DOI:10.3969/j.issn.1672-5565.201705007

自动编码器方法的蛋白质二级结构预测

张帅燕,刘毅慧
(齐鲁工业大学 信息学院,济南 250353)

摘 要:蛋白质二级结构预测是进行蛋白质三级结构研究的重要基础,氨基酸的编码方式对二级结构预测有一定的影响。本文应用了一种新的组合编码方式,即将基团编码与位置特异性打分矩阵(PSSM)进行组合的编码方式。本文中提出的基团编码是针对氨基酸的一种新的编码方式,基团编码是根据氨基酸内部组成来进行编码的,由42位属性组成。本文选取位置特异性打分矩阵(PSSM)中的Blosum62进化矩阵和新的基团编码进行组合,形成新的编码方式。然后对CB513和25pdb两组数据分别进行实验。本文中将采用贝叶斯分类器与自动编码器两种方法来对这种新的编码方式进行实验,然后比较这两种方法得到的两组数据的结果。可以很明显的发现采用自动编码器的实验结果要比使用贝叶斯分类器的结果要高出1.65%。在本文的实验中,可以提取特征的自动编码器的预测准确率更好。

关键词:蛋白质二级结构预测;基团编码;PSSM;贝叶斯分类器;自动编码器
中图分类号:TP391 **文献标志码**:A **文章编号**:1672-5565(2018)01-036-07

The prediction of protein secondary structure using auto encoder

ZHANG Shuaiyan, LIU Yihui
(*School of Information, Qilu University of Technology, Jinan* 250353, *China*)

Abstract: The secondary structure prediction is the basis of tertiary structure of protein, and the encoding method has influence on the prediction of secondary structure. A new encoding method composed of radical group encoding and position-specific scoring matrix(PSSM) is proposed. The radical group encoding contains 42 features, which is generated according to amino acids composition. A new encoding method was generated by combining the radical group encoding and the evolution matrix Blosum62. The Bayes classifier and auto encoder are used to predict the secondary structure for CB513 and 25pdb datasets. According to the comparison of the accuracy, the accuracy of auto encoder is higher 1.65% than the accuracy of Bayes classifier. In the experiment, the auto encoder extracting features can achieve higher accuracy.

Keywords: Protein secondary structure prediction; Radical group encoding; PSSM; Bayes classifier; Auto-encoder

随着蛋白质数据量的不断扩大,如何从蛋白质序列中提取出有用的生物信息是目前比较紧急重要的任务。蛋白质二级结构[1]是蛋白质形成稳定构象的重要基础,是研究蛋白质序列的重要前提。进而为研究蛋白质的功能以及相互作用模式提供基础,蛋白质研究的进一步发展有利于新药的研发,所以蛋白质的二级结构预测是目前的重要工作。

在蛋白质二级结构预测方面,已有许多的研究方法。基于残基构象性的Chou-Fasman算法[2]和以信息论作为基础的GOR算法[3],都是单序列预测方法。Qian[4]于1988年用人工神经网络进行预测,得到的准确率在60%以上。吴玉明[5]提取氨基酸的理化特征和倾向因子,采用支持向量机(SVM)进行预测,得到的预测准确率为70.7%。利用kb-prossp-nn算法,对数据CB396的预测准确率为82%[6]。文献[7]中,作者组合特定位置打分矩阵信息和深度学习网络架构,得到的预测准确率为80.7%。

收稿日期:2017-05-22;**修回日期**:2017-10-22.
基金项目:国家自然科学基金项目(61375013);山东省自然科学基金项目(ZR2013FM020).
作者简介:张帅燕,女,硕士研究生,研究方向:生物医学信息处理、智能计算;E-mail:1593730838@qq.com.
*通信作者:刘毅慧,女,教授,研究方向:生物计算,智能信息处理;E-mail: yxl@qlu.edu.cn.

文献[8]中,作者运用动态贝叶斯分类器来进行蛋白质二级结构预测。并且引入了动态贝叶斯分类器稀疏参数算法,得到的预测准确率为 76.3%。文献[9]中,作者将支持向量(SVM)和贝叶斯分类器结合起来进行蛋白质二级结构预测,实验数据为 RS126 数据集。文献[10]中,作者对 SARS 数据集进行实验,用支持向量机(SVM)进行蛋白质二级结构预测,得到的 OvO 准确率为 89.27%。作者用支持向量机(SVM)对位置特异性打分矩阵(Position specific scoring matrix, PSSM)进行蛋白质二级结构预测,实验数据为 CB513,得到的预测准确率为 76.11%[11]。

本实验中,将基团编码与 PSSM 数据组合一起得到的新的编码方式,分别进行自动编码器和贝叶斯分类器进行预测。

1　新的编码方式

1.1　基团编码

本文按照含有氨基酸内部组成结构信息的基团表,对氨基酸进行编码(见表 1),基团编码由 42 个属性组成。

表 1　氨基酸名称简写表

Table 1　Short names of amino acids

氨基酸名称	编码简写	氨基酸名称	编码简写
丙氨酸	A	甲硫氨酸	M
半胱氨酸	C	天冬酰胺	N
天冬氨酸	D	脯氨酸	P
谷氨酸	E	谷氨酰胺	Q
苯丙氨酸	F	精氨酸	R
甘氨酸	G	丝氨酸	S
组氨酸	H	苏氨酸	T
异亮氨酸	I	缬氨酸	V
赖氨酸	K	色氨酸	W
亮氨酸	L	酪氨酸	Y

表 2 中的 1 表示带环基团的编号,共 37 个。表中 2 表示各个基团中含有的原子名称。表中 3 表示含有对应基团的氨基酸名称简写。表格 1 中第 3 列中的 I1 类表示是 L、P、F、W、Y、E、R、K、H、M、N、D、Q 等 13 个氨基酸。“起始”是指蛋白质序列中的第一个氨基酸,“终止”是指蛋白质序列中的终止氨基酸。

表 3 中的 1 表示剩余不带环基团的编号,2 表示含有这个基团的氨基酸名称,3 表示基团中含有的原子名称。

表 2　不带环基团信息表

Table 2　Information of the radical group without band

1	2	3	1	2	3
1	N,A,C,2HA	G	20	B,CG,2CD,HG	L
2	N,A,C,B,HA	除 G	21	B,CG,CD,2HGRQKE	
3	2HN,N,A	起始	22	B,CG,SD,2HG	M
4	H,N,A	所有	23	CG,CD1,3HD	L
5	A,C,O1,O2	终止	24	CG,CD2,3HD	LI
6	A,C,O	所有	25	CG,SD,CE	M
7	A,B,3HB	A	26	CG,CD,NE1,OE2,	Q
8	A,B,CG,2HB	I1 类	27	CG,CD,2OE	E
9	A,B,OG,2HB	S	28	CG,CD,NE,2HD	R
10	A,B,SG,2HB	C	29	CG,CD,CE,2HD	R
11	A,B,2CG,HB	VI	30	SD,CE,3HE	M
12	A,B,OG1,CG2,HB	T	31	CD,NE,CZ,HE	R
13	B,CG1,3HG	VI	32	CD,CE,NZ,2HE	K
14	B,CG2,3HG	VT	33	NE,CZ,2NH	R
15	B,CG2,CD,2HG	I	34	CE,NZ,3HZ	K
16	B,OG1,HG	T	35	CZ,NH1,2HH	R
17	B,CG,ND1	N	36	CZ,NH2,3HH	R
18	B,CG,OD1	D	37	CZ,CH,HH5	Y
19	B,CG,OD2	ND			

表 3　带环基团信息表

Table 3　Information of the radical group with band

1	2	3
38	P	N,A,C,B,CG,CD,HA,HB1,HB2,HG1,HG2,HD1,HD2
39	H	B,CG,CD1,ND2,NE1,CE2,HD1,HE1,HE2
40	F	B,CG,CD1,CD2,CE1,CE2,CZ,HD1,HD2,HE1,HE2,HZ
41	W	B,CG,CD1,CD2,NE1,CE2,CE3,CZ1,CZ2,CH,HD1,HE1,HE2,HZ1,HZ2,HH
42	Y	B,CG,CD1,CD2,CE1,CE2,CZ,OH,HD1,HD2,HE1,HE2

1.2　基团编码方式

本文对氨基酸的编码方式:氨基酸 A,根据已知的基团信息表,氨基酸 A 含有的基团包括 2,4,6,7 几个位置,那么 42 位编码中需要在 2,4,6,7 几个位置用 1 表示,剩余的位置则用 0 表示,42 位基团编码 010101100000000000000000000000000000000000 可以表示氨基酸 A。编码氨基酸 C,根据基团信息表可以发现,出现氨基酸 C 的基团编号是 2,4,6,10 等 4 个位置,在 42 位编码中 2,4,6,10 几个位置上用 1 来表示,其余位置用 0 来表示,那么可以用编码 010101000100000000000000000000000000000000 表示氨基酸 C。

依照这种方法依次对 20 种氨基酸进行编码,可以总结得到基团编码(见表 4)。

就其中一条蛋白质来说,如果序列中含有 m 个氨基酸,在对基团编码进行滑动窗口设置时,在数据前后补 0。将滑动窗口数选为 15 时,需要在氨基酸

序列前后分别补上 7 个 0,,依次对序列从前到后进行滑动窗口选取数据,生成 42 * 15 维的数据,则这条蛋白质可以表示为 $m * 630$。

表 4 氨基酸-基团编码对应表

Table 4 the radical group encoding method

氨基酸	基因编码
A	0101011000000000000000000000000000000000000000
R	0101010100000000000010000001101010110000000
N	0101010100000000101000000000000000000000000
D	0101010100000000011000000000000000000000000
C	0101010001000000000000000000000000000000000
Q	0101010100000000000010000100000000000000000
E	0101010100000000000010000010000000000000000
G	1001010000000000000000000000000000000000000
H	0101010100000000000000000000000000000001000
I	0101010000101010000000010000000000000000000
L	0101010100000000000100110000000000000000000
K	0101010100000000000010000000000101000000000
M	0101010100000000000010010000100000000000000
F	0101010100000000000000000000000000000000100
P	0101010100000000000000000000000000000010000
S	0101010010000000000000000000000000000000000
T	0101010000010101000000000000000000000000000
W	0101010100000000000000000000000000000000010
Y	0101010100000000000000000000000000000100001
V	0101010000101100000000000000000000000000000

1.3 PSSM

PSI-BLAST 是 BLAST 的变种[12],并且通过蛋白质的功能特征搜索蛋白质序列数据库中的相关序列,并且这些相关序列的相似性太远是 BLAST 搜索不到的。

位置特异性打分矩阵(Position specific scoring matrix, PSSM)是通过 PSI-Blast 搜索 nr 数据库后,计算联配中每个位点的新得分建立的,含有丰富的生物进化信息。在本文试验中,选用 Blosum 62 进化矩阵来进行实验。在实验时,选取 Blosum62 矩阵的前 20 列,这样假设其中一条蛋白质长度为 m,则 PSSM 矩阵表示为$m * 20$。对 PSSM 数据取滑动窗口为 13 时,需要在蛋白质序列前后分别补上 6 个 0,依次对序列从前到后进行滑动窗口选取数据,生成 $20 * 13$ 维的数据,则这条蛋白质可以表示为 $m * 260$。将基团编码与位置特异性得分矩阵组合在一起,得到新的编码方式。如果某一条蛋白质长度为 m,如果取滑动窗口数为 13 时,42 位基团编码数据是$m * 546$,位置特异性打分矩阵(PSSM)数据为 $m * 260$。将两组数据组合在一起组成新的蛋白质序列编码为$m * 806$。

1.4 DSSP 结构

DSSP 结构划分方法在蛋白质二级结构预测中使用的较为广泛,DSSP[13] 由 Wolfgang Kabsch 于 1983 年提出的, DSSP 含有 8 种结构状态,分别是 B(β 桥)、C(非 B,E,G,H,I,S 或 T)、E(β 折叠)、G(3_{10} 螺旋)、H(α 螺旋)、I(π 螺旋)、S(转角)、T(β 转角)。在实际试验中会将这 8 种结构状态进行简化,比较常用的简化划分方法有 5 种,(1)H 是 H, E 是 E,其它都是 C;(2)H 是 H,B、E 是 E,其它都是 C;(3)G、H 是 H,E 是 E,其它都是 C;(4)G、H 是 H,B、E 是 E,其它都是 C;(5)G、H、I 是 H,B、E 是 E,其它都是 C。

在本文试验中,将选取 G、H、I 是 H,B、E 是 E,其它都是 C 的结构状态划分方法。

1.5 预测指标

实验中选取的蛋白质准确率预测的定义指标是 Q_3 和 Q_i (i 分别是 C、E、H 这 3 个类别)[14]:

$$Q_3 = \frac{TP_C + TP_E + TP_H}{T} \tag{1}$$

$$Q_i = \frac{TP_i}{TP_i + FP_i} \tag{2}$$

其中,i 为 C、E、H 中的某一类,可以得到每个类别的准确率。TP_i 表示某个状态被准确预测出的残基数,T 表示蛋白质序列中含有的残基总数。

2 自动编码器

深度学习是具有人工神经网络特点的学习方法,作为在深度学习中广泛运用的自动编码器,自动编码器具有层次结构的特点,是一种多层前传神经网络。自动编码器由 Rumelhart 于 1986 年提出,用于处理高维复杂的数据,可以对高维的实验数据进行降维处理,得到低维特征向量[15-16],然后将得到的特征向量送入分类器中。

自动编码器[17] 包括输入层、隐含层、输出层。输入层和输出层含有相同的维度,都是 m 维,隐含层则为 n 维。隐含层神经元数要比输入层神经元数多,则需要隐含层神经元具有稀疏性。需要加入稀疏性限制,确保隐含层的大部分是被抑制的状态。

针对原始数据 X,经过学习、调整参数、映射之后得到新的数据 Y。在编码过程中,得到隐含层的表达:

$$Z = s_e(W^1 X + b^1) \tag{3}$$

其中, s_e 是编码过程中的激活函数,通常选用 sigmoid 函数。W^1 为权值向量, b^1 为偏差量。

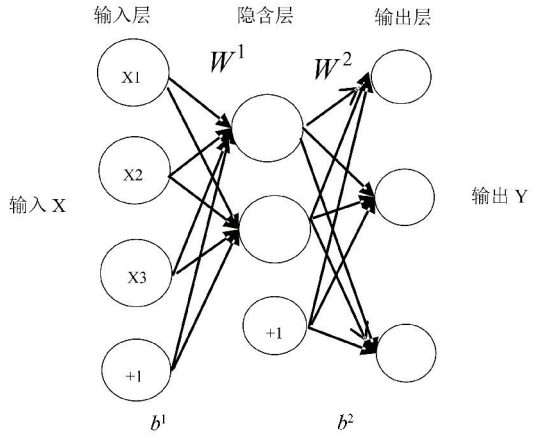

图 1　单层自动编码器结构图

Fig.1　The structure of single-layer auto encoder

再对映射后的表达 Z 进行反向加权映射,即对隐含层得到的表示进行重构。在解码过程中,可以得到与原始输入具有相同维度的新的表达:

$$Y = s_d(W^2 Y + b^2) \tag{4}$$

其中, s_d 为解码过程中的激活函数,选用 sigmoid 函数。通常 W^1 与 W^2 互为转置,可以表示为 $W^{1T} = W^2$。

W^2 为权重向量, b^2 为偏差量,自动编码器在训练过程中寻找参数 $\theta = \{W, b^1, b^2\}$。

重构得到的新的表达 Y 相当于在已知 Z 表达的条件下,对输入数据 X 的预测,这样的重构存在误差。自动编码器中的学习过程就是使重构后得到的输出尽量还原输入层的数据,即使表达 Z 与表达 X 尽量相同。重构错误有其方法可以量化,重构中的误差函数选择均方误差函数:

$$E = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{5}$$

其中,N 是样本总数。

期望值(稀疏度值)可以用 ρ 表示[18],每个神经元 i 的平均激活度可以表示为:

$$\hat{\rho}_i = \frac{1}{N}\sum_{j=1}^{N} y_i^{\,1}(x_j) = \frac{1}{N}\sum_{j=1}^{N} s_e(w_i^{\,1T} x_j + b_i^{\,1}) \tag{6}$$

其中,N 是训练样本总数, X_j 是第 j 个训练样本, $b_i^{\,1}$ 是 b^1 偏差量中的第 i 个偏差量。当 $\hat{\rho}_i$ 和 ρ 的值相差太大时,需要用到 Kullback - Leibler divergence 函数进行惩罚。

$$KL(\rho \,||\, \hat{\rho}_i) = \rho\log\left(\frac{\rho}{\hat{\rho}_i}\right) + (1-\rho)\log\left(\frac{1-\rho}{1-\hat{\rho}_i}\right) \tag{7}$$

KL 函数是衡量这两个数值之间距离大小关系的函数,它随着 $\hat{\rho}_i$ 与 ρ 之间的距离增大而增大,当 $\hat{\rho}_i$ 和 ρ 的取值相同时,函数值为 0。要是使损失函数达到最小化,要使 $\hat{\rho}_i$ 和 ρ 的值尽可能相近。ρ 是稀疏度值,在实验中取值为 0.06。

$$\sigma = \frac{1}{2}\sum_{l}^{L}\sum_{j=1}^{N}\sum_{c}^{p}(w_{jc}^{\,l})^2 \tag{8}$$

其中,L 表示隐含层数,在本文实验中隐含层设置为 1 层。N 是样本总数,p 表示训练中的变量数,即输入样本的维度 m。稀疏编码器中的损失函数可以表示为:

$$J = E + \lambda * \sigma + \beta\sum_{j=1}^{m} KL(\hat{\rho}_j \,||\, \rho) \tag{9}$$

其中 β 为控制稀疏性正则化的系数,在本文实验中的取值为 4。λ 为控制权重正则化系数,本文实验中取值为 0.06。ρ 是稀疏度值,试验时取值为 0.06。在经过自动编码器得到特征向量后,将特征向量送入到贝叶斯分类器中得到预测准确率。在实验中,选择了单层编码器来获得特征向量。由于在实验中,将自动编码器的隐含层数设置为了1 500,所以送入贝叶斯分类器的输入数据为1 500维。

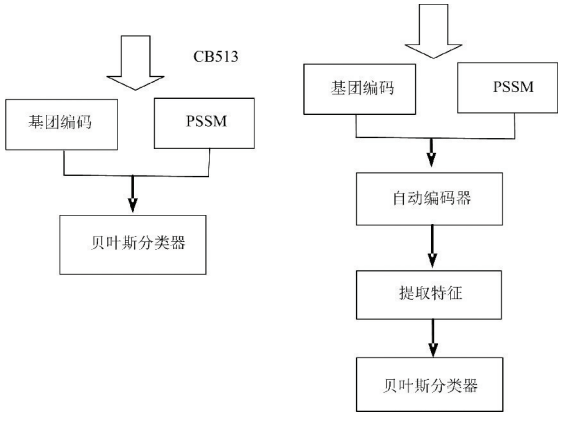

图 2　数据的处理方法图

Fig.2　The data processing process

实验中,对数据进行两种方法进行预测,一种是将基团编码与 PSSM 矩阵组合后直接送入贝叶斯分类器中进行预测,输入的维度就是实际数据的维度。如果选取 13 个滑动窗口,其中一条蛋白质长度为 m,则输入贝叶斯分类器中的维度为$m * 806$。一种是将基团编码与 PSSM 矩阵组合后的数据送入自动编码器中,提取特征,如果隐含层设置的神经元数为 1 500,则提取到的特征数为$m * 1\,500$,然后将提取到的特征送入贝叶斯分类器中进行预测。

3 贝叶斯分类器

贝叶斯分类器是用来分类的贝叶斯网络,它的原理[19]通过先验概率、条件概率和贝叶斯公式得到后验概率。贝叶斯分类器处理一个 C 分类问题,在本文实验中,蛋白质的二级结构是一个 3 分类问题,即 C 值 $C = 1,2,3$。如果每个类别的先验概率是已知的,它是 $P(X_i)(i = 1,2,3)$,条件概率密度 $P(Y \mid X_i)(i = 1,2,3)$,用贝叶斯公式,可以计算出的后验概率:

$$P(X_i \mid Y) = \frac{P(Y \mid X_i)P(X_i)}{\sum P(Y \mid X_j)P(X_j)} \qquad (10)$$

4 实验数据

试验中选用的数据集为 CB513[20] 数据集和 25PDB[21] 数据集。CB513 数据集中共含有 513 条蛋白质,其中蛋白质序列的相似度小于 25%。25PDB 数据集是典型的非冗余数据集,其序列相似度小于 25%。25PDB 数据集共包括 1 673 条蛋白质序列,其中 43 条蛋白质序列为空,最后的实验过程中采用 1 640 条蛋白质序列。在实验过程中需要将蛋白质数据进行 3 折交叉验证,即将所有数据分为 3 份,选取其中任意 2 份为训练集,则剩余 1 份作为测试集。然后将实验数据分别用自动编码器和贝叶斯分类器进行实验,得到两组数据。

4.1 结果分析

4.1.1 CB513 数据结果

1) 在对 CB513 数据进行实验,将 513 条蛋白质共包括 84 119 个氨基酸,随机进行 3 折交叉验证,其中训练样本含有 342 条蛋白质,测试样本含有 171 条蛋白质。当滑动窗口数为 13 时,由基团编码 42 * 13,PSSM 编码 20 * 13,组合生成新的编码方式,即 546+260 = 806,则输入向量的维度为 806 维,送入贝叶斯分类器中得到实验结果(见表 5)。

表 5 CB513 贝叶斯分类器结果

Table 5 The accuracy of Bayes classifier %

窗口	Q_3	Q_C	Q_E	Q_H
13	70.56	76.80	56.13	72.32
15	70.85	76.80	56.79	72.70
17	70.95	76.82	56.98	72.84
19	70.94	76.68	57.35	72.76
21	70.98	76.54	57.58	72.94

2) 用单层自动编码器试验时,对基团编码与 PSSM 组合形成的新的编码数据进行 3 折交叉验证。当滑动窗口为 13 时,单个隐含层神经元设置为 1 500,控制稀疏性正则化系数为 4,权重正则化系数为 0.06,稀疏度值为 0.06。得到的特征向量数据为 1 500维,然后送入贝叶斯分类器中(见表 6)。

表 6 CB513 自动编码器结果

Table 6 The accuracy of single-layer auto encoder %

窗口数	Q_3	Q_C	Q_E	Q_H
13	71.95	77.20	59.16	73.86
15	71.67	77.07	58.28	73.80
17	71.70	77.89	57.49	73.33
19	71.22	77.89	56.34	72.80
21	71.18	77.07	57.52	72.87

4.1.2 25PDB 数据结果

1) 对 25PDB 数据进行实验时,数据共包括 1 640 条蛋白质序列。进行 3 折交叉验证,当滑动窗口数为 15 时,由基团编码维度 42 * 15,PSSM 编码维度 20 * 15,组合生成新的编码方式,维度为 630+300 = 930 维,则输入向量的维度为 930 维,送入贝叶斯分类器中得到实验结果(见表 7)。

表 7 25PDB 贝叶斯分类器结果

Table 7 The accuracy of Bayes classifier %

窗口数	Q_3	Q_C	Q_E	Q_H
13	71.54	76.84	59.19	73.31
15	71.73	76.81	59.74	73.49
17	71.79	76.75	59.95	73.68
19	71.85	76.66	60.10	73.83
21	71.92	76.80	60.11	73.86

2) 用单层自动编码器试验时,对基团编码与 PSSM 组合形成的新的编码数据进行 3 折交叉验证。单个隐含层神经元设置为 1 500,控制稀疏性正则化系数为 4,权重正则化系数为 0.06,稀疏度值为 0.06。得到的特征向量数据为1 500维,然后送入贝叶斯分类器中(见表 8)。

表 8 25PDB 自动编码器结果

Table 8 The accuracy of single-layer auto encoder %

窗口数	Q_3	Q_C	Q_E	Q_H
13	73.19	79.26	61.52	73.67
15	73.08	79.04	61.60	73.53
17	73.04	78.33	61.70	74.21
19	72.92	78.21	61.66	74.02
21	72.28	77.86	60.35	73.48

4.2　结果分析

根据结果分析 CB513 数据(见图 3),发现当滑动窗口数不同时,贝叶斯分类器结果在70.56%~70.98%。使用自动编码器时,准确率在71.18%~71.95%。比较两组分类器结果,当滑动窗口数为 13 时,自动编码器的准确率比贝叶斯分类器要高出 1.39%。分析数据25pdb时,贝叶斯分类器的结果在 71.54%~71.92%,自动编码器准确率在73.04%~73.19%。当滑动窗口数为 13 时,自动编码器的预测准确率比贝叶斯分类器要高1.66%。随着蛋白质序列数量的增加,25pdb的贝叶斯分类器结果要比 CB513 数据的结果高出0.98%。25pdb数据的自动编码器结果要比 CB513 的结果高出 1.7%(见图 4)。由于在实验中,用到的分类器都是贝叶斯分类器,可以发现,经过自动编码器得到特征向量后用贝叶斯分类器进行分类的结果要比单独经过贝叶斯分类器的结果要高 1.66%。而且随着蛋白质序列的增加,二级结构预测准确率也会得到提升。

与文献[11]相比,作者使用支持向量机预测特异性位置打分矩阵(PSSM)得到的预测率更高一些。本文首先在 DSSP 结构划分上不一样,选取的划分方法是是所有五种划分结构中准确率最低的,导致预测结果有所下降。其次我们用自动编码器提取特征之后用贝叶斯分类器进行预测,贝叶斯分类器的处理速度比支持向量机更快。

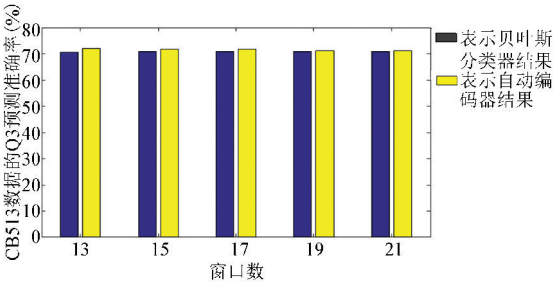

图 3　CB513 数据的实验直方图

Fig.3　Histogram of CB513

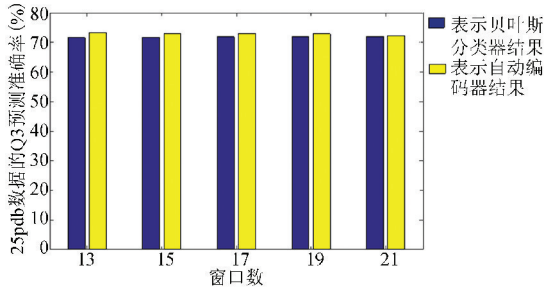

图 4　25pdb 数据的实验直方图

Fig.4　Histogram of 25pdb

5　结　论

通过两组数据的直方图可以发现:

1)本文提出组合基团编码与 PSSM 数据的新的编码方式,然后选取不同的分类器进行实验。

2)选择单层自动编码器和贝叶斯分类器对这种新的编码方式进行预测分类,其中自动编码器的稀疏度参数选取的比较小,自动编码器的预测准确率比贝叶斯分类器结果要高出 1.65%。

3)比较 CB513 和 25pdb 两组可以发现,随着蛋白质数据量的增加,无论是贝叶斯分类器还是自动编码器,预测准确率得到了提高。

在本文的实验中,选取的是含单个隐含层的自动编码器来进行分类预测的,设置的稀疏度比较小,运行时间比较长。那么在接下来的实验中,首先进行参数的调整,降低运行时间。堆叠的自动编码器含有更多的隐含层,并且需要设置更多的隐含层神经元数。接下来我们会选取堆叠自动编码器来进行预测实验。

参考文献(References)

[1]邵建林, 徐东, 王兰州,等. 一种新的预测蛋白质二级结构的模型-贝叶斯神经网络[J]. 计量学报, 2006, 27(3):281-285.DOI: 10.3321/j.issn:1000-1158.2006.03.020.
SHAO Jianlin, XU Dong, WANG Lanzhou, et al. A new model for predicting protein two level structure-Bayesian neural network[J]. Metrology Journal, 2006, 27(3): 281-285.DOI: 10.3321/j.issn:1000-1158.2006.03.020.

[2]CHOU P Y, FASMAN G D. Prediction of protein conformation[J]. Biochemistry, 1974, 13(2): 222-245. DOI: 10.1021/bi00699a002.

[3]GARNIER J, OSGUTHORPE D J, ROBSON B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins.[J]. Journal of Molecular Biology, 1978, 120(1): 97-120. DOI: 10.1016/0022-2836(78)90297-8.

[4]QIAN N, SEJNOWSKI T J. Predicting the secondary structure of globular proteins using neural network models[J]. Journal of Molecular Biology, 1988, 202(4): 865-884. DOI:10.1016/0022-2836(88)90564-5.

[5]吴玉明. 蛋白质二级结构预测的一种新的编码方式[J]. 工业控制计算机, 2015,(4): 109-110, 113. DOI: 10.3969/j.issn.1001-182X.2015.04.048.
WU Yuming. A new coding method for prediction of protein. secondary structure[J]. Industrial Control Computer, 2015, (4): 109-110, 113. DOI: 10.3969/j.issn.1001-182X.2015.

04.048.

[6]PATEL M S, MAZUMDAR H S. Knowledge base and neural network approach for protein secondary structure prediction [J]. Journal of Theoretical Biology, 2014, 361:182-189. DOI:10.1016/j.jtbi.2014.08.005.

[7]SPENCER M, EICKHOLT J, CHENG J. A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2015, 12(1):103-112.DOI:10.1109/TCBB.2014.2343960.

[8]ZAFER A, AJIT S, JEFF B. Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure[J]. BMC Bioinformatics, 2011, 12(1):154.

[9]王宝文, 王水星, 刘文远,等. 结合支持向量机和贝叶斯方法进行蛋白质二级结构预测[J]. 生物信息学, 2010, 8(1):75-77. DOI:10.3969/j.issn.1672-5565.2010.01.018.

WANG Baowen, WANG Shuixing, LIU Wenyuan, et al. Combining support vector machines and Bayesian methods to predict protein two structure prediction, [J].Chinese Journal of Bioinformatics, 2010, 8 (1): 75-77. DOI:10.3969/j.issn.1672-5565.2010.01.018.

[10]吴琳琳, 徐硕. 基于 SVM 的蛋白质二级结构预测[J]. 生物信息学, 2010, 08(3):187-190. DOI:10.3969/j.issn.1672-5565.2010.03.001.

Wu Linlin, XU Shuo. SVM based protein two class structure prediction [J]. Chinese Journal of Bioinformatics, 2010, 08 (3): 187-190. DOI:10.3969/j.issn.1672-5565.2010.03.001.

[11]WANG Y C, CHENG J Y, LIU Y H, et al. Prediction of Protein Secondary Structure using Support Vector Machine with PSSM Profiles——2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference [C].Chongqing:IEEE, 2016.502-505.DOI: 10.1109/ITNEC.2016.7560411.

[12]泽瓦勒贝 M (ZVELEBIL M)著.李亦学, 郝沛主译 理解生物信息学[M].北京:科学出版社, 2012.

[13]KABSCH W,SANDER C, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features [J]. Biopolymers, 1983, 22 (12): 2577-2637. DOI:10.1002/bip.360221211.

[14]张海霞,唐焕文,张立震,等. 蛋白质二级结构预测方法的评价[J].计算机与应用化学,2003,20(6):735-740. DOI: 10.3969/j.issn.1001-4160.2003.06.005.

ZHANG Haixia, TANG Huanwen,ZHANG Lizhen, et al. Evaluation of protein two grade structure prediction method [J].Computer and Applied Chemistry,2003, 20(6): 735-740. DOI: 10.3969/j.issn.1001-4160.2003.06.005.

[15]张开旭, 周昌乐. 基于自动编码器的中文词汇特征无监督学习[J]. 中文信息学报, 2013, 27(5):1-7,92.DOI: 10.3969/j.issn.1003-0077.2013.05.001.

ZHANG Kaixu, ZHOU Changle. Unsupervised feature Learning for chinese Lexion based on auto-encoder[J]. Journal of Chinese Information Processing,2013, 27(5): 1-7,92.DOI: 10.3969/j.issn.1003-0077.2013.05.001.

[16]RUMELHART D E, HINTON G E, EILLIAMS R J. Learning representations by back-propagating errors[J]. Readings in Cognitive Science, 1988 , 323 (6088) :399-421.DOI:10.1038/323533a0.

[17]邓俊锋, 张晓龙. 基于自动编码器组合的深度学习优化方法[J]. 计算机应用, 2016, 36(3):697-702.DOI: 10.11772/j.issn.1001-9081.2016.03.697.

DENG Junfeng, ZHANG Xiaolong. Depth learning optimization method based on automatic encoder combination[J]. computer application, 2016, 36(3): 697-702. DOI: 10.11772/j.issn.1001-9081.2016.03.697.

[18]OLSHAUSEN, B A, FIELD D J. Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?[J].Vision Research, 1997, 37(23): 3311-3325. DOI: 10.1016/S0042-6989(97)00169-7.

[19]THEODORIDIS S, KOUTROUMBAS K. Pattern Recognition, Third Edition[J].Encyclopedia of Information Systems, 2003:459-479. DOI:10.1016/B0-12-227240-4/00132-5.

[20]CUFF J A, BARTON G J.Application of multiple sequence alignment Profiles to improve protein secondary structure prediction[J]. Proteins: Structure, Function and Genetics, 2000, 40 (3): 502 - 511. DOI: 10. 1002/1097 - 0134 (20000815)40:33.0.CO.

[21]KEDARISETTI K D, KURGAN L A, DICK S. Classifier ensembles for protein structural class prediction with varying homology[J]. Biochemical Biophysical Research Communications, 2006, 348 (3): 981 - 988. DOI: 10.1016/j.bbrc.2006.07.141.