法对 DNA 数据进行表征，从不同的维度表示数据信息，通过 3 种不同的神经网络模型，从不同维度充分提取更有区分度的特征，根据集成策略融合 3 个模型输出的最终结果。相比上述传统机器学习模型，本文方法不需要先验知识构建人工特征，且预测准确度比上述传统机器学习模型更高。针对上述深度学习模型均使用单一编码表征 DNA 序列，均使用单一模型提取特征的局限性，本文方法从多维度对 DNA 序列数据进行表征，尽可能地保留更多的关键信息，利用多个模型分别提取特征，进而提高预测准确度。本文提出的核小体定位预测的集成学习模型在 4 个公开数据集上进行了 20 倍交叉验证的实验，在 Accuracy，Precision，F1-Score 和 AUC 这 4 种指标方面，相比现有方法，本文模型有效提高了核小体定位预测的性能。

2 核小体定位预测集成模型

2.1 编码方式

为了充分提取 DNA 序列中的特征，本文采用了 3 种 DNA 序列编码，从不同维度来表示 DNA 序列。这些编码包括：1）K-mer 结合 BOW 编码；2）K-mer 结合 Word2vec 编码；3）One-hot 编码。

K-mer 编码将一条 DNA 长序列分割成若干条长度为 K 的子序列，并保留着序列单体分组信息和序列顺序信息两个特征。设 L 为序列的长度，K 为子序列的长度，S 为步长，则可以得到 N 个子序列，其中 $N=\lfloor(L-K)/S\rfloor+1$，可见最多有 4^K 种长度为 K 的子序列。使 $K=5$，$S=1$，序列{“AGACCTGATCG”}的 K-mer 编码为{“AGACC”}，{“GACCT”}，{“ACCTG”}，{“CCTGA”}，{“CTGAT”}，{“TGATC”}，{“GATCG”}。BOW 编码是自然语言处理中常用的方法，它不考虑词与词的上下文联系，只考虑词的权重。K-mer 结合 BOW 编码首先将 DNA 序列划分为长度为 K 的子序列，然后统计该子序列在所有子序列中出现的频数。如将序列{“ATCTC”}以步长 $S=1$ 划分 $K=2$ 的子序列{“AT”，“TC”，“CT”，“TC”}，统计的频数如表 1 所列，得到特征向量“0001000000010020”。

表 1 K-mer 结合 BOW 编码

Table 1 K-mer combined with BOW coding

AA	AG	AC	AT	GA	GG	GC	GT
0	0	0	1	0	0	0	0
CA	CG	CC	CT	TA	TG	TC	TT
0	0	0	1	0	0	2	0

Word2vec 编码使用一个连续的稠密向量来刻画一个字词的特征，并将意思相近的词映射到向量空间中相近的位置，使词与词之间的相似度刻画得简单直接，还可以建立一个从向量到概率的函数模型，使相似的词向量可以得到相似的概率空间映射。K-mer 结合 Word2vec 编码将 DNA 序列 L 划分为若干个 K 长度的子序列 $L=\{l_1,l_2,l_3,\cdots,l_n\}$。根据 K 构造词空间 $S=\{S_1,S_2,S_3,\cdots,S_{4^k}\}$（一共有 4^k 种），U 为词空间 S 的索引，其中 $U=\{1,2,3,\cdots,4^k\}$，根据 S，U 将 L 转化为对应的正整数特征向量 $\boldsymbol{D}_s=\{d_1,d_2,d_3,\cdots,d_s\}$，利用 Word2vec 将正整数特征向量 $\boldsymbol{D}$ 投影到高维空间，从而得到二维特征表示 X_{sj}，其中 s 表示 $\boldsymbol{D}_s$ 中的第 s 个词，j 表示映射的维度。

One-hot 编码是处理 DNA 序列常用的编码。DNA 序列只包含 4 种核苷酸，即 $X=\{A,G,C,T\}$，其中 A 表示为“1000”，G 表示为“0100”，C 表示为“0010”，T 表示为“0001”。例如，{“ATGAA”}可以表示为{“1000”，“0001”，“0100”，“1000”，“1000”}，该编码能够保存 DNA 序列的上下文关系。

2.2 独立模型介绍

核小体定位预测指基于给定相关 DNA 序列数据预测

DNA 双螺旋相对于组蛋白的位置。核小体的相对位置与核苷酸序列局部和远程相互作用密切相关。为了更好地提取 DNA 序列的特征，本文构建了 3 种特征提取模型。第一种称为 KBNup，该模型将利用 K-mer 结合 BOW 编码得到的特征向量作为模型的输入，然后利用一维卷积提取特征向量局部特征，局部特征的学习方式如下：

$$y_t^0 = f_{conv}(\chi_t) = \sum_{\kappa=1}^{K} \omega_k x_{t-\kappa+1} \tag{1}$$

$$y_t^3 = D(y_t^2), y_t^2 = P(y_t^1), y_t^1 = h(y_t^0) \tag{2}$$

其中，ω_1，ω_2，…为滤波器或者卷积核，x_1，x_2，…为输入向量，$h(\cdot)$为激活函数，本文选用了线性整流函数(Rectified Linear Unit，Relu)作为激活函数，$P(\cdot)$代表池化操作，本文选用了最大池化，$D(\cdot)$代表 Dropout 操作，防止过拟合。为了简单起见，从$x_t \to y_t^2$的特征提取过程记为 $y = Conv(x)$，则该模型的计算式为：

$$f(x) = D(W(D(Conv(D(Conv(x)))))+b) \tag{3}$$

$$Output1 = sigmoid(f(x)) \tag{4}$$

其中，W 表示分类器权重，b 表示神经元偏置，用 $sigmoid() = \frac{1}{1-e^{-z}}$ 输出预测结果。该神经网络模型的具体结构如图 1 所示。

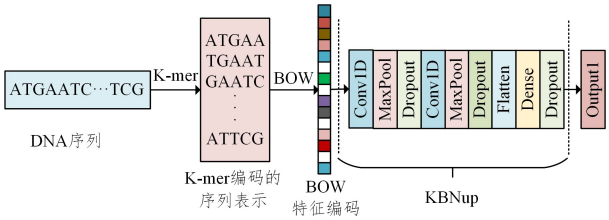

图 1　KBNup 模型

Fig. 1　Model of KBNup

第二种称为 KWNup，该模型利用 K-mer 编码表示 DNA 序列，将得到的短序列进行 Word2vec 编码，用高维向量表示短序列之间的关系。该模型利用卷积神经网络(Convolutional Neural Networks，CNN)操作来提取高维向量局部特征，再用长短时记忆神经网络(Long Short Time Memory，LSTM)来提取全局特征，该神经网络模型的具体结构如图 2 所示，LSTM 的计算式如下：

$$f_t = \gamma((W_f \cdot [h_{t-1}, x_t]) + b_f) \tag{5}$$

$$i_t = \gamma((W_i \cdot [h_{t-1}, x_t] + b_i)) \tag{6}$$

$$C_t' = \vartheta(W_C \cdot [h_{t-1}, x_t] + b_c) \tag{7}$$

$$C_t = f_t * C_{t-1} + i_t * C_t' \tag{8}$$

$$O_t = \gamma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{9}$$

$$h_t = O_t * \vartheta(C_t) \tag{10}$$

其中，f_t，i_t，O_t分别表示输入门、遗忘门和输出门的操作状态，γ代表激活函数 sigmoid，ϑ代表激活函数 tanh，C_t'表示输出层的待更新数据，C_t表示更新后的单元状态，h_t表示单元的最终输出。LSTM 是循环神经网络的一个变体，可有效解决简单的循环神经网络梯度爆炸和梯度消失的问题。为了简单起见，从$x_t \to h_t$过程记为 $y = LSTM(x)$，则该模型的计算式为：

$$f(x_t) = D(W(LSTM(Conv(D(x_t))))+b) \tag{11}$$

$$Output2 = \text{sigmoid}(f(x_t)) \tag{12}$$

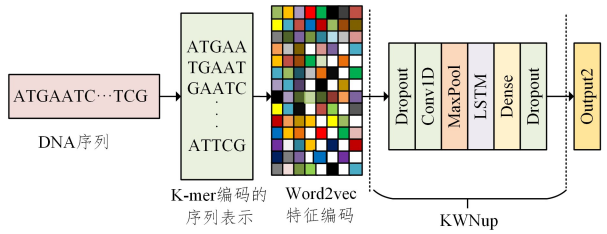

图 2　KWNup 模型

Fig. 2　Model of KWNup

第三种称为 CORENup，该模型是目前预测核小体定位综合性能最优的模型，利用 One-hot 编码表示 DNA 序列，并行使用 CNN 和 LSTM 来提取 DNA 序列的局部特征和远程特征。该神经网络模型的具体结构如图 3 所示，先用一层一维 CNN 提取序列的局部特征，并行使用一层 LSTM 提取 DNA 序列的远程特征，然后再用一层一维 CNN 进一步提取 DNA 序列的局部特征，将两者提取的特征合并放入分类器进行分类。为了防止过拟合，使用 Dropout 正则化处理，该模型的计算式为：

$$X_{tl} = D(LSTM(D(Conv(x_t)))) \tag{13}$$

$$X_{tc} = D(Conv(D(Conv(x_t)))) \tag{14}$$

$$X_t = X_{tl} \oplus X_{tc} \tag{15}$$

$$f(X_t) = w_2(D(w_1(X_t) + b_1)) + b_2 \tag{16}$$

$$Output = \text{sigmoid}(f(X_t)) \tag{17}$$

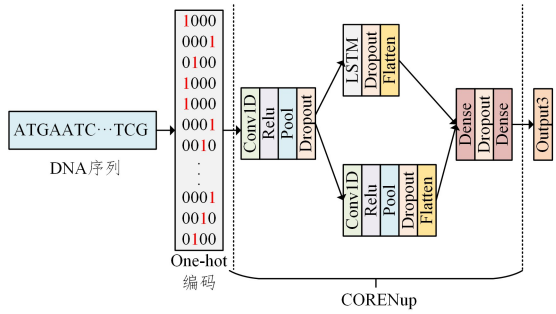

图 3　CORENup 模型

Fig. 3　Model of CORENup

2.3　集成策略

本文采用带权多数表决法的集成策略，假设有 N 个独立模型，将这 N 个独立模型训练后得到预测准确度分别为：ω_1，ω_2，…，ω_N $(0 \leqslant \omega \leqslant 1)$，对于二分类预测的结果 γ_1，γ_2，…，γ_N $(\gamma \in \{0,1\})$，最终预测结果 Y 的计算式为：

$$\begin{cases} \omega_1\gamma_1 + \omega_2\gamma_2 + \cdots + \omega_N\gamma_N > \frac{1}{2}(\omega_1 + \omega_2 + \cdots + \omega_N), \\ \qquad Y = 1 \\ \omega_1\gamma_1 + \omega_2\gamma_2 + \cdots + \omega_N\gamma_N \leqslant \frac{1}{2}(\omega_1 + \omega_2 + \cdots + \omega_N), \\ \qquad Y = 0 \end{cases} \tag{18}$$

本文利用 3 种不同的编码方式将 DNA 序列转化为特征序列，从不同维度保留 DNA 的关键特征，利用 3 个独立的神经网络模型，依据带权多数表决集成策略进行融合，从不同的维度自动提取特征。集成模型 FuseENup 的具体结构如图 4 所示。

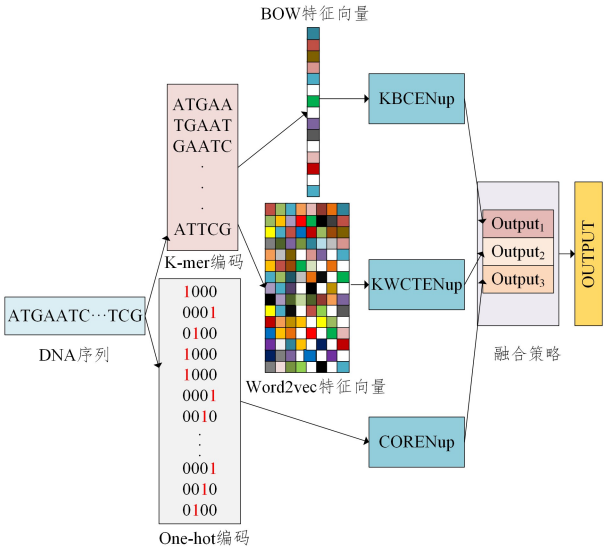

图 4 FuseENup 模型

Fig. 4 Model of FuseENup

3 实验及结果分析

本实验环境的主要参数如下：处理器为 AMD 3800X CPU 3.90GHz；图形加速卡为 NVIDIA GeForce GTX 1070Ti 8GB；操作系统为 Window10 64 bit；内存为 16GB。本文基于开源深度学习库 Keras 构建神经网络模型进行训练和测试，Keras 使用 TensorFlow 作为后端。

为了评估该融合模型的性能，在人类(HS)、果蝇(DM)、线虫(E)、酵母菌(Y)这 4 个公开数据集上使用 20 倍交叉验证训练和测试，数据的具体分布如表 2 所列。本文将核小体 DNA 视为正样本并标记为 1，将连接区 DNA 视为负样本并标记为 0。

表 2 数据分布

Table 2 Data distribution

	Link	Nucleosome	Total	Train	Test
HS	2300	2273	4573	4351	222
DM	2850	2900	5750	5472	278
E	2608	2567	5175	4921	254
Y	1740	1880	3620	3439	181

本文采用的 4 个数据集分别含有 4573，5750，5175，3620 条 DNA 序列数据，正样本与负样本的数量之差绝对值占总样本的比值分别为 0.6%，0.9%，0.8%和 3%，因此这 4 个数据集可以视为平衡数据集。

针对核小体定位预测问题，依据融合策略将 3 个独立模型融合成模型 FuseENup，独立模型的参数配置如表 3 所列。本文设计了 6 个实验，采用 *Accuracy*，*Precision*，*Recall*，*F1 score*，*AUC* 这 5 个指标来评价实验结果。

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{19}$$

$$Precision=\frac{TP}{TP+FP} \tag{20}$$

$$Recall=\frac{TP}{TP+FN} \tag{21}$$

$$F1\ score=1-\frac{FN+FP}{2TP+FN+FP} \tag{22}$$

$$AUC=\frac{\sum_{Ins_i\in pc} rank_{Ins_i}-\frac{M(M+1)}{2}}{M\times N} \tag{23}$$

其中，TP 代表真实类别为正例、预测类别为正例的样本数量，FN 代表真实类别为正例、预测类别为负例的样本数量，FP 代表真实类别为负例、预测类别为正例的样本数量，TN 代表真实类别为负例、预测类别为负例的样本数量，Ins_i 表示第 i 个样本，pc 表示正样本集合，$rank_{Ins_i}$ 表示第 i 条样本的序号(按概率得分从小到大排序)，M 表示正样本数，N 表示负样本数。

表 3 模型参数配置

Table 3 Parameter configuration model

KBNup		KWNup		CORENup	
参数	取值	参数	取值	参数	取值
迭代次数	65	迭代次数	12	迭代次数	75
批大小	128	批大小	128	批大小	128
优化器	Adam	优化器	Adam	优化器	Adam
学习率	0.001	学习率	0.001	学习率	0.0003
Dropout	0.5	Dropout	0.5	Dropout	0.5
CNN1	(100,5)	CNN	(50,7)	CNN1	(50,5)
CNN2	(75,5)	LSTM	50	CNN2	(50,10)
—	—	—	—	LSTM	50

实验 1 在 KBNup 模型中，一条 DNA 序列数据经过 K-mer 编码得到若干条长度为 K 的短序列，这些短序列经过 BOW 编码得到一条长度为 4^K 的特征向量，由此可见，K 的值直接影响 DNA 序列的特征向量表示，为了找出合适的 K 值，该实验将 K 分别取值为 3，4，5，6，7，在上述 4 个数据集中进行 20 倍交叉验证，分别得出预测的准确度，比较结果如图 5 所示。

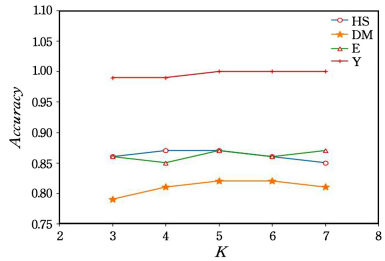

图 5 不同的 K 值下 KBNup 模型的准确度

Fig. 5 Accuracy of KBNup model under different K values

结果显示，不同的 K 值对 KBNup 模型的预测准确度有影响，在 20 倍交叉验证的情况下，当正整数 K 依次取值为 3 到 7 时，在 HS 数据集中的准确度分别为 0.86，0.87，0.88，0.86，0.85；在 DM 数据集中的准确度分别为 0.79，0.81，0.82，0.82，0.81；在 E 数据集中的准确度分别为 0.86，0.85，0.87，0.86，0.87；在 Y 数据集中的准确度分别为 0.99，0.99，0.99，1.00，1.00。考虑到模型的准确率和训练时间，本文将 K 值设置为 5。

实验 2 在 KWNup 模型中，同样地，一条 DNA 序列数据经过 K-mer 编码得到若干条长度为 K 的短序列，该短序列经过 Word2vec 编码得到长度为 100 的特征向量，为了探索在 20 倍交叉验证的情况下，不同的 K 值对模型预测准确度的影响，将正整数 K 依次取值为 3 到 7，在 HS，DM，E，Y 这 4 个数据集上进行验证，结果如图 6 所示。实验结果表明，不同的

K 会对 KWNup 模型的预测准确度产生影响，当正整数 K 依次取值为 3 到 7 时，在 HS 数据集上的准确度分别为 0.83，0.84，0.86，0.85，0.82；在 DM 数据集上的准确度分别为 0.75，0.80，0.79，0.80，0.79；在 E 数据集上的准确度分别为 0.85，0.88，0.89，0.85，0.87；在 Y 数据集上的准确度分别为 0.98，0.98，0.99，0.99，0.99。K 取 5 时模型的预测性能不是最优的，例如，在 DM 数据集上 K 取 5 时，模型的准确度为 0.79，而 K 取 6 时模型的准确率为 0.80。综合考虑训练时间和预测准确度，本文将 K 值设置为 5。

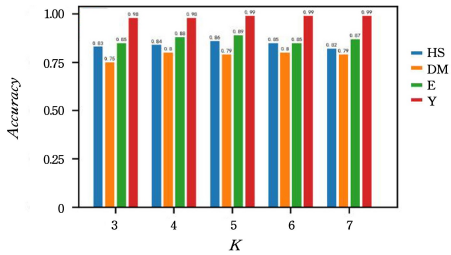

图 6　不同 K 值下 KWNup 模型的准确度

Fig. 6　Accuracy of KWNup model under different K values

实验 3　为了检验该集成模型相对于独立模型的有效性，将 KBNup，KWNup 和 CORENup 这 3 个独立模型作为基准模型与集成模型进行对比。模型融合时依据上述的融合策略和在 HS 数据集上的表现，ω_1，ω_2，ω_3 分别取 88，86，86。3 个基准模型以及集成模型在 HS，DM，E，Y 这 4 个数据集上，通过 20 倍交叉验证得到的 *Accuracy*，*Precision*，*Recall*，*F1 score*，*AUC* 这 5 个指标的均值和方差，如表 4 所列。从表 4 可以看出，集成模型 FuseENup 在数据集 HS，DM 和 E 上经过 20 倍交叉验证后，其指标 *Accuracy*，*Precision*，*F1 score* 和 *AUC* 相比基准模型均有所提高，且整体稳定性更强，但指标 *Recall* 表现欠佳；在数据集 Y 上，其指标 *Accuracy*，*Precision* 和 *AUC* 相比基准模型均有所提高，且整体稳定性更强。综上所述，集成模型相比独立模型而言整体性能和稳定性更好。

表 4　集成模型 FuseENup 和独立模型的效果对比

Table 4　Comparison of FuseENup and independent model

		KBNup	KWNup	CORENup	FuseENup
HS	*Accuracy*	0.88(0.02)	0.86(0.02)	0.86(0.02)	**0.89(0.02)**
	Precision	0.86(0.03)	0.83(0.03)	0.83(0.03)	**0.92(0.02)**
	Recall	0.89(0.02)	0.89(0.04)	**0.91(0.03)**	0.85(0.02)
	F1 score	0.87(0.02)	0.86(0.02)	0.87(0.02)	**0.88(0.02)**
	AUC	0.92(0.02)	0.90(0.01)	0.92(0.02)	**0.93(0.02)**
DM	*Accuracy*	0.82(0.02)	0.79(0.03)	0.87(0.02)	**0.89(0.02)**
	Precision	0.84(0.04)	0.80(0.05)	0.86(0.03)	**0.92(0.02)**
	Recall	0.81(0.04)	0.79(0.05)	**0.89(0.03)**	0.85(0.03)
	F1 score	0.82(0.02)	0.79(0.03)	0.87(0.02)	**0.88(0.02)**
	AUC	0.88(0.03)	0.83(0.01)	0.94(0.01)	**0.95(0.02)**
E	*Accuracy*	0.87(0.02)	0.89(0.02)	0.90(0.02)	**0.91(0.02)**
	Precision	0.85(0.03)	0.86(0.03)	0.87(0.03)	**0.91(0.03)**
	Recall	0.89(0.03)	0.92(0.03)	**0.94(0.02)**	0.90(0.03)
	F1 score	0.87(0.02)	0.89(0.02)	0.90(0.02)	**0.92(0.02)**
	AUC	0.92(0.02)	0.92(0.01)	0.96(0.01)	**0.96(0.01)**
Y	*Accuracy*	0.99(0.00)	1.00(0.01)	1.00(0.01)	**1.00(0.00)**
	Precision	0.99(0.00)	0.99(0.01)	0.99(0.01)	**1.00(0.00)**
	Recall	0.99(0.00)	0.99(0.01)	**1.00(0.00)**	**1.00(0.00)**
	F1 score	0.99(0.00)	0.99(0.00)	**1.00(0.01)**	**1.00(0.01)**
	AUC	0.99(0.00)	0.99(0.01)	0.99(0.01)	**1.00(0.01)**

实验 4　为了验证该集成模型相对于传统机器学习方法的有效性，本实验采用支持向量机（Support Vector Machine，SVM）、随机森林（Random Forest，RF）、贝叶斯分类器（Bayesian Decision Theory，BD）和 Adaboost 作为基准方法与集成模型 FuseENup 进行对比，使用随机编码将 DNA 序列数据数字化作为传统机器学习方法的输入。其中，SVM 的 kernel 配置为 rbf，gamma 配置为 scale，C 配置为 1.0。RF 的 n_estimators 设为 100，criterion 设为 gini。AdaBoost 的 max_depth 配置为 2，min_samples_split 配置为 20，min_samples_leaf 配置为 5。BD 的 var_smoothing 配置为 1×10^{-9}。*Accuracy*，*Precision*，*Recall*，*F1 score*，*AUC* 这 5 个指标的均值和方差的实验结果如表 5 所列，实验结果表明，集成模型 FuseENup 的性能明显优于传统机器学习方法。

表 5　集成模型 FuseENup 和传统机器学习方法的效果对比

Table 5　Comparison of FuseENup and traditional machine learning methods

		SVM	RF	Adaboost	BD	FuseENup
HS	*Accuracy*	0.70(0.03)	0.67(0.02)	0.65(0.03)	0.65(0.03)	**0.89(0.02)**
	Precision	0.68(0.03)	0.68(0.03)	0.65(0.03)	0.64(0.03)	**0.92(0.02)**
	Recall	0.74(0.05)	0.63(0.04)	0.66(0.05)	0.65(0.06)	**0.85(0.02)**
	F1 score	0.71(0.03)	0.65(0.03)	0.65(0.03)	0.65(0.04)	**0.88(0.02)**
	AUC	0.70(0.03)	0.67(0.02)	0.65(0.03)	0.65(0.03)	**0.93(0.02)**
DM	*Accuracy*	0.68(0.03)	0.66(0.02)	0.68(0.03)	0.64(0.03)	**0.89(0.02)**
	Precision	0.70(0.03)	0.66(0.02)	0.68(0.03)	0.65(0.03)	**0.92(0.02)**
	Recall	0.64(0.04)	0.63(0.04)	0.67(0.06)	0.64(0.04)	**0.85(0.03)**
	F1 score	0.64(0.03)	0.65(0.03)	0.68(0.04)	0.64(0.03)	**0.88(0.02)**
	AUC	0.68(0.03)	0.65(0.02)	0.68(0.03)	0.64(0.03)	**0.95(0.02)**
E	*Accuracy*	0.76(0.03)	0.80(0.03)	0.82(0.02)	0.76(0.02)	**0.91(0.02)**
	Precision	0.74(0.03)	0.81(0.03)	0.82(0.03)	0.75(0.02)	**0.91(0.03)**
	Recall	0.81(0.04)	0.78(0.04)	0.82(0.04)	0.76(0.04)	**0.90(0.03)**
	F1 score	0.77(0.03)	0.79(0.03)	0.82(0.02)	0.75(0.03)	**0.92(0.02)**
	AUC	0.77(0.03)	0.80(0.03)	0.82(0.02)	0.76(0.02)	**0.96(0.01)**
Y	*Accuracy*	0.96(0.01)	0.96(0.01)	0.97(0.02)	0.96(0.02)	**1.00(0.00)**
	Precision	0.94(0.02)	0.94(0.02)	0.95(0.03)	0.95(0.03)	**1.00(0.00)**
	Recall	0.99(0.01)	0.98(0.01)	0.98(0.02)	0.99(0.01)	**1.00(0.00)**
	F1 score	0.96(0.01)	0.96(0.01)	0.97(0.01)	0.97(0.02)	**1.00(0.01)**
	AUC	0.96(0.01)	0.96(0.01)	0.97(0.02)	0.96(0.02)	**1.00(0.01)**

实验 5　为了检测本文集成模型相对于其他深度学习模型的有效性，以 LSTM 模型[24]、ConVNet[22] 模型和 LeNup 模型[25] 为基准模型。*Accuracy*，*Precision*，*Recall*，*F1 score*，*AUC* 这 5 个指标的均值和方差的实验结果如表 6 所列。

表 6 集成模型 FuseENup 和基准深度学习模型的效果对比

Table 6 Comparison of FuseENup and benchmark deep learning model

		LSTM	ConVNet	LeNup	FuseENup
HS	*Accuracy*	0.84(0.02)	0.86(0.02)	0.87(0.02)	**0.89(0.02)**
	Precision	0.84(0.03)	0.86(0.03)	0.91(0.02)	**0.92(0.02)**
	Recall	0.89(0.03)	0.85(0.05)	0.84(0.02)	0.85(0.02)
	F1 score	0.86(0.02)	0.85(0.02)	0.87(0.02)	**0.88(0.02)**
	AUC	0.90(0.03)	0.91(0.03)	0.92(0.02)	**0.93(0.02)**
DM	*Accuracy*	0.85(0.02)	0.85(0.02)	0.87(0.02)	**0.89(0.02)**
	Precision	0.85(0.03)	0.83(0.03)	0.90(0.02)	**0.92(0.02)**
	Recall	**0.88(0.02)**	**0.88**(0.04)	0.87(0.03)	0.85(0.03)
	F1 score	0.87(0.01)	0.86(0.02)	0.88(0.02)	**0.88(0.02)**
	AUC	0.93(0.01)	0.92(0.02)	0.93(0.02)	**0.95(0.02)**
E	*Accuracy*	0.90(0.02)	0.90(0.02)	0.91(0.04)	**0.91(0.02)**
	Precision	0.88(0.03)	0.90(0.02)	0.90(0.02)	**0.91(0.03)**
	Recall	**0.93(0.02)**	0.89(0.03)	0.88(0.02)	0.90(0.03)
	F1 score	0.90(0.02)	0.89(0.02)	**0.92(0.02)**	**0.92(0.02)**
	AUC	0.96(0.02)	0.96(0.02)	0.96(0.02)	**0.96(0.01)**
Y	*Accuracy*	0.99(0.03)	0.99(0.04)	**1.00**(0.01)	**1.00(0.00)**
	Precision	0.99(0.02)	0.99(0.03)	0.99(0.01)	**1.00(0.00)**
	Recall	0.99(0.02)	0.99(0.04)	**1.00**(0.00)	**1.00(0.00)**
	F1 score	0.99(0.00)	0.99(0.02)	**1.00(0.01)**	**1.00(0.01)**
	AUC	0.99(0.00)	0.99(0.00)	**1.00(0.01)**	**1.00(0.01)**

从表中可以看出，在数据集 HS，DM 和 E 中，本文模型的 *Accuracy*，*Precision*，*F1 score* 和 *AUC* 相比基准模型有所提高，且整体稳定性更好，对于 Recall 指标，LSTM 模型表现更好，在数据集 Y 中，该集成模型的整体性能优于其他 3 种基准模型。对于 FuseENup 模型与 LeNup 模型的结果差距较小的原因可能有以下两点：1）上文已经说明 DNA 序列信息是核小体定位的重要影响因素，但不起决定性作用，影响核小体定位的程度至今仍不确定，而 LeNup 模型是近年来预测核小体定位的权威模型，该模型或许比较接近预测核小体定位的峰值，因此本文模型相比 LeNup 模型的结果差距较小；2）理论上如果提取的特征相互独立，则本文模型会有很高的预测准确度，例如，在 HS 数据集上，理论上预测准确度会达到 95%，而实际上为 89%，这说明提取的特征具有一定程度的相关性，这可能导致本文模型相比 LeNup 模型的结果差距较小。总体而言，集成模型 FuseENup 的整体性能优于其他 3 种基准模型。

实验 6 为了探究是否将编码方式对应的神经网络交换也能取得很好的效果，本文在保持其他自变量不变的情况下，通过交换构造了以下 5 种集成模型，分别是：编码 1）结合 KBNup 模型，编码 2）结合 CORENup 模型，编码 3）结合 KWNup 模型，记为 Model_A；编码 1）结合 KWNup 模型，编码 2）结合 KBNup 模型，编码 3）结合 CORENup 模型，记为 Model_B；编码 1）结合 KWNup 模型，编码 2）结合 CORENup 模型，编码 3）结合 KBNup 模型，记为 Model_C；编码 1）结合 CORENup 模型，编码 2）结合 KBNup 模型，编码 3）结合 KWNup 模型，记为 Model_D；编码 1）结合 CORENup 模型，编码 2）结合 KWNup 模型，编码 3）结合 KBNup 模型，记为 Model_E。*Accuracy*，*Precision*，*Recall*，*F1 score*，*AUC* 这 5 个指标的均值和方差的实验结果如表 7 所列，可以看出，这 5 种集成模型在 HS，DM 和 E 数据集上与基准模型以及 FuseENup 模型相比准确率的差距分别不超过 0.02，0.04 和 0.01，在其他指标上也具有非常相近的表现 。此外，与 FuseENup 模型一样，这 5 种集成模型在 Y 数据集中的各项指标均取得了 100%的效果。实验进一步验证了，从多个维度表征数据和利用不同神经网络从多个维度提取特征方法的有效性。

表 7 编码方式与神经网络的不同组合实验的效果

Table 7 Experimental results of different combinations of coding methods and neural networks

		Model_A	Model_B	Model_C	Model_D	Model_E
HS	*Accuracy*	0.88(0.02)	0.87(0.02)	0.87(0.02)	0.87(0.03)	0.88(0.02)
	Precision	0.91(0.03)	0.91(0.03)	0.89(0.03)	0.92(0.02)	0.91(0.03)
	Recall	0.84(0.03)	0.82(0.04)	0.85(0.03)	0.82(0.05)	0.85(0.03)
	F1 score	0.87(0.02)	0.87(0.03)	0.87(0.02)	0.87(0.03)	0.88(0.02)
	AUC	0.93(0.02)	0.93(0.02)	0.93(0.01)	0.93(0.02)	0.94(0.01)
DM	*Accuracy*	0.85(0.02)	0.86(0.02)	0.86(0.02)	0.85(0.02)	0.87(0.02)
	Precision	0.85(0.03)	0.85(0.03)	0.84(0.03)	0.85(0.03)	0.87(0.03)
	Recall	0.85(0.03)	0.88(0.03)	0.87(0.04)	0.86(0.04)	0.86(0.03)
	F1 score	0.85(0.02)	0.86(0.02)	0.86(0.02)	0.85(0.02)	0.86(0.02)
	AUC	0.89(0.02)	0.94(0.01)	0.93(0.01)	0.90(0.02)	0.93(0.01)
E	*Accuracy*	0.89(0.03)	0.90(0.02)	0.90(0.02)	0.89(0.03)	0.90(0.02)
	Precision	0.91(0.04)	0.94(0.02)	0.94(0.02)	0.92(0.04)	0.93(0.03)
	Recall	0.87(0.04)	0.86(0.03)	0.86(0.04)	0.85(0.06)	0.88(0.03)
	F1 score	0.89(0.03)	0.90(0.02)	0.90(0.02)	0.88(0.03)	0.90(0.02)
	AUC	0.94(0.02)	0.96(0.01)	0.96(0.01)	0.94(0.02)	0.96(0.01)
Y	*Accuracy*	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
	Precision	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
	Recall	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
	F1 score	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
	AUC	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)

结束语 核小体定位预测准确度的提高对全面了解某些疾病的发病机制和探索有效的治疗方法以及生物遗传物质进化具有重大意义。针对核小体定位预测很大程度上受 DNA 影响，本文以 DNA 序列数据为输入，采用不同的编码方式将 DNA 序列数据转化为特征向量，利用 3 种不同的模型分别提取特征，依据集成策略将 3 种不同的模型转化为一个端到端的集成预测模型 FuseENup。相比基准模型，本文的 FuseENup 模型有效地提升了核小体定位的预测性能。实验结果表明，集成策略可以有效提高核小体定位的预测性能，具有良好的稳定性，这也预示着集成的神经网络模型可以用于改善核小体定位以外的预测问题。分析本文模型取得较好效果的原因可能是，本文模型综合利用 K-mer，BOW，Word2vec 和 One-hot 这 4 种编码方式来表征 DNA 序列，利用 3 种不同的深度学习模型提取特征。K-mer 结合 BOW 的编码方式突出

了 DNA 子序列对预测核小体定位的作用,该子序列对预测核小体定位的作用越大,其权重就越大,就更有利于深度学习模型学习提取关键子序列。K-mer 结合 Word2vec 的编码方式突出了 DNA 子序列之间的关联性,作用越相近的子序列在坐标系中的距离越近,这种性质有利于深度学习模型对 DNA 子序列和 DNA 序列进行更加泛化的分析。One-hot 编码保留了 DNA 序列的顺序信息,因此该模型可以更充分地表征 DNA 序列。这 3 种深度学习模型提取特征的能力可以相互补充,例如,假设这 3 种模型的提取特征相互独立,且每种模型的预测准确度均为 80%,则根据经典概率模型集成模型的预测准确度为89.6%,因此该模型可以更好地预测核小体定位。本文模型使用 DNA 数据进行核小体预测,但是核小体定位受多种因素影响,因此融合更多类型的数据进行预测核小体定位将是今后改进的方向。

参 考 文 献

[1] RIDGWAY P,ALMOUZNI G. Chromatin assembly and organization[J]. Journal of Cell Science,2001,114(15):2711-2712.

[2] MASKELL D P,RENAULT L,SERRAO E,et al. Structural basis for retroviral integration into nucleosomes[J]. Nature,2015,523(7560):366-369.

[3] TABERLAY P C,STATHAM A L,KELLY T K,et al. Reconfiguration of nucleosome-depleted regions at distal regulatory elementsaccompanies DNA methylation of enhancers and insulators in cancer[J]. Genome Research,2014,24(9):1421-1432.

[4] COLE H A,CUI F,OCAMPO J,et al. Novel nucleosomal particles containing core histones and linker DNA but no histone H1[J]. Nucleic Acids Research,2016,44(2):573-581.

[5] SHAHBAZIAN M D,GRUNSTEIN M. Functions of Site-Specific Histone Acetylation and Deacetylation[J]. Annual Review of Biochemistry,2007,76:75-100.

[6] SCHNITZLER G R. Control of nucleosome positions by DNA sequence and remodeling machines[J]. Cell Biochemistry and Biophysics,2008,51(2/3):67-80.

[7] ZHENG D S,TRYNDA J,SUN Z F,et al. NUCLIZE for quantifying epigenome:generating histone modification data at single-nucleosome resolution using genuine nucleosome positions[J]. Bmc Genomics,2019,20(1):541-544.

[8] BUITRAGO D,CODO L,ILLA R,et al. Nucleosome Dynamics:a new tool for the dynamic analysis of nucleosome positioning[J]. Nucleic Acids Research,2019,47(18):9511-9523.

[9] SATCHWELL S C,DREW H R,TRAVERS A A. Sequence Periodicities in Chicken Nucleosome Core DNA[J]. Journal of Molecular Biology,1986,191(4):659-675.

[10] DREW H R,TRAVERS A A. DNA Bending and Its Relation to Nucleosome Positioning[J]. Journal of Molecular Biology,1985,186(4):773-790.

[11] LOWMAN H,BINA M. Correlation between Dinucleotide Periodicities and Nucleosome Positioning on Mouse Satellite DNA[J]. Biopolymers,1990,30(9/10):861-876.

[12] ZIVKOVIĆ V,STANKOVIĆ A,TATJANA C,et al. Anti-dsDNA,Anti-Nucleosome and Anti-C1q Antibodies as Disease Activity Markers in Patients with Systemic Lupus Erythematosus[J]. Srpski Arhiv za Celokupno Lekarstvo,2014,142:431-436.

[13] YANG J F,XU Z Z,SUI M S,et al. Co-Positivity for Anti-dsDNA,-Nucleosome and-Histone Antibodies in Lupus Nephritis is Indicative of High Serum Levels and Severe Nephropathy[J]. Plos One,2015,10(10):0140441.

[14] OLIVEIRA R C,OLIVEIRA I S,SANTIAGO M B,et al. High Avidity dsDNA Autoantibodies in Brazilian Women with Systemic Lupus Erythematosus:Correlation with Active Disease and Renal Dysfunction[J]. Journal of Immunology Research,2015,2015:814748.

[15] CHENG J,CHEN H,MEN J L. Correlation between anti-nucleosome antibodies and systemic lupus erythematosus[J]. Anhui Medicine,2019,23(1):83-86.

[16] ZHANG D F,MA Q Y,YIN T M. Third-generation sequencing technology and its application[J]. Chinese Journal of Bioengineering,2013,33(5):125-131.

[17] XING Y Q,LIU G Q,ZHAO X J,et al. An analysis and prediction of nucleosome positioning based on information content[J]. Chromosome Research,2013,21(1):63-74.

[18] STRUHL K,SEGAL E. Determinants of nucleosome positioning[J]. Nature Structural & Molecular Biology,2013,20(3):267-273.

[19] LIELEG C,KRIETENSTEIN N,WALKER M,et al. Nucleosome positioning in yeasts:methods,maps,and mechanisms[J]. Chromosoma,2015,124(2):131-151.

[20] MEHER P K,SAHU T K,RAO A R. Identification of species based on DNA barcode using k-mer feature vector and Random forest classifier[J]. Gene,2016,592(2):316-324.

[21] CHEN W,FENG P M,DING H,et al. Using deformation energy to analyze nucleosome positioning in genomes[J]. Genomics,2016,107(2/3):69-75.

[22] BOSCO G L,RIZZO R,FIANNACA A,et al. A Deep Learning Model for Epigenomic Studies[C]// International Conference on Signal-image Technology & Internet-based Systems. 2017.

[23] GANGI M A D ,GAGLIO S,BUA C L,et al. A Deep Learning Network for Exploiting Positional Information in Nucleosome Related Sequences[J]. Bioinformatics and Biomedical Engineering,2017,10209(4):524-533.

[24] DI G M,LO B G,RIZZO R. Deep learning architectures for prediction of nucleosome positioning from sequences data[J]. BMC Bioinformatics,2018,19(Suppl 14):418.

[25] ZHANG J,PENG W,WANG L. LeNup:learning nucleosome positioning from DNA sequences with improved convolutional neural networks[J]. Bioinformatics,2018,34(10):1705-1712.

CHEN Wei, born in 1997, postgraduate. His main research interests include deep learning and bioinformatics.

LI Wei-hua, corresponding author, born in 1977, Ph.D, associate professor. Her main research interests include data mining and bioinformatics.

(责任编辑:柯颖)