

MIMLRBF 预测谷物蛋白质功能方法的改进

刘 静¹, 崔双龙¹, 曹洪伟², 管 骁^{*2}

(1. 上海海事大学 信息工程学院, 上海 201306; 2. 上海理工大学 医疗器械与食品学院, 上海 200093)

摘要: 随着人们对营养与保健功能的关注, 谷物蛋白质功能预测已经成为当前研究热点。面对大量已完成测序的谷物蛋白质基因组数据, 利用计算方法来预测谷物蛋白质功能已经成为主流。从谷物蛋白质结构域序列出发, 首次将 MIMLRBF 算法运用到蛋白质功能预测, 并在此算法基础上, 提出了多种改进后的谷物蛋白质功能预测模型。其中, 针对平均 Hausdorff 距离削弱了两种蛋白质之间最短结构域距离所起作用的问题, 为平均 Hausdorff 距离引入一个自动调节系数来计算蛋白质之间的相似性。同时, 为提高 MIMLRBF 算法的预测效果, 利用改进后的混合径向基核函数进行激活, 得到了改进后的 MIMLRBF 算法模型。最终利用主流的评价标准对预测结果进行评价, 可以发现改进后的 MIMLRBF 比传统的预测效果更好, 证明了所建模型的优越性。

关键词: 谷物; 蛋白质功能预测; 多示例多标记; Hausdorff 距离; 核函数

中图分类号: Q 816 文章编号: 1673-1689(2021)04-0036-08 DOI: 10.3969/j.issn. 1673-1689.2021.04.005

Improvement of MIMLRBF Algorithm for Predicting Function of Grain Proteins

LIU Jing¹, CUI Shuanglong¹, CAO Hongwei², GUAN Xiao^{*2}

(1. College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China; 2. School of Medical instrument and Food engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: With people's attention to nutrition and health care functions, the prediction of grain protein function has become a research hotspot. Faced with large amounts of sequenced cereal protein genome data, the use of computational methods to predict grain protein function has become the mainstream. For the first time, the MIMLRBF algorithm was applied to protein function prediction from the grain protein domain sequence. Based on this algorithm, several improved grain protein function prediction models were proposed. Among them, an automatic adjustment coefficient for the average Hausdorff distance was introduced to calculate the similarity between proteins aiming to solve the problem that the average Hausdorff distance weakened the shortest domain distance between two proteins. At the same time, in order to improve the prediction effect, the improved MIMLRBF algorithm model was obtained using the improved hybrid radial basis kernel function to

收稿日期: 2019-12-26

基金项目: 国家自然科学基金项目(31701515); 上海市曙光计划项目(19SG45); 上海市国内科技合作项目(19395800200)。

作者简介: 刘 静(1979—), 女, 博士, 副教授, 主要从事生物信息、信息技术与食品功能交叉领域的研究。E-mail: jingliu@shmtu.edu.cn

* 通信作者: 管 骁(1979—), 男, 教授, 博士研究生导师, 主要从事谷物加工与营养研究。E-mail: gnxo@163.com

activate. Finally, the improved MIMLRBF was confirmed to be better than the traditional prediction based on the evaluation of prediction results using mainstream evaluation criteria, which proved the superiority of the proposed model.

Keywords: grain, protein function prediction, multi-instance multi-label, Hausdorff distance, kernel function

谷物食品是人类营养摄入的主要来源,在我们的膳食中有举足轻重的地位^[1]。在我国,谷物品种主要包括:大麦、水稻、高粱、玉米、燕麦等,它们都是良好的主食食品原料^[2]。随着人们对营养与保健功能的关注,谷物蛋白质功能预测已经成为当前研究的热点问题之一^[3]。面对大量的已完成测序的谷物蛋白质数据,其所对应的生物学功能仍然未知。采用传统的手工注释费时费力,已经无法满足需求,而以计算方法来预测蛋白质的生物学功能已经渐渐成为了主流。

近年来,国内外对运用计算智能技术解决此类问题已经建立了较为有效的方法。2007年,Zhou等^[4]最早提出了多示例多标记(multi-instance multi-label learning, MIML)框架,随后又基于退化的思想提出了 MIMLBOOST 和 MIMLSVM^[5]算法应用于图像场景的识别。在生物信息学方面,Li等^[6]提出了 MIMLSVM⁺算法能够很好地解决果蝇基因表达模式注释的问题。Wu等^[7]在 MIMLNN 算法的基础上,提出了 En-MIMLNN 算法用于全基因组的蛋白质功能预测,通过在 En-MIMLNN 算法中加入径向基函数来激活神经网络,使得其在主流的评价标准上预测结果有显著提升。Zhang等^[8]也提出了一种基于 RBF 神经网络的 MIMLRBF 算法应用于图像场景的识别,该方法选择平均 Hausdorff 距离度量图像之间的相似性,从而在图像场景识别的应用中优于传统方法。在多示例多标记框架的基础上,衍生出了多种方法应用到各个不同的领域中,同时也在各种算法的基础上,主要通过对 Hausdorff 距离进行改进来度量各种场景中物体的相似性以及采用不同的径向基函数进行激活来提升算法的预测结果。

在本文中,首次将 MIMLRBF 算法应用于谷物

蛋白质功能预测,并在此基础上,提出了多种改进的谷物蛋白质功能预测模型。首先,针对平均 Hausdorff 距离削弱了两种蛋白质之间最短结构域距离所起作用的问题,为平均 Hausdorff 距离引入一个自动调节系数^[9]。同时,为提高 MIMLRBF 算法的预测效果,采用改进后的混合径向基核函数进行激活。通过所搭建模型分别在大麦、水稻两种谷物蛋白质数据上进行蛋白质功能预测,并使用三种主流的评价标准进行度量,可以发现,改进后的 MIMLRBF 算法比传统的预测效果更好。再将作者提出的改进算法与传统的 MIML 算法(En-MIMLNN、MIMLSVM)进行比较,可以发现,作者提出的改进后的算法在 Hamming Loss、Macro-F1、Micro-F1 这 3 项度量指标下具有较好的预测结果。

1 蛋白质功能预测原理

在 1961 年,Anfinsen^[10]提出蛋白质序列、结构、功能之间存在某种相互作用关系,即通过蛋白质氨基酸序列特征可以决定其三维结构特征,同时在结构特征的基础上,可以确定所具有的某种功能。蛋白质功能预测在此理论的基础上,不断地发展和完善,提出了一系列蛋白质功能预测的方法。

在蛋白质中,其结构和功能与结构域密切相关^[11],通过对氨基酸序列信息进行特征提取,所得到的由结构域信息所组成的特征向量与蛋白质的功能之间存在某种映射关系。因此可以将蛋白质功能预测转化成一个多示例多标记问题,即一个蛋白质样本往往可以由多个结构域(示例)进行表示,同时具有多种基因本体(GO)生物学功能(标记)。其预测原理见图 1。

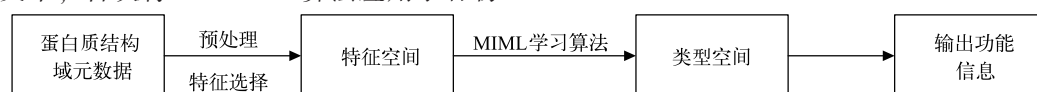


图 1 蛋白质功能预测原理

Fig. 1 Principles of protein function prediction

2 实验方法

2.1 RBF 神经网络的建立

RBF(径向基函数)神经网络^[12]是一种基于径向基函数的多层前向神经网络模型,其网络结构可以看成是输入层空间到隐含层空间的一种非线性映射,以及隐含层空间到输出层的线性映射。从其网络结构(见图2)可以看出,隐含层神经元数目、RBF的中心、RBF的宽度和隐含层与输出层之间的权值矩阵等,在RBF神经网络中都是非常重要的学习参数。

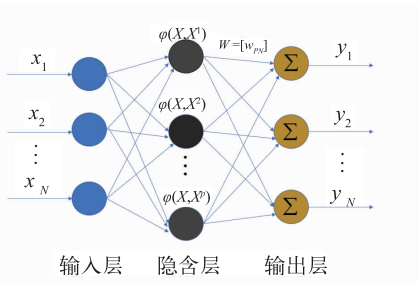


图2 RBF神经网络结构图

Fig. 2 RBF neural network structure diagram

而在隐含层中,通常采用径向基核函数进行激活,常用的径向基核函数有:

高斯(Gauss)核函数:

$$\varphi(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\delta^2}\right) \quad (1)$$

多元二次核函数:

$$\varphi(\mathbf{x}, \mathbf{y}) = \sqrt{\|\mathbf{x} - \mathbf{y}\|^2 + \delta^2} \quad (2)$$

逆多元二次核函数:

$$\varphi(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{\|\mathbf{x} - \mathbf{y}\|^2 + \delta^2}} \quad (3)$$

其中, δ 为核函数中围绕中心点的宽度, δ 越小, 则核函数宽度越大, 函数的选择性就越大。

2.2 基于 RBF 神经网络 MIML 算法

MIMLRBF (multi-instance multi-label radial basis function) 神经网络是从传统 RBF 神经网络派生而来的, 用于解决多示例多标记问题的一种神经网络的方法。而蛋白质功能预测恰恰可以转化为一个多示例多标记问题。通过从已知蛋白质氨基酸序列信息出发, 经过特征提取, 将一个蛋白质的结构由所提取多个结构域(示例)信息表示, 然后和已知蛋白质中多个 GO 生物学功能(标记)进行关联。

MIMLRBF 算法^[8]使用双层架构进行训练。在第一层中, 通过结合包间平均 Hausdorff 距离, 利用 k-Medoids 算法^[13]将训练样本进行聚类, 保留每个聚类簇的中心点。在第二层中, 通过径向基函数神经网络计算样本和中心点之间的基函数, 然后利用最小化平方和误差函数所获取的一、二层之间的权重乘以基函数 $\varphi(\cdot)$, 从而得到模型的输出, 其中输出值大于等于零的标记设为正标记, 否则为负标记。

令 $S = \{(\mathbf{X}_i, \mathbf{Y}_i) | 1 \leq i \leq N\}$ 为训练样本集, 其中 N 为蛋白质样本的数目, $\mathbf{X}_i$ 为样本集中的第 i 个蛋白质, 用多个结构域转化而成的特征向量所表示, 也就是一个包含多个示例的包 (bag), $\mathbf{Y}_i$ 表示对应于 $\mathbf{X}_i$ 的生物学功能, 由其基因本体生物学功能方面的 GO 编号表示, 也就是一个含有多个标记的集合。用 $U_l = \{\mathbf{X}_i | (\mathbf{X}_i, \mathbf{Y}_i) \in S, l \in \mathbf{Y}_i\}$ 表示拥有第 l 个标记的蛋白质的集合, 其中 S 为所有 U_l 的合集, 其标记总数为 m 。

为了测算蛋白质之间的相似性, 通过计算由包组成的蛋白质之间的平均 Hausdorff 距离来表示。即对包 $A = \{a_1, a_2, \dots, a_{n_1}\}$ 和包 $B = \{b_1, b_2, \dots, b_{n_2}\}$, 两包之间的平均 Hausdorff 距离^[14]为:

$$\text{Havg}(A, B) = \frac{\sum_{a \in A} \min_{b \in B} \|a - b\| + \sum_{b \in B} \min_{a \in A} \|b - a\|}{|A| + |B|} \quad (4)$$

其中 $| \cdot |$ 为集合中的元素数目, $\| \cdot \|$ 为结构域(示例)之间的欧氏距离 (Euclidean distance)。

对于集合 U_l , 利用 k-Medoids 算法将其聚类为 M_l 个互斥的类簇 $G_j^l (1 \leq j \leq M_l)$, 其中 $M_l = \alpha |U_l|$, α 为分数参数。在第一层中总包数为 $M = \sum_{l=1}^m M_l$ 。

而每个类簇 G_j^l 中心点 C_j^l 定义如下:

$$C_j^l = \arg \min_{A \in G_j^l} \sum_{B \in G_j^l} \text{Havg}(A, B) \quad (1 \leq l \leq m, 1 \leq j \leq M_l) \quad (5)$$

接着在第二层中, 对于 $\mathbf{X}_i$ 的第 j 个径向基核函数的激活定义如下:

$$\varphi_j^l(\mathbf{X}_i) = \exp\left(-\frac{H_{\text{avg}}(\mathbf{X}_i, C_j^l)^2}{2\delta_j^2}\right) \quad (1 \leq i \leq N, 1 \leq l \leq m, 1 \leq j \leq M_l) \quad (6)$$

其中额外的核函数 $\varphi_0(\mathbf{X}_i)$ 设定初始值为 1。而对于所有的 δ_j 取相同的标准偏差 δ , 其为每对中心点包

间平均 Hausdorff 距离的平均值的 u 倍。

$$\delta = u \times \frac{\sum_{k=1}^{M-1} \sum_{q=k+1}^M H_{\text{avg}}(C_k, C_q)}{M(M-1)/2} \quad (7)$$

其中 u 为缩放系数。

接着通过两层之间的权值矩阵 $W=[w_j^l]$, 使用式 (8) 结合径向基核函数得到输出:

$$y_l(X_i) = \sum_{j=0}^{M_l} w_j^l \varphi_j(X_i) \quad (8)$$

W 通过最小化平方和误差函数来优化:

$$E = \frac{1}{2} \sum_{i=1}^N \sum_{l=1}^m (y_l(X_i) - t_l^i)^2 \quad (9)$$

其中 t_l^i 是样本 X_i 在对于第 l 个功能的标记 (ground-truth), 若 $l \in Y_i$ 取 +1, 否则取 -1。

于是, 对于一个样本 X , 其最终预测结果 Y 表示为:

$$Y = \left\{ l | y_l(X) = \sum_{j=0}^M w_j^l \varphi_j(X) > 0, l \in y \right\} \quad (10)$$

即对于一个给定的蛋白质, 首先对氨基酸序列信息进行特征提取, 得到由结构域信息所组成的特征向量; 再通过 MIMLRBF 算法模型进行分析, 最终可以预测其可能拥有的 GO 生物学功能。

2.3 基于 RBF 神经网络的 MIML 算法改进

2.3.1 包间距离的度量方式及改进 在本文中, 蛋白质之间的相似性转化为求解两种蛋白质包与包之间的距离, 而选取不同的距离公式所计算出的距离将直接决定整个算法的学习性能的好坏。在以往的研究中, Wu 等^[7]在 MIMLNN 算法的基础上, 通过利用集成的 Hausdorff 距离进行改进, 提出了 En-MIMLNN 算法用于预测全基因组的蛋白质功能。而集成的 Hausdorff 距离就是以包之间 3 种 Hausdorff 距离 (最大 Hausdorff^[15]、最小 Hausdorff^[15] 和平均 Hausdorff^[14]) 的平均值作为包间距离。其中最大 Hausdorff 距离对外围的噪声点比较敏感, 而最小 Hausdorff 距离不会受到噪声点的影响, 但是只考虑了最近结构域之间的距离。平均 Hausdorff 距离对最大 Hausdorff 距离和最小 Hausdorff 距离进行了修正, 充分考虑了噪声点以及最近结构域之间的距离的影响。在本文中, MIMLRBF 算法求解包之间的距离就是采用了平均 Hausdorff 距离。

然而, 从公式 (4) 中可以看出, 平均 Hausdorff 距

离主要是一个蛋白质中每个结构域与其它蛋白质中最近的结构域的平均值。而一个蛋白质有可能由多个结构域所组成, 个别较远的结构域可能增大了两种蛋白质之间的包间距离, 从而削弱了两种蛋白质之间最近结构域的距离。在本文中, 在原有的平均 Hausdorff 距离上进行改进, 通过引用一个自动调节参数^[9], 将两种蛋白质之间所有结构域间的最短距离考虑进来, 其定义如下:

$$W = 1 - \exp \left(-\lambda * \min_{a \in A} \min_{b \in B} \|a - b\| \right) \quad (11)$$

其中, $\min_{a \in A} \min_{b \in B} \|a - b\|$ 为蛋白质之间最短结构域间距离, $1 \leq \lambda \leq 5$ 。 W 可以随两种蛋白质之间最短结构域距离变化而变化。

则蛋白质之间的相似性通过蛋白质包 A 与包 B 之间距离求解, 其定义如下:

$$H(A, B) = \frac{\sum_{a \in A} \min_{b \in B} \|a - b\| + \sum_{b \in B} \min_{a \in A} \|b - a\|}{|A| + |B|} \times W \quad (12)$$

即两种蛋白质之间最近结构域间的距离越小, 则 W 值越小, 蛋白质包间距离也就相对较小。

2.3.2 径向基核函数的改进 在本文中, MIMLRBF 算法模型中隐含层激活函数通常采用径向基核函数, 而不同的核函数的内推和外推能力不同。因此, 选取不同的核函数, 所搭建的算法模型的学习能力各有优劣。而为了得到一个更具有泛化能力的核函数, 可以将不同的核函数进行混合得到新的核函数, 即:

$$\phi(x, y) = \sum_{i=1}^n a_i \varphi_i(x, y) \quad (13)$$

其中 a_i 为各个核函数在混合核函数中所占的权重, n 为核函数的数量。

当选取的单个核函数 φ_i 满足文献[16]提出的函数逼近定理, 则混合核函数 $\phi(x, y)$ 一定满足, 因此, 可以用新的混合核函数在 MIMLRBF 算法模型隐含层中进行激活。而且混合核函数 $\phi(x, y)$ 是由多个核函数 φ_i 所张成的空间, 所以其函数总体的逼近能力显然要强于其中任一单个核函数的逼近能力。

在当前的研究中, 核函数的类型有许多, 但是归结起来, 可以分为全局性核函数与局部性核函数。在本文中, 通过选取高斯核函数 (局部性核函数) 与多元二次核函数 (全局性核函数) 进行混合^[17], 得到新的混合径向基核函数 $\phi(x, y)$, 即:

$$\phi(\mathbf{x}, \mathbf{y}) = a \cdot \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\delta^2}\right) + (1-a) \sqrt{\|\mathbf{x} - \mathbf{y}\|^2 + \delta^2} \quad (14)$$

其中: a 为权重系数, $0 \leq a \leq 1$ 。

将新的混合核函数代替原有的 MMLRBF 算法中的核函数进行激活, 通过所搭建的模型对蛋白质功能进行预测。

3 实验与结果

3.1 实验配置

在本文中, 从 UniProt 蛋白质生物数据库中获取大麦 (*Hordeum vulgare*)、水稻 (*Oryza sativa* subsp. *indica*) 两种谷物蛋白质数据集。对于每个生物体, 通过表 1 关键词进行检索, 分别从 UniProtKB/Swiss-Prot (包含检查过的、手工注释的条目) 和 UniProtKB/TrEMBL (包含未校验的、自动注释的条目) 中得到蛋白质氨基酸序列元数据以及对应的基因本体 (GO) 生物学功能^[18] (包括分子功能、生物学过程、细胞组分) 信息 (2018 年 2 月发布)。其中, 通过关键词“molecular function”限定, 使得所检索出的蛋白质必然包含分子功能方面相关信息。

然后, 通过对所获取的氨基酸序列进行特征提取^[19-20], 可以得到由结构域信息所组成的 216 维的特征向量来表示任意一段氨基酸序列。而对于一个蛋白质往往具有一个或多个结构域, 因此对于任意蛋白质可以由特征向量组成的示例所表示, 同时对所有的蛋白质的 GO 生物学功能标记求合集, 确定每个蛋白质的标记向量, 从而得到了一个多示例多标记样本库。

作者选取了 3 种常见的多标记学习评价标准: Hamming Loss (HL)、Macro-F1 (maF1)、Micro-F1 (miF1)^[21-23]。其中, HL 指标通过计算预测的标记结果与实际样本标记之间的差距来衡量蛋白质功能模型的性能, 其值越小则预测性能越好。其计算方法定义如下:

$$HL(h) = \frac{1}{N} \sum_{i=1}^N \frac{|h(x_i) \oplus y_i|}{y} \quad (15)$$

其中 $\oplus$ 表示两个集合的对称差分, N 为样本的数量, y 为标记向量, x_i 为预测值, y_i 为真实值。maF1、miF1 指标表示分别对 F1 值 (F1 Measure) 应用宏平均 (macro average) 和微平均 (micro average)。其中, maF1 是基于统计量求得在各个类上的分类性能,

然后把所有类上的均值作为最终结果。而 miF1 首先将各个类上的统计量相加, 然后再将求得分类性能作为最终结果, 两者指标值越大, 其预测效果越好。两者的计算方法定义如下:

$$\left\{ \begin{aligned} \text{maF1}(h) &= \frac{\frac{1}{N} \sum_{i=1}^N y_i \cdot h_l(x_i)}{\sum_{i=1}^N y_i + \sum_{i=1}^N h_l(x_i)} \\ \text{miF1}(h) &= \frac{2 * \sum_{i=1}^N \langle h(x_i), y_i \rangle}{\sum_{i=1}^N |h(x_i)| + \sum_{i=1}^N y_i} \end{aligned} \right. \quad (16)$$

其中 $\langle \rangle$ 为数量积, N 为样本的数量, x_i 为预测值, y_i 为真实值。

作者采用 10 折交叉验证, 即通过 UniProtKB/TrEMBL 蛋白质中获取的样本数据进行 10 折交叉验证, 获得了所建功能模型的内部测试结果。为了更好地说明所建模型的泛化能力, 将 UniProtKB/Swiss-Prot 中的数据替换上述 10 折交叉的验证集进行验证, 得到了所建功能模型的外部测试结果, 更进一步证明所建模型的优越性。且每个算法都运行 10 次, 计算“均值标准方差”作为最终结果。

表 1 UniProt 检索条件

Table 1 UniProt search criteria

检索条件	关键词
annotation	type:"positional domain"
keyword	molecular function [KW-9992]
reviewd	YES or NO
organism	name

3.2 参数选取与算法性能

在 MIMLRBF 蛋白质功能预测模型中, 存在两个关键参数, 分别是分数参数 a 和缩放系数 u 。在本文中, 利用网格化方法进行搜索, 其中 a 以步长 0.02 在区间 [0.02, 0.1] 变化, 而 u 则以步长 0.2 在区间 [0.2, 1] 变化。获取所对应参数中 Hamming Loss、maF1、miF1 三个评价指标值的最优值, 其最优结果见表 2。对于大麦, 可以发现, 单独使用 Hausdorff 距离改进与核函数改进都取得了较好的结果, 通过将两者组合进行改进, 其改进效果比单独改进取得了更好的结果。而对于水稻, 两者组合改进的效果只

表 2 改进前后 MIMLRBF 性能比较

Table 2 Performance comparison of MIMLRBF before and after improvement

数据集	方法	内部测试			外部测试		
		HL ↓	maF1 ↑	miF1 ↑	HL ↓	maF1 ↑	miF1 ↑
大麦 (<i>Hordeum vulgare</i>)	改进前	0.005 3±0.000 0	0.539 1±0.064 5	0.703 1±0.055 8	0.011 3±0.000 0	0.231 1±0.023 2	0.325 7±0.031 4
	距离改进	0.005 2±0.000 0	0.561 6±0.061 0	0.716 6±0.051 5	0.011 3±0.000 0	0.237 2±0.018 2	0.330 9±0.025 8
	核函数改进	0.005 2±0.000 0	0.558 9±0.056 7	0.717 5±0.051 8	0.011 2±0.000 0	0.237 9±0.022 8	0.327 3±0.029 4
	两者一起改进	0.005 1±0.000 0	0.574 0±0.060 8	0.723 2±0.053 3	0.011 2±0.000 0	0.244 6±0.021 5	0.330 8±0.029 3
水稻 (<i>Oryza sativa</i> subsp. <i>indica</i>)	改进前	0.001 6±0.000 0	0.467 4±0.019 1	0.600 0±0.026 8	0.003 4±0.000 0	0.225 6±0.006 5	0.287 3±0.006 8
	距离改进	0.001 5±0.000 0	0.492 3±0.017 2	0.611 1±0.021 3	0.003 4±0.000 0	0.229 8±0.005 7	0.295 3±0.005 4
	核函数改进	0.001 6±0.000 0	0.468 8±0.022 2	0.601 4±0.023 7	0.003 4±0.000 0	0.227 2±0.008 4	0.291 5±0.009 3
	两者一起改进	0.001 6±0.000 0	0.491 0±0.018 0	0.608 5±0.021 8	0.003 4±0.000 0	0.229 6±0.005 6	0.292 6±0.007 6

注:“↑”表示值越大越好,“↓”表示值越小越好。

是优于单独的某一种改进。为了更加合理地说明算法模型的可靠,采用了 Swiss-Prot 数据库(手工注释)中的数据进行外部测试,其改进的效果与内部测试所得的效果相符合,由于在外部测试中,其验证数据的数据量较小,所以改进幅度有所下降。总之,在本文中,通过对 Hausdorff 距离引入一个自动调节系数,然后采用改进后的混合径向基函数进行激活,所搭建的 MIMLRBF 蛋白质模型取得了更好的结果。

3.3 与其他算法性能比较

在本文中,将改进 MIMLRBF 算法与两种常见的 MIML 算法(即 En-MIMLNN、MIMLSVM)进行性能比较,且算法的参数值都取最优值。将 En-MIMLNN 算法中 a 设为 0.1, u 设为 0.8。而在 MIMLSVM 算法中,使用宽度为 0.2 的高斯核。实验结果见表 3,可以发现:使用两种方法一起改进后的 MIMLRBF 算法,其性能在谷物蛋白质功能预测方面表现得比其他的 MIML 算法更好一些。

表 3 各种算法性能比较

Table 3 Performance comparison of various algorithms

数据集	方法	内部测试			外部测试		
		HL ↓	maF1 ↑	miF1 ↑	HL ↓	maF1 ↑	miF1 ↑
大麦 (<i>Hordeum vulgare</i>)	两者协同/共同改进 MIMLRBF	0.005 1±0.000 0	0.574 0±0.060 8	0.723 2±0.053 3	0.011 2±0.000 0	0.244 6±0.021 5	0.330 8±0.029 3
	En-MIMLNN	0.005 4±0.000 0	0.530 6±0.051 9	0.697 9±0.053 4	0.011 4±0.000 0	0.209 6±0.021 6	0.304 9±0.029 1
	MIMLSVM	0.007 3±0.000 0	0.492 0±0.050 7	0.583 0±0.058 5	0.014 2±0.000 0	0.168 8±0.020 1	0.184 5±0.025 8
水稻 (<i>Oryza sativa</i> subsp. <i>indica</i>)	两者协同/共同改进 MIMLRBF	0.001 6±0.000 0	0.491 0±0.018 0	0.608 5±0.021 8	0.003 4±0.000 0	0.229 6±0.005 6	0.292 6±0.007 6
	En-MIMLNN	0.001 6±0.000 0	0.472 6±0.026 5	0.600 5±0.021 9	0.003 5±0.000 0	0.217 0±0.006 2	0.284 3±0.007 0
	MIMLSVM	0.002 4±0.000 0	0.282 2±0.019 4	0.327 4±0.021 4	0.003 9±0.000 0	0.209 7±0.003 7	0.199 7±0.005 0

注:“↑”表示值越大越好,“↓”表示值越小越好。

3.4 预测功能与真实功能对比及分析

在表 4 中,选取了其中一种基于距离和核函数共同改进的 MIMLRBF 算法,将其对大麦蛋白质进行外部测试所得出的预测结果和真实功能比较。在 UniProtKB/Swiss-Prot 数据库中,MYB3_HORVU (P20027) 蛋白质通过手工注释已经发现具有 GO:0003677、GO:0005634、GO:0006351、GO:0006355 功能。其中,GO:0003677 主要是其分子功能,表示基因产物与 DNA 有选择性的结合。GO:0005634 是

其细胞组分属性,表示膜上一种细胞器,染色体可以在其中容纳和复制。而 GO:0006351 和 GO:0006355 是其参与的生物学过程,表示 DNA 模板上 RNA 合成以及调控转录频率及范围。通过对 MYB3_HORVU (P20027) 蛋白质进行预测,可以发现,GO:0006355 功能没有预测出来,可能 GO:0006351 与 GO:0006355 功能相近,无法做出正确预测。同样地,对于 IF1C_HORVU (A1E9M5) 蛋白质,其手工注释的功能有 GO:0003743、GO:0009507 及

GO:0019843。其中,GO:0003743,GO:0019843 主要对应其分子功能,表示在 mRNA 转化为多肽的过程中起作用以及与 rRNA 有选择性的结合。而 GO:0009507 对应其细胞组分属性,表示主要是一种含有叶绿素的质体。在预测的过程中,还预测出了 GO:0043022 分子功能,即表现出与核糖体有选择性相互作用。而在文献[24]中,已经明确地表示出 GO:0043022 是其核心功能之一。

表 4 大麦蛋白质预测结果举例

Table 4 Examples of barley protein prediction results

蛋白质	真实功能	预测结果
P20027	GO:0003677 GO:0005634 GO:0006351 GO:0006355	GO:0003677 • GO:0005634 • GO:0006351 •
A1E9M5	GO:0003743 GO:0009507 GO:0019843	GO:0003743 • GO:0009507 • GO:0019843 • GO:0043022

注:预测成功的功能用 • 标出。

4 结 语

随着计算智能技术的发展,对谷物及其营养的研究越来越被重视。而在谷物食品开发的过程中,存在一系列亟待解决的关键技术难题,其中,对谷物蛋白质功能预测的研究目前还较为缺乏。在本文中,首次将 MIMLRBF 算法运用到谷物的蛋白质功能预测,然后在此基础上,提出了多种改进后的 MIMLRBF 算法模型。其中,针对平均 Hausdorff 距离削弱了两种蛋白质之间最短结构域距离所起作用的问题,为平均 Hausdorff 距离引入一个自动调节系数来计算蛋白质之间的相似性。同时,为提高 MIMLRBF 算法的预测能力,采用改进后的混合径向基核函数进行激活,建立了改进后的谷物蛋白质功能模型。通过使用交叉验证以及利用主流的评价标准进行评价,最后可以发现,作者提出的改进的 MIMLRBF 算法综合预测效果最佳,即所建模型具有较好的泛化能力。

参考文献:

- [1] 李倩楠,贾思依,张正涵. 全谷物营养食品发展[J]. 现代食品,2017,(1):22-24.
- [2] 郭顺堂. 我国全谷物食品的开发及存在的问题[J]. 北京工商大学学报(自然科学版),2012,30(5):11-15.
- [3] 梁彬霞,朱豪. 谷物蛋白的结构功能及开发特性的研究进展[J]. 农产品加工(学刊),2013,(12):63-65.
- [4] ZHOU Z H,ZHANG M L,HUANG S J,et al. Multi-instance multi-label learning[J]. **Artificial Intelligence**,2012,176(1):2291-2320.
- [5] ZHOU Z H,ZHANG M L. Multi-instance multi-label learning with application to scene classification[C]// International Conference on Neural Information Processing Systems. Cambridge,MA:MIT Press,2006:1609-1616.
- [6] LI Y X,JI S,KUMAR S,et al. Drosophila gene expression pattern annotation through multi-instance multi-label learning[J]. **IEEE/ACM Transactions on Computational Biology & Bioinformatics**,2011,9(1):98-112.
- [7] WU J,HUANG S J,ZHOU Z H. Genome-wide protein function prediction through multi-instance multi-label learning[J]. **IEEE/ACM Transactions on Computational Biology & Bioinformatics**,2014,11(5):891-902.
- [8] ZHANG M L,WANG Z J. MIMLRBF:RBF neural networks for multi-instance multi-label learning[J]. **Neurocomputing**,2009,72(16):3951-3956.
- [9] 杨素燕. 基于多示例多标记学习的自然场景图像分类[D]. 武汉:武汉理工大学,2015.
- [10] ANFINSEN C B,HABER E,SELA M,et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain[J]. **Proceedings of the National Academy of Sciences of the United States of America**,1961,47(9):1309-1314.
- [11] JANIN J,CHOTHIA C. Domains in proteins:definitions,location,and structural principles[J]. **Methods in Enzymology**,1985,115(4):420-430.
- [12] ZHANG Z,LIU G,LIU F. Radial basis function networks based on adaptive projective learning algorithm and its applications[J]. **Acta Electronica Sinica**,2000,28(9):120-122.
- [13] PARK H S,JUN C H. A simple and fast algorithm for K-medoids clustering[J]. **Expert Systems with Applications**,2009,36(2):3336-3341.

- [14] ZHANG M L,ZHOU Z H. Multi-instance clustering with applications to multi-instance prediction[J]. **Applied Intelligence**,2009, 31(1):47-68.
- [15] WANG J,ZUCKER J D. Solving the multiple instance problem:a lazy learning approach[C]// Proceedings of International Conference on Machine Learning, Stanford, CA: Stanford University,2000:1119-1126.
- [16] 蒋传海. 神经网络中的逼近问题[J]. 数学年刊:中文版,1998,(3):295-300.
- [17] 付丽华,李宏伟,张猛. 尺度可调的混合核 RBF 网络[J]. 电子学报,2011,39(1):184-189.
- [18] ASHBURNER M,BALL C A,BLAKE J A,et al. Gene ontology:tool for the unification of biology[J]. **Nature Genetics**,2000, 25(1):25-29.
- [19] WU J,HU D,XU X,et al. A novel method for quantitatively predicting non-covalent interactions from protein and nucleic acid sequence[J]. **Journal of Molecular Graphics & Modelling**,2011,31(11):28-34.
- [20] SHEN J,ZHANG J,LUO X,et al. Predicting protein-protein interactions based only on sequences information[J]. **Proceedings of the National Academy of Sciences of the United States of America**,2007,104(11):4337-4341.
- [21] GHAMRAWI N,MCCALLUM A. Collective multi-label classification[C]// ACM International Conference on Information & Knowledge Management. New York:ACM,2005:195-200.
- [22] ROGATI M,YANG Y. High-performing feature selection for text classification[C]// Eleventh ACM International Conference on Information and Knowledge Management. New York:ACM,2002:659-661.
- [23] YANG S J,JIANG Y,ZHOU Z H. Multi-instance multi-label learning with weak label[C]// The International Joint Conference on Artificial Intelligence. Palo Alto,CA:AAAI Press,2013:1862-1868.
- [24] VOGEL J P,GARVIN D F,MOCKLER T C,et al. Genome sequencing and analysis of the model grass *Brachypodium distachyon* [J]. **Nature**,2010,463(7282):763-768.

科 技 信 息

欧盟批准艾蒿酊剂作为所有动物的饲料添加剂

据欧盟官方公报消息,2021 年 3 月 9 日,欧盟委员会发布法规(EU)2021/421 号条例,根据欧洲议会和理事会法规(EC) No 1831/2003,批准艾蒿酊剂(a tincture derived from *Artemisia vulgaris* L. (Mugwort tincture))作为所有动物的饲料添加剂。

根据附件中规定的条件,这种添加剂被授权作为动物添加剂的所属类别为“感官添加剂”,功能组别为“调味化合物”。授权结束日期为 2031 年 3 月 30 日。

[信息来源] 食品伙伴网. 欧盟批准艾蒿酊剂作为所有动物的饲料添加剂 [EB/OL]. (2021-3-10). <http://news.foodmate.net/2021/03/586996.html>