中国科学: 信息科学 2017 年 第 47 卷 第 11 期: 1538–1550
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
机器学习专题 · 论文

基于多示例多标记迁移学习的蛋白质功能预测

胡海峰[1,2], 郑茂[1], 吴伟坚[3], 王俊[4], 吴建盛[4*]

1. 南京邮电大学通信与信息工程学院, 南京 210003
2. 南京信息工程大学江苏省优势学科、江苏省大气环境与装备技术协同创新中心, 南京 210044
3. 河海大学计算机与信息学院, 南京 211100
4. 南京邮电大学地理与生物信息学院, 南京 210023
* 通信作者. E-mail: jansen@njupt.edu.cn

收稿日期: 2017–05–14; 接受日期: 2017–06–30; 网络出版日期: 2017–11–13
国家自然科学基金 (批准号: 61571233, 61271082)、国家重点基础研究发展计划 (973) (批准号: 2011CB302903)、江苏省高校自然科学研究重大项目 (批准号: 14KJA510003)、江苏省重点研发计划 (批准号: BE2015700) 和南京信息工程大学 PAPD 与 CICAEET 资助项目

摘要 随着各种基因组测序计划的推出, 不断有很多物种被新测序完成, 需要对这些物种的蛋白质功能进行注释. 这些物种中已知功能的蛋白质数量少, 可以考虑使用亲缘关系近、已知功能蛋白质数量多的物种来帮助这些物种进行蛋白质功能预测. 本文把这个任务抽象为多示例多标记迁移学习问题, 并提出了第一个多示例多标记迁移学习框架 TR-MIML 来解决此任务. TR-MIML 通过最小化投影空间上加权源域样本中心点与目标域样本中心点的距离, 给源域样本赋予不同权值, 并基于目标域和源域样本训练多示例多标记学习模型. 在两个新完成测序物种上, 实验结果证明了迁移学习有助于它们的蛋白质功能预测. 另外, 亲缘关系越近的物种作为源域进行迁移学习帮助越大.

关键词 新测序物种, 蛋白质功能预测, 迁移学习, 多示例多标记学习, 样本加权

1 引言

随着高通量测序技术的发展和各种基因组测序计划的推出, 不断有很多物种被测序完成, 对这些新完成测序物种的蛋白质功能进行注释非常重要也很迫切. 生物实验方法可以准确得到蛋白质的生物学功能, 但耗时耗力, 很难应用于这种大规模的蛋白质功能预测工作中. 近年来, 计算学方法已成功应用于蛋白质的生物学功能及相关的预测工作中, 其中基于机器学习的方法更是获得了很好的预测性能 [1,2], 有基于传统二分类方法的 [3~5], 也有基于多标记学习方法的 [6~9].

新完成测序物种中已知功能蛋白质数量少, 很难利用传统机器学习方法构建一个好的预测模型. 我们发现, 目前已有很多物种, 其蛋白质功能注释信息丰富. 我们是否可以使用在进化上亲缘关系接

引用格式: 胡海峰, 郑茂, 吴伟坚, 等. 基于多示例多标记迁移学习的蛋白质功能预测. 中国科学: 信息科学, 2017, 47: 1538–1550, doi: 10.1360/N112017-00090
Hu H F, Zheng M, Wu W J, et al. Protein function prediction through multi-instance multi-label transfer learning (in Chinese). Sci Sin Inform, 2017, 47: 1538–1550, doi: 10.1360/N112017-00090

近、蛋白质功能注释丰富的物种来帮助新完成测序物种的蛋白质功能预测? 不同物种的蛋白质样本数据分布往往不一样, 如果直接使用亲缘物种的蛋白质样本数据, 利用传统机器学习方法, 将不满足机器学习训练样本和测试样本要求的 "独立同分布假设", 很难得到好的新完成测序物种蛋白质功能预测模型.

近年来, 迁移学习作为一种新的机器学习方法, 得到了广泛的关注和研究 [10]. 迁移学习是利用一个领域已有的知识帮助解决不同但相关领域问题的机器学习方法. 它无需完全局限于传统机器学习中的两个基本假设: (1) 用于学习的训练数据与新的测试数据满足独立同分布的条件; (2) 需要有足够的训练样本才可以得到不错的分类模型 [10,11]. 迁移学习已成功应用到蛋白质功能的预测任务中. Mei 等 [12] 提出了一种利用同源蛋白质进行迁移学习的方法来预测蛋白质的生物学功能. 最近, Xu 等 [13] 提出了一种多示例度量迁移学习方法, 将每个物种自身的蛋白质样本划分为目标域和源域, 并通过优化源域样本权值来使源域与目标域样本分布一致, 并构建距离度量学习模型进行基因组水平的蛋白质功能预测. 虽然迁移学习已成功应用于蛋白质功能预测任务中, 但由于新完成测序物种中已知功能的蛋白质样本数量少, 目前还没有看到迁移学习应用于新完成测序物种基因组水平蛋白质功能预测方面工作的报道.

在以前的研究中, 蛋白质功能预测被抽象为多示例多标记学习问题 [14], 即一个蛋白质样本往往包含多个结构域 (示例), 同时具有多种 GO 生物学功能 (标记). 本文把利用亲缘物种来帮助新完成测序物种的蛋白质功能预测任务抽象为多示例多标记迁移学习问题, 并提出了第一个多示例多标记迁移学习框架 (TR-MIML) 来解决此任务. TR-MIML 学习框架首先将源域和目标域数据集中的多示例样本转化为单示例样本; 然后, 最小化投影空间上加权源域样本中心点与目标域样本中心点的距离, 给源域样本赋予不同权值; 最后, 基于目标域样本和源域样本, 得到多示例多标记学习模型.

在实验中, 以目前最好的 3 种多示例多标记学习算法 (MIMLfast [15], MIMLNN [16], MIML-SVM [16,17]) 为基分类器来各自实现本文的多示例多标记迁移学习算法. 以两个新完成测序物种 (Geobacter sulfurreducens 和 Azotobacter vinelandii) 作为目标域, 以 3 种已知功能蛋白质数量多的模式生物 (Rattus norvegicus, Mus musculus 和 Saccharomyces cerevisiae) 作为源域, 证明了迁移学习有助于新完成测序物种的蛋白质功能预测. 再以物种 Rattus norvegicus 作为目标域, 以亲缘关系不同的 5 种物种作为源域, 证明了利用亲缘关系越近的物种进行迁移学习越有助于目标域物种的蛋白质功能预测.

总之, 本文的主要贡献如下: (1) 把利用亲缘物种来帮助新完成测序物种的蛋白质功能预测任务抽象为多示例多标记迁移学习问题, 并提出了第一个多示例多标记迁移学习框架来解决此任务; (2) 实验证明了迁移学习有助于新完成物种的蛋白质功能预测, 而且亲缘关系越近的物种帮助越大.

2 方法

2.1 问题抽象

本文为了解决目前大量新完成测序物种蛋白质功能预测问题, 拟利用生命系统发生树上蛋白质功能注释信息丰富的亲缘物种, 解决新完成测序物种已知功能蛋白质样本不足的问题, 得到好的蛋白质功能预测模型. 迁移学习为解决此问题提供了一条很好的途径, 新完成测序物种可认为是目标域, 而亲缘物种可认为是源域.

在以前的研究中, 蛋白质功能预测被抽象为多示例多标记学习问题 [14], 即一个蛋白质样本往往包含多个结构域 (示例), 同时具有多种 GO 生物学功能 (标记). 对新完成测序物种, 即目标域, 假

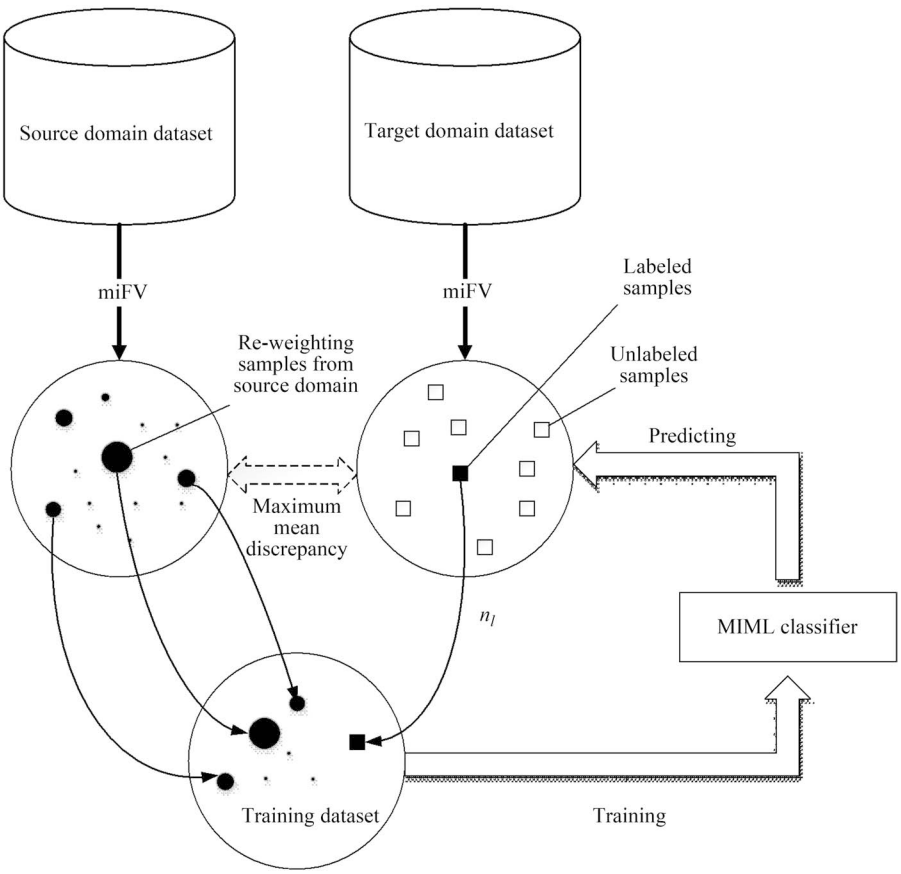


图 1　多示例多标记迁移学习框架 (TR-MIML), 包括源域样本加权阶段和分类模型学习阶段

Figure 1　Multi-instance multi-label transfer learning framework (TR-MIML), including the re-weighting data samples from source domain stage and the classification model construction stage

设 χ^T 代表其示例空间而 Y^T 代表其标记空间. 目标域包含很多未标记样本 $D_u^T = X_i^T|_{i=1}^{n_u}$, 还可能有少部分标记样本 $D_l^T = (X_i^T, Y_i^T)|_{i=1}^{n_l}$, n_u 和 n_l 分别表示目标域中标记样本和未标记样本的个数. 那么, 目标域数据集 D^T 可表示为 $D_l^T \cup D_u^T$, 而且 $n_T = n_l + n_u$. 在此, $X_i^T \subseteq \chi^T$ 为目标域中第 i 个蛋白质样本, 含有 z_i 个示例 $\{x_1^{(i)}, x_2^{(i)}, \ldots, x_{z_i}^{(i)}\}$; 而 $Y_i^T \subseteq Y^T$ 为 X_i^T 所对应的一组 GO 标记 $\{y_1^{(i)}, y_2^{(i)}, \ldots, y_{l_i}^{(i)}\}$. 对亲缘物种, 即源域, 假设 χ^S 代表其示例空间而 Y^S 代表其标记空间; 将数据集 D^S 表示为 $\{(X_1^S, Y_1^S), (X_2^S, Y_2^S), \ldots, (X_{n_S}^S, Y_{n_S}^S)\}$, 其中, $X_i^S \subseteq \chi^S$ 为源域中第 i 个蛋白质样本, 含 b_i 个示例 $\{x_1^{(i)}, x_2^{(i)}, \ldots, x_{b_i}^{(i)}\}$; 而 $Y_i^S \subseteq Y^S$ 为 X_i^S 所对应的一组 GO 标记 $\{y_1^{(i)}, y_2^{(i)}, \ldots, y_{t_i}^{(i)}\}$. 因此, 本文的新完成测序物种的蛋白质功能预测任务可以抽象为多示例多标记迁移学习问题, 需要设计一个多示例多标记迁移学习框架来解决此问题.

2.2　多示例多标记迁移学习框架

本文提出了第一个多示例多标记迁移学习框架 (TR-MIML) (图 1), 框架由两个阶段组成: 源域样本加权阶段和分类模型学习阶段.

2.2.1 源域样本加权

源域样本加权阶段可以分为 (图 1): 首先, 将源域和目标域中的多示例样本转化为单示例样本; 然后, 通过最小化投影空间上加权源域样本中心点与目标域样本中心点的距离, 来给源域样本赋予不同权值.

与传统迁移学习样本不同, 本文中源域和目标域中的蛋白质样本都是由多个示例组成的样本包, 我们使用 miFV 方法 [18] 先将多示例样本转化为单示例样本.

在给源域样本赋予不同权值的过程中, 本文借助 maximum mean discrepancy (MMD) 准则 [19], 最小化再生核 Hilbert 空间 (RKHS) 上加权源域样本中心点与目标域样本中心点的距离:

$$\min_{\beta}\left\|\frac{1}{n_S}\sum_{i=1}^{n_S}\beta_i\phi\left(f_i^S\right)-\frac{1}{n_T}\sum_{j=1}^{n_T}\phi\left(f_j^T\right)\right\|_H^2, \tag{1}$$

其中, β 为需要求解的源域样本的权重向量, n_S 和 n_T 分别为源域和目标域中样本个数, f_i^S 和 f_j^T 分别表示源域样本与目标域样本通过 miFV 方法 [18] 得到的单示例样本, $\phi(\cdot)$ 为核函数, 将样本映射高维特征空间.

可将式 (1) 重写为 [20]

$$\min_{\beta}\ \frac{1}{2}\beta^{\mathrm{T}}K\beta-\kappa^{\mathrm{T}}\beta\qquad \text{s.t.}\ \ \beta_i\in[0,B],\ \frac{\left|\sum_{i=1}^{n_S}\beta_i-n_S\right|}{n_s}\leqslant\varepsilon, \tag{2}$$

其中, B 与 ε 均为常量, B 是 β 的上界 (本文取值 1000), $\varepsilon=(\sqrt{n_S}-1/\sqrt{n_S})$, 而 $\kappa_i=n_S/n_T\sum_{j=1}^{n_T}k(f_i,f_j)$, $\forall i=1,\ldots,n_S$, 而 K 为 kernel Gram matrix [21]

$$K=\begin{pmatrix}K_{S,S} & K_{S,T}\\ K_{T,S} & K_{T,T}\end{pmatrix}, \tag{3}$$

其中, K 矩阵中的元素 $K_{i,j}=k\left(f_i,f_j\right)$.

本文选择的是 Gauss 核函数

$$k\left(f_i,f_j\right)=\exp\left(-\sigma\left\|f_i-f_j\right\|^2\right), \tag{4}$$

其中, σ 为 Gauss 核半径.

式 (2) 是一个标准的二次优化问题, 本文使用 Matlab 优化工具箱的二次规划函数 quadprog 进行求解, 得到源域样本的权重向量 β.

2.2.2 分类模型学习

在得到源域样本的权重向量 β 后, 随机挑选出少量目标域中的标记样本, 加上源域样本, 生成多示例多标记学习训练样本数据集, 通过求解式 (5) 得到分类模型 $\widehat{h}$:

$$\widehat{h}=\arg\min_{h}\ \mu\sum_{i=1}^{n_S}\beta_i\frac{1}{n_S}\ell\left(h\left(X_i^S\right),Y_i^S\right)+\sum_{j=1}^{n_l}\frac{1}{n_l}\ell\left(h\left(X_j^T\right),Y_j^T\right), \tag{5}$$

其中, μ 是平衡因子 [22], 用于平衡加权后的源域样本与目标域样本对 MIML 分类模型的影响; β_i 为上节得到的第 i 个源域样本的权值; $\ell(\cdot)$ 为损失函数; n_l 为从目标域中随机挑选的标记样本个

数; $h(\cdot)$ 为多示例多标记学习分类器, 例如最新的多示例多标记学习算法 MIMLfast [15], 将包 X 在第 l 个标记上的预测值定义为它的所有示例在这个标记上预测的最大值, 其具体形式为: $h_l(X) = \max_{x\in X} \max_{k=1,\ldots,K} h_{l,k}(x)$, 其中, x 为包 X 中的一个示例, $h_{l,k}(x)$ 为示例 x 在第 l 个标记的第 k 个子概念上分类器. 在对式 (5) 求解过程中, 使用多示例多标记学习分类器类似的求解方法, 如 $h(\cdot)$ 为 MIMLfast 算法时, 使用随机梯度下降方法来更新变量 W_0^{t+1}, $w_{y,k}^{t+1}$ 和 $w_{y,\bar{k}}^{t+1}$ [15], 如果随机抽到的是源域样本, 则更新的变量值乘以 β_i. TR-MIML 学习框架伪代码见算法 1.

算法 1 Pseudo code of TR-MIML learning framework

$\widehat{h} =$ **TR-MIML**(D^T, D^S)

Input: D^T : Target domain dataset; D^S : Source domain dataset.

Output: $\widehat{h}$: Classifier

Steps:

1. **for** X_i^S in D^S **do**
2. $\quad f_i^S = \mathrm{miFV}(X_i^S)$;
3. **end for**
4. **for** X_i^T in D^T **do**
5. $\quad f_i^T = \mathrm{miFV}(X_i^T)$;
6. **end for**
7. Compute β by solving (2);
8. Learn the classifier $\widehat{h}$ by solving (5).

3 实验与结果

3.1 数据与实验设置

本文的实验数据包括 7 种真实物种的全基因组蛋白质样本数据集, 含两种新完成测序物种: 硫还原地杆菌 (Geobacter sulfurreducens)、棕色固氮菌 (Azotobacter vinelandii) 和 5 种蛋白质功能注释齐全的物种: 小家鼠 (Mus musculus)、褐家鼠 (Rattus norvegicus)、人 (Homo sapiens)、拟南芥 (Arabidopsis thaliana) 以及酿酒酵母 (Saccharomyces cerevisiae).

首先, 从 UniProt-GOA ftp 站点[1) [23] 下载 gene_association.goa_ref_uniprot 文件, 然后通过物种的 Taxon ID 号 (如: 人是 9606) 得到所有蛋白质的 Uniprot ID 号和基因本体学 (GO) ID 号 (剔除 evidence code 为 IEA 的 GO ID 号). 基因本体学从 3 个方面上来描述蛋白质生物学功能: 分子功能 (molecular function)、生物学过程 (biological process) 以及细胞组分 (cellular component) [24]. 本文考虑分子功能.

然后, 通过上面得到的蛋白质 Uniprot ID 号从 Universal Protein Resource (UniProt) 数据库 [25] 中下载得到 FASTA 格式的蛋白质序列文件. 对每一个物种, 将下载得到的全基因组 FASTA 格式的蛋白质序列文件上传到 NCBI 的 Batch CD-Search servers 服务器 [26], 得到蛋白质的保守结构域信息. 对蛋白质的每个结构域, 根据其序列计算三联体出现频率 (216 维) [27,28] 来作为特征向量. 每个蛋白质往往具有多个结构域, 因此每个蛋白质样本将被表示为由多个特征向量组成的示例包.

最后, 对上面得到的蛋白质分子功能 GO ID 号, 采用同样的策略 [14], 基于从基因本体学网站[2)下

1) ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/.

2) http://geneontology.org/page/download-ontology.

表 1 实验数据统计信息

Table 1 Experimental dataset statistics

Species	Proteins	GO terms	Domains per protein (Mean±std.)	GO terms per protein (Mean±std.)
Geobacter sulfurreducens	379	320	3.20±1.21	3.14±3.33
Azotobacter vinelandii	407	340	3.07±1.16	4.00±6.97
Mus musculus	11676	3065	2.76±1.84	44.64±50.27
Rattus norvegicus	5991	2600	2.53±1.71	39.51±44.78
Homo sapiens	13773	3311	2.98±4.30	55.81±126.63
Arabidopsis thaliana	8986	1811	2.02±1.46	27.68±70.37
Saccharomyces cerevisiae	3509	1566	1.86±1.36	15.89±11.52

载的 go.obo 文件 (更新时间 2017-01-21) 中描述的 GO 分子功能 ID 号有向无环图结构关系来得到每一个物种中蛋白质样本的 GO 标记向量.

在以前的基于多示例多标记学习的蛋白质功能预测工作 [14] 中, 每个蛋白质样本被表示为一个由多个示例组成的样本包, 其中每个示例为一个结构域, 并且该蛋白质样本含有多个标记, 其中每个标记为一个 GO 功能术语. 表 1 总结了本文中用到的 7 个物种实验数据的统计信息. 例如, 在 Geobacter sulfurreducens 这个新完成测序的物种数据集中, 共含有 379 个蛋白质样本, 320 个 GO 分子功能术语, 每个蛋白质样本平均包含 3.20 ± 1.21 个结构域和 3.14 ± 3.33 个 GO 分子功能术语 (其数量远少于 5 种蛋白质功能注释齐全物种中每个蛋白质平均含有的 GO 分子功能术语数量).

本研究主要目的是想利用蛋白质功能注释齐全的物种 (源域) 来帮助新完成测序物种 (目标域) 进行蛋白质功能预测. 新完成测序物种中已知功能的蛋白质数量少, 而且蛋白质 GO 功能预测天然存在类不平衡问题, 即对大多数 GO 功能标记, 每个物种中含有此标记的蛋白质样本数量远少于不含此标记的蛋白质样本数量. 因此, 本文选用源域中样本数量最多的前 20 个 GO 标记来帮助构建多示例多标记迁移学习模型, 以更好地评价迁移学习对帮助新完成测序物种蛋白质功能预测的贡献. 为了评估算法的性能, 本文选用了 5 种最常见的多示例多标记学习评价指标, Average Precision (AP), Coverage (CV), Hamming Loss (HL), One Error (OE), Ranking Loss (RL) [16]. AP 指标衡量的是按照预测值排序的标记序列中, 被排在相关标记之前的标记仍然是相关标记的情况, 其取值越大则性能越好 [16]. CV 指标计算样本的相关标记中, 根据预测值被排在最靠后的那个标记所处的排名, 其取值越低则性能越好 [16]. HL 指标计算的是所有样本在所有标记上的错误率, 其取值也在 0 到 1 之间, 值越小表示性能越好 [16]. OE 指标计算预测为最相关的标记实际上与该样本不相关的情况在测试集中占的比例, 其取值介于 0 到 1 之间, 值越小性能越好 [16]. RL 指标计算相关与无关标记对排序出现错误的比例, 取值越小越好 [16]. 为了客观评价算法的性能, 本文选用了 3 折交叉验证方法, 并且所有实验重复 10 次, 计算 “均值 ± 标准方差” 作为最终结果.

3.2 迁移学习对帮助新完成测序物种蛋白质功能预测的贡献

为了证明迁移学习能够有效帮助新完成测序物种的蛋白质功能预测工作, 本文选用了两种新完成测序物种 Geobacter sulfurreducens 和 Azotobacter vinelandii 作为目标域, 3 种蛋白质功能注释齐全的模式生物 Rattus norvegicus, Mus musculus, Saccharomyces cerevisiae 作为源域. 另外, 为了充分说明迁移学习的有效性, 本文选用了 3 种目前最好的多示例多标记学习方法 MIMLfast [15], MIMLNN [16],

表 2　以物种 Geobacter sulfurreducens 为目标域, 不同多示例多标记学习方法在有迁移学习和无迁移学习时的结果比较, 证明了在此任务上迁移学习可明显提高蛋白质功能预测性能 [a)]

Table 2 Performance comparison on Geobacter sulfurreducens (target domain) by multiple multi-instance multi-label learning methods with or without transfer learning. Results demonstrate that transfer learning can improve the performance of protein function prediction on this task

Source domain	Method	AP (↑)	CV (↓)	HL (↓)	OE (↓)	RL (↓)
Mus musculus	TR-MIMLfast	**0.58±0.02**	**4.04±0.07**	**0.15±0.00**	**0.55±0.04**	**0.32±0.01**
	MIMLfast	0.44±0.02	4.76±0.03	0.20±0.01	0.71±0.03	0.43±0.01
	TR-MIMLNN	0.56±0.01	4.28±0.20	0.21±0.01	0.58±0.02	0.36±0.03
	MIMLNN	0.54±0.01	4.51±0.19	0.24±0.01	0.58±0.01	0.37±0.01
	TR-MIMLSVM	**0.53±0.02**	**4.16±0.02**	0.18±0.0	**0.61±0.01**	0.37±0.09
	MIMLSVM	0.44±0.01	4.62±0.05	0.19±0.02	0.67±0.01	0.40±0.01
Rattus norvegicus	TR-MIMLfast	**0.56±0.02**	**4.02±0.17**	**0.16±0.00**	**0.58±0.03**	**0.33±0.02**
	MIMLfast	0.43±0.05	5.22±0.07	0.23±0.03	0.75±0.08	0.42±0.06
	TR-MIMLNN	**0.53±0.01**	**4.11±0.06**	0.16±0.02	**0.58±0.01**	0.38±0.01
	MIMLNN	0.48±0.01	4.74±0.09	0.19±0.00	0.66±0.03	0.41±0.01
	TR-MIMLSVM	0.53±0.03	**4.29±0.03**	0.17±0.01	**0.60±0.03**	0.39±0.01
	MIMLSVM	0.52±0.01	5.10±0.08	0.17±0.01	0.66±0.01	0.35±0.02
Saccharomyces cerevisiae	TR-MIMLfast	0.53±0.04	4.42±0.14	**0.17±0.00**	0.62±0.05	0.35±0.01
	MIMLfast	0.50±0.02	4.50±0.14	0.22±0.02	0.66±0.04	0.35±0.02
	TR-MIMLNN	0.53±0.01	4.58±0.09	0.14±0.04	0.61±0.09	**0.37±0.01**
	MIMLNN	0.52±0.02	4.72±0.11	0.16±0.01	0.62±0.02	0.41±0.02
	TR-MIMLSVM	0.54±0.02	4.61±0.07	0.17±0.01	0.57±0.01	0.40±0.03
	MIMLSVM	0.53±0.01	4.72±0.04	0.18±0.01	0.60±0.01	0.43±0.01

a) 粗体表示有迁移学习的结果要显著好于无迁移学习的结果 (基于置信度为 95% 的配对样本 t 检验).

MIML-SVM [16] 作为基分类器, 来比较有迁移学习和无迁移学习时的预测结果. MIMLfast 方法首先将复杂的原始空间映射到一个标记共享的低维子空间, 并通过随机梯度下降技术 (SGD) 来快速优化排序损失 (rank loss) [15]. MIMLNN 方法利用平均 Hausdorff 距离和 k-MEDOIDS 算法将多示例多标记样本转化为多标记样本, 最小化神经网络模型平方和损失, 并通过奇异值分解 (SVD) 方法求解 [16]. MIMLSVM 方法通过最大 Hausdorff 距离和 k-MEDOIDS 算法将多示例多标记样本转化为多标记样本, 然后使用 ML-SVM 算法进行求解 [16]. 考虑到新完成测序物种中已知功能蛋白质数量少的实际情况, 在多示例多标记迁移学习实验中, 对每个 GO 功能标记, 我们在目标域中只随机挑选含此 GO 功能的一个蛋白质标记样本. 在多示例多标记学习基分类器的实验中, 使用所有目标域样本构建模型. 在所有实验中, 3 种多示例多标记学习方法均使用参考文献 [15, 16] 中的默认参数.

迁移学习对 Geobacter sulfurreducens 和 Azotobacter vinelandii 两种新完成测序物种的蛋白质功能预测的实验结果比较如表 2 和 3 所示. 其中, ↑ (↓) 表示评价指标值越大 (小) 性能越好, 粗体数值表示有迁移学习的结果明显优于无迁移学习的结果 (基于置信度为 95% 的配对样本 t 检验). 结果表明, 在两种新完成测序的物种上, 本文的多示例多标记迁移学习模型在绝大多数情况下都要优于无迁移学习的基分类器模型 (表 2 和 3). 这表明, 迁移学习有助于新完成测序物种的蛋白质功能预测. 新完成测序物种中已知功能的蛋白质样本数量少, 难以了解样本整体空间的分布, 不足以学习到一个好的分类器模型. 本文通过迁移学习, 通过引入亲缘物种中分布相似的加权蛋白质样本信息, 丰富了新完成

表 3 以物种 Azotobacter vinelandii 为目标域, 不同多示例多标记学习方法在有迁移学习和无迁移学习时的结果比较, 证明了在此任务上迁移学习可明显提高蛋白质功能预测性能 [a)]

Table 3 Performance comparison on Azotobacter vinelandii (target domain) by multiple multi-instance multi-label learning methods with or without transfer learning. Results demonstrate that transfer learning can improve the performance of protein function prediction on this task

Source domain	Method	AP (↑)	CV (↓)	HL (↓)	OE (↓)	RL (↓)
Mus musculus	TR-MIMLfast	**0.55±0.00**	4.30±0.40	**0.15±0.00**	**0.58±0.01**	0.34±0.02
	MIMLfast	0.49±0.03	4.55±0.15	0.21±0.02	0.69±0.05	0.38±0.02
	TR-MIMLNN	0.52±0.00	4.35±0.0	**0.22±0.03**	0.68±0.02	0.39±0.01
	MIMLNN	0.48±0.01	4.79±0.03	0.27±0.01	0.65±0.00	0.41±0.00
	TR-MIMLSVM	0.50±0.0	4.55±0.12	**0.22±0.04**	0.63±0.01	0.41±0.02
	MIMLSVM	0.47±0.01	4.67±0.12	0.28±0.01	0.64±0.02	0.44±0.02
Rattus norvegicus	TR-MIMLfast	**0.54±0.02**	4.64±0.23	**0.17±0.00**	0.64±0.02	0.39±0.04
	MIMLfast	0.49±0.02	5.03±0.25	0.25±0.00	0.67±0.03	0.40±0.02
	TR-MIMLNN	0.50±0.03	4.93±0.20	0.20±0.01	**0.66±0.02**	**0.40±0.01**
	MIMLNN	0.46±0.01	5.36±0.13	0.23±0.01	0.71±0.02	0.46±0.02
	TR-MIMLSVM	0.52±0.01	4.73±0.05	0.21±0.02	**0.65±0.01**	**0.42±0.01**
	MIMLSVM	0.49±0.00	4.82±0.03	0.25±0.01	0.70±0.01	0.46±0.01
Saccharomyces cerevisiae	TR-MIMLfast	**0.62±0.02**	4.52±0.10	**0.18±0.00**	**0.53±0.02**	0.37±0.02
	MIMLfast	0.55±0.03	4.87±0.28	0.24±0.02	0.60±0.03	0.35±0.02
	TR-MIMLNN	**0.59±0.01**	5.02±0.10	0.19±0.04	**0.60±0.01**	**0.43±0.02**
	MIMLNN	0.51±0.00	5.49±0.10	0.19±0.01	0.64±0.00	0.46±0.00
	TR-MIMLSVM	0.52±0.01	4.77±0.03	0.19±0.01	0.62±0.03	**0.40±0.01**
	MIMLSVM	0.49±0.01	4.75±0.00	0.20±0.01	0.66±0.01	0.46±0.02

a) 每个评价指标上最好的结果用粗体表示.

测序物种用于构建蛋白质功能预测模型的样本信息, 有助于更好了解样本整体空间的分布情况, 帮助构建更为鲁棒, 泛化能力更强的预测模型.

在迁移学习中, 有多种对样本重新加权的方法, 为了分析样本重新加权的方法对结果的影响, 本文比较了 TrAdaBoost 方法 (有迁移学习) [29] 和 AdaBoost 方法 (无迁移学习) [30] 的预测结果, 还比较了 Domain Adaptive Logistic Regression (DALR) 方法 (有迁移学习) [31] 和 Logistic Regression (LR) 方法 (无迁移学习) 的预测结果. 因为这 4 种方法针对的都是传统的单示例单标记学习问题, 而本文针对的是多示例多标记学习问题, 为了对比实验的公平性, 均采用 miFV 方法 [18] 先将多示例样本转化为单示例样本, 然后对每个标记分别构建模型. 结果表明, 在两种新完成测序的物种 Geobacter sulfurreducens 和 Azotobacter vinelandii 上, 有迁移学习的 TrAdaBoost 和 DALR 方法在绝大多数情况下都要优于对应的无迁移学习方法 (表 4 和 5). 结果表明, 迁移学习中不同样本重新加权方法都有助于新完成测序物种的蛋白质功能预测. 本文通过对源域中的样本加权进行迁移学习, 不同的加权方法对单个样本的选择及其权重可能会有较大影响, 但对由目标域样本和选择的加权源域样本组成的样本空间的整体分布影响往往较小, 因此迁移学习中不同的样本重新加权方法一般有助于新完成测序物种的蛋白质功能

表 4　以物种 Geobacter sulfurreducens 为目标域, AdaBoost 和 Logistic Regression (LR) 方法在有迁移学习和无迁移学习时的结果比较, 证明了在此任务上迁移学习可明显提高蛋白质功能预测性能 [a)]

Table 4　Performance comparison on Geobacter sulfurreducens (target domain) by AdaBoost and Logistic Regression (LR) learning methods with or without transfer learning. Results demonstrate that transfer learning can improve the performance of protein function prediction on this task

Source domain	Method	AP (↑)	CV (↓)	HL (↓)	OE (↓)	RL (↓)
Mus musculus	TrAdaBoost	**0.58±0.01**	**4.10±0.03**	**0.23±0.01**	**0.61±0.01**	**0.33±0.01**
	AdaBoost	0.47±0.01	4.51±0.04	0.28±0.01	0.69±0.01	0.41±0.01
	DALR	**0.44±0.02**	4.24±0.02	0.26±0.00	**0.66±0.00**	**0.34±0.02**
	LR	0.35±0.03	4.32±0.05	0.27±0.02	0.82±0.01	0.45±0.02
Rattus norvegicus	TrAdaBoost	**0.45±0.03**	4.23±0.02	0.27±0.01	**0.65±0.02**	**0.33±0.01**
	AdaBoost	0.36±0.02	4.31±0.03	0.28±0.03	0.83±0.03	0.45±0.02
	DALR	**0.45±0.01**	4.11±0.03	0.9±0.01	**0.78±0.01**	0.46±0.04
	LR	0.32±0.01	4.41±0.05	0.30±0.01	0.88±0.00	0.54±0.01
Saccharomyces cerevisiae	TrAdaBoost	**0.44±0.01**	4.11±0.02	0.28±0.01	**0.77±0.01**	**0.46±0.04**
	AdaBoost	0.32±0.01	4.41±0.05	0.30±0.01	0.88±0.01	0.54±0.01
	DALR	**0.56±0.08**	4.35±0.01	0.28±0.02	0.61±0.00	0.32±0.01
	LR	0.47±0.02	4.48±0.00	0.33±0.01	0.67±0.01	0.34±0.00

a) 每个评价指标上最好的结果用粗体表示.

表 5　以物种 Azotobacter vinelandii 为目标域, AdaBoost 和 Logistic Regression (LR) 方法在有迁移学习和无迁移学习时的结果比较, 证明了在此任务上迁移学习可明显提高蛋白质功能预测性能 [a)]

Table 5　Performance comparison on Azotobacter vinelandii (target domain) by AdaBoost and Logistic Regression (LR) learning methods with or without transfer learning. Results demonstrate that transfer learning can improve the performance of protein function prediction on this task

Source domain	Method	AP (↑)	CV (↓)	HL (↓)	OE (↓)	RL (↓)
Mus musculus	TrAdaBoost	0.52±0.00	**4.07±0.01**	**0.28±0.00**	**0.55±0.00**	0.40±0.00
	AdaBoost	0.50±0.00	4.77±0.03	0.36±0.00	0.69±0.00	0.42±0.00
	DALR	**0.53±0.00**	4.47±0.02	**0.26±0.01**	**0.55±0.00**	0.38±0.00
	LR	0.48±0.01	4.79±0.05	0.31±0.02	0.67±0.03	0.41±0.02
Rattus norvegicus	TrAdaBoost	**0.57±0.03**	4.25±0.00	0.27±0.01	0.62±0.01	0.33±0.01
	AdaBoost	0.46±0.02	4.49±0.01	0.34±0.02	0.66±0.02	0.35±0.02
	DALR	**0.53±0.01**	**4.15±0.01**	0.17±0.00	0.60±0.01	0.36±0.09
	LR	0.44±0.00	4.61±0.05	0.19±0.01	0.67±0.00	0.39±0.00
Saccharomyces cerevisiae	TrAdaBoost	0.49±0.01	4.43±0.01	0.30±0.00	0.60±0.01	0.41±0.04
	AdaBoost	0.45±0.01	4.69±0.09	0.37±0.01	0.66±0.01	0.48±0.01
	DALR	**0.44±0.02**	4.24±0.02	0.26±0.00	**0.66±0.00**	**0.34±0.01**
	LR	0.35±0.01	4.33±0.07	0.28±0.00	0.84±0.01	0.44±0.01

a) 每个评价指标上最好的结果用粗体表示.

预测, 但性能会存在一些差异.

3.3　利用亲缘关系不同物种进行迁移学习对蛋白质功能预测的影响

为了研究利用亲缘关系不同物种进行迁移学习对目标域物种蛋白质功能预测的影响, 本文选择

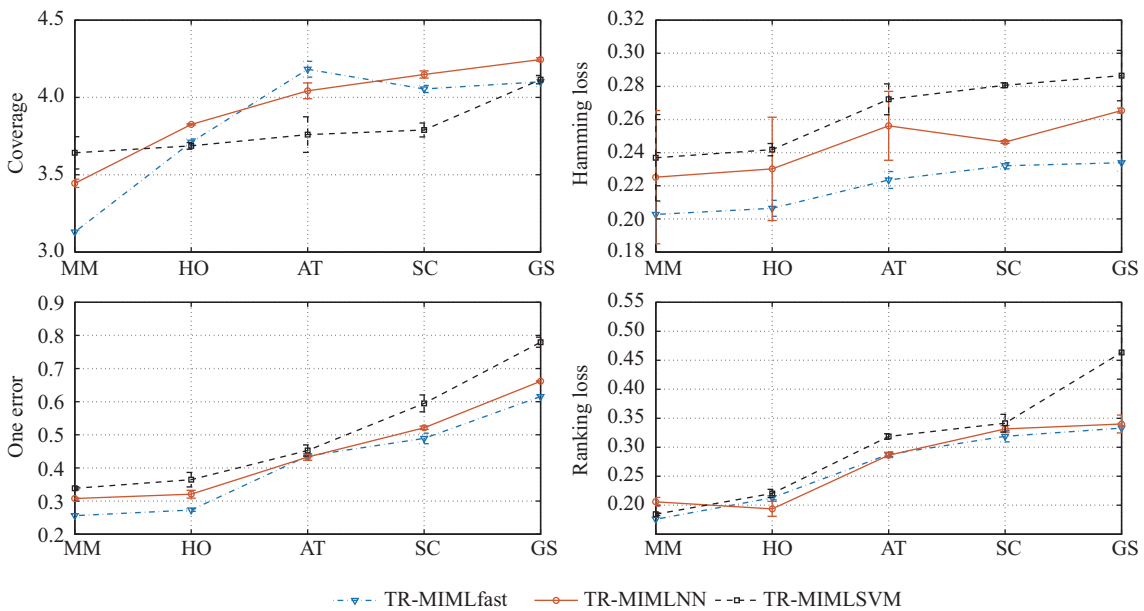


图 2 (网络版彩图) 利用亲缘关系不同的物种进行迁移学习对物种 Rattus norvegicus 蛋白质功能预测的影响

Figure 2 (Color online) Effect on protein function prediction of Rattus norvegicus by transfer learning using five species with different phylogenetic relationship

Rattus norvegicus 作为目标域, 而 Mus musculus, Homines, Arabidopsis thaliana, Saccharomyces cerevisiae 与 Geobacter sulfurreducens 分别作为源域. 选择 Rattus norvegicus 作为目标域是因为比较好界定它与多种源域物种的亲缘关系远近, Rattus norvegicus 与它们的亲缘关系由近到远排序为 Mus musculus, Homines, Arabidopsis thaliana, Saccharomyces cerevisiae 与 Geobacter sulfurreducens. Mus musculus 同属鼠科, Homines 同属哺乳动物, Arabidopsis thaliana 同属多细胞真核生物, Saccharomyces cerevisiae 同属真核生物, Geobacter sulfurreducens 同属生物.

为了客观评价利用亲缘关系不同物种进行迁移学习对目标域物种蛋白质功能预测的影响, 本文也选用了 MIMLfast [15], MIMLNN [16], MIMLSVM [16] 3 种方法作为基分类器来设计多示例多标记迁移学习算法. 为了显示的需要, 在图 2 中没有列出在 Average Precision 上的比较结果. 这里考虑的 4 个指标, Coverage, Hamming loss, One error 和 Ranking loss, 其值越小表示性能越好. 图 2 横坐标上的物种 MM, HO, AT, SC 和 GS 分别表示 Mus musculus, Homines, Arabidopsis thaliana, Saccharomyces cerevisiae 与 Geobacter sulfurreducens, 其与 Rattus norvegicus 的亲缘关系在横坐标上由左到右依次降低. 结果显示, 在绝大多数情况下, 利用亲缘关系越近的物种, 进行迁移学习帮助目标域物种进行蛋白质功能预测, 可以取得越好的性能 (图 2). 这表明, 利用亲缘关系近的物种进行迁移学习更有助于新完成测序物种的蛋白质功能预测. 亲缘关系近的物种往往含有更多相似的同源蛋白质, 而且亲缘关系近的物种蛋白质样本空间分布差异往往更小, 容易得到更多的源域样本, 有助于学习到性能优、泛化能力强的蛋白质功能预测模型.

4 结束语

新完成测序物种中已知功能的蛋白质数量少, 可以使用亲缘关系近、已知功能蛋白质数量多的

物种来帮助其进行蛋白质功能预测. 本文把这个任务抽象为多示例多标记迁移学习问题, 并提出了第一个多示例多标记迁移学习框架来解决此任务. 在两个新完成测序物种上, 实验结果证明了迁移学习有助于它们的蛋白质功能预测. 另外, 利用亲缘关系越近的物种作为源域进行迁移学习越有帮助. 在以后的研究中, 可以在更多的新完成测序物种上、在生物学过程、细胞组分等更多的蛋白质生物学功能上, 研究迁移学习的贡献. 另外, 还可以考虑利用更多的亲缘物种, 使用多源域迁移的方法, 进一步提升新完成测序物种的蛋白质功能预测性能. 本文算法的代码和实验数据可以在 https://github.com/njuptml/MIMLTR 进行下载.

参考文献

1 Jiang Y, Oron T R, Clark W T, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol, 2016, 17: 184

2 Radivojac P, Clark W T, Oron T R, et al. A large-scale evaluation of computational protein function prediction. Nature Meth, 2013, 10: 221–227

3 Wei L, Xing P, Shi G, et al. Fast prediction of protein methylation sites using a sequence-based feature selection technique. IEEE/ACM Trans Comput Biol Bioinform, 2017. doi: 10.1109/TCBB.2017.2670558

4 Li S, Li D, Zeng X, et al. nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. BMC Bioinform, 2014, 15: 298

5 Zou Q, Wan S X, Ju Y, et al. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. BMC Syst Biol, 2016, 10: 114

6 Wu Q, Ye Y, Zhang H, et al. ML-Tree: a tree-structure-based approach to multilabel learning. IEEE Trans Neural Netw Learn Syst, 2015, 26: 430–443

7 Wu Q, Ng M K, Ye Y, et al. Multi-label collective classification via Markov chain based learning method. Knowl-Based Syst, 2014, 63: 1–14

8 Wu Q, Ng M K, Ye Y. Markov-Miml: a markov chain-based multi-instance multi-label learning algorithm. Knowl Inf Syst, 2013, 37: 83–104

9 Wu Q, Tan M, Song H, et al. ML-Forest: a multi-label tree ensemble method for multi-label classification. IEEE Trans Knowl Data Eng, 2016, 28: 2665–2680

10 Pan S J, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng, 2010, 22: 1345–1359

11 Zhuang F Z, Luo P, He Q, et al. Survey on transfer learning research. J Softw, 2015, 26: 26–39 [庄福振, 罗平, 何清, 等. 迁移学习研究进展. 软件学报, 2015, 26: 26–39]

12 Mei S, Wang F, Zhou S. Gene ontology based transfer learning for protein subcellular localization. BMC Bioinform, 2011, 12: 44

13 Xu Y, Min H, Wu Q, et al. Multi-instance metric transfer learning for genome-wide protein function prediction. Sci Rep, 2017, 7: 41831

14 Wu J S, Huang S J, Zhou Z H. Genome-wide protein function prediction through multi-instance multi-label learning. IEEE Trans Comput Biol Bioinform, 2014, 11: 891–902

15 Huang S J, Gao W, Zhou Z H. Fast multi-instance multi-label learning. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI'14), Quebec City, 2014. 1868–1874

16 Zhou Z H, Zhang M L, Huang S J, et al. Multi-instance multi-label learning. Artif Intel, 2012, 176: 2291–2320

17 Zhou Z H, Zhang M L. Multi-instance multi-label learning with application to scene classification. In: Proceedings of the 20th Annual Conference on Neural Information Processing Systems, Vancouver, 2006

18 Wei X S, Wu J, Zhou Z H. Scalable algorithms for multi-instance learning. IEEE Trans Neural Netw Learn Syst, 2017, 28: 975–987

19 Borgwardt K M, Gretton A, Rasch M J, et al, Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics, 2006, 22: 49–57

20 Huang J, Smola A J, Gretton A, et al. Correcting sample selection bias by unlabeled data. In: Proceedings of the 19th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2006. 601–608

21 Chattopadhyay R, Fan W, Davidson I, et al. Joint transfer and batch-mode active learning. In: Proceedings of the

30th International Conference on Machine Learning, Atlanta, 2013. 253–261
22 Sun Q, Chattopadhyay R, Panchanathan S, et al. A two-stage weighting framework for multi-source domain adaptation. In: Proceedings of Advances in Neural Information Processing Systems, Lake Tahoe, 2012. 505–513
23 Camon E, Magrane M, Barrell D, et al. The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology. Nucleic Acids Res, 2004, 32: 262–266
24 Ashburner M, Ball C A, Blake J A, et al. Gene ontology: tool for the unification of biology. Nat Genet, 2000, 25: 25–29
25 Bairoch A, Apweiler R, Wu C H, et al. The universal protein resource (UniProt). Nucleic Acids Res, 2007, 35: 154–159
26 Marchlerbauer A, Lu S, Anderson J B, et al. CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res, 2011, 39: 225–229
27 Wu J, Hu D, Xu X, et al. A novel method for quantitatively predicting non-covalent interactions from protein and nucleic acid sequence. J Mol Graph Model, 2011, 31: 28–34
28 Shen J W, Zhang J, Luo X M, et al. Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci USA, 2007, 104: 4337–4341
29 Dai W, Yang Q, Xue G R, et al. Boosting for transfer learning. In: Proceedings of the 24th International Conference on Machine Learning, Corvalis, 2007. 193–200
30 Freund Y, Schapire R E. A desicion-theoretic generalization of on-line learning and an application to boosting. In: Proceedings of the 2nd European Conference on Computational Learning Theory. Berlin: Springer, 1995. 23–37
31 Jiang J, Zhai C X. Instance weighting for domain adaptation in NLP. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, 2007. 264–271

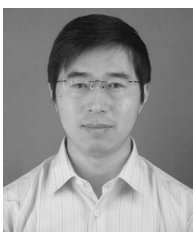
Protein function prediction through multi-instance multi-label transfer learning

Haifeng HU[1,2], Mao ZHENG[1], Weijian WU[3], Jun WANG[4] & Jiansheng WU[4*]

1. *School of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing* 210003*, China*;
2. *A Project Funded by the Priority Academic Program Development of Jiangsu Higer Education Institutions, Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science & Technology, Nanjing* 210044*, China*;
3. *College of Computer and Information, Hohai University, Nanjing* 211100*, China*;
4. *School of Geography and Biological Information, Nanjing University of Posts and Telecommunications, Nanjing* 210023*, China*
* Corresponding author. E-mail: jansen@njupt.edu.cn

Abstract With the release of various genome sequencing projects, there are many species whose genomic sequences have been recently completed. It is essential to annotate the protein functions of these species. Owing to the lack of proteins with known functions, it is important to exploit their relative species with a large number of proteins whose functions are known to assist in predicting the protein functions of these species. In this paper, we treat this task as a multi-instance multilabel transfer learning problem and propose the first multi-instance multilabel transfer learning framework to perform this task. Experiments on two newly completed sequencing species demonstrate that transfer learning contributes to protein function prediction. Moreover, the closer the polygenetic relationship between the source domain species and target domain species, the better the performance of transfer learning.

Keywords new sequencing-completed species, protein function prediction, transfer learning, multi-instance multi-label learning, sample reweighting



Haifeng HU received his B.S. degree in radio engineering from Anhui University, Hefei, China, and his M.S. and Ph.D. degrees in signal processing from Nanjing University of Posts and Telecommunications in 2002 and 2008, respectively. He is currently an associate professor at Nanjing University of Posts and Telecommunications. His research interests include large-scale similarity search, wireless sensor networks, wireless networking, and distributed systems.

Jiansheng WU received his B.S., M.S., and Ph.D. degrees in bioengineering, ecology, biomedical engineering from Nanchang University, East China Normal University, Southeast university, China, in 2000, 2004, and 2009, respectively. He joined the School of Geography and Biological Information of Nanjing University of Posts and Telecommunications, China, in 2009. His research interests include machine learning and bioinformatics.

Jun WANG was born in 1973. He received his Ph.D. degree in acoustics from Nanjing University in 2003. He is currently a professor at Nanjing University of Posts and Telecommunications. His research interests include biomedical information processing.