

基于多核 LSSVM 的谷物蛋白质二级结构预测与优化

梁俊¹ 刘静^{1*} 管晓² 陈滢滢¹

(1. 上海海事大学信息工程学院, 上海 201306; 2. 上海理工大学健康科学与工程学院, 上海 200093)

摘要:蛋白质的二级结构对其空间结构和功能有着极其重要的影响, 利用机器学习方法进行谷物蛋白质二级结构预测是生物和食品领域的重要研究内容。作者在现有蛋白质数据库中选取玉米、小麦、大豆的谷物蛋白质, 使用多特征融合方式对蛋白质序列进行特征提取, 提出将多核学习与最小二乘支持向量机 (LSSVM) 相结合, 以多个核函数的线性加权组合代替传统单一核函数, 利用核权重调整融合效果, 构建多核 LSSVM 模型预测谷物蛋白质二级结构。使用粒子群优化算法 (PSO) 对模型超参数进行优化, 寻找最佳超参数组合提升模型预测性能。研究结果表明, 多核 LSSVM 模型能够改善单一核函数高维映射的局限性, 融合各核函数优势, 通过 PSO 算法获取最佳超参数组合。该模型结合多特征提取方式显著提高了谷物蛋白质二级结构预测的 Q_3 准确率。

关键词: 谷物; 蛋白质二级结构; 多核; 最小二乘支持向量机; 粒子群算法

Research on Prediction and Optimization of Cereal Protein Secondary Structure Based on Multi-Kernel LSSVM

LIANG Jun¹ LIU Jing^{1*} GUAN Xiao² CHEN Yingying¹

(1. School of Information Engineering, Shanghai Maritime University, Shanghai 201306; 2. School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093)

Abstract: The secondary structure of protein has an extremely important impact on their spatial structure and function. Using machine learning methods to predict the secondary structure of grain proteins is one of the important topics in the fields of biology, food, etc. In this experiment, three grain proteins of corn, wheat, and soybeans were selected from the existing protein data bank, using multi-feature fusion method to extract features of protein sequence, and propose to combine multi-core learning with least square support vector machine classifier, the linear weighted combination of multiple kernel functions is used to replace the traditional single kernel function, and the kernel weight is used to adjust the fusion effect to construct a multi-core LSSVM prediction model to predict the secondary structure of grain proteins. The particle swarm optimization algorithm is used to optimize the model hyper parameters and find the best hyper parameter combination to improve the prediction performance of the model. The experimental results prove that the multi-kernel LSSVM algorithm proposed in this paper can make up for the limitations of single kernel function high-dimensional mapping, integrate the advantages of each kernel function, and find the best hyperparameter combination through the PSO algorithm, combine multiple feature extraction methods to significantly improve the Q_3 accuracy of grain protein secondary structure prediction.

Keywords: cereals; protein secondary structure; multi-kernel; least square support vector machine; particle swarm optimization

基金项目: 国家自然科学基金项目 (32172247); 上海市科技兴农项目 (2021-02-08-00-12-F00780)。

通信作者: 刘静 (1979—), 女, 博士, 副教授, 硕士研究生导师, 主要从事信息技术与食品安全交叉领域、生物信息、机器学习、智能算法等研究。Email: jingliu@shmtu.edu.cn

收稿日期: 2021-12-21 改回日期: 2022-03-10

谷物是人类赖以生存的主要食物来源。玉米、小麦、大豆等都是我国食用量较大的谷物,其提供的纤维素、微量营养素等成分具有良好的营养保健作用^[1-2],同时其富含的膳食纤维在预防疾病方面有一定成效^[3]。《中国居民膳食指南》提出了“食物多样,谷类为主,粗细搭配”的原则,可见谷物在国人膳食中的重要地位。谷物的功效由其包含的蛋白质所提供,蛋白质的功能决定其在生物体中发挥的作用。蛋白质的结构信息蕴含了分子层面的功能信息^[4],作者对谷物蛋白质结构和功能进行了分析,为谷物膳食营养均衡^[5]、营养品质调控提供参考。

随着高通量测序技术的出现,存在大量未知结构的已测序谷物蛋白质,传统生物学中常使用晶体衍射 X 射线、核磁共振等技术测定谷物蛋白质结构,但耗时耗力,无法广泛使用。伴随着生物信息学和计算技术的发展,应用计算方法预测谷物蛋白质结构更加便捷高效^[6],可利用已知的氨基酸序列及结构信息预测谷物蛋白质相关结构。

蛋白质二级结构是氨基酸序列进化为蛋白质三级结构的桥梁,是氨基酸序列在多肽链中的局部空间构象^[7],常用于研究蛋白质突变、蛋白质晶体结构解析等^[8]。蛋白质二级结构预测一直备受国内外学者的关注,目前研究进展主要分为 3 个进程。初期利用统计学方法提取氨基酸信息,如著名的 Chou-Fasman 方法^[9];第二阶段考虑氨基酸残基间的相互作用和影响,对局部信息进行提取,如 GOR 算法^[10],基于信息论解决氨基酸作用力信息传递问题;现阶段主要利用计算技术对蛋白质二级结构进行预测,围绕特征提取与智能算法两个研究层面进行相关蛋白质二级结构的预测。

作者基于谷物蛋白质数据进行蛋白质二级结构预测,在有效提取氨基酸序列特征的基础上,提出多特征融合获取氨基酸序列信息,并使用滑动窗口机制实现局部信息的提取。考虑谷物数据大小的局限性,神经网络易出现过拟合,首次提出基于多核最小二乘支持向量机建立预测模型。针对模型超参数对模型的预测准确率、稳定性、泛化能力等的影响,作者应用粒子群算法^[11]并借助智能搜索策略优化模型超参数。通过最优超参数的分类模型提取蛋白质序列中的多特征关系,进一步提

高模型性能,为谷物蛋白质二级结构预测提供参考。

1 数据与方法

1.1 数据来源

从 PDB 数据库中选取了 3 种谷物蛋白质,分别是小麦、玉米和大豆,见表 1。获取已测序并已知其结构的蛋白质序列,提取氨基酸残基和对应的二级结构信息,利用该信息作为数据集的样本和标签进行蛋白质的二级结构预测。

表 1 PDB 数据库检索条件

Table 1 PDB database search conditions

谷物	物种
玉米	<i>Zea mays</i>
小麦	<i>Triticum aestivum</i>
大豆	<i>Glycine</i>

蛋白质二级结构按照 DSSP 分类方法一般分为 α -螺旋(H)、 β -2 桥(B)、折叠(E)、螺旋-3(G)、螺旋-5(I)、转角(T)、卷曲(S)和环(L)等 8 种类型,谷物蛋白质二级结构预测常用三类结构划分^[12],通过官方定义的 8 种结构形态,按照形态划分成 3 种二级结构,分别对应螺旋、折叠和无规卷曲。作者主要研究谷物蛋白质在三类结构上的预测,按照上述划分说明,将 8 种结构的 G、H、I 替换成 H, B、E 替换成 E,其余都替换成 C。

1.2 特征提取

特征提取是数据预处理中的关键步骤,优秀的特征提取方式能够有效提取氨基酸序列的总体与局部信息,进而提高预测准确率。作者提出多特征融合方式对相关特征进行处理,氨基酸类别信息是氨基酸残基的基本特征,生物学中目前已定义的氨基酸有 20 种,分别用 20 个大写字母表示,在特征提取时常用五位编码或正交编码表示^[13],本研究使用的是正交编码。每种氨基酸用 20 维的向量表示,20 种氨基酸由 20 个向量表示,向量之间两两正交,见表 2。该方式计算速度较快,但仅使用正交编码有其局限性,如过于稀疏,导致信息量过少而不够密集^[14]。

氨基酸理化性质在氨基酸进化过程中扮演着重要角色^[15]。不同的理化性质对其结构域功能有着重要影响^[16]。常用的理化性质较多,作用于蛋白质的不同方面。在氨基酸进化过程中,部分理

表2 氨基酸的正交编码

氨基酸	编码向量
丙氨酸	10000000000000000000
半胱氨酸	01000000000000000000
天冬氨酸	00100000000000000000
谷氨酸	00010000000000000000
⋮	⋮

化性质对其结构有一定的改变和促进作用。如氨基酸的疏水性质与等电点对多肽链的局部构象有显著影响^[17],疏水性氨基酸常折叠在蛋白质内部,创造催化反应的疏水环境,有利于氨基酸残基间的理化反应。以氨基酸等电点及电离平衡常数等特性为基础的电泳及沉淀等方法,常用于氨基酸和蛋白质的分离和提纯等过程^[17],其性质也代表蛋白质所带电荷在电场中的移动方向,对蛋白质二级结构的形成有重要影响。氨基酸残基脱水缩合形成肽链的过程中,其相对分子质量对于空间结构的形成有较大影响,蛋白质相对分子质量也是由其包含的氨基酸相对分子质量所决定,在其局部构象以及空间排列中起着重要的作用。因此,作者在分析氨基酸理化性质对结构的影响后,提出将相对分子质量、等电点、亲水指数以及电离平衡常数等4个性质作为其理化性质,见表3。上述特征经标准化后作为其强化特征信息,进而提高蛋白质二级结构预测的准确率。

位置特异性得分矩阵^[18]是目前最常用的氨基酸序列特征表示方法,利用蛋白质序列相似性获取可能相似的同源结构,挖掘结构相似信息进而对比现有结构,通过量化氨基酸进化突变可能性作为其特征表示。使用 PSI-BLAST 工具在数据库中对氨基酸序列进行多序列对比,获取其位置特异性矩阵,序列中每个氨基酸由 20 维的向量表示该氨基酸在进化过程中突变为另一种氨基酸的概率得分。PSSM 矩阵具有丰富的氨基酸进化信息,相比氨基酸组成成分更能体现二级结构局部构象的进化过程,在预测蛋白质结构中表现优异。

基于上述 3 种特征提取方式,作者结合多特征融合技术提取氨基酸序列信息,从氨基酸种类到其理化性质,再至全局生物进化信息,将其组合连接为一个特征向量,作为一个氨基酸的特征表示,其编码见图 1。

表3 氨基酸表示及常用理化性质

Table 3 Amino acid representation and commonly used physicochemical properties

氨基酸	英文简写	相对分子质量	亲水指数	等电点	电离平衡常数
丙氨酸	A	89.09	1.8	6.00	2.4
半胱氨酸	C	121.16	2.5	5.07	1.9
天冬氨酸	D	133.10	-3.5	2.77	2.0
谷氨酸	E	147.13	-3.5	3.22	2.1
苯丙氨酸	F	165.19	2.8	5.48	2.2
甘氨酸	G	75.07	-0.4	5.97	2.4
组氨酸	H	155.16	-3.2	7.59	1.8
异亮氨酸	I	131.18	4.5	6.02	2.3
赖氨酸	K	146.19	-3.9	9.74	2.2
亮氨酸	L	131.18	3.8	5.98	2.3
蛋氨酸	M	149.21	1.9	5.74	2.1
天冬酰胺	N	132.12	-3.5	5.41	2.1
脯氨酸	P	115.13	-1.6	6.30	2.0
谷氨酰胺	Q	146.15	-3.5	5.65	2.2
精氨酸	R	174.20	-4.5	10.76	1.8
丝氨酸	S	105.09	-0.8	5.68	2.2
苏氨酸	T	119.12	-0.7	6.16	2.1
缬氨酸	V	117.15	4.2	5.96	2.2
色氨酸	W	204.23	-0.9	5.89	2.4
酪氨酸	Y	181.19	-1.3	5.66	2.2

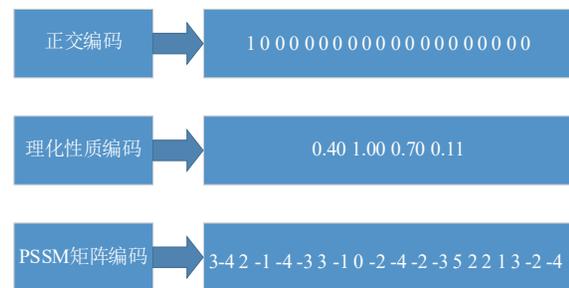


图1 氨基酸特征表示

Fig. 1 Amino acid characterization

氨基酸序列局部相互作用在实际的蛋白质二级结构预测中有着重要影响,利用滑动窗口机制能够有效解决获取局部信息困难的问题。作者研究了不同滑动窗口对预测精度的影响,以每个目标氨基酸为中心,通过在序列上滑动来提取窗口内所包含的氨基酸。对其两端超出序列范围的窗口位置,用零向量代替其特征。若滑动窗口为 15,即一个氨基酸残基由 660 维特征(15×44 (20PSSM+4PP+20AAC))表示。该多特征融合

方式所构建的特征提取模型不仅包含了蛋白质序列的基本信息,也包含了氨基酸的生物特征信息和进化信息,有效提高了蛋白质二级结构预测中的特征提取效率,为后续模型预测奠定了基础。

1.3 结构预测模型

支持向量机(SVM)是基于统计学习的一种高效机器学习方法,能够将样本数据从低维空间映射至高维空间进而线性可分,通过结构风险最小化原则,减少以神经网络算法为代表的易产生的过拟合现象,在相关非线性复杂问题的解决中有较好的应用。在蛋白质结构预测中,基于多特征融合的特征提取增加了模型的计算复杂度。利用最小二乘支持向量机作为基本分类模型,将最小二乘线性系统作为其损失函数,用误差的二范数表示。在LSSVM中,将不等式约束转化为等式约束,原SVM模型中二次规划问题转换成线性方程组的求解问题,优化问题变为:

$$\min J(\omega, e) = \frac{1}{2} \|\omega\|^2 + \frac{1}{2} c \sum_{i=1}^n e_i^2 \quad (1)$$

$$\text{s.t. } y_i = \omega^T \varphi(x_i) + b + e_i, i = 1, 2, \dots, n$$

其中 (x_i, y_i) 代表数据样本集, J 为损失函数, c 为惩罚系数, ω 为权值向量, e_i 为误差变量。

建立拉格朗日方程:

$$L(\omega, b, e, \alpha) = J(\omega, e) - \sum_{i=1}^n \alpha_i [\omega^T \varphi(x_i) + b + e_i - y_i] \quad (2)$$

利用KKT条件可得:

$$\begin{cases} \frac{\partial L}{\partial \omega} = 0 \Rightarrow \omega = \sum_{i=1}^n \alpha_i \varphi(x_i) \\ \frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^n \alpha_i = 0 \\ \frac{\partial L}{\partial e_i} = 0 \Rightarrow \alpha_i = c e_i, i = 1, 2, \dots, n \\ \frac{\partial L}{\partial \alpha_i} = 0 \Rightarrow \omega^T \varphi(x_i) + b + e_i - y_i = 0, i = 1, 2, \dots, n \end{cases} \quad (3)$$

通过消除 ω 和 e , 最终得到分类函数:

$$y(x) = \text{sign} \left[\sum_{i=1}^n \alpha_i y_i K(x, x_i) + b \right] \quad (4)$$

其中 $K(x, x_i)$ 为LSSVM的核函数,该函数为半正定矩阵,直接计算出特征向量映射至高维空间的内积,便于解决因维度过高导致计算能力不足的问题。常用的核函数有以下4种:

$$\text{线性核函数: } K_{\text{Linear}}(X_i, X_j) = x_i^T x_j \quad (5)$$

$$\text{高斯核函数: } K_{\text{Rbf}}(X_i, X_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (6)$$

$$\text{多项式核函数: } K_{\text{Poly}}(X_i, X_j) = (x_i^T x_j + 1)^d \quad (7)$$

$$\text{Sigmoid核函数: } K_{\text{Sigmoid}}(X_i, X_j) = \tanh(\gamma x_i^T x_j + \theta) \quad (8)$$

上述4种核函数根据其作用范围可分为全局核函数和局部核函数,局部核函数在模型中的学习能力较强,但泛化能力较弱,高斯核能够描述局部样本数据的变化,而多项式核函数可利用高等幂函数建立高维映射,对总体样本信息有准确的把握。相关研究证明^[19-20],单一核函数的SVM模型表达能力有限,将全局和局部核函数结合形成混合核函数能有效提高模型的预测准确率及泛化能力。在本研究中,主要建立如下的组合核函数:

$$K_{\text{multiple}}(x_i, x_j) = \sum_{i=1}^N \gamma_i K(x, x_i) \quad (9)$$

$$\sum_{i=1}^N \gamma_i = 1, 0 < \gamma_i < 1, i = 1, 2, \dots, N$$

其中 γ_i 为各独立核函数的权重系数,组合核函数由单核函数线性组合而成,权重之和为1,满足Mercer定理。利用权重大小调整各核函数占比大小,将核函数选择问题转换为核函数权值求解问题^[21]。使用梯度下降法迭代生成最佳权重,主要步骤如下:

步骤一:初始化每个核函数的权重,利用公式(10)初始化权重。

$$\gamma_i = \frac{1}{N} \quad (10)$$

其中 N 为基本核函数个数。

步骤二:使用混合核函数的线性加权公式(11)替换目标函数的核函数。

$$\sum_{i=1}^N \gamma_i K_i(x_i, x_j) \quad (11)$$

计算目标函数。

$$y(x) = \text{sign} \left[\sum_{i=1}^n \alpha_i y_i \sum_{i=1}^N \gamma_i K_i(x_i, x_j) + b \right] \quad (12)$$

步骤三:对 γ_i 进行求导,确定梯度方向与步长。

步骤四:利用梯度更新公式更新 γ_i 。

$$\gamma_{i+1} = \gamma_i + r_i D_i \quad (13)$$

其中 r_i 为最优步长, D_i 为梯度方向。

步骤五:若满足终止条件,则停止运算,若不满足则重复步骤二至步骤四。

1.4 优化算法

上述多核 LSSVM 模型建立后,模型优化一直是构建最优模型的重要环节,通过调整超参数的值可以提高模型的学习及泛化能力,在相关谷物蛋白质结构中提高预测性能。粒子群优化算法(PSO)是一种常用的全局优化进化算法。通过模拟鸟群捕食行为,所有目标问题的可能解均处于 D 维空间中,每个解称之为“粒子”。根据鸟类间的合作与竞争关系,将粒子运动表现出鸟类觅食的特性。粒子当前所处位置的好坏由适应度函数确定,可以得出其适应值(fitness value)。

在上述多核模型中,惩罚因子 c 表示误差的容忍度, c 越低,易欠拟合, c 越高,易过拟合,导致泛化能力变差。除该参数外,核参数也是模型的重要参数,如多项式核中的 degree 项,高斯核与 Sigmoid 核中的 gamma 项等,均对模型训练产生重要影响。使用 PSO 算法寻找最佳参数组合,将 MKLSSVM 模型的分类准确率作为适应度函数,在进行适应度评价时,准确率越高则粒子位置越优异。设粒子群规模为 m , D 维空间的粒子位置可由向量 $X_i = (X_{i1}, X_{i2}, \dots, X_{iD})$, $i=1,2,3, \dots, m$, 每个粒子的运行速度为 $V_i = (V_{i1}, V_{i2}, \dots, V_{iD})$, 通过目标函数确定各粒子所处的最优位置 P_{best} 和整个粒子群所能搜索到的最优位置 g_{best} , 通过式(14)更新粒子的速度和位置。

$$\begin{cases} V_{i+1} = wV_i + c_1r_1(p_{besti} - x_i) + c_2r_2(g_{besti} - x_i) \\ x_{i+1} = x_i + v_{i+1} \end{cases} \quad (14)$$

其中 w 为惯性权重参数; r_1 和 r_2 为 0~1 的随机数, c_1 和 c_2 为学习因子。

PSO 算法能够解决大部分全局最优解问题,并有结构简单、收敛速度快等优点。在对目标问题的优化过程中,复杂参数较少,相较于其他智能优化算法来说优点突出^[21]。

试验流程见图 2。

1.5 评价标准

Q_3 准确率是目前蛋白质二级结构预测中最常用的评价指标。 Q_3 准确率为正确预测的氨基酸数占所有氨基酸的比例,见式(15)。

$$Q_3 = \frac{Q_H + Q_E + Q_C}{S} \times 100\% \quad (15)$$

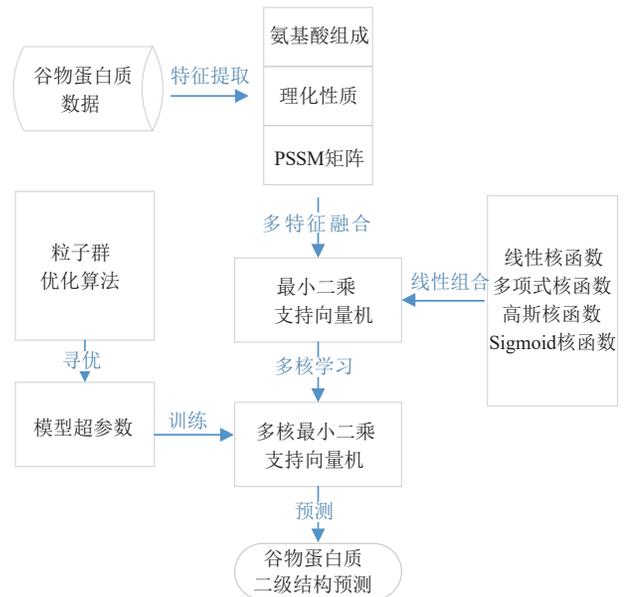


图 2 试验流程

Fig. 2 Experimental process

从残基层面的角度评价, Q_H 为准确预测 H 类蛋白质结构数目, Q_E 为准确预测 E 类蛋白质结构数目, Q_C 为准确预测 C 类蛋白质结构数目, S 为总的氨基酸数。

2 结果与分析

2.1 LSSVM 中不同特征提取方式及核函数对谷物蛋白质二级结构的预测

利用上述标准对 LSSVM 算法预测谷物蛋白质二级结构进行评价,基于氨基酸组成成分逐步融合其他特征,作者尝试使用 4 种不同的核函数作为 LSSVM 的核函数进行结构预测,使用五折交叉验证,结果见表 4。

将 LSSVM 作为基本预测模型时,在对 3 种谷物蛋白质二级结构预测中,将氨基酸组成成分作为基本特征,将理化性质和位置特异性矩阵作为特征融合到氨基酸序列特征提取中时, Q_3 准确率逐步提升,尤其在加入 PSSM 矩阵后, Q_3 准确率大幅提升,表明 PSSM 矩阵中包含的进化信息对预测蛋白质结构非常重要。同时证明本研究使用的多特征融合提取方式能够更加全面地提取序列信息,提高二级结构的预测准确率,见图 3。

在核函数选择中,4 种核函数分别是线性核函数、多项式核函数、高斯核函数和 Sigmoid 核函数。高斯核函数是典型的局部核函数,能够有效提取

表 4 不同特征提取方式及核函数的谷物蛋白质二级结构预测结果

Table 4 Prediction results of grain protein secondary structure based on different feature extraction methods and kernel functions

数据	核函数	Q_3 准确率/%		
		AAC	AAC+PP	AAC+PP+PSSM
玉米 (<i>Zea mays</i>)	Liner	60.65	62.34	67.38
	Poly	61.72	64.56	70.34
	RBF	63.45	65.47	72.42
	Sigmoid	61.02	63.28	71.12
小麦 (<i>Triticum aestivum</i>)	Liner	59.72	61.05	66.71
	Poly	60.99	62.89	69.78
	RBF	62.83	64.46	71.48
	Sigmoid	61.78	62.82	70.05
大豆 (<i>Glycine</i>)	Liner	60.23	62.09	65.21
	Poly	61.97	63.29	69.32
	RBF	63.04	64.80	70.89
	Sigmoid	62.03	62.41	69.96

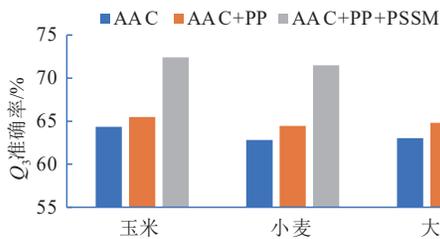


图 3 基于 RBF 核函数不同编码方式的蛋白质二级结构预测结果

Fig. 3 Prediction results of protein secondary structure based on different coding modes in RBF kernel function

局部特征信息,刻画样本的局部特性,在多特征融合中有着非常好的表现。其余 3 种为全局核函数,相较局部核函数插值能力较弱,善于提取样本的全局特性。由表 4 可以看出,多项式核函数和 Sigmoid 核函数比线性核函数在二级结构预测中表现良好。

分析 3 种谷物蛋白质在 LSSVM 的表现可知,3 种特征相结合的方法比单个特征更能表现序列特征信息,核函数中高斯核函数比其他 3 种核函数更适用于谷物蛋白质二级结构预测。但在 LSSVM 模型中,利用线性加权将单一核函数组合形成新的混合核函数,结合不同核函数的特性,使其在建模中增强了适用性和解释性,提升模型性能。

2.2 LSSVM 中混合核函数对谷物蛋白质二级结构的预测

由上述试验可知,多项式核函数和 Sigmoid 核函数在混合核函数中只占少量权重,主要在特征向量的高维映射中提供局部信息。根据上述核函数的预测效果,选择表现较好的 Poly 核函数、Rbf 核函数和 Sigmoid 核函数,尝试上述核函数的线性加权组合,利用各核函数的性质及优势,寻找适合上述特征提取方式的最优核函数。利用滑动窗口机制^[20]可以有效提取氨基酸局部信息,通过窗口大小调整特征维度,一定程度上可以提高预测精度。

由表 5 可以发现,将基本核函数的加权线性组合作为 LSSVM 的核函数,对模型预测有重要的影响。在 3 种谷物蛋白质二级结构预测中, Poly、Rbf 与 Sigmoid 核函数的加权线性组合效果最优,整体 Q_3 准确率达到 80% 以上。而两两组合的混合核函数预测效果欠佳, Poly 核函数与 Sigmoid 核函数的线性加权组合效果提高不明显,表明在谷物蛋白质特征提取中,其选择的多特征融合向量在该混合核函数的高维空间映射不能很好分割。高斯核函数在多特征融合的特征提取方式中表现优异。其在样本映射至高维空间时,能够有效处理特征与标签之间的非线性关系,相较于多项式核函数参数较少,能够减少因参数个数及取值对模型精度和复杂度的影响。图 4 也证明本研究构建的高斯核函数、多项式核函数与 Sigmoid 核函数线性加权形成的混合核函数,能够综合基本核函数的优点,具备更强的学习和泛化能力。

表 6 中的权重系数也证明在多特征提取氨基酸序列信息中,混合核矩阵中高斯核矩阵占较高权重参与训练,而多项式核矩阵和 Sigmoid 核矩阵只占有少量权重,在特征向量高维映射中提供局部作用。可以看出,融合多个不同的核矩阵能够提高 LSSVM 的训练效果,通过梯度下降法迭代寻找多核分类器中的最优权重,利用核权重调整核矩阵占比,提高分类器的性能。

对于小麦和大豆来说,滑动窗口为 15 时,预测效果最佳;而对于玉米而言,滑动窗口为 13 时,预测效果最佳。因此,为了进一步提高预测性能,进行谷物蛋白质序列预测时可调整滑动窗口的大小。

表 5 不同核函数组合与滑动窗口对谷物蛋白质二级结构预测的影响

Table 5 The influence of different kernel function combinations and sliding windows on the prediction of grain protein secondary structure

谷物	核函数组合	Q_3 准确率/%					
		7	9	11	13	15	17
玉米(<i>Zea mays</i>)	Poly+RBF	76.14	78.05	80.53	81.23	81.14	81.06
	Poly+Sigmoid	75.42	77.25	78.89	79.93	79.42	78.49
	Rbf+Sigmoid	78.93	80.49	81.71	82.16	82.14	81.85
	Poly+RBF+Sigmoid	79.21	81.23	82.96	83.78	83.24	82.13
小麦(<i>Triticum aestivum</i>)	Poly+RBF	77.28	78.23	78.93	79.49	80.83	80.24
	Poly+Sigmoid	76.46	77.24	77.83	78.20	78.46	78.14
	RBF+Sigmoid	77.83	78.44	79.09	80.51	80.91	80.16
	Poly+RBF+Sigmoid	78.17	79.42	80.26	81.49	82.47	82.12
大豆(<i>Glycine</i>)	Poly+RBF	77.22	78.29	79.12	79.67	80.07	79.79
	Poly+Sigmoid	76.29	77.13	77.86	78.66	79.45	79.21
	RBF+Sigmoid	77.38	78.53	79.58	80.15	80.73	80.29
	Poly+RBF+Sigmoid	78.66	79.53	80.27	81.18	81.68	81.39

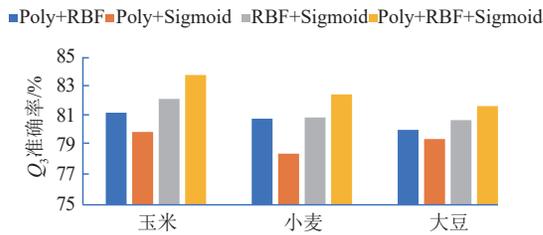


图 4 最佳窗口下不同组合核函数的蛋白质结构预测结果

Fig. 4 Protein structure prediction results of different combination kernel functions under the optimal window

表 6 最佳多核 LSSVM 模型核函数权重系数

Table 6 Multi-kernel LSSVM model kernel function weight coefficient

谷物	权重系数		
	RBF	Poly	Sigmoid
玉米(<i>Zea mays</i>)	0.643	0.125	0.232
小麦(<i>Triticum aestivum</i>)	0.779	0.058	0.163
大豆(<i>Glycine</i>)	0.728	0.093	0.179

表 5 数据说明随着滑动窗口大小的增加,单个氨基酸的特征向量维数增加,其包含的局部信息越广泛,预测效果越佳,但其提高程度逐渐下降,达到某个阈值后, Q_3 准确率达到最高,继续增加窗口大小无法提高预测的准确率,说明在预测中若特征向量维度无限制增加,无法取得理想的预测效果。

2.3 不同模型对谷物蛋白质二级结构的预测

在上述构建的多核 LSSVM 模型中,存在众多超参数,利用粒子群或其变种算法能够有效寻找支持向量机的最佳参数^[22]。作者利用 PSO 算法对该模型进行优化,寻找最优的超参数组合,优化预测模型,结果见图 5。

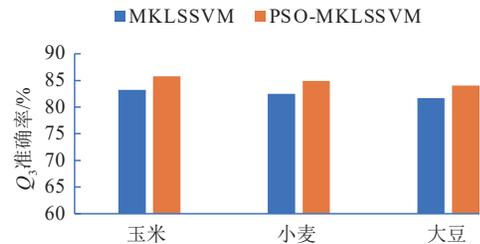


图 5 PSO 算法优化 MKLSSVM 模型前后结果对比

Fig. 5 Comparison of results before and after optimization of MKLSSVM model by PSO algorithm

可以看出,PSO 优化后的多核 LSSVM 模型比原模型表现优秀,在多核 LSSVM 模型中,存在众多超参数,如惩罚系数 C 、各核函数参数等,对其模型的泛化能力有其重要的影响。利用 PSO 智能优化算法将 Q_3 准确率作为适应度函数,寻找最佳超参数组合。试验证明利用该优化算法能够有效提高模型性能,提高预测准确率。

目前存在众多蛋白质二级结构预测在线服务器,为了验证本算法的有效性,作者选择 Jpred4、

Psipred 和 RaptorX 共 3 种服务器对谷物蛋白质二级结构进行预测,利用 Q_3 指标对模型预测效果进行评估,见表 7。

表 7 本研究模型与其他预测算法对比结果

Table 7 Comparison results of the model in this paper and other prediction algorithms

模型	Q_3 准确率/%		
	玉米	小麦	大豆
Jpred4	79.28	80.12	79.47
Psipred	80.05	80.79	82.51
RaptorX	84.43	83.17	84.53
本研究算法	85.76	84.91	84.04

结果表明,使用多核学习方法优化基本核分类器能够提高分类效果,针对谷物蛋白质二级结构,多特征融合向量与每个核相关联,集成不同内核的加权组合,有效提高了准确性。同时与其他扩展算法相比, Q_3 准确率也高于或接近其他模型的预测结果,融合相关特征信息并使用多核学习方法能够显著提高模型的分类性能。

3 结论

在谷物蛋白质二级结构预测中,针对氨基酸序列的特征提取方式,将氨基酸成分作为基本特征,继而加入氨基酸理化性质特征,最后融合 PSSM 矩阵作为最终的特征表示。随着相关特征逐渐融合,二级结构预测效果更好,表明多特征融合提取方式比单一特征提取方式能够更加全面捕获氨基酸序列信息。

在谷物蛋白质二级结构预测模型中,作者基于 LSSVM 基本分类模型,提出多核学习方法优化模型性能,构建谷物蛋白质二级结构预测模型。研究结果表明,在预测模型中多核 LSSVM 比单核 LSSVM 在二级结构预测中更有优势,其中将 Poly、RBF 和 Sigmoid 等 3 种核函数构成的混合核函数作为最终核函数效果最好,结合滑动窗口机制和 PSO 算法优化模型后,3 种谷物(玉米、小麦、大豆)的蛋白质二级结构预测 Q_3 准确率分别达到了 85.76%、84.91% 和 84.04%,比相关预测模型具有更大的优势。

参考文献

[1] 谭斌,谭洪卓,刘明,等. 粮食(全谷物)的营养与健康[J].

中国粮油学报,2010,25(4):100-107.

TAN B, TAN H Z, LIU M, et al. The grain, the wholegrain: nutrition and health benefits[J]. Journal of the Chinese Cereals and Oils Association, 2010, 25(4): 100-107. (in Chinese)

[2] 屈凌波. 谷物营养与全谷物食品的研究开发[J]. 粮食与食品工业,2011,18(5):7-9.

QU L B. Grain nutrition and research and development of whole grain food[J]. Cereal & Food Industry, 2011, 18(5): 7-9. (in Chinese)

[3] 鲍王璐,孟婷婷,佟恩杰,等. 整粒小麦加工全麦脆片前后营养成分的变化[J]. 食品与生物技术学报,2020,39(7):67-73.

BAO W L, MENG T T, TONG E J, et al. Changes of nutrients in whole wheat grains before and after processed into whole wheat crisps[J]. Journal of Food Science and Biotechnology, 2020, 39(7):67-73. (in Chinese)

[4] MERLINO A, PICONE D, ERCOLE C, et al. Chain termini cross-talk in the swapping process of bovine pancreatic ribonuclease[J]. Biochimie, 2012, 94(5): 1108-1118.

[5] 杨淑雅,管彤,陆奕成,等. 氨基酸均衡谷物营养粉的开发与研究[J]. 食品与发酵科技,2021,57(5):42-48.

YANG S Y, GUAN T, LU Y C, et al. Development and research of special dietary cereal nutritious powder with balanced amino acid[J]. Food and Fermentation Sciences & Technology, 2021, 57(5):42-48. (in Chinese)

[6] ZHI X W. The current situation and prospect of protein structure prediction[J]. Chemistry of Life, 1998, 18(6):19-22.

[7] FRISHMAN D, ARGOS P. Knowledge-based protein secondary structure assignment[J]. Proteins: Structure, Function, and Bioinformatics, 1995, 23(4):566-579.

[8] 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.

ZHANG H X, TANG H W, ZHANG L Z, et al. Evaluation on prediction methods of protein secondary structure[J]. Computers and Applied Chemistry, 2003, 20(6): 735-740. (in Chinese)

[9] CHOU P Y, FASMAN G D. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins[J]. Biochemistry, 1974, 13(2):211-222.

[10] KLOCZKOWSKI A, TING K L, JERNIGAN R L, et al. Protein secondary structure prediction based on the GOR algorithm incorporating multiple sequence alignment information[J]. Polymer, 2002, 43(2):441-449.

[11] 张利彪. 基于粒子群和微分进化的优化算法研究[D]. 长春:吉林大学,2007.

- [12] BERRAR D C. Encyclopedia of bioinformatics and computational biology[M]. Cambridge: Academic Press, 2019.
- [13] CHEN C, CHEN L X, ZOU X Y, et al. Predicting protein structural class based on multi-features fusion[J]. Journal of Theoretical Biology, 2008, 253(2): 388-392.
- [14] 梁珩琳. 基于集成学习的蛋白质二级结构预测研究[D]. 广州: 华南理工大学, 2020.
- [15] 朱臣臣, 赵熙强. 基于氨基酸的理化性质和位置信息的蛋白质序列相似性分析方法[J]. 中国海洋大学学报(自然科学版), 2021, 51(S1): 95-100.
- ZHU C C, ZHAO X Q. Protein sequence similarity analysis method based on physical and chemical properties and position information of amino acids[J]. Periodical of Ocean University of China, 2021, 51(S1): 95-100. (in Chinese)
- [16] 刘静, 崔双龙, 曹洪伟, 等. MIMLRBF 预测谷物蛋白质功能方法的改进[J]. 食品与生物技术学报, 2021, 40(4): 36-43.
- LIU J, CUI S L, CAO H W, et al. Improvement of MIMLRBF algorithm for predicting function of grain proteins[J]. Journal of Food Science and Biotechnology, 2021, 40(4): 36-43. (in Chinese)
- [17] 朱永春, 张宝发, 李玉侠, 等. 氨基酸等电点分离方法的回收率与分离度[J]. 大学化学, 1997, 12(4): 51-52.
- ZHU Y C, ZHANG B F, LI Y X, et al. Recovery and resolution of isoelectric point separation method for amino acids[J]. University Chemistry, 1997, 12(4): 51-52. (in Chinese)
- [18] JONES D C. Protein secondary structure prediction based on position-specific scoring matrices[J]. Journal of Molecular Biology, 1999, 292(2): 195-202.
- [19] 刘斌, 温雪岩. 优化多核 SVM 的蛋白质二级结构预测[J]. 现代电子技术, 2020, 43(8): 139-142.
- LIU B, WEN X Y. Protein secondary structure prediction based on optimized multi-kernel SVM[J]. Modern Electronics Technique, 2020, 43(8): 139-142. (in Chinese)
- [20] REDDY K S S, BINDU C, STREAM S W. A density-based approach for clustering data streams over sliding windows[J]. Measurement, 2019, 144: 14-19.
- [21] 程国建, 郭瑞华. PSO-LSSVM 分类模型在岩性识别中的应用[J]. 西安石油大学学报(自然科学版), 2010, 25(1): 96-99.
- CHENG G J, GUO R H. Application of PSO-LSSVM classification model in logging lithology recognition[J]. Journal of Xi'an Shiyou University (Natural Science Edition), 2010, 25(1): 96-99. (in Chinese)
- [22] GUAN X, LIU J, HUANG Q R, et al. Assessing the freshness of meat by using quantum-behaved particle swarm optimization and support vector machine[J]. Journal of Food Protection, 2013, 76(11): 1916-1922.

(责任编辑:李春丽)