DOI:10.3969/j.issn.1672-5565.201712004

基于 CNN 与 LSTM 模型的蛋白质二级结构预测

王 剑[1],成金勇[1*],赵志刚[2],鹿文鹏[1]

(1.齐鲁工业大学(山东省科学院) 信息学院,济南 250353;
2.齐鲁工业大学(山东省科学院)山东省计算中心(国家超级计算济南中心),济南 250101)

摘 要:蛋白质结构的预测在理解蛋白质结构组成和蛋白质的生物学功能有重要意义,而蛋白质二级结构预测是蛋白质结构预测的重要环节。当 PSSM 位置特异性进化矩阵被广泛应用于将蛋白质初级结构序列编码作为输入样本后,每个残基可以被表示成二维空间的数据平面,由此文中尝试利用卷积神经网络对其进行训练。文中还设计了另一种卷积神经网络,利用长短记忆网络感知了 CNN 最后卷积特征面的横向特征和纵向特征后连同卷积神经网络的全连接共同完成分类,最后用 ensemble 方法对两类卷积神经网络模型进行了整合,最终 ensemble 方法中包含两类卷积神经网络的六个模型,在 CB513 蛋白质数据集测得的 Q3 结果为 77.2。

关键词:卷积神经网络;长短记忆网络;蛋白质二级结构预测;Ensemble 方法
中图分类号:TP391.4 **文献标志码**:A **文章编号**:1672-5565(2018)02-130-07

Protein secondary structure prediction based on CNN and LSTM models

WANG Jian[1], CHENG Jinyong[1*], ZHAO Zhigang[2], LU Wenpeng[1]

(1. *College of Information, Qilu University of Technology (Shandong Academy of Sciences), Jinan* 250353, *China*,
2. *Shandong Computer Science Center (National Supercomputer Center in Jinan) Qilu University of Technology (Shandong Academy of Sciences), Jinan* 250101, *China*)

Abstract: The prediction of protein structure is of great significance in understanding the structure and the biological function of proteins. The prediction of protein secondary structure is an important part of protein structure prediction. When PSSM position-specific evolution matrix is widely used to encode the primary sequence of a protein, and used as input sample, each residue can be represented as a two-dimensional data plane. Therefore, a convolutional neural network can be adopted as a model to train them. In this paper, we also designed another type of CNN in which LSTM were used to perceive the features of CNN last convolution feature maps both horizontally and vertically, and completed classification collaboratively with the fully-connected neural elements of convolution model. Finally, an ensemble method was adopted to integrate these two types of CNN models. This designed ensemble method includes six models of these two types of CNN. The Q3 accuracy obtained from CB513 is 77.2.

Keywords: CNN; LSTM; Protein secondary structure prediction; Ensemble method

蛋白质二级结构是由蛋白质初级结构链折叠形成的一些通用结构,蛋白质初级结构是指构成蛋白质的氨基酸序列,蛋白质的组成包含 20 种天然的氨基酸,为了定义氨基酸序列折叠形成的结构,一种流行的算法是 DSSP[1](Define secondary structure of proteins)算法,DSSP 定义了八种蛋白质二级结构(β 桥,β 折叠,3_{10} 螺旋,α 螺旋,π 螺旋,转角,β 转角,其他的结构如螺旋),并将这八种状态结构合并成三种,分别为螺旋(用 H 表示),链(用 E 表示)和卷曲(用 C 表示)。

收稿日期:2017-12-20;**修回日期**:2017-3-21.
基金项目:国家自然科学基金(61375013);山东省自然科学基金(ZR2013FM020).
作者简介:王剑,男,硕士研究生,研究方向:机器学习、生物信息学;E-mail:857863876@qq.com.
*通信作者:成金勇,男,副教授,研究方向:机器学习、生物信息学;E-mail:cjy@qlu.edu.cn.

在生物信息学中,为了预测蛋白质二级结构,蛋白质的氨基酸序列携带的蛋白质结构信息是重要的样本信息来源[2]。PSSM(置特异性打分矩阵)作为一种蛋白质序列的编码方法,PSSM 矩阵可以对蛋白质数据库中蛋白质序列进行比对和打分,将蛋白质原始序列编码后可以包含蛋白质残基序列的相对完整的信息,将 PSSM 应用于蛋白质二级结构预测开始于文献[3]。而许多深度学习应用于蛋白质二级结构预测的文献也都采用 PSSM 来表示蛋白质残基,文献[3,6]采用基于深度学习的模型应用于蛋白质二级结构预测,原残基序列样本被编码成 PSSM 当作输入特征矩阵。

本文中对 PSSM 预测蛋白质二级结构做了研究和试验,在研究过程中,本文采用 CNN(卷积神经网络)作为预测模型,考虑到卷积神经网络的预测结果不稳定,所以文中最终采用 ensemble(集成方法)对实验结果进行整合,并且为了提高预测效果,本文对原本设计的卷积神经网络进行了改进,将 LSTM(长短时记忆神经网络)与 CNN 结合,设计了两类 CNN 神经网络,经过试验,预测结果虽然没有显著提高,稳定性得到了提升。

1　模型原理与方案提出

1.1　卷积神经网络简单结构

根据文献[7],CNN 基本包含卷积层和池化层(在文献[7]称作下采样层)单元,根据文献[7]这些卷积层和池化层单元的级联构成了卷积神经网络隐层,在文献[7]表明卷积神经网络的分类层之前要有全连接层。

如图 1 所示的 CNN 中,只包含一个卷积层一个下采样层和一个全连接层,最后是输出层,假设输入层与卷基层设置 α 个卷积核,则卷积层的输出的第 α 个特征面 $h_{1,\alpha} = f(x * W^{1,\alpha} + b^{1,\alpha})$,其中 $W^{1,\alpha}$ 表示 x 样本的第 α 个卷积核,$b^{1,\alpha}$ 表示对应卷积核偏置。完成下采样后的特征面为 $h_{2,\alpha} = g(\beta_2 \operatorname{down}_{\lambda_2,\tau_2}(h_{1,\alpha}) + \gamma_2)$,那么全连接层可以表示为($h_{\alpha_1} \cdots h_{\alpha_n}$),如果有多个卷积层,假设第 k 层的下采样面是 $h_{k,\alpha_i}(i \in \{1,2 \cdots n\})$,第 k+1 层表示卷积层,卷积核个数设置为 w,则该卷积层构造的第 w 个卷积面 $h_{k+1,w} \rightarrow h_{k+1,w} = f(\sum_{i=1}^{n} h_{k,\alpha i} * W_{\alpha_i}^{k+1,w} + b^{k+1,w})$ 中 $W_{\alpha_i}^{k+1,w}$ 表示第 α_i 个下采样面对应的第 w 个卷积核,$b^{k+1,w}$ 为对应偏置(*代表卷积计算)。

1.2　LSTM 神经网络简单原理

LSTM 神经网络是一种循环神经网络,是在文献[9]中,为了解决其他循环网络消耗时间过长,梯度消失等不足而提出的,后来文献[10]在文献[9]基础上为 LSTM 加入了记忆扩展结构,使隐层记忆状态计算依赖于遗忘门中的权值矩阵,发展成了现代 LSTM 网络。

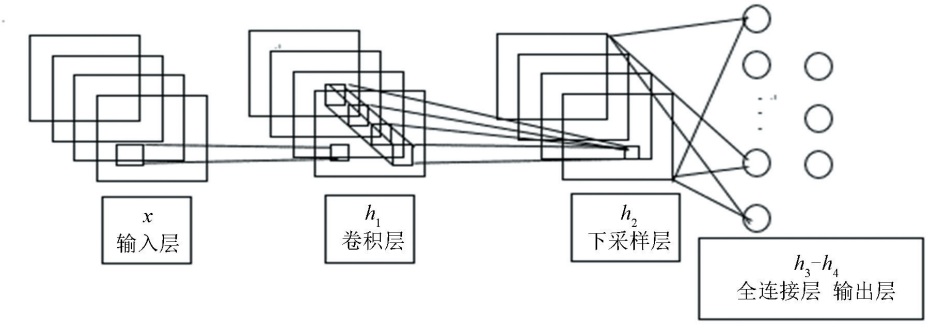

图 1　卷积神经网络简单结构图

Fig.1　Simple architecture of convolutional neural network

如图 2 所示,在标准的 LSTM 模型中,t 时刻样本通过输入门与 $t-1$ 时刻 LSTM 模型的输出共同计算出输入 $i_t \rightarrow i_t = \sigma(W^{xi} x_t + W^{hi} h_{t-1} + b_i)$,同时通过输入挤压单元与 $t-1$ 时刻 LSTM 的输出计算出新的记忆 $g_t \rightarrow g_t = tanh(W^{xg} x_t + W^{hg} h_{t-1} + b_g)$,再次通过遗忘门单元与 $t-1$ 时刻 LSTM 的输出计算出决定记忆的状态因子 $f_t \rightarrow f_t = \sigma(W^{xf} x_t + W^{hf} h_{t-1} + b_f)$,最后新的记忆 g_t 与 t 时刻的输入 i_t 的点乘结果结合 $t-1$ 时刻的记忆状态 S_{t-1} 与 t 时刻记忆状态因子 f_t 的点乘结果共同更新了 t 时刻的最终记忆状态 $S_t \rightarrow S_t = S_{t-1} \circ f_t + i_t \circ g_t$,LSTM 的输出是 t 时刻的样本输入与 LSTM 模型 $t-1$ 时刻的输出通过输出门单元计算出的 t 时刻输出因子 $o_t \rightarrow o_t = \sigma(W^{xo} x_t + W^{ho} h_{t-1} + b_o)$,之后与 t 时刻 LSTM 模型的最终记忆状态的激活函数值点乘计算出的结果 $h_t \rightarrow h_t = o_t \circ tanh(S_t)$。

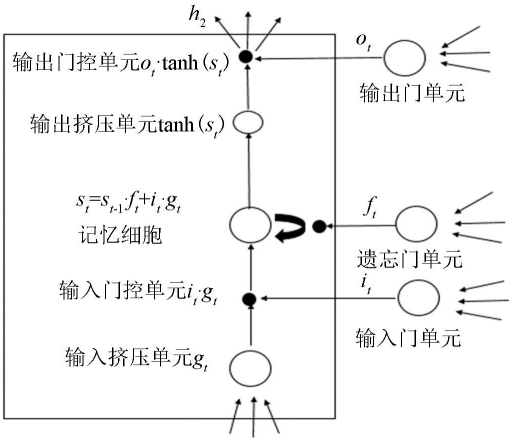

图 2 LSTM 简单原理图

Fig.2 Simple schematic of LSTM

LSTM 的模型的输入包括样本大小维度,是指 LSTM 输入的样本个数,特别的包含一个时间维度,此维度代表时间片的大小,体现在图 1 中就是 t 的大小,而最后的数据维度代表当前时间片的数据内容。

1.3 将 LSTM 与 CNN 组合

在本文提出的方案中,在 CNN 的 Flatten 层之前(在图 1 中为全连接层 h_3),假设 h_3 是 $a \times b \times n$ 的卷积特征面输出,其中 a 和 b 分别为卷积特征面的长度和宽度,n 为卷积特征面的个数,为了训练两个 LSTM 模型,分别将卷积特征面纵向排列将卷积面的宽度作为时间维度输入 LSTM 神经网络,再将卷积特征面横向排列将卷积特征面的长度作为时间维度输入一个 LSTM 神经网络,最后将两个 LSTM 的输出连同卷积特征面共同经过 Concatenate 操作经过 softmax 完成分类(如图 3)。

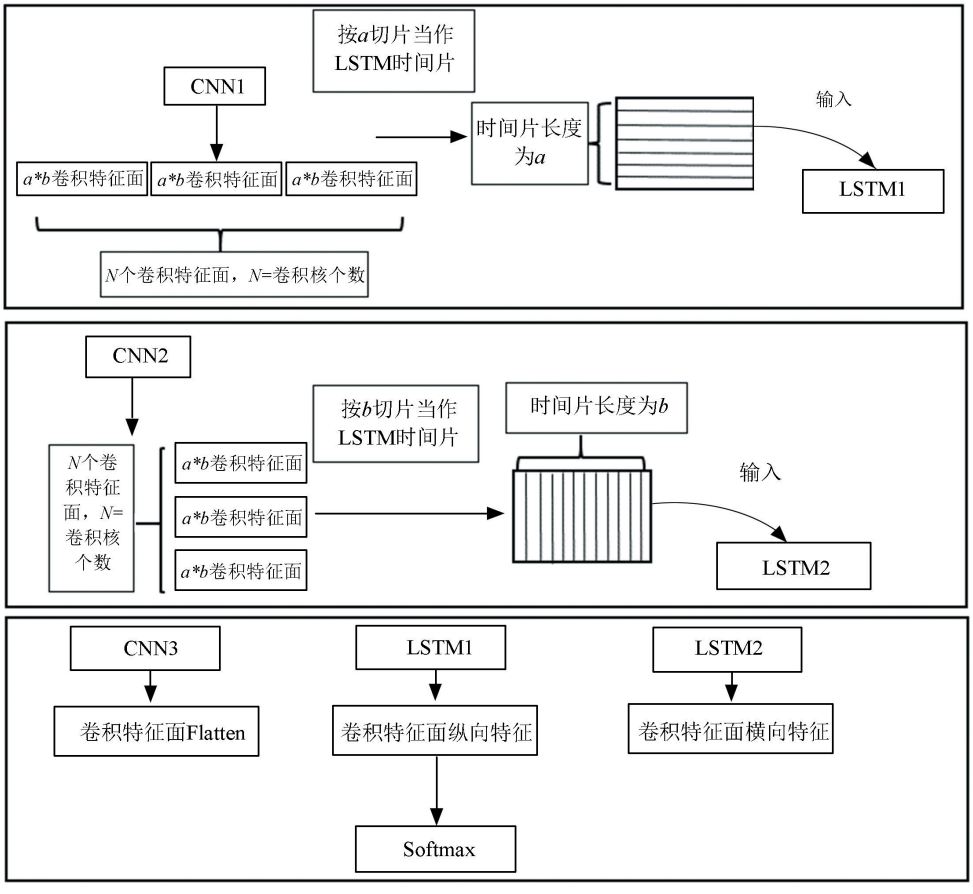

图 3 LSTM 与 CNN 组合原理图

Fig.3 Schematics of combination of LSTM and CNN

2 模型设计

2.1 CNN 的设计

首先本文中 CNN 的设计原则是 CNN 的隐层越多越好,按照文献[11]中的描述,卷积神经网络的层数越多,每层的计算任务就越简单,因为每层的输入和输出相差就不会很大,简单的层模型有利于整个模型构建。在文献[12]中,作者说明在分类前的卷积特征面要足够小,同时卷积核要设置的足够大有利于模型的构建。由于 PSSM 组成的二维数据平面的大小的限制,本文 CNN 设计成了三层,每层设置 40 个卷积核,为了避免 CNN 训练过程中的过拟合,本文在 CNN 的设计中采用了 Dropout[13] 技术。

本文 CNN 的结构图如图 4。

2.2　LSTM 与 CNN 组合模型的设计

在上述 CNN 设计完成的基础上,我们将 CNN 的最后计算的卷积特征面经过适当处理,分别作为两个 LSTM 模型的输入,模型设计结构如图 5。

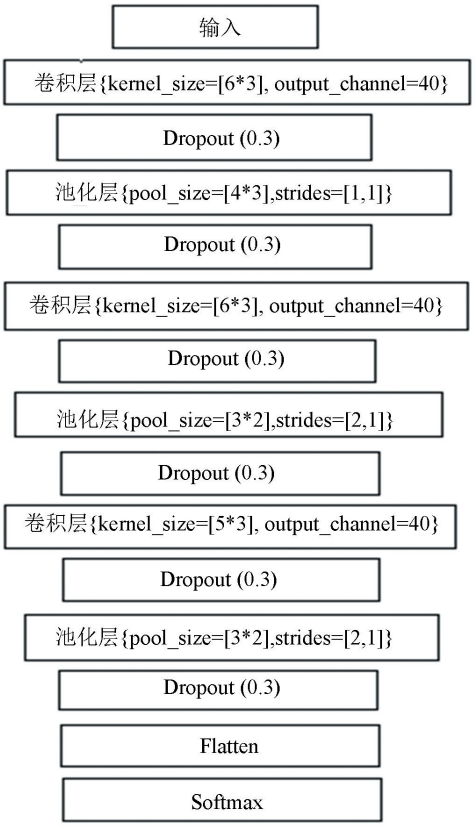

图 4　CNN 结构图

Fig.4　Structural design of CNN

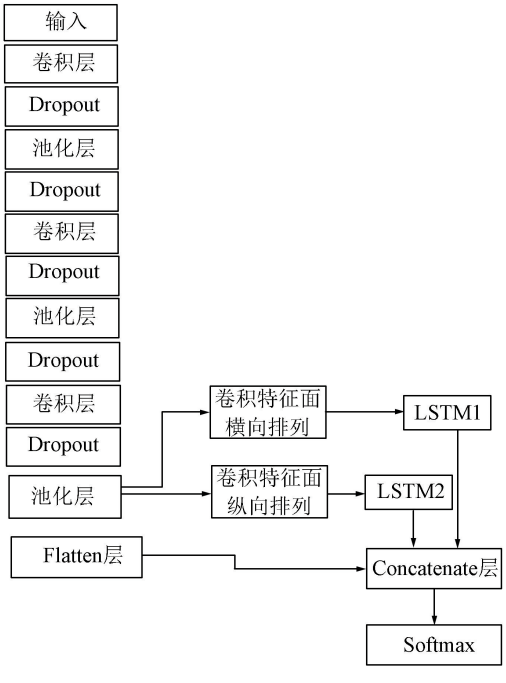

图 5　CNN+LSTM 结构图

Fig.5　Structural design of CNN+LSTM

2.2.1　LSTM 参数的设置

LSTM 中也加入了 dropout 技术,同时设置 LSTM 的输出为整个时间片对应的序列,LSTM 隐层个数设置为 20 个。

2.3　Ensemble 方法的设计

由于两类 CNN 网络产生的预测结果会不稳定,模型之间预测结果会存在差集,根据文献[14]本文采用了 ensemble 方法来解决这个问题,在多个分类器存在预测结果差异的情况下,ensemble 方法会将多个分类器的结果结合在一起会取得更好的结果[14]。

本文 ensemble 方法中包含两类模型,一种是 CNN 模型,另一种模型是将 LSTM 与 CNN 组合的模型,这两类模型分别训练三个,一共六个模型,最后 ensemble 方法的输出为六个模型每类均值最大值。

关于 Ensemble 方法的实现,假设一共训练好 N 个预测模型,将预先训练好的 N 个模型分别预测数据样本,将预测结果进行整合,假设 $Model_n$ 在第 i 类二级结构的预测结 $\text{Output}_n^i (i \in \{1,2,3\}$,$n$ 表示模型下标),Ensemble 模型的输出为 $\text{Output}_{\text{ensemble}} = \max\left(\left[\sum_{n=1}^{N} \text{Output}_n^i\right]/N\right)$。

3　数据与样本处理

3.1　蛋白质样本的编码

本文中将蛋白质序列取窗口大小为 m 来表示单个残基,蛋白质每个残基表示成 20 维的 PSSM 信息和 20 维的正交编码,每个残基被编码成了 40 维的信息,取窗口后,每个残基被编码成 $m*40$ 的 2 维数据平面作为每个氨基酸残基的模型输入,如图 6 所示。本文中采用 CB513[15] 蛋白质序列样本作为数据样本,CB513 中包含 513 个蛋白质氨基酸序列,同源相似性小于 25%。PSSM 用 BLAST 工具生成,二级结构标签的定义采用 DSSP 的三类定义螺旋(H),折叠(E),卷曲(C),数据集中 513 个蛋白质样本一共84 119个氨基酸残基经 PSSM 编码后作为训练样本,螺旋(H)有29 097个, 折叠(E)有17 897个,卷曲(C)有37 125个。

3.2　评估方法

在试验中,采用 Q3 作为预测结果的评估方法:

$$Q_3 = \frac{Q_H + Q_E + Q_C}{\text{残基数量}} * 100$$,其中 Q_H , Q_E , Q_C 分别是 H 类 E 类 C 类残基预测正确的残基数量。

3.3　PSSM 的产生方法

根据文献[16],利用 PSI-BLAST(https://blast.

ncbi.nlm.nih.gov/Blast.cgi)工具调用 3 次迭代,检测进化矩阵设置为 BLOSUM62 矩阵,E-value 设置为 0.00 1搜索 CB513 数据集中每个蛋白质样本生成相应的 PSSM 矩阵,在搜索蛋白质数据库是选择非冗余蛋白质序列库(nr),PSI-BLAST 有本地化工具只需将 nr(ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz)序列库下载到本地即可。

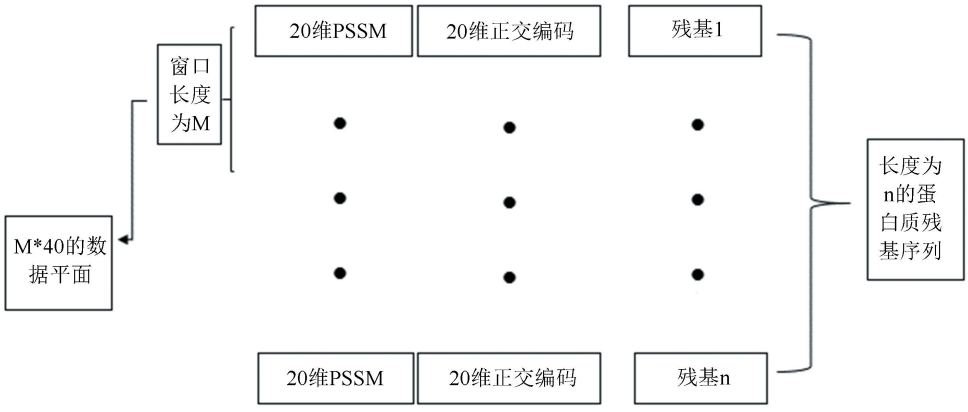

图 6 蛋白质样本的编码示意图

Fig.6 Encoding diagram of protein sample

4 结果与讨论

在本文提出的方案的实现过程中,以 keras[17] 平台框架作为工具,以 Python3.6 作为开发环境。在实验过程中我们采用三折交叉验证方法对试验结果进行了统计,将数据样本分为三份,2/3 作为训练集 1/3 作为测试集。

因为 ensemble 需要三个模型,本文对 CNN 的设计一共训练了三个 CNN 模型。

对于 CNN 与 LSTM 组合的模型,对于 Ensemble 方法设计需要三个。

表 1 CNN 模型交叉验证实验结果

Table 1 Cross validation test results for CNN %

模型	Q3	QC	QE	QH
模型 1	76.7	83.7	61.2	77.3
模型 2	76.8	83.1	63.6	76.8
模型 3	76.6	83.2	63.7	76.2

表 2 (CNN+LSTM 试验交叉验证结果)

Table 2 Cross validation test results for CNN+LSTM %

模型	Q3	QC	QE	QH
模型 1	76.0	85.0	64.5	71.5
模型 2	76.5	85.2	62.2	74.2
模型 3	76.2	86.6	61.0	72.3

图 7 为前两个模型三次试验结果的对比折线图。

在图 10 中,横向观察,不论是只用 CNN 还是 LSTM 与 CNN 的组合,蛋白质二级结构的预测结果都不很稳定,有一定的起伏;纵向观察,可以明显的观察出 CNN 模型中 QH 的结果要比 LSTM 与 CNN 的组合的 QH 结果高,而 LSTM 与 CNN 的组合的 QC 结果要高,所以两类模型有差集。文中最后采用了 ensemble 方法将每类这三个模型整合在一起取均值后作为 ensemble 模型的预测输出。

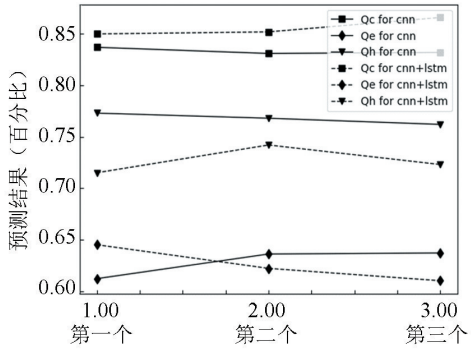

图 7 CNN 与 CNN+LSTM 的对比结果

Fig.7 Comparison between CNN and CNN+LSTM

Ensemble 方法的实验结果如表 3。

通过对比可以发现,两类模型成员经过集成方法,原本的预测结果并没有缩减,证明 ensemble 是可以增加模型的稳定性,并且集成方法的两类模型预测结果的差集通过 ensemble 可以很好的得到整合起到提高预测结果的作用。

其他方法与本文中方法的对比(CB513 的实验结果)。

表 3　Ensemble 方法的实验结果

Table 3　Test result of Ensemble method　　%

模型	Q3	QC	QE	QH
Ensemble	77.2	85.4	63.2	75.5
Ensemble(cnn+lstm only)	76.6	86.1	62.7	73.0
Ensemble(cnn only)	76.9	83.7	63.0	77.1

表 4　CB513 数据集预测结果对比

Table 4　Comparison of prediction results for CB513 dataset by different methods　　%

方法	Q3	QC	QE	QH
ELM[18]	71.2	76.9	63.1	67.6
SVM[19]	73.5	79.0	60.0	75.0
本文 Ensemble	77.2	85.4	63.2	75.5
打分函数法[20]	80.5	81.2	80.9	79.8

由上表可以看出,Ensemble 方法卷曲(C)的预测结果最高,Ensemble(CNN+LSTM)模型的卷曲预测结果达到了 86.1%,并且 Ensemble(CNN only)的螺旋(H)的预测结果 77.1%与最高结果79.8%相比,差距很小。

5　结　论

本文通过将 CNN(卷积神经网络)应用到蛋白质二级结构预测,并且将 LSTM 与 CNN 组合之后,将两类 CNN 模型通过 ensemble 方法整合在一起,从实验结果来看 ensemble 方法的应用能够综合多个模型的输出结果并使结果有所提高,另一方面将 LSTM 应用到 CNN 之后,CNN 模型预测结果会与原 CNN 预测结果产生差集,这使得将两类 CNN 模型整合在一起使结果更加稳定和有提高的空间。

参考文献(References)

[1]KABSCH W, SANDER C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features[J]. Biopolymers, 1983, 22(12):2577-2637.

[2]DING Y S, ZHANG T L, CHOU K C. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network[J]. Protein & Peptide Letters, 2007, 14(8):811-815.

[3]JONES D T. Protein secondary structure prediction based on position-specific scoring matrices[J]. Journal of Molecular Biology,1999, 292(2):195-202.

[4]LI Z, YU Y. Protein secondary structure prediction using cascaded convolutional and recurrent neural networks[A]. Proceedings of the twenty-fifth international joint conference on artificial intelligence[C]. AAAI Press, 2016: 2560-2567.

[5]SPENCER M, EICKHOLT J, CHENG J. A deep learning network approach to ab initio protein secondary structure prediction[J]. Computational Biology & Bioinformatics IEEE/ACM Transactions,2015,12(1):103-112.

[6]WANG S, PENG J, MA J, et al.Protein secondary structure prediction using deep convolutional neural fields[J]. Scientific Reports,2016,6:18962,DOI: 10.1038/srep18962.

[7]DUMOULIN V, VISIN F. A guide to convolution arithmetic for deep learning[EB/OL]. https://pdfs.semanticscholar.org/7918/2aab186f0b68a8432540d8695e1646338479.pdf, 2016.

[8]李彦冬, 郝宗波, 雷航. 卷积神经网络研究综述[J].计算机应用,2016,36(9):2508-2515. DOI: 10.11772/j.issn.1001-9081.2016.09.2508.
LI Yandong, HAO Zongbo, LEI Hang. Convolutional neural network research summary[J]. Journal of Computer Applications,2016,36(9):2508-2515.DOI: 10.11772/j.issn.1001-9081.2016.09.2508.

[9]HOCHREITER S, SCHMIDHUBER J.Long short-term memory[J]. Neural Computation,1997, 9(8):1735-1780.

[10]GERS F. A, SCHMIDHUBER J, CUMMINS F.Learning to forget: continual prediction with LSTM[J]. Neural Compu-

tation,2000 , 12 (10) :2451-2471.

[11]SMITH L N, TOPIN N. Deep convolutional neural network design patterns [EB/OL]. https://arxiv. org/abs/1611.00847,2017.

[12]CAO X. A practical theory for designing very deep convolutional neural networks[EB/OL].https://pdfs.semanticscholar.org/7922/2fad9f671be142bd7e42cd785a2cb06a1d30.pdf,2015.

[13]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.

[14]MACLIN R, OPITZ D. Popular ensemble methods: an empirical study [J]. Journal of Artificial Intelligence Research,2011,11:169-198.DOI: 10.1613/jair.614.

[15]CUFF J A, BARTON G J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction[J]. Proteins-structure Function & Bioinformatics, 1999 , 34 (4) :508.

[16]CUFF J A, BARTON G J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction[J]. Proteins-structure Function & Bioinformatics,2000,40(3):502-511.

[17]CHOLLET. Keras [EB/OL]. https://github. com/fchollet/keras,2015.

[18]WANG G, ZHAO Y, WANG D. A protein secondary structure prediction framework based on the Extreme Learning Machine[J]. Neurocomputing, 2008, 72(1-3):262-268.

[19]HUA S, SUN Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach.[J]. Journal of Molecular Biology, 2001, 308(2):397-407.

[20]冯永娥. 基于打分函数的蛋白质二级结构的识别[J]. 生物数学学报, 2016(4):455-460.

FENG Yonge. Identification of protein secondary structure based on scoring function[J]. Journal of Biomathematics, 2016(4):455-460.