

基于卷积神经网络的基因剪接位点预测

李国斌¹, 杜秀全^{2,3}, 李新路¹, 吴志泽¹

- (1. 合肥学院 人工智能与大数据学院, 安徽 合肥 230601;
2. 安徽大学 计算智能与信号处理教育部重点实验室, 安徽 合肥 230601;
3. 安徽大学 计算机科学与技术学院, 安徽 合肥 230601)

摘要:研究剪接位点可以更深入地探索剪接机制和基因预测方法,准确预测剪接位点至关重要。基于深度学习技术提出一种新的预测方法,无需人工提取样本特征,以基因序列的 K-MER 编码向量作为输入,采用训练后的卷积神经网络(CNN)模型进行预测。基于人类基因 HS3D 供体数据集,与传统机器学习方法进行预测比较,结果表明预测模型的主要性能指标,包含马修斯相关系数(MCC)、灵敏度(SN)均超过传统的机器学习方法。

关键词:深度学习;卷积神经网络;剪接位点预测;K-MER 编码

中图分类号:TP183.1

文献标志码:A

文章编号:1671-5322(2020)02-0020-05

随着 DNA 序列检测需求的快速增长,真核生物基因的检测技术越来越得到人们的重视。通常真核基因组中的基因是不连续的,它由一组分离的外显子和内含子组成^[1],外显子和内含子之间的界限称为剪接位点。为了检测真核生物的基因,必须准确预测其剪接位点。

大多数机器学习方法预测剪接位点通常包括3个主要步骤:特征提取、特征选择(可选)和分类。数字向量作为机器学习分类器的输入,而DNA序列是字符串,因此,将基因序列编码为数字向量进行特征提取是剪接点位置预测的首要任务。零阶马尔可夫模型(MM0)、一阶马尔可夫模型(MM1)、二阶马尔可夫模型(MM2)和马尔可夫链模型(MCM)等概率编码方法是目前比较著名的、使用率很高的特征提取方法^[2-4],这些方法均表明核苷酸位置信息和相互依赖关系在剪接位点预测中起着关键作用。

在分类步骤中,可使用机器学习(ML)方法检测剪接位置,但大多数方法如支持向量机(SVM)^[5-6]、人工神经网络(ANN)、随机森林(RF)^[7]等均需要手工提取特征,同时分类不稳

定、泛化性差。

虽然传统的机器学习方法可用于预测,并达到较高的精度水平,但卷积神经网络(CNN)方法提供了更具吸引力的性能,并已成为生物信息学计算的主要解决方案。本文基于卷积神经网络(CNN)设计了一种新的剪接位点预测方法,通过实验分析给出其网络设计层数和其他结构参数,并从对比实验中找出最合适的网络设计和网络参数。该模型在 HS3D 公共基因数据集上进行的一系列对比实验,结果表明:结合 K-MER 编码方法的 CNN 模型可以充分利用深度神经网络的自学习能力,寻找精确的预测函数,对基因剪接位点的预测达到更高的精度和性能。

1 基于 CNN 的预测方法

1.1 卷积神经网络

大多数卷积神经网络由卷积层、池化层和全连接层等堆叠组成。为了识别剪接位点,本文设计了多层的 CNN 网络结构,即在全连接层(Fully Connected Layers)之前参照文献[8]重复使用几个网络层进行堆叠(本文分别取堆叠块数 $L=1$ 、

收稿日期:2019-12-24

基金项目:安徽省自然科学基金青年项目(1908085QF);安徽高校自然科学研究重点项目(KJ2019A0835);合肥学院科研发展项目基金(18ZR19ZDA)。

作者简介:李国斌(1981—),男,安徽和县人,实验师,硕士,主要研究方向为生物计算、机器学习。

2、3,依次构造浅度、中度、深度 3 种神经网络模型),每一个网络层由 1 个卷积运算和 1 个最大池

化运算组成,如图 1 所示。

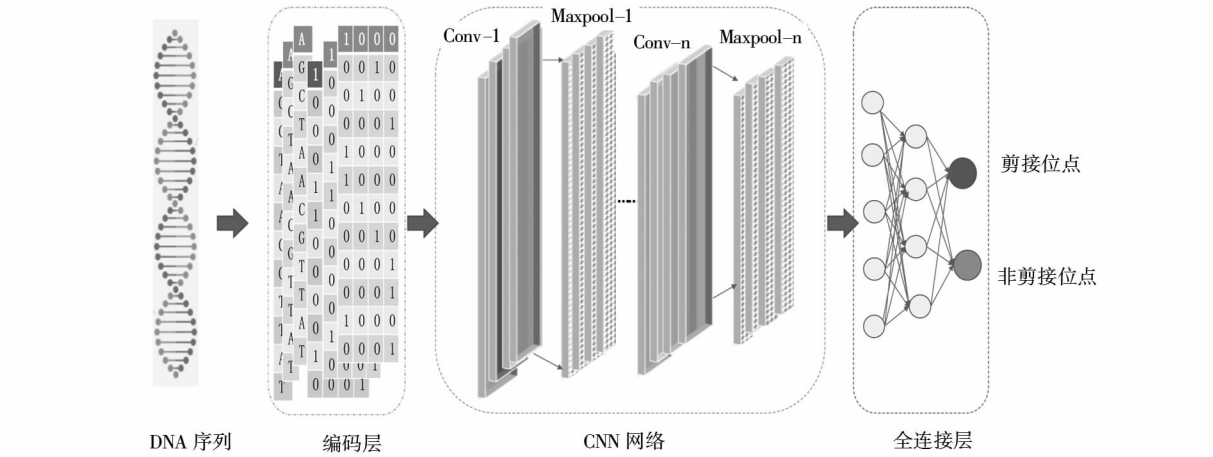


图 1 CNN 模型识别剪接位点结构

Fig 1 Architecture of the CNN model for splice site prediction

实验中,CNN 网络每层由 1 个卷积操作、1 个最大池化操作和 1 个 DropOut 操作组成,全连接层由 1 个 Flatten 操作、2 个稠密层(dense layer)和 1 个 DropOut 操作组成。为提升神经网络模型的表达能力,通常使用非线性激活函数如 Sigmoid 函数、tanh 函数和 Relu 函数等,本模型采用 Relu 激活函数,并引入 DropOut 参数以提升网络模型的健壮性^[9]。

1.2 编码方法

基因序列由 A、C、G、T 组成,代表 DNA 分子中的 4 种碱基。在进行剪接位点位置预测时,这些

序列必须通过编码方案转换为数字向量才能输入到多层 CNN 网络中,最后通过一个全连接层块输出预测结果进行网络训练和验证。

1.2.1 独热编码(One-hot)

独热编码是将分类数据表示为二进制向量的方法,经常运用于基因数据的准备中。对于 DNA 分子中的 4 种碱基 A、T、G、C,其独热编码可以用 $A = [1, 0, 0, 0]$ 、 $T = [0, 1, 0, 0]$ 、 $G = [0, 0, 1, 0]$ 、 $C = [0, 0, 0, 1]$ 表示。因此,DNA 序列 ATAGC 可以编码成如图 2 所示的向量。

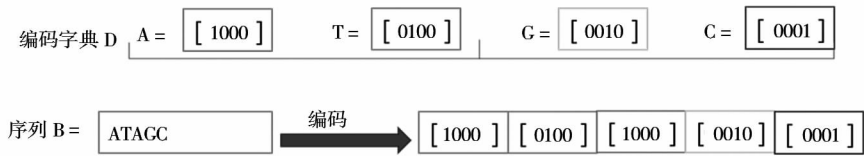


图 2 独热编码(One-hot)方案

Fig 2 Schematic diagram of One-hot encoding

1.2.2 K-MER 编码

编码 K-MER 的 DNA 序列特征可以看作是 One-hot 编码器的延伸(One-hot 编码实质是 K-MER 编码($K = 1$)的特殊形式)。K-MER 将读取序列划分为包含 k 个碱基的字符串,长度 m 的序列可分为 $(m - k + 1)$ 个 K-MER 编码。例如:对于 DNA 序列 ATAGC,如果 K-MER 编码参数 k 为 3,可以将其分解为 ATA-TAG-AGC,再把这些 K-

MER 编码拼接成一个完整的数字向量。使用 K-MER 编码可以看作是增强基因序列局部特征的一种方法。

2 实验

2.1 实验准备及数据集

实验使用人类基因供体数据集,其真剪接位点序列和假剪接位点序列均来自人类基因数据集

HS3D^[10]。数据集中正样本集合和负样本集合分别包含 2 796、271 938 个序列,每个序列长度都是 140,供体序列的二核苷酸 G 和 T 分别位于第 71 和 72 位。实验时将正负样本按照 1:1 比例组成平衡数据集,并基于 Ubuntu Linux 环境,使用 NVIDIA 1080Ti 显卡,通过 Python 语言实现。

2.2 实验指标度量

剪接位点预测属于典型的二分类问题,而混淆矩阵包含二分类器的实际和预测类的信息,其中 TP 、 TN 、 FP 、 FN 含义如表 1 所示。实验使用 10 折交叉验证方法,充分利用显卡的高性能计算能力,对基因数据进行预处理,并预测剪接位点。实验结果以准确度 ACC 、特异性 SP 、灵敏度 SN 、马修斯相关系数 MCC 等性能指标进行定量分析。

表 1 混淆矩阵
Table 1 Confusion Matrix

	预测正确	预测错误
正样本	TP	FP
负样本	TN	FN

各指标计算公式如下:

$$ACC = \frac{(TP + TN)}{(TP + NP + FP + FN)} \tag{1}$$

$$SN(\text{recall}) = \frac{TP}{(TP + FN)} \tag{2}$$

$$SP = \frac{TN}{(TN + FP)} \tag{3}$$

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{4}$$

3 实验结果及分析

3.1 多种网络结构预测结果比较

基于 K-MER 编码 ($K = 1, 2, 3$) 对不同 CNN 网络层数 $L(L = 1, 2, 3)$ 的预测结果如图 3 所示。图 3 表明网络结构复杂度同预测效果并非正相关。因此,网络结构不宜过于复杂,过深的网络结构意味着更多的超参数需要迭代学习并通过反向传播进行优化,使得计算时间和空间复杂度显著提高。

在 K-MER 编码方案中,如果 $K > 3$,将需要一次准备更多的内存读入编码向量,从而可能产生远远超过 GPU 硬件的计算负荷。实验中 K 选取 1、2、3, L 选取 1、2、3,经实验对比与综合考察,选取性能最优的网络结构参数组合 ($K = 3, L = 1$),在此基础上同传统机器学习方法进行比较。

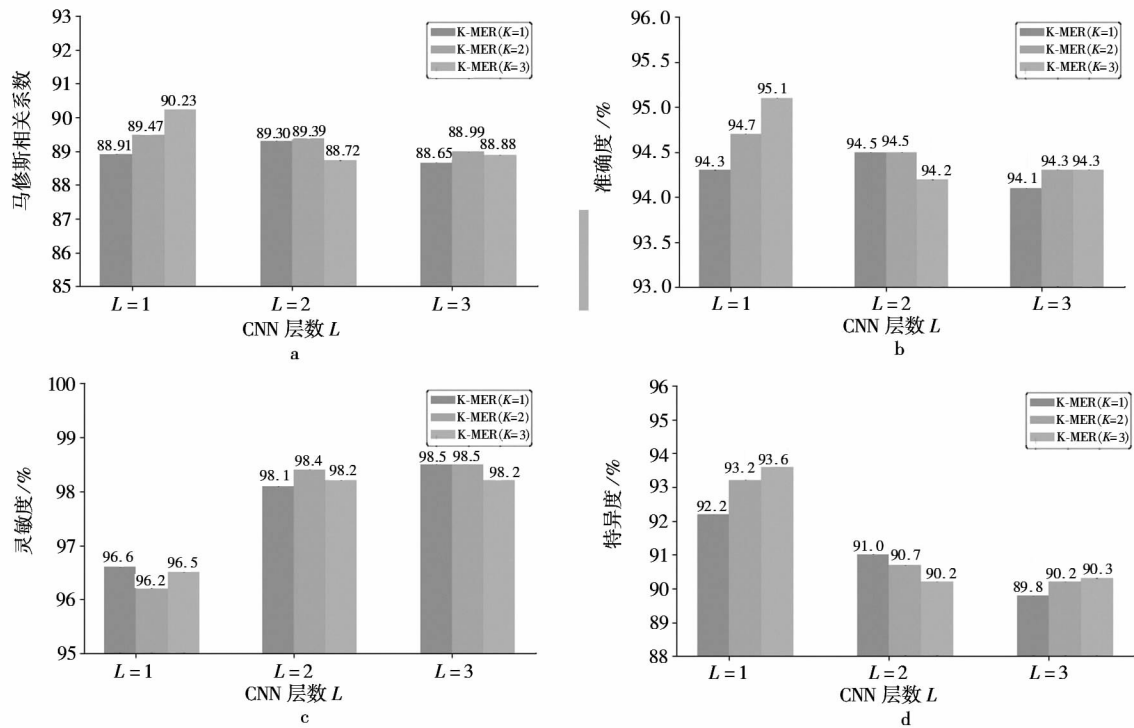


图 3 不同组合 CNN 模型的性能比较

Fig 3 Performance comparison of CNN models with different structures

3.2 同传统机器学习方法的比较

基于相同的 HS3D 供体数据集,选取基于 K-MER 编码的 CNN 模型方法与其他基于特征选择的传统机器学习方法如支持向量机(SVM)、人工神经网络(ANN)、随机森林(RF)和 AdaBoost 分

类器等方法的预测结果进行比较。传统机器学习预测结果见文献[8],从中选取最优结果与本文基于 K-MER 编码的 CNN 模型方法的预测结果一起列于表 2。

表 2 CNN 方法同传统机器学习方法预测结果对比

Table 2 Comparison of performance results between CNN method with other traditional machine learning methods

预测方法	灵敏度 <i>SN</i>	特异度 <i>SP</i>	<i>MCC</i> 系数
随机森林(TN-FDTF + RF) ^[8]	96.24	92.38	88.71
支持向量机(TN-FDTF + SVM) ^[8]	95.21	93.67	88.91
人工神经网络(TN-FDTF + ANN) ^[8]	95.17	93.35	88.55
AdaBoost 分类器 (TN-FDTF-AdaBoost) ^[8]	95.85	93.88	89.77
卷积神经网络 CNN(<i>K</i> = 3, <i>L</i> = 1)	96.50	93.60	90.20

表 2 表明:基于卷积神经网络算法的基因剪接位点预测方法,通过设计合理的编码方案和网络结构,无需手工提取特征,其预测结果的性能指标除特异度 *SP* 稍低,灵敏度 *SN* 和 *MCC* 系数指标均超过传统机器学习方法,取得了较好的效果。

4 结论

基于 CNN 的 K-MER 编码方案的基因预测方

法既利用了 K-MER 编码增强基因序列的局部特征,又充分挖掘了深度学习算法的自学习优势,其预测结果的灵敏度 *SN* 和 *MCC* 系数指标均超过传统机器学习方法,获得了很好的预测效果。该方法对于其他生物基因序列的检测和预测都具有较好的借鉴意义。

参考文献:

[1] MEHER P K, SAHU T K, RAO A R. Prediction of donor splice sites using random forest with a new sequence encoding approach[J]. BioData Mining, 2016,9:4.

[2] PASHAEI E, YILMAZ A, OZEN M, et al. A novel method for splice sites prediction using sequence component and hidden Markov model[C]//38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2016:3076-3079.

[3] HERNDON N, CARAGEA D. A study of domain adaptation classifiers derived from logistic regression for the task of splice site prediction[J]. IEEE Transactions on NanoBioscience, 2016,15(2):75-83.

[4] ZHANG Q W, PENG Q K, ZHANG Q, et al. Splice sites prediction of human genome using length-variable Markov model and feature selection[J]. Expert Systems with Applications, 2010,37(4):2771-2782.

[5] GOEL N, SINGH S, ASERI T C. An improved method for splice site prediction in DNA sequences using support vector machines[J]. Procedia Computer Science, 2015,57:358-367.

[6] MEHER P K, SAHU T K, RAO A R, et al. Identification of donor splice sites using support vector machine:a computational approach based on positional, compositional and dependency features[J]. Algorithms for Molecular Biology, 2016, 11:16.

[7] PASHAEI E, AYDIN N. Frequency difference based DNA encoding methods in human splice site recognition[C]//2017 International Conference on Computer Science and Engineering (UBMK), Antalya, 2017:586-591.

[8] DU X Q, YAO Y, DIAO Y Y, et al. DeepSS:exploring splice site motif through convolutional neural network directly from DNA sequence[J]. IEEE Access, 2018,6:32958-32978.

[9] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout;a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014,15(1):1929-1958.

[10] POLLASTRO P, RAMPONE S. Hs3d, a dataset of homo sapiens splice regions, and its extraction procedure from a major public database[J]. International Journal of Modern Physics C, 2002,13(8):1105-1117.

Prediction of Gene Splicing Sites Based on Convolution Neural Network

LI Guobin¹, DU Xiuquan^{2,3}, LI Xinlu¹, WU Zhize¹

1. School of Artificial Intelligence and Big Data, Hefei University, Hefei Anhui 230601, China;

2. Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei Anhui 230601, China;

3. School of Computer Science and Technology, Anhui University, Hefei Anhui 230601, China

Abstract:The study of splicing sites can further explore splicing mechanisms and gene prediction methods. It is very important to predict splice sites accurately. Based on the deep learning technology, a new prediction method is proposed. This method does not need to extract sample features manually. The K-MER coding vector of gene sequence is used as input, and the convolutional neural network (CNN) model after training is used for prediction. Based on the human gene HS3D donor data set, this model was compared with the traditional machine learning methods for prediction. The results showed that the main performance indicators of the prediction model, including Matthews correlation coefficient (MCC) and sensitivity (SN), exceeded the traditional machine learning methods.

Keywords:deep learning; convolutional neural network; splice site prediction; K-MER encoding

(责任编辑:李华云)