

CL-RBF:一种基于改进 ML-RBF 的 蛋白质亚细胞多点定位预测算法

薛 卫¹, 洪晓宇¹, 胡雪娇¹, 陈行健¹, 张 梁^{*2}

(1. 南京农业大学 信息科学技术学院, 江苏 南京 210095; 2. 江南大学 粮食发酵工艺与技术国家工程实验室, 江苏 无锡 214122)

摘要: 综合考虑标记内和标记间的聚类结果对多目标学习径向基神经网络算法 (RBF Neural Networks for Multi-Label Learning, ML-RBF) 的影响, 提出 CL-RBF 算法并应用到蛋白质亚细胞多点定位预测中。通过引入轮廓系数(Silhouette Coefficient)对 ML-RBF 隐层中心的个数进行优化, 并通过分析标记间聚类结果的关系, 对小于某一阈值的标记间的聚类中心重新聚类, 使用梯度下降算法进行参数调整, 最后依据测试样本与标记 L 的隐层中心和不属于标记 L 的样本生成的聚类中心的欧式距离差调整预测结果。在 10 折交叉验证下, 采用词袋模型(Bag of Words)和氨基酸组合法(Amino acid composition, AAC)结合的方式提取特征向量, 选取另外 4 种多目标学习算法作对比实验, 根据不同评价指标的结果, 得出 CL-RBF 算法在 4 个多标记数据集上的综合性能最优的结论。本研究预测算法通过网站 https://njau.applinzi.com/homepage_final.jsp 实现。
关键词: ML-RBF; 亚细胞定位; 轮廓系数; 词袋模型

中图分类号: TP391.4 文章编号: 1673-1689(2020)02-0066-08 DOI: 10.3969/j.issn. 1673-1689.2020.02.009

CL-RBF: An Improved ML-RBF Method for Prediction of Protein Subcellular Location

XUE Wei¹, HONG Xiaoyu¹, HU Xuejiao¹, CHEN Xingjian¹, ZHANG Liang^{*2}

(1. School of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China; 2. National Engineering Laboratory for Cereal Fermentation Technology, Jiangnan University, Wuxi 214122, China)

Abstract: CL-RBF algorithm was proposed to predict the protein subcellular localization, which is considered about cluster results within one label and between different labels of the ML-RBF method. Silhouette coefficient was introduced to get the optimal number of centroids on hidden layer. The previous approach only considered optimization of clustering algorithms within the same label. In this paper, larger distance between two centroids which were generated from two labels was taken into account, when there were less samples covering these two labels. Besides, gradient descent

收稿日期: 2018-01-04

基金项目: 国家重点研发计划项目(2017YFD0800204); 国家“十二五”科技支撑计划项目(2015BAK36B05); 江苏省自然科学基金项目(BK2012363); 中央高校基本科研业务费专项资金项目(Y0201600175)。

作者简介: 薛卫(1979—), 男, 博士, 副教授, 硕士研究生导师, 主要从事生物信息、模式识别方面的研究。E-mail: xwsky@njau.edu.cn

* 通信作者: 张梁(1978—), 男, 博士, 教授, 博士研究生导师, 主要从事代谢工程方面的研究。E-mail: zhangl@jiangnan.edu.cn

algorithm was used to adjust the parameters. The final adjustment was made by analyzing the distance between train samples, the hidden centers obtained by label L and the clustering centers not belonging to label L. Bag of words and AAC method were employed to extract the feature of protein sequence. Compared with the methods which have been introduced previously for bacterial protein subcellular localization prediction via 10-fold cross-validation test, the new predictor performed more powerful and flexible on four different multi-label datasets. The prediction server was available on https://njau.applinzi.com/homepage_final.jsp.

Keywords: ML-RBF, protein subcellular localization, silhouette coefficient, Bag of Words

蛋白质是生命体内功能最丰富的生物大分子,在必需的生命活动中,发挥着关键性作用。其功能的正常发挥依赖其所在的亚细胞位置,只有处于特定的细胞器上,蛋白质才能正常发挥作用,从而保证生命体的正常运转。因此蛋白质亚细胞定位预测在识别未知功能的蛋白质序列和确定基因组标注中都有重要的意义和作用,还可以极大地提高药物靶点的识别^[1]。随着生物数据快速更新和大量积累,使用人工实验去获取蛋白质的位置已远不能达到科研需要,从而促使了机器学习在蛋白质亚细胞定位预测中的发展。1991年, Nakai 和 Kanehisa^[2]在对革兰氏阴性菌蛋白质进行亚细胞位置识别时,整理出了一系列选择判别规则,将机器学习的方法首次应用到蛋白质亚细胞定位预测中。2000年, Cai 和 Chou^[3]采用了基于人工神经网络的自组织模型来预测原核和真核生物蛋白质的亚细胞位置。2013年,曹隽喆^[4]建立了一种不平衡权重的多标签尺 KNN 模型,以消除蛋白质数据集分布不均衡的情况。Yang 等人^[5]在 2014 年提出了一种基于可能性的 SVM 模型被用来识别人类蛋白质亚细胞位置。2015 年, Liu 和 Tao 等人^[6]将基于最大间隔原理的 SVM-RFE 算法与基于 PSSM (Position-Specific Score Matrix) 的 Tri-gram 编码方式结合,在 3 个不同的数据集上都取得了较好的实验结果。Bendtsen 等人^[7]在 SignalP 方法的基础上进行了改进,提出了 SignalP 3.0。Rahman 等人^[8]将蛋白质的多种特征融合,提出了 AAIDPAAC 和 PPMPAAC 特征,与 SVM 结合得到了较高的预测准确率。

在人工神经网络应用的早期, BP 神经网络、概率神经网络和 SOM 使用较多,但后来逐渐被径向基(Radial Basis Function, RBF)神经网络所取代^[9]。为了解决多目标分类问题, Zhang^[10]提出了一种基于

RBF 神经网络的多目标学习方法 ML-RBF。该算法的训练一般分为两步:第一步在不同的目标下,对属于该目标的训练样本进行 K-means 聚类得到隐层中心,将原型向量与隐层中心关联,计算径向基函数;第二步通过最小化误差平方和来获得最优化权重。因此对 ML-RBF 算法的改进也主要集中在:结构设计,即隐藏层节点设计;参数优化,包括基函数的数据中心及扩展常数、输出节点的权值等模型参数。作者分别从这两点出发并结合实验对象特点对 ML-RBF 算法进行改进,提出 CL-RBF 算法。实验结果显示, CL-RBF 算法在不同的数据集上都显示出了较高的性能。

1 材料与方法

1.1 数据集

Xiao 等人^[11]指出细菌蛋白质在生命体活动中是一把“双刃剑”,既存在巨大危害,又发挥着不可忽视的积极作用,因此被认为是最有研究价值的蛋白质之一,故选取细菌作为实验对象。目前数据集构建方法可分为 3 种:一是直接选用已有数据集;二是选取已有数据集并进行更新;三是重新构建数据集。直接用前人提出的数据集进行实验的弊端是忽略了时间间隔内的数据更新问题,因此选取后两种方式进行数据集构建。原始数据集都来自于 UniProt 数据库,通过 CC(comment or notes)和 OC (organism classification)检索框来搜索革兰氏阴性菌和革兰氏阳性菌的蛋白质序列。所有数据必须严格按照下列标准进行筛选:

1) 革兰氏阴性菌:在 OC 检索框中输入“proteobacteria”;革兰氏阳性菌:在 OC 检索框中输入“firmicutes”和“actinobacteria”。

2) 在 CC 检索框中选取 subcellular location

term, 由于实验针对的是多目标, 因此不选取具体的亚区间, 以 * 代替, 蛋白质序列只选取经过实验验证的, 即“experimental assertion”。

3) 蛋白质序列的长度规定在 50~10 000 区间, 并且不能是片段。

通过筛选后得到原始的蛋白质序列集合, 由于过高的相似度会增加实验的准确率, 为了使训练出的模型更加客观, 利用 CD-HIT 设置相似度阈值为 30%, 剔除相似度过高的蛋白质序列。再采用网络爬虫技术到 UniProt 数据库中抓取亚细胞信息, 具体抓取条件如下:

1) 尽管表述方式不同, 但是多个关键字可能代表同个细胞区域。关键字“extracell”, “extracellular”和“secreted”可以相互替换; 对于革兰氏阴性菌, “cytoplasm”和关键字“cytoplasmic”具有同样的意义。对于革兰氏阳性菌, 当检索蛋白质亚细胞时, 关键字“plasma membrane”, “integral membrane”, “multi-pass membrane”和“single-pass membrane”都归类于亚细胞“cell membrane”。

2) 当蛋白质序列被模糊或者不确定项标注时, 例如“potential”, “probable”, “probably”, “maybe”, 或者“by similarity”则会被剔除。

经过处理后得到革兰氏阴性菌数据集革兰氏阴性菌 833 和革兰氏阳性菌数据集革兰氏阳性菌 464^[12-13]。为了维持样本的均衡性, 选择革兰氏阴性菌和革兰氏阳性菌蛋白质序列分布较广的亚细胞^[14]。革兰氏阴性菌 833 数据集包含 5 个亚细胞, 分别是细胞内膜 (Cell inner membrane)、细胞外膜 (Cell outer membrane)、细胞质 (Cytoplasm)、细胞外基质 (Extracellular) 和细胞周质 (Periplasm); 革兰氏阳性菌 464 数据集包括细胞膜 (Cell membrane)、细胞壁 (Cell wall)、细胞质 (Cytoplasm) 和细胞外基质 (Extracellular) 4 个亚细胞, 数据集的具体分布情况见表 1-2。

表 1 革兰氏阴性菌 833 数据集构成

Table 1 Distribution of gram-negative 833 dataset

亚细胞区间	序列数量	序列总数
细胞内膜	337	833
细胞外膜	67	
细胞质	208	
细胞外基质	114	
细胞周质	144	

表 2 革兰氏阳性菌 464 数据集构成

Table 2 Distribution of gram-positive 464 dataset

亚细胞区间	序列数量	序列总数
细胞膜	167	464
细胞壁	67	
细胞质	139	
细胞外基质	177	

此外, 作者还对两个已经构建好的革兰氏阴性菌数据集^[15]和革兰氏阳性菌数据集^[16]进行筛选和重构。剔除革兰氏阴性菌中样本数量较少的亚细胞, 即菌毛 (Fimbrium), 鞭毛 (Flagellum) 和拟核 (Nucleoid)。利用网络爬虫技术重新获取亚细胞信息, 从而使更新后的数据集能反应最新的序列信息。亚细胞抓取和处理过程同数据集革兰氏阴性菌 833 和革兰氏阳性菌 464。重构后的数据集分别为革兰氏阴性菌 1392 和革兰氏阳性菌 519, 数据集的具体分布见表 3-4。

表 3 革兰氏阴性菌 1392 数据集构成

Table 3 Distribution of gram-negative 1392 dataset

亚细胞区间	序列数量	序列总数
细胞内膜	561	1392
细胞外膜	110	
细胞质	394	
细胞外基质	136	
细胞周质	194	

表 4 革兰氏阳性菌 519 数据集构成

Table 4 Distribution of gram-positive 519 dataset

亚细胞区间	序列数量	序列总数
细胞膜	165	519
细胞壁	22	
细胞质	186	
细胞外基质	130	

1.2 序列特征编码

对蛋白质序列进行特征提取是蛋白质亚细胞定位预测中的重要环节, 有效的特征提取方法可以增加预测准确率。赵南等人^[17]将词袋模型应用到蛋白质特征提取当中, 并与传统的基于氨基酸组成的序列表示方法相结合, 提出了词袋特征这个概念。实验证明, 词袋特征能有效增加分类器的预测准确率。因此本研究采用词袋模型结合 AAC 方法表示

一条蛋白质序列。

1.2.1 氨基酸组合法 Nakashima 和 Nishikawa^[18]最早将氨基酸的组成和蛋白质亚细胞位置联系起来,提出 AAC 编码方式,统计每个氨基酸在蛋白质序列中出现的频率。

AAC 方法定义如公式(1)所示:

$$P=[f_1 f_2 f_3 \cdots f_u \cdots f_{20}]^T \quad (1)$$

$f_u(u=1,2,3,\cdots,20)$ 为氨基酸 u 在序列中出现的频率; f_u 的求解方法如公式(2)所示:

$$f_u = \frac{1}{L} \sum_{i=1}^L F_i, F_i = \begin{cases} 1, & \text{if } R_i = A(u) \\ 0, & \text{if } R_i \neq A(u) \end{cases} \quad (2)$$

其中, L 表示一条蛋白质序列的长度,即包含的所有氨基酸残基的总数目。首先要对 20 种氨基酸从 1 到 20 进行编号。

1.2.2 词袋特征 通过 AAC 算法计算蛋白质序列 P 的序列单词特征,可得一个片段特征矩阵,如公式(3)所示:

$$\begin{bmatrix} v_{11} & v_{1n} & \cdots & v_{1n} \\ v_{21} & \cdots & \cdots & v_{2n} \\ \vdots & \cdots & \cdots & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mn} \end{bmatrix} \quad (3)$$

式(3)中, v 表示蛋白质片段中氨基酸出现的频率; m 表示一条蛋白质序列切割成的片段条数; n 为经过特征提取算法处理后的特征维度。具体的求解方法见文献[17]。

通过滑动窗口分割法和基于类内方差和最小的聚类方法得出一个最优的词袋字典,字典包含 i 种不同的类别,每一类都是一个 n 维的向量,具体如公式(4)所示:

$$\begin{bmatrix} d_{11} & d_{1n} & \cdots & d_{1n} \\ d_{21} & \cdots & \cdots & d_{2n} \\ \vdots & \cdots & \cdots & \vdots \\ d_{i1} & d_{i2} & \cdots & d_{in} \end{bmatrix} \quad (4)$$

式(4)中, $d_{i1} d_{i2} d_{in}$ 表示第 i 类的词袋特征,式(5)为词袋特征中第 q 维的计算方法:

$$\sum_{j=1}^m e_q = \frac{\text{bow}_j}{m} (q=1,2,\cdots,i) \quad (5)$$

其中, bow_j 表示片段 j 是否属于字典中的 q 类。具体解决方法如下:

$$\text{bow}_j = \begin{cases} 1, & \text{片段 } j \text{ 属于字典中的 } q \text{ 类} \\ 0, & \text{片段 } j \text{ 不属于字典中的 } q \text{ 类} \end{cases}$$

通过计算欧氏距离判断片段是否属于第 q 类,

如果片段 j 与 q 类的欧式距离最小,则片段 j 属于 q 类。最后得到完整的词袋特征,如公式(6)所示:

$$E=(e_1, e_2, \cdots, e_i) \quad (6)$$

1.3 基于改进 ML-RBF 的分类预测算法

1.3.1 基于标记内的聚类方法改进策略 聚类是一种无监督的学习方法,理想的聚类结果一般是类内间距最小,类间间距最大。Kaufman 等人^[19]基于上述思想,提出了轮廓系数 S 以及个体轮廓系数 s_i ,具体表述方式如下:

$$S = \frac{1}{n} \sum_{i=1}^n s_i \quad (7)$$

$$s_i = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \quad (8)$$

$$a(i) = \frac{1}{n_c - 1} \sum_{j \in C_c, i \neq j} d(i, j) \quad (9)$$

$$b(i) = \min_{p, p \neq c} \left[\frac{1}{n_p} \sum_{i \in C_c, j \in C_p} d(i, j) \right] \quad (10)$$

其中, n 表示样本个数。假设样本 i 属于聚簇 c , $a(i)$ 表示样本 i 和同属于聚簇 c 的其他样本距离的平均值。 $b(i)$ 选取样本 i 和不属于聚簇 c 的每个聚簇中所有样本的平均距离的最小值。

ML-RBF 神经网络的预测效果与隐层中心有密切的关系,依次对每个类别标记,采用传统 K-means 算法对所属样本进行聚类,计算不同聚类中心个数下的轮廓系数。轮廓系数越大,说明对应类别标记选取的聚类中心个数越合理。具体的改进步骤如下:

1)对于每一个 $l \in L$,设置聚类个数 k 的搜索范围为 $k \in [k_{\min}, k_{\max}]$ 。

2)for $k_{\min}$ to $k_{\max}$ 。

●调用传统 K-means 算法对属于类别标记 l 的样本进行聚类;

●利用公式(8)计算个体轮廓系数 s_i 。

●利用公式(7)计算轮廓系数 S 。

3)选取最大轮廓系数对应的聚类个数,作为类别标记 l 的最佳聚类个数 k_{best} 。

1.3.2 基于标记间的聚类方法改进策略 以往对 ML-RBF 中隐层中心的改进主要针对同一类别标记内的聚类结果进行优化,从而忽略了不同类别标记间的隐层中心之间的相互影响。作者提出了一种基于类别标记间的聚类方法改进策略。该方法将标记间(Inter-Label)隐层中心距离与一个预设的阈值 γ 进行对比,从而调整不同聚簇中的样本,进而重新

计算聚簇中心(隐层中心)。

阈值 γ 的公式如下:

$$\gamma = \alpha * \mu \quad (11)$$

其中, α 为一个常数; μ 为比例因子。

针对不同的隐层中心对 (c_i^{lm}, c_j^{ln}) , 其中 l 为类别标记, 而且 $l_m \neq l_n$, 有集合 $Z_{two} = \{(x_i, Y_i) | 1 \leq i \leq m, l_m \in Y_i, l_n \in Y_i\}$ 。集合 Z_{two} 的元素个数记为 $K_{[lm, ln]}$, 总的训练样本个数为 $K_{total} \circ K_{[lm, ln]} / K_{total}$ 的比重越大, 则说明类别标记 l_m 和 l_n 之间的关联性越高, 因此阈值 γ 的设置可以较其他的类别对低。比例因子由 Sigmoid 函数经过变换得来, 公式如下:

$$\mu = \frac{1}{1 + e^{10 * (x - 0.5)}} \quad (12)$$

$$x = K_{[lm, ln]} / K_{total} \quad (13)$$

当 x 很小时, 比例因子的下降速度很慢, 比例因子接近于 1, 阈值 γ 无限接近于 α 。

改进的具体流程如下:

- 1) 计算不同类别标记间的隐层中心距离 $dist$ 。
- 2) 将隐层中心之间的距离从小到大排序, 并生成隐层中心对集合 $D_{two} = \{(c_i^p, c_j^q) | dist(c_i^p, c_j^q) < \alpha\} p, q \in L$
- 3) 从集合 D_{two} 选取一对隐层中心 (c_i^{lm}, c_j^{ln}) 进行调整, 同时从集合 D_{two} 将其删除。
- 4) 利用公式 (11)~(13) 计算阈值 γ 。
- 5) for $dist(c_i^{lm}, c_j^{ln}) < \gamma$
 - a. 计算样本分别距离集合 $\{c_1^l, \dots, c_{k_l}^l\}$ 和 $\{c_1^l, \dots, c_{k_l}^l\}$ 的距离, k_l 是属于类别标记 l 的隐层中心个数, 得出属于聚簇 c_i^l 和 c_j^l 的样本集合 X_l 和 X_l 。
 - b. 计算集合 X_l 中的元素和集合 X_l 中元素的距离, 从集合 X_l 和 X_l 中删除距离最近的一对训练样本。
 - c. 分别加上 c_i^l 和 c_j^l , 统计集合 X_l 和 X_l 的平均值, 以此作为新的隐层中心, 更新 c_i^l 和 c_j^l 的值。
- 6) 直到 $D_{two} = \emptyset$, 算法结束, 否则跳转到 3)。

1.3.3 自适应梯度下降调整参数 为了进一步缩小误差, 使用梯度下降的学习方法对平滑指数 σ_i 以及隐藏层与输出层的链接权重 W 进行参数调整。由于求解权重 W 产生的误差较大, 初始学习率设置 $\eta_1 > \eta_2$, η_1 为调节权重的学习率, η_2 为调节平滑指数的学习率。通过在革兰氏阳性菌和革兰氏阴性菌数据集上反复试验, 发现两者采用不同步的调整策略能更快地逼近最小值。当代价函数增大时, 减小权重步长; 当代价函数减小时, 增加平滑指数步长。

1.3.4 基于聚类优化的结果集调整策略 $X = R^d$ 为输入空间, 给定一个多标记训练集合 $D = \{(x_i, Y_i) | 1 \leq i \leq m\}$, 其中 $x_i \in X$ 表示一个训练示例, $Y_i \subseteq L$ 表示与示例 x_i 关联的类别标记集合。

针对某个类别标记 l , 分别有 $X_{positive} = \{x_i | (x_i, Y_i) \in D, l \in Y_i\}$, $x_i \in X$ 表示训练样本 x_i 属于类别标记 l , 该训练样本构成集合 $X_{positive}$; $X_{negative} = \{x_j | (x_j, Y_j) \in D, l \notin Y_j\}$, $x_j \in X$ 表示训练样本 x_j 不属于类别标记 l , 该训练样本构成集合 $X_{negative}$ 。并且有 $X_{positive} \cap X_{negative} = \emptyset$, $X_{positive} \cup X_{negative} = X$ 。

在传统的 ML-RBF 训练中, 第一步仅考虑对集合 $X_{positive}$ 进行聚类, 而忽略了对 $X_{negative}$ 进行聚类, 分析两者的关系。对 $X_{positive}$ 和 $X_{negative}$ 两种样本分别进行聚类, 当待测样本与 $X_{positive}$ 的聚类中心距离更近时, 说明测试样本与 l 类的样本 $X_{positive}$ 相似度更高, 预测值应该增大; 反之应该减小。对 $X_{positive}$ 通过基于标记内和标记间改进的 K-means 聚类得到 $k_{positive}$ 个隐层中心, 构成集合 $C_{positive} = \{c_1^p, c_2^p, \dots, c_{k_{positive}}^p\}$ 。对 $X_{negative}$ 同样采用引进个体轮廓系数的方法进行聚类, 生成 $k_{negative}$ 个聚类中心, 构成集合 $C_{negative} = \{c_1^n, c_2^n, \dots, c_{k_{negative}}^n\}$ 。在调整了类别标记间隐层中心距离的前提下, 当测试样本 x 与隐层中心 c_i^p 的距离 $dist(x, c_i^p)$ 越小, 则 x 越可能被认为属于类别标记 l , 然而此时若有 $dist(x, c_j^n) < dist(x, c_i^p)$, 则说明 x 与不能归类于类别标记 l 的训练样本的聚簇中心更近; 反之亦然。调整过程如下:

$$dist_{positive} = \min[dist(x, c_i^p)] \quad i = 1, 2, 3, \dots, k_{positive} \quad (14)$$

$$dist_{negative} = \min[dist(x, c_j^n)] \quad j = 1, 2, 3, \dots, k_{negative} \quad (15)$$

$$y_{l_{new}}(x) = y_l(x) - \mu \cdot (dist_{positive} - dist_{negative}) \quad (16)$$

$dist_{positive}$ 为 x 到 $C_{positive}$ 中每一个中心的最小距离, $dist_{negative}$ 为 x 到 $C_{negative}$ 中每一个中心的最小距离。 μ 为比例因子, 本研究设置为 1, 当 $dist_{positive} < dist_{negative}$, $y_{l_{new}}(x)$ 输出结果增大, 反之减小。

1.3.5 基于 CL-RBF 的蛋白质亚细胞定位预测器 将 CL-RBF 应用到蛋白质亚细胞定位预测中, 具体流程如下:

1) 计算数据集中所有蛋白质序列的词袋特征, 把词袋特征集合作为样本集合。将所有序列的词袋特征分为 n 组互不相交的子集合, 依次取出一个子集作为测试集, 其余 $n-1$ 个子集构成训练集。

2) 针对某个亚细胞 i , 将训练样本分为属于亚

细胞 i 的训练样本和不属于亚细胞 i 的训练样本,分别对两种训练样本使用基于标记内改进的 K_{means} 算法进行聚类,得到两组聚类中心集合 C_{in} 和 C_{out} 。对于每一个亚细胞都可以生成一组这样的集合。

3) 标记内改进的 K-means 算法中常数 α 设定为 0.15, 进一步优化 C_{in} , 组成 ML-RBF 的隐层中心, 所有的 C_{out} 将用于结果调整。

4) 通过隐层中心和属于亚细胞 i 的训练样本计算出 ML-RBF 模型的权重。采用梯度下降算法调整权重和平滑指数, 最后得到一个 CL-RBF 分类器。

5) 将测试样本特征向量作为输入送入分类器, 分类器会给出一个或者多个预测结果, 采用基于聚类优化的结果集调整策略以及调整预测结果。

6) 若预测结果与实际的亚细胞位置相同, 则预测正确, 否则预测错误。

7) 预测完毕后将测试样本放回样本集合中并取出下一组子集作为测试样本, 其余样本作为训练样本, 再次训练出一个 CL-RBF 分类器, 并用测试样本进行测试。以此类推直至所有子集测试完毕。

2 结果与分析

对比算法选取了 ML-KNN 算法、LIFT 算法^[20]、MLNB 算法^[21]和 ML-RBF 算法。不同的分类预测算法在不同的数据集上单独运行 50 次, 采用十折交叉实验, 模型每运行完 10 次, 数据集会重新分成 10 个子集, 评价指标取平均值。

我们使用多标记学习当中经常用到的 5 个评价算法, 汉明损失 (Hamming loss)、单错误 (one-error)、覆盖范围 (coverage)、排位损失 (ranking loss) 和平均查准率 (average precision)。其中汉明损失、单错误、覆盖范围和排位损失这 4 个评价指标越小, 说明算法的效果越好。而平均查准率, 我们期望越大越好。由于 5 种评价指标的考察方向不同, 所以不能保证算法在所有的指标上都有很好的结果。

表 5 为 5 种多目标学习算法在革兰氏阳性菌 464 上的实验结果。可以看出, ML-RBF 算法在革兰氏阳性菌 464 数据集上的性能较好。除了汉明损失这个指标, CL-RBF 算法在其他 4 个指标上都最优秀, 覆盖范围上比 ML-RBF 算法降低了两个百分点, 平均查准率达到了 85.37%。

表 5 不同多目标算法在革兰氏阳性菌 464 上的性能比较

Table 5 Performance of different multi-learning algorithms on gram-positive464 dataset

算法	汉明损失	单错误	覆盖范围	排位损失	平均查准率
CL-RBF	0.206 4	0.302 0	0.716 8	0.185 2	0.853 7
ML-RBF	0.204 7	0.313 6	0.739 1	0.192 7	0.806 7
ML-KNN	0.224 7	0.345 7	0.807 6	0.218 7	0.785 8
LIFT	0.234 0	0.363 6	0.744 7	0.187 5	0.789 8
MLNB	0.223 4	0.325 7	0.776 9	0.212 3	0.798 2

根据表 6 的评价指标显示, 在 Gram-positive515 数据集上, LIFT 和 ML-RBF 的 5 个评价指标也比较不错, 其中 LIFT 算法出现单错误的概率最低, 为 24.41%。CL-RBF 和 ML-RBF 算法在其他 4 个指标

上都要优于其他的比对算法, 而 LIFT 在这 4 个指标上的表现紧随其后。MLNB 在前 4 种指标中都达到了最大值, 并且在平均查准率上最小, 因此 MLNB 在该数据集上的表现是最差的。

表 6 不同多目标算法在革兰氏阳性菌 515 上的性能比较

Table 6 Performance of different multi-learning algorithms on gram-positive515 dataset

算法	汉明损失	单错误	覆盖范围	排位损失	平均查准率
CL-RBF	0.153 6	0.247 0	0.321 9	0.131 2	0.879 8
ML-RBF	0.156 9	0.247 9	0.331 3	0.135 9	0.853 0
ML-KNN	0.165 0	0.270 0	0.352 1	0.145 8	0.840 5
LIFT	0.155 1	0.244 1	0.329 9	0.134 2	0.854 9
MLNB	0.166 8	0.270 3	0.380 5	0.155 2	0.838 1

从表 7 可知,ML-RBF 预测算法在数据集革兰氏阴性菌 833 上的评价指标同样优于 ML-KNN、LIFT 和 MLNB3 种算法,而且经过改进后,CL-RBF 的指标

数据都得到了提升,达到了最优值。ML-KNN 算法在革兰氏阴性菌 833 上的表现最差,在单错误、覆盖范围和排位损失都是最大的,而在平均查准率上是最小的。

表 7 不同多目标算法在革兰氏阴性菌 833 上的性能比较

Table 7 Performance of different multi-learning algorithms on gram-negative833 dataset

算法	汉明损失	单错误	覆盖范围	排位损失	平均查准率
CL-RBF	0.139 4	0.332 8	0.617 7	0.165 2	0.842 8
ML-RBF	0.142 6	0.338 5	0.631 0	0.169 4	0.789 8
ML-KNN	0.157 9	0.381 7	0.748 3	0.201 1	0.757 0
LIFT	0.146 9	0.358 3	0.716 1	0.191 5	0.770 7
MLNB	0.171 6	0.359 4	0.666 6	0.176 5	0.775 1

从表 8 可知,ML-RBF 在该数据集上的表现仍旧很不錯,除了 LIFT 在汉明损失这个指标上最优秀,CL-RBF 在这 4 个评价指标上都达到了最优,

ML-RBF 次之。ML-KNN 和 MLND 算法在 5 个评价指标上的表现都较差。

表 8 不同多目标算法在革兰氏阴性菌 1392 上的性能比较

Table 8 Performance of different multi-learning algorithms on gram-negative1392 dataset

算法	汉明损失	单错误	覆盖范围	排位损失	平均查准率
CL-RBF	0.116 2	0.273 2	0.492 1	0.127 8	0.877 3
ML-RBF	0.116 4	0.273 7	0.497 3	0.129 1	0.832 7
ML-KNN	0.123 3	0.299 1	0.579 9	0.151 3	0.812 9
LIFT	0.115 4	0.287 4	0.542 4	0.140 9	0.821 3
MLNB	0.148 7	0.296 9	0.578 2	0.148 2	0.815 3

3 结 语

本研究在 ML-RBF 算法的基础上,提出 CL-RBF 算法。算法对隐层中心选取、参数计算和结果集处理分别进行了调整。通过轮廓系数优化隐层中心的个数,采用传统的 K-means 聚类计算隐层中心的值,针对不同标记间隐层中心距离较近的情况进行了处理。采用自适应的梯度下降算法对平滑指数和链接权重进行调整,充分考虑待测样本与该标记的隐层中心、不属于该标记的训练样本的聚类中心之间的关系,来进一步调整最终的预测值。选取一个基于词袋模型和 AAC 相结合的方法作为本研究的特征提取方法。实验选取了 LIFT、MLNB、ML-KNN

和 ML-RBF 4 种算法与 CL-RBF 算法在 4 个多标记数据集上进行对比实验。从实验结果可以看出,在革兰氏阳性菌和革兰氏阴性菌的蛋白质亚细胞定位预测上,LIFT 算法、ML-RBF 算法和 CL-RBF 算法不论是在汉明损失、单错误、覆盖范围、排位损失还是平均准确率上,表现都要优于 ML-KNN 算法和 MLNB 算法。三者的性能排序是 LIFT 算法<ML-RBF 算法<CL-RBF 算法,ML-RBF 经过改进后能在绝大部分评价指标上获得比较好的结果。对比其他算法,改进的 ML-RBF 算法在 4 个多标记数据集上都取得了最优的覆盖范围、排位损失和平均查准率。这些结果证明 CL-RBF 是一种较为有效的蛋白质亚细胞定位预测算法。

参考文献:

- [1] WU xiaohong, XUE Wei, ZHANG Liang, et al. Auto-refinement of genome-scale metabolic network model[J]. **Journal of Food Science and Biotechnology**, 2017, 36(9): 982-989. (in Chinese)
- [2] NAKAI K, KANEHISA M. Expert system for predicting protein localization sites in gram-negative bacteria [J]. **Proteins**:

- Structure, Function, and Bioinformatics**, 1991, 11(2):95-110.
- [3] CAI Y D, CHOU K C. Using neural networks for prediction of subcellular location of prokaryotic and eukaryotic proteins[J]. **Molecular Cell Biology Research Communications**, 2000, 4(3): 172-173.
- [4] 曹隽喆. 基于机器学习的多定位点蛋白质亚细胞定位预测方法研究[D]. 大连: 大连理工大学, 2013.
- [5] YANG F, XU Y Y, WANG S T, et al. Image-based classification of protein subcellular location patterns in human reproductive tissue by ensemble learning global and local features[J]. **Neurocomputing**, 2014, 131(9): 113-123.
- [6] LIU T, TAO P, LI X, et al. Prediction of subcellular location of apoptosis proteins combining tri-gram encoding based on PSSM and recursive feature elimination[J]. **Journal of Theoretical Biology**, 2015, 366: 8-12.
- [7] BECERRA S C, ROY D C, SANCHEZ C J, et al. An optimized staining technique for the detection of gram positive and gram negative bacteria within tissue[J]. **BMC Research Notes**, 2016, 9(1): 1-10.
- [8] RAHMAN J, MONDAL M N, ISLAM M K, et al. Feature fusion based SVM classifier for protein subcellular localization prediction[J]. **Journal of Integrative Bioinformatics**, 2016, 13(1): 23-33.
- [9] LIU Bingjie, GUO Hong. Predicting subcellular localization of protein based on PSSM and GO features[J]. **Journal of Fuzhou University**. 2017, 45(1): 16-24. (in Chinese)
- [10] ZHANG M L. ML-RBF: RBF neural networks for multi-label learning[J]. **Neural Processing Letters**, 2009, 29(2): 61-74.
- [11] XIAO X, WU Z C, CHOU K C. A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites[J]. **Plos One**, 2011, 6(6): e20592.
- [12] CHOU K C, SHEN H B. Large-scale predictions of gram-negative bacterial protein subcellular locations[J]. **Journal of Proteome Research**, 2006, 5(12): 3420-3428.
- [13] SHEN H B, CHOU K C. Gpos-PLoc: an ensemble classifier for predicting subcellular localization of gram-positive bacterial proteins[J]. **Protein Engineering Design & Selection**, 2007, 20(1): 39-46.
- [14] YU N Y, WAGNER J R, LAIRD M R, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes[J]. **Bioinformatics**, 2010, 26(13): 1608-1615.
- [15] SHEN H B, CHOU K C, Gneg-mPLoc: A top-down strategy to enhance the quality of predicting subcellular localization of gram-negative bacterial proteins[J]. **Journal of Theoretical Biology**, 2010, 264: 326-333.
- [16] SHEN H B, CHOU K C. Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of gram-positive bacterial proteins[J]. **Protein and Peptide Letters**, 2009, 16: 1478-1484.
- [17] ZHAO Nan, ZHANG Liang, XUE Wei, et al. Application of bag of words model in the prediction of protein subcellular location[J]. **Journal of Food Science and Biotechnology**, 2017, 36(3): 296-301. (in Chinese)
- [18] NAKASHIMA H, NISHIKAWA K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies[J]. **Journal of Molecular Biology**, 1994, 238(1): 54-61.
- [19] ZHANG Jing, DUAN Fu. Improved k-means algorithm with meliorated initial centers[J]. **Computer Engineering and Design**, 2013, 34(5): 1691-1694.
- [20] ZHANG M L, Peña J M, ROBLES V. Feature selection for multi-label naive bayes classification[J]. **Information Sciences and International Journal**, 2009, 179(19): 3218-3229.