

一种新的蛋白质亚细胞定位预测训练集构造方法

曹隽喆¹, 顾宏^{*1}, 贺建军²

(1. 大连理工大学 控制科学与工程学院, 辽宁 大连 116024;

2. 大连民族学院 信息与通信工程学院, 辽宁 大连 116602)

摘要: 设计了一种新的蛋白质亚细胞定位预测训练集构造方法. 该方法针对传统预测方法缺乏足够的实验标记数据的问题, 基于主动学习策略从非实验标记蛋白质数据中主动选择有效数据, 并与原有的实验标记数据共同训练预测模型, 以提高基准分类器的预测精度. 结合支持向量机分类器, 该方法在病毒蛋白质独立测试集上进行了预测实验, 测试结果表明, 该方法能够有效地提高基准分类器的预测能力, 性能优于现有的病毒蛋白质预测系统.

关键词: 生物信息学; 机器学习; 蛋白质亚细胞定位; 非实验标记数据; 主动学习
中图分类号: TP181 **文献标志码:** A

0 引言

蛋白质亚细胞定位是蛋白质组学研究和蛋白质功能研究的重要内容, 它在推断蛋白质生物学功能, 验证蛋白质相互作用以及认识疾病机理等方面具有非常重要的实际意义和研究价值. 随着基因组学和蛋白质组学研究的不断深入, 新的蛋白质序列数量急剧膨胀, 传统人工实验方法的高成本、低效率的缺点日趋明显, 因此利用计算方法进行蛋白质亚细胞定位预测以其方便快捷并且成本较低的特点, 越来越受到人们的关注, 目前已成为生物信息学的一个重要研究热点.

基于计算方法的蛋白质亚细胞定位预测算法最早出现在 1991 年, 由 Nakai 等^[1]提出, 他们利用 N 端分选信号设计 IF-THEN 规则对 4 类革兰阴性菌蛋白质位置进行亚细胞定位预测, 并取得 83% 的精度. 此后, 许多蛋白质亚细胞定位预测方法陆续出现, 这些方法主要的不同之处在于: (1) 蛋白质序列特征的提取方式不同, 主要包括利用分选信号信息^[2]、氨基酸序列信息^[3]、模体信息^[4]和 GO 注释信息^[5]等; (2) 设计的预测模型不同, 主要包括 K 近邻^[6]、支持向量机^[7]、神经网络^[8]和贝叶斯方法^[9]等. 尽管许多计算方法已经

被用来预测如人类、植物、细菌、真菌等不同物种的蛋白质亚细胞位置, 并在某些蛋白质数据集上取得很不错的效果, 预测的精度甚至超过了一些实验方法, 但遗憾的是, 针对病毒的预测方法目前却很少. 病毒作为一种特殊的寄生性非细胞型微生物, 其本身没有完整的细胞结构, 其蛋白质亚细胞定位是指要确定病毒蛋白质在宿主的亚细胞位置. 由于病毒蛋白质转运到宿主细胞的亚细胞位置不同, 其对被感染细胞的影响和破坏性也有所不同, 因此研究病毒蛋白质的亚细胞定位信息对于研究病毒蛋白质功能、作用机理和药物研制等方面具有重要的意义.

与其他物种相比, 目前国际上对病毒蛋白质的亚细胞定位预测方法研究较晚, 2007 年第一个针对病毒蛋白质的预测器 Virus-PLoc^[10]由 Shen 等提出, 他们结合伪氨基酸序列组分和 GO 注释信息, 利用优化证据理论 K 近邻 (optimized evidence-theoretic KNN) 分类算法对病毒蛋白质进行了预测, 但该方法只能预测单定位点蛋白质亚细胞位置. Shen 等^[11]在 Virus-PLoc 的基础上提出了可以预测多定位点病毒蛋白质亚细胞位置的预测系统 Virus-mPLoc, 并用留一法测试得到

收稿日期: 2012-03-15; 修回日期: 2012-09-10.

基金项目: 国家自然科学基金资助项目(61174027); 国家科技重大专项资助项目(2010ZX04007-011-5); 辽宁省自然科学基金资助项目(20102025).

作者简介: 曹隽喆(1984-), 男, 博士生; 顾宏^{*}(1961-), 男, 教授, 博士生导师, E-mail: guhong@dlut.edu.cn.

60.3%的预测成功率. Xiao 等^[12]提出了基于多标签 K 近邻算法的病毒蛋白质预测系统 iLoc-Virus, 将同一个数据集的预测精度提高到 78.2%, 预测成功率依然不高. 上述预测系统提供给分类器进行学习的数据均来自于实验方法标记的蛋白, 而能用于训练预测模型的此类病毒蛋白质数量却很少, 这势必造成预测效果不佳. 本文针对上述问题, 利用主动学习策略在非实验标记蛋白质数据中挑选可以作为训练数据的具有较高可信度的数据, 以弥补训练集不足的问题, 并结合支持向量机分类模型对病毒蛋白质的 5 个亚细胞位置进行预测.

1 实验数据

目前蛋白质亚细胞定位预测所用的数据主要来自于著名的蛋白质数据库 UniProt/SwissProt, 人们通常利用该数据库中具有亚细胞注释信息并且是由实验方法标记的蛋白质来构造数据集. 但是数据库中可用于帮助预测的病毒蛋白质的数量却很少, 文献^[11,12]所用的病毒蛋白质数据集只有 207 个训练数据, 且随着数据库的更新其中的一些数据已经被删改不合适继续使用. 其实, 在 UniProt 数据库中还存在另一类具有亚细胞注释信息的蛋白质, 它们的注释信息是由非实验方法得到的, 并且数量十分可观. 在 UniProt 数据库 (2012-01-25 发布) 具有亚细胞位置注释的全部 313 678 个蛋白质数据中, 具有实验标记信息的数据有 70 552 个, 而具有非实验标记信息的数据则有 243 126 个, 约占全部数据的 77.5%. 虽然这类蛋白质的亚细胞注释信息没有经过实验验证, 存在一定的不准确性, 但是这类注释也是根据大量有力的非实验性证据给出的, 因而大部分数据还是具有较高的可信度. 如果能够将如此多的非实验标记数据加以充分利用, 无疑可以弥补训练数据不足的问题.

在本文中, 选取文献^[11,12]所用的数据集, 并进一步根据最新版本的 UniProt 数据库 (2012-01-25) 的注释信息将病毒蛋白质进行重新筛选和标记, 删除和更正了原数据集中一部分有误的数据, 得到 197 个病毒蛋白质数据组成基本训练集. 又从同一数据库中收集了 236 个非实验标记病毒蛋白质数据作为备挑选的辅助训练集. 另外, 另行

收集了 188 个新的实验标记病毒蛋白质作为独立测试集, 共包括 5 个亚细胞位置, 在该测试集中, 只含有一个亚细胞位置的蛋白质有 158 个, 同时含有两个位置的有 28 个, 同时含有 3 个位置的有 2 个, 没有含有 4 个或 4 个以上位置的蛋白质. 为了避免同源性偏差, 在收集上述新的实验标记和非实验标记病毒蛋白质时, 全部蛋白质均使用在线序列相似性比对工具 PISCES^[13]进行了处理, 以保证所用到的全部数据中任意两个蛋白质样本的序列相似性都不超过 25%. 蛋白质数据在各亚细胞位置数量分布如表 1 所示.

表 1 病毒蛋白质数据集的分布

Tab. 1 Distribution of the viral protein datasets

标记序号	亚细胞位置	蛋白质数量		
		基本训练集	辅助训练集	独立测试集
1	宿主细胞膜	36	11	76
2	宿主内质网	18	18	35
3	宿主细胞质	87	116	51
4	宿主细胞核	91	113	53
5	分泌蛋白	21	29	5
数据集中蛋白质数据总数		197	236	188
数据集中亚细胞位置总数		253	287	220

2 原理和方法

2.1 蛋白质序列特征信息的提取

对于蛋白质序列特征提取, 本文采用被广泛使用的伪氨基酸序列组分分析 (PseAAC) 方法^[14], 该方法将氨基酸残基的物理化学性质与氨基酸序列按一定排序方法结合起来以表示蛋白质序列特征. 使用 PseAAC 网络服务器提供的 Type2 型伪氨基酸组分, 选取其中的亲水性、疏水性和侧链相对分子质量 3 种物化性质提取特征, 得到了蛋白质序列的如下序列特征表示:

$$\mathbf{x}_k = (x_{k,1} \quad x_{k,2} \quad \cdots \quad x_{k,20} \quad x_{k,21} \quad \cdots \quad x_{k,20+3\lambda})$$

(1)

式中

$$x_{k,n} = \begin{cases} \frac{f_n}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{3\lambda} \tau_j}; & 1 \leq n \leq 20 \\ \frac{w\tau_n}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{3\lambda} \tau_j}; & 20+1 \leq n \leq 20+3\lambda \end{cases}$$

(2)

式中: $f_n(n = 1, 2, \cdots, 20)$ 为蛋白质序列的 20 个固有氨基酸的出现频率, τ_j 为第 j 个序列相关因子, ω 为权重因子, 文献 [14] 中详细介绍了 PseAAC 的具体用法及参数. 最终对样本 $(\mathbf{x}_k \ \mathbf{y}_k)$ 提取到的蛋白质特征 $\mathbf{x}_k$ 为一个 $20 + 3\lambda$ 维的向量, 其对应的亚细胞位置标记为 $\mathbf{y}_k = (y_{k1} \ y_{k2} \ \cdots \ y_{km})^\top$, 其中 m 为病毒蛋白质亚细胞位置的种类个数, 这里 $m = 5$, 若该蛋白质具有第 t 个亚细胞位置, 则对应的 $y_{kt} = 1$, 否则 $y_{kt} = -1$.

2.2 基于主动学习技术的非实验标记蛋白质挑选

主动学习是一种利用未标记标签信息的示例进行学习的方法, 其算法流程是: (1) 利用已标记样本训练一个基准分类器; (2) 为未标记示例制定一个反映其信息量的评估函数, 根据评估函数从未标记样本中主动选择信息量大的样本, 并对其给出合理的标签; (3) 将选出的未标记样本添加到训练集中, 并在新的训练集上对分类器进行更新以完成一次循环; (4) 多次进行该循环直至满足预设条件为止, 从而挑选出对分类最有价值的那部分未标记样本以改善分类器的性能.

本文使用上述主动学习的思想从非实验标记的蛋白质数据中选取对预测最有帮助的那部分数据来帮助训练分类器, 但备选对象不是未标记样本而是非实验标记样本. 非实验标记的蛋白质数据与普通的未标记样本有很大区别, 它并不是完全没有标签信息, 而是它的标签信息可能不准确, 但是仍然具有很高的可信性和参考性, 因此这里需要建立一种新的评估函数, 使得可以利用到样本的非实验标记信息. 以最小最大化主动学习策略为基础, 假设基准分类器模型为 f , 则对一个非实验标记样本 $\mathbf{x}_s$ 设计如下的评估函数:

$$E(\mathbf{D}_l, \mathbf{x}_s) = \max_{\mathbf{y}_s} \arg \min_{f \in \mathcal{H}} \left(\frac{\epsilon}{2} \|\mathbf{f}\|_{\mathcal{H}}^2 + \sum_{\mathbf{x}_k \in \mathbf{D}_l} L_1(\mathbf{y}_k, f(\mathbf{x}_k)) + L_2(\hat{\mathbf{y}}_s, \mathbf{y}_s, f(\mathbf{x}_s)) \right) \tag{3}$$

该评估函数考察了将非实验样本 $\mathbf{x}_s$ 加入训练集后, 对分类器模型复杂度和风险误差造成的影响, 估计值越小说明对分类模型的帮助越大. 其中, $\mathbf{D}_l$ 是基本训练集, $|\mathbf{D}_l| = n_l; \mathbf{y}_k = (y_{k1} \ y_{k2} \ \cdots$

$y_{km})^\top$ 表示训练样本 $\mathbf{x}_k$ 的标签; $\hat{\mathbf{y}}_s = (\hat{y}_{s1} \ \hat{y}_{s2} \ \cdots \ \hat{y}_{sm})^\top$ 和 $\mathbf{y}_s = (y_{s1} \ y_{s2} \ \cdots \ y_{sm})^\top$ 表示非实验标记样本 $\mathbf{x}_s$ 的已知标签和模型预测标签; $\frac{\epsilon}{2} \|\mathbf{f}\|_{\mathcal{H}}^2$ 是正则化项, $\mathbf{H}$ 表示一个由核函数 $k^f(\cdot, \cdot)$ 确定的再生核希尔伯特空间; $\sum_{\mathbf{x}_k \in \mathbf{D}_l} L_1(\mathbf{y}_k, f(\mathbf{y}_k))$ 为 f 在训练样本集上的损失, $L_2(\hat{\mathbf{y}}_s, \mathbf{y}_s, f(\mathbf{x}_s)) = \frac{1}{2} \sum_{j=1}^m p(y_{sj} | \mathbf{x}_s, \hat{y}_{sj}) (y_{sj} - f_j(\mathbf{x}_s))^2$ 表示 f 在非实验标记样本 $\mathbf{x}_s$ 上的损失, 损失越小意味着这个样本对由训练样本确定的分类模型的帮助越大. 其中权重系数 $p(y_{sj} | \mathbf{x}_s, \hat{y}_{sj})$ 表示当 $\mathbf{x}_s$ 具有非实验标记 $\hat{y}_{sj}$ 时, 其真实标签为 y_{sj} 的概率, 可以由非实验标记样本的先验知识和训练数据共同计算得到, 它正好将非实验标记样本的标签信息引入到了评估函数中. 在估计后验概率 $p(y_{sj} | \mathbf{x}_s, \hat{y}_{sj})$ 时, 采用 Parzen-Window 密度估计方法, 令 $\mathbf{E}_1 = \{\mathbf{x}_p | \mathbf{x}_p \in \mathbf{D}_l, y_{pj} = 1\}$, $\mathbf{E}_2 = \{\mathbf{x}_q | \mathbf{x}_q \in \mathbf{D}_l, y_{qj} = -1\}$, $N_1 = |\mathbf{E}_1|$, $N_2 = |\mathbf{E}_2|$, 则有

$$p(\mathbf{x}_s | y_{sj} = 1, \hat{y}_{sj}) = \frac{1}{N_1} \sum_{\mathbf{x}_p \in \mathbf{E}_1} \phi_1(\mathbf{x}_s) \tag{4}$$

$$\phi_1(\mathbf{x}_s) = \frac{1}{\alpha \sqrt{2\pi}} \exp \left\{ -\frac{(\mathbf{x}_s - \mathbf{x}_p)^\top (\mathbf{x}_s - \mathbf{x}_p)}{2\alpha} \right\} \tag{5}$$

α 为固定的参数, 同样的方法可以计算

$$p(\mathbf{x}_s | y_{sj} = -1, \hat{y}_{sj}) = \frac{1}{N_2} \sum_{\mathbf{x}_q \in \mathbf{E}_2} \phi_2(\mathbf{x}_s) \tag{6}$$

因此根据贝叶斯定理后验概率估计 $p(y_{sj} | \mathbf{x}_s, \hat{y}_{sj})$ 定义为

$$p(y_{sj} | \mathbf{x}_s, \hat{y}_{sj}) = p(\mathbf{x}_s | y_{sj} = 1, \hat{y}_{sj}) \times p(y_{sj} = 1 | \hat{y}_{sj}) / [p(\mathbf{x}_s | y_{sj} = 1, \hat{y}_{sj}) \times p(y_{sj} = 1 | \hat{y}_{sj}) + p(\mathbf{x}_s | y_{sj} = -1, \hat{y}_{sj}) \times p(y_{sj} = -1 | \hat{y}_{sj})] \tag{7}$$

由于非实验标记信息按可信程度由高到低分为 Probable、Potential 和 By Similarity 3 类, 先验概率 $p(y_{sj} = 1 | \hat{y}_{sj})$ 作为估计参数, 其大小与非实验标记的类别有关, 即 Probable 标记的先验概率最大, Potential 次之, By Similarity 最小.

显然,有了上述的评估函数之后,最优的非实验标记样本应满足

$$\mathbf{x}_s^* = \arg \min_{\mathbf{x}_s \in \mathbf{D}_u} |f^*(\mathbf{x}_s)| \tag{8}$$

其中

$$\begin{aligned} f^* = \arg \min_{f \in \mathbf{H}} & \left(\frac{\epsilon}{2} |f|_{\mathbf{H}}^2 + \frac{1}{2} \sum_{i=1}^{n_l} \sum_{j=1}^m (y_{ij} - \right. \\ & \left. f_{ij}(\mathbf{x}_i))^2 + \frac{1}{2} \sum_{j=1}^m p(y_{sj} | \mathbf{x}_s, \hat{y}_{sj}) \times \right. \\ & \left. (y_{sj} - f_{sj}(\mathbf{x}_s))^2 \right) \end{aligned} \tag{9}$$

因此计算估计函数的值实际上就是一个最小最大化最优问题:

$$\begin{aligned} \min_{\mathbf{x}_s \in \mathbf{D}_u} \max_{y_s} \min_{f \in \mathbf{H}} & \left(\frac{\epsilon}{2} |f|_{\mathbf{H}}^2 + \frac{1}{2} \sum_{i=1}^{n_l} \sum_{j=1}^m (y_{ij} - \right. \\ & \left. f_{ij}(\mathbf{x}_i))^2 + \frac{1}{2} \sum_{j=1}^m p(y_{sj} | \mathbf{x}_s, \hat{y}_{sj}) \times \right. \\ & \left. (y_{sj} - f_{sj}(\mathbf{x}_s))^2 \right) \end{aligned} \tag{10}$$

正则项中的 $|f|_{\mathbf{H}}^2 = \mathbf{f}^T \mathbf{K}^{-1} \mathbf{f}$,故由计算可知

$$\begin{aligned} \min_{f \in \mathbf{H}} & \left(\frac{\epsilon}{2} |f|_{\mathbf{H}}^2 + \frac{1}{2} \sum_{i=1}^{n_l} \sum_{j=1}^m (y_{ij} - \right. \\ & \left. f_{ij}(\mathbf{x}_i))^2 + \frac{1}{2} \sum_{j=1}^m p(y_{sj} | \mathbf{x}_s, \hat{y}_{sj}) \times \right. \\ & \left. (y_{sj} - f_{sj}(\mathbf{x}_s))^2 \right) = \frac{1}{2} \mathbf{Y}^T \mathbf{L} \mathbf{Y} \end{aligned} \tag{11}$$

其中 $\mathbf{Y} = (\mathbf{y}_1 \ \mathbf{y}_2 \ \cdots \ \mathbf{y}_{n_l} \ \mathbf{y}_s)^T, \mathbf{L} = (\epsilon^{-1} \mathbf{K} +$

$$\mathbf{A}^{-1})^{-1}, \mathbf{K} = \begin{pmatrix} k_{1,1} \mathbf{I} & k_{1,2} \mathbf{I} & \cdots & k_{1,n_l+1} \mathbf{I} \\ k_{2,1} \mathbf{I} & k_{2,2} \mathbf{I} & \cdots & k_{2,n_l+1} \mathbf{I} \\ \vdots & \vdots & & \vdots \\ k_{n_l+1,1} \mathbf{I} & k_{n_l+1,2} \mathbf{I} & \cdots & k_{n_l+1,n_l+1} \mathbf{I} \end{pmatrix},$$

$\mathbf{I}$ 为 $m \times m$ 的单位矩阵, $k_{i,j}$ 为核矩阵 $\tilde{\mathbf{K}} = (k^f(\mathbf{x}_i, \mathbf{x}_j))_{(n_l+1) \times (n_l+1)}$ 中第 i 行第 j 列对应的元素,

$$\mathbf{A} = \begin{pmatrix} \mathbf{I} & & & \\ & \ddots & & \\ & & \mathbf{I} & \\ & & & \mathbf{P} \end{pmatrix},$$

$$\mathbf{P} = \begin{pmatrix} p(y_{s1} | \mathbf{x}_s, \hat{y}_{s1}) & & & \\ & p(y_{s2} | \mathbf{x}_s, \hat{y}_{s2}) & & \\ & & \ddots & \\ & & & p(y_{sm} | \mathbf{x}_s, \hat{y}_{sm}) \end{pmatrix}.$$

令 $\mathbf{L} = \begin{pmatrix} \mathbf{L}_{ll} & \mathbf{L}_{ls} \\ \mathbf{L}_{sl} & \mathbf{L}_{ss} \end{pmatrix}$, 则

$$\mathbf{Y}^T \mathbf{L} \mathbf{Y} = \mathbf{Y}_l^T \mathbf{L}_{ll} \mathbf{Y}_l + \mathbf{y}_s^T \mathbf{L}_{ss} \mathbf{y}_s + \mathbf{y}_s^T \mathbf{L}_{sl} \mathbf{Y}_l + \mathbf{Y}_l^T \mathbf{L}_{ls} \mathbf{y}_s \tag{12}$$

式(12)中估计值的计算与分类器 f 无关,并且除了 $\mathbf{y}_s$ 外的各部分均可由已知条件算出,因此只需代入每个可能的 $\mathbf{y}_s$,就可得到所有的 $\mathbf{Y}^T \mathbf{L} \mathbf{Y}$,从而求出 $\mathbf{x}_s$ 的估计值 $\mathbf{E}(\mathbf{D}_l, \mathbf{x}_s) = \max_{y_s} \mathbf{Y}^T \mathbf{L} \mathbf{Y}$. 利用上述方法对全部非实验数据进行评估并按评估值排序,则可按需要选出评估值最小的那部分数据加入训练集以帮助训练分类器.

3 结果与讨论

本文选择支持向量机(SVM)作为基准分类器,用 MATLAB7.0 软件和 LibSVM 支持向量机工具包进行数值实验以观察本文提出的主动学习样本选择的效果. 由于有部分病毒蛋白质同时存在多个亚细胞位置,该预测问题从机器学习的观点来看是一种多标记学习问题,所以采取“一对其余”的方法将 5 个二分类支持向量机组合起来以处理该多标记分类问题. 在实验中,选择 PseAAC 特征提取的参数为 $\lambda=5, w=0.2$; 主动学习的评估函数参数是 $\epsilon=1, \alpha=0.5$, Probable、Potential 和 By Similarity 3 类非实验标记对应的先验概率依次设为 0.9、0.8 和 0.7; 支持向量机的核函数选用径向基函数(RBF),核函数参数采用 LibSVM 工具包默认值. 表 2 给出了利用该主动学习方法从辅助训练集选取不同比例的最优非实验标记蛋白质样本加入基本训练集后,分类器在独立测试集上的预测成功率.

从表 2 中可以看出,当开始使用挑选出的最优样本参与训练分类器时,分类器的总体预测成功率逐渐增大,当挑选出的样本的比例达到 40% 时达到最高,这说明由于增加了训练集的规模,使得分类器学习到了更丰富的信息,预测效果越来越好;而当挑选出的样本的比例再继续增大时,总体预测成功率却逐渐下降,这是因为非实验标记样本的亚细胞位置信息存在不准确性,当信息较为准确的那部分最优数据加入训练集之后,再继续挑选的就是注释信息不太准确的非实验标记样本了,这样随着越来越多的不准确样本逐渐参与训练分类器,预测成功率也就随之下降了. 因此,可以看出本文提出的方法并不是仅仅依靠训练数据数量

上的增加,而是确实能够将对提高预测精度最有帮助的数据挑选出来从而提高分类器的性能.

进一步地,将本文提出的主动学习方法(挑选比例为 40%)与独立的 SVM 基准分类器以及两个最新的预测系统 Virus-mPLoc^[11] 和 iLoc-Virus^[12]在独立测试集上的表现进行了对比,具

体表现如表 3 所示.从表中可以看出本文的方法在除了 Hamming loss 之外其余的全部评价指标都优于 Virus-mPLoc 和 iLoc-Virus 预测系统,且全部指标都好于 SVM 基础分类器,这说明该方法能有效地对病毒蛋白质亚细胞位置进行高精度的预测,具有很强的实用性.

表 2 本文的主动学习方法在病毒蛋白质独立测试集上的预测成功率

Tab. 2 Prediction success rates of the proposed active learning approach on the independent dataset

亚细胞位置	利用主动学习方法挑选出最优样本占全部非实验标记病毒蛋白质样本的比例										
	0	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
宿主细胞膜	0.618 4	0.618 4	0.618 4	0.618 4	0.618 4	0.605 3	0.579 0	0.565 8	0.513 2	0.460 5	0.421 1
宿主内质网	0.228 6	0.285 7	0.285 7	0.285 7	0.285 7	0.285 7	0.285 7	0.228 6	0.200 0	0.200 0	0.200 0
宿主细胞质	0.568 6	0.588 2	0.666 7	0.686 3	0.686 3	0.686 3	0.666 7	0.705 9	0.705 9	0.647 1	0.627 5
宿主细胞核	0.547 2	0.547 2	0.547 2	0.547 2	0.584 9	0.603 8	0.622 6	0.584 9	0.584 9	0.603 8	0.660 4
分泌蛋白	0.800 0	0.800 0	0.800 0	0.800 0	0.800 0	0.800 0	0.800 0	0.800 0	0.800 0	0.600 0	0.600 0
总计	0.531 8	0.545 5	0.563 6	0.568 2	0.577 3	0.577 3	0.568 2	0.554 5	0.531 8	0.500 0	0.513 6

表 3 本文的主动学习方法和其他几种方法的表现

Tab. 3 Performances of the proposed active learning approach and some other methods

方法	I				
	Success rate	Hamming loss	One-error	Coverage	Ranking loss
Virus-mPLoc	0.540 9	0.187 1	0.377 7	1.659 6	0.468 1
iLoc-Virus	0.550 0	0.166 7	0.404 3	1.611 7	0.466 4
SVM	0.531 8	0.232 3	0.398 9	1.739 4	0.505 0
本文方法	0.577 3	0.218 8	0.356 4	1.579 8	0.458 7

注:Success rate 越大性能越好,其他 4 个指标越小性能越好

4 结 语

本文根据主动学习策略提出一个新的蛋白质亚细胞定位预测训练集构造方法,并通过实验证明了该方法的合理性和实用性.基于主动选择思想设计的评估函数能有效地对非实验标记样本的价值进行评估,从而挑选出最能够帮助改善分类器性能的那部分数据加入原训练集中,以弥补缺乏训练数据的不足.

值得注意的是,本文设计的主动学习方法并不依赖基准分类器的类型,实验中选择一个较为简单的支持向量机分类器就已经能够有较好的表现,如果基准分类器的性能更出色,会使得预测效果进一步提升,因此该方法具有很大的发展潜力.另外,本文主要针对的是病毒蛋白质数据集,对于其他数据较少物种的蛋白质定位预测问题,该方法也有进一步研究的价值.

参考文献:

[1] 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.

[2] Emanuelsson O, Nielsen H, Brunak S, *et al.* Predicting subcellular localization of proteins based on their N-terminal amino acid sequence [J]. **Journal of Molecular Biology**, 2000, **300**(6):1005-1016.

[3] Khan A, Majid A, Hayat M. CE-PLoc: An ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition [J]. **Computational Biology and Chemistry**, 2011, **35**(4):218-229.

[4] Lin Tien-ho, Murphy R, Ziv B. Discriminative motif finding for predicting protein subcellular

- localization [J]. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, 2011, **8**(2):441-451.
- [5] Chou Kuo-chen, CAI Yu-dong. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology [J]. **Biochemical and Biophysical Research Communications**, 2003, **311**(3):743-747.
- [6] SHEN Hong-bin, Chou Kuo-chen. Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites [J]. **Biochemical and Biophysical Research Communications**, 2007, **355**(4):1006-1011.
- [7] HUA Su-jun, SUN Zhi-rong. Support vector machine approach for protein subcellular localization prediction [J]. **Bioinformatics**, 2001, **17**(8):721-728.
- [8] CAI Yu-dong, ZHOU Guo-ping. Prediction of protein structural classes by neural network [J]. **Biochimie**, 2000, **82**(8):783-785.
- [9] Briesemeister S, Rahnenfuhrer J, Kohlbacher O. Going from where to why — interpretable prediction of protein subcellular localization [J]. **Bioinformatics**, 2010, **26**(9):1232-1238.
- [10] SHEN Hong-bin, Chou Kuo-chen. Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells [J]. **Biopolymers**, 2007, **85**(3):233-240.
- [11] SHEN Hong-bin, Chou Kuo-chen. Virus-mPLoc: A fusion classifier for viral protein subcellular location prediction by incorporating multiple sites [J]. **Journal of Biomolecular Structure & Dynamics**, 2010, **28**(2):175-186.
- [12] XIAO Xuan, WU Zhi-cheng, Chou Kuo-chen. iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites [J]. **Journal of Theoretical Biology**, 2011, **284**(1):42-51.
- [13] WANG Guo-li, Dunbrack R. PISCES: a protein sequence culling server [J]. **Bioinformatics**, 2003, **19**(12):1589-1591.
- [14] Chou Kuo-chen. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology [J]. **Current Proteomics**, 2009, **6**(4):262-274.

A novel method of training set construction for predicting protein subcellular localization

CAO Jun-zhe¹, GU Hong^{*1}, HE Jian-jun²

(1. School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China;

2. College of Information & Communication Engineering, Dalian Nationalities University, Dalian 116602, China)

Abstract: A novel method of training set construction for predicting protein subcellular localization is designed. The proposed approach aims at the shortage of valid data with experimental annotations in the traditional methods, and employs active learning strategy to actively select some effective data with non-experimental annotations which train the prediction model together with the original data, and also improves the prediction accuracy of the benchmark classifier. Combined with a support vector machine classifier, the approach is tested on an independent testing dataset of viral proteins. The test results indicate that the presented method can effectively improve the predicting capability of the benchmark classifier and can achieve a better performance than the existing predicting systems for viral proteins.

Key words: bioinformatics; machine learning; protein subcellular localization; data with non-experimental annotations; active learning