doi:10.11887/j.cn.202203004 http://journal.nudt.edu.cn

面向蛋白质功能预测中有向无环图标记结构的多示例多标记学习*

吴建盛[1],唐诗迪[1],梅德进[2],朱燕翔[3],刁业敏[4]

(1. 南京邮电大学 地理与生物信息学院,江苏 南京 210023;2. 南京邮电大学 通信与信息工程学院,江苏 南京 210003;
3. 南京仁面集成电路技术有限公司,江苏 南京 210088;4. 南京叁角加文化发展中心,江苏 南京 210005)

摘 要:在多示例多标记学习问题中,标记之间往往是相互关联的,其中有向无环图结构是一种常见的层次关联结构,可见于蛋白质的基因本体学生物学功能预测的应用场景中。针对其标记间的有向无环图结构,提出了一种新的多示例多标记学习算法。算法从原始数据的特征空间训练出所有标记共享的低维子空间,通过随机梯度下降方法来降低模型排序损失,并融入标记间有向无环图结构关系对预测标记进行优化。将该算法应用于多个数据集的蛋白质功能预测中,实验结果表明,该算法具有更高的效率及预测性能。

关键词:多示例多标记学习;蛋白质功能预测;有向无环图标记结构;标记相关性

中图分类号:TP301.6 **文献标志码**:A **文章编号**:1001-2486(2022)03-023-08

Multi-instance multi-label learning for labels with directed acyclic graph structures in protein function prediction

WU Jiansheng[1], *TANG Shidi*[1], *MEI Dejin*[2], *ZHU Yanxiang*[3], *DIAO Yemin*[4]

(1. School of Geographic and Biological Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;
2. School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
3. Nanjing Renmian Integrated Circuit Technology Limited Company, Nanjing 210088, China;
4. Nanjing Triangular Plus Culture Development Centre, Nanjing 210005, China)

Abstract: In MIML (multi-instance multi-label learning) tasks, labels are often correlated with each other, and DAG (directed acyclic graph) is a common hierarchically structure which often occurs in the prediction of gene ontology biological functions of proteins. Considering the labels with directed acyclic graph structures in MIML, a novel algorithm named MIMLDAG (multi-instance multi-label directed acyclic graph) was proposed. MIMLDAG trained a low-dimensional subspace of shared labels from the feature space of original datasets, minimized the rank loss by a stochastic gradient descent method, and then incorporated the inner DAG hierarchical structure of labels for optimizing the output labels. MIMLDAG was applied to predict the protein functions in multiple datasets, and the results show that MIMLDAG possesses higher efficiency and predictive performance.

Keywords: multi-instance multi-label learning; protein function prediction; labels with directed acyclic graph structure; label relationship

在经典监督学习中,一个对象仅由一个示例来表示,且仅有一个对应标记。实际上,一个对象可由多个示例来表示,并属于多个类别标记[1],其对应的学习框架叫作多示例多标记学习(multi-instance multi-label learning, MIML)。

过去十几年来,研究者们提出了许多基于MIML的算法,例如MIMLBoost[2]将MIML问题分解为多个单标记多示例子问题单独求解,而MIMLSVM[3]将其分解为多个单示例多标记问题;Yang等提出了一种基于全连接卷积网络的多示例多标记学习方法MIML-FCN+[4];Nguyen等提出了基于深度神经网络的DeepMIML模型[5],为MIML生成表示示例的同时,自动识别与标记相关联的关键示例;Zhu等后来又提出了DMNL[6],用一种高效的增强型拉格朗日优化方法预测隐藏的新颖标记;Hu等提出了一种利用标记相关性的度量多示例多标记学习算法MI(ML)^{2}kNN[7]。

近年来,多示例多标记学习被应用于众多场景中,例如Song等提出了MMCNN-MIML[8],将卷积神经网络(convolutional neural network,CNN)模型引入MIML的图片分类问题中;Xu等将MIML应用于预测肝癌细胞的基因突变问题[9];Li等提出基于卷积神经网络的层次性多示例多标记学习方法HMIML来对果蝇胚胎发育图像进行分类[10];Li

* **收稿日期**:2021-06-22
基金项目:国家自然科学基金资助项目(61872198,61971216);江苏省科技厅基础研究计划面上资助项目(BK20201378)
作者简介:吴建盛(1979—),男,江西临川人,副教授,博士,硕士生导师,E-mail:jansen@njupt.edu.cn

等提出的 AC-MIMLLN[11] 模型将 MIML 应用于情感分析任务;Mercan 等基于 MIML 方法来对乳腺组织病理图像进行多分类研究[12];Zhang 等利用 MIML 学习并基于时空预修剪技术来对原始视频中的动作进行识别与定位[13];Li 等开发了多示例多标记学习网络并应用于情感类别预测[14];Pan 等将 MIML 应用于雷达信号的识别中[15]。

在 MIML 学习问题里,标记之间往往是相互关联的,其中有向无环图(directed acyclic graph, DAG)是一种常见的层次关联结构,它从任意一个顶点出发均不能经过若干条边回到该顶点。蛋白质常包含多个结构域,也同时拥有多种生物学功能。蛋白质中每个结构域可以独立或与周边结构域相互协作完成其生物学功能。蛋白质生物学功能预测也可表示为多示例多标记学习问题,其中每个蛋白质表示为多 MIML 学习中的一个样本对象,而每个结构域表示为一个示例,每个生物学功能表示为一个标记[16]。蛋白质的生物学功能有多种描述方式,其中基因本体学(gene ontology, GO)使用最为广泛[17],其中的基因功能本体就是一个 DAG 结构的典型例子。目前也有不少研究利用 GO 结构辅助进行生物学功能学习。Li 等使用层次聚类的方法,针对 GO 的结构,在经典的层次聚类模型中加入了新的聚类条件对 GO 生物学功能进行了预测[18];Zhang 等使用了深度神经网络,结合蛋白质序列以及蛋白质 - 蛋白质结合(protein-protein interaction, PPI)网络,对蛋白质的 GO 生物学功能进行预测[19];Zhao 等引入基于转换器的双向编码表征(bidirectional encoder representation from transformers, BERT)模型从蛋白质的 GO 标记及其序列中提取特征,对配体 - 受体结合亲和力(drug-target binding affinity, DTA)进行预测[20]。

目前还没有有效的算法可针对基于 DAG 结构的 MIML 问题进行学习。因此,本文提出新的基于有向无环图结构的多示例多标记学习(MIML based on directed acyclic graph, MIMLDAG)算法,通过训练标记共享低维子空间,降低模型排序损失,并融入标记间 DAG 间层次结构关系对样本预测标记进行优化,提升了算法的学习性能。

1 MIMLDAG 算法

提出面向标记之间有向无环图结构的多示例多标记学习算法。算法分三层构建模型,其中前两层充分借鉴了 MIMLfast 算法[21] 的思想。首先,算法从原始数据集的特征空间训练出一个被所有标记共享的低维子空间;然后,训练标记线性模型并通过随机梯度下降方法来优化排序损失;最后,在层次性结构中找到一个子图[22],通过合并子节点来构建与有向无环图结构层次一致的多标记信息,融入标记间的有向无环图结构,得到未见示例样本的层次性标记集合。MIMLDAG 算法的整体框架如图 1 所示,算法伪代码见算法 1。

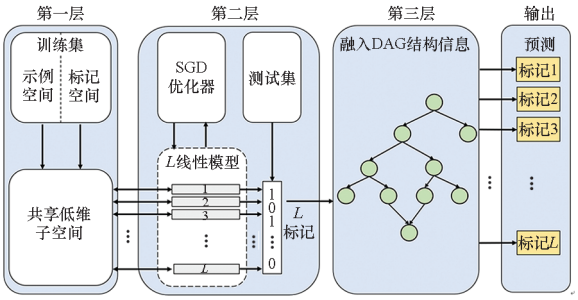


图 1 MIMLDAG 算法框架

Fig. 1 Framework of the MIMLDAG algorithm

算法 1 MIMLDAG 算法伪代码

Alg. 1 Pseudo code of the MIMLDAG algorithm

输入:训练集 $\boldsymbol{D}=\{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\mid 1\leqslant i\leqslant N\}$,样本包数目 N,对应示例数 n_t,标记个数 L

输出:优化后的有向无环图结构的标记集合 $\hat{y}^*$

/ * 训练阶段 * /

repeat

  for $i=1,\cdots,L$

    随机抽取三元组$(\boldsymbol{X},y,\bar{y})$

    依据两层模型训练 $\boldsymbol{W}_0,\boldsymbol{w}_l$

  end for

until 达到收敛准则

由测试样本集X_t 得到标记集合 $\hat{y}=\sum_{j=1}^{n_t}\sum_{l=1}^{q}\boldsymbol{w}_l^{\mathrm{T}}\boldsymbol{W}_0\boldsymbol{x}$

/ * 融入有向无环图信息 * /

初始化所有标记为超节点

初始化 $H=1$,$\Psi(i)=0$,并按照 SNV 排序

while $H<L$

  挑选 Ψ 未分配且具有最大 SNV 的超节点 S^*

  if $\Psi(pa(S^*))=1$

    $\Psi(S^*)\leftarrow\min\{1,(L-H)/n(S^*)\}$

    $H\leftarrow H+n(S^*)$

  else

  从所有 S^* 的未分配的父结点中,找到对应 SNV 最小的父节点 $\tilde{S}$

  合并 S^* 和 $\tilde{S}$ 为一个新的超节点

  end if

end while

/ * 输出结果 * /

输出有向无环图结构的标记集合 $\hat{y}^*$

1.1　训练线性模型

1.1.1　共享低维子空间

在第一层中,给定 MIML 数据集 $\boldsymbol{D}=\{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\mid 1\leqslant i\leqslant N\}$,样本包 $\boldsymbol{X}_i=\{\boldsymbol{x}_1^i,\boldsymbol{x}_2^i,\cdots,\boldsymbol{x}_{n_i}^i\}$,包含 n_i 个样本示例;$Y_i=\{y_{i,1},\cdots,y_{i,L}\}$ 为第 i 个样本包的 L 个标记集合。初始化一个待学习的 $m\times d$ 维($m\ll d$)共享矩阵 $\boldsymbol{W}_0$,目的在于将原始 d 维示例 $\boldsymbol{x}$ 映射到 m 维的低维共享空间,即 $\boldsymbol{W}_0\boldsymbol{x}$,减少内存消耗与计算复杂度。

1.1.2　优化排序损失

在第二层中,利用上一层的共享矩阵 $\boldsymbol{W}_0$ 可以将任意样本包 $\boldsymbol{X}_i$ 的示例 $\boldsymbol{x}$ 在第 l 个标记上的分类器定义为:

$$f_l(\boldsymbol{x})=\boldsymbol{w}_l^{\mathrm{T}}\boldsymbol{W}_0\boldsymbol{x} \tag{1}$$

式中,$\boldsymbol{w}_l$ 是第 l 个标记的 m 维权重向量。

接着,通过随机梯度下降方法来优化模型排序损失。具体地,如果用$(\boldsymbol{X}_i,y,\bar{y})$表示在第 i 轮随机梯度下降(stochastic gradient descent, SGD)方法中抽取的内容,其中 $\boldsymbol{X}_i$ 表示当前抽取的样本包;y、$\boldsymbol{x}$ 和 k 分别表示相关标记、其关键示例与对应的子概念;同理,$\bar{y}$、$\bar{\boldsymbol{x}}$ 和 $\bar{k}$ 分别表示无关标记、其关键示例与对应的子概念。抽取内容对应的损失函数为:

$$L(\boldsymbol{X}_i,y,\bar{y})=\varepsilon(\boldsymbol{X}_i,y)\mid 1+f_{\bar{y}}(\boldsymbol{X}_i)\mid_+ \tag{2}$$

式中,

$$\varepsilon(\boldsymbol{X}_i,y)=\sum_{i=1}^{R(\boldsymbol{X}_i,y)}\frac{1}{i} \tag{3}$$

$R(\boldsymbol{X}_i,y)$ 表示相关标记排在无关标记后面的个数。改用一个近似值 $|\bar{Y}|/v$ 可以快速估算出 $R(\boldsymbol{X}_i,y)$,其中 v 表示在第 v 次抽样中出现了第一个无关标记排在相关标记的前面[23]。那么式(2)可以重写为:

$$L(\boldsymbol{X}_i,y,\bar{y})\approx\begin{cases}0, & f_y(\boldsymbol{X}_i)>f_{\bar{y}}(\boldsymbol{X}_i)\\ S_{\bar{y},v}[1+(\boldsymbol{w}_{\bar{y},\bar{k}}^t)^{\mathrm{T}}\boldsymbol{W}_0^{\mathrm{T}}\bar{\boldsymbol{x}}-(\boldsymbol{w}_{y,k}^t)^{\mathrm{T}}\boldsymbol{W}_0^{\mathrm{T}}x], & \text{其他}\end{cases} \tag{4}$$

其中,$\boldsymbol{w}_{y,k}$ 对应标记 y 的第 k 个子概念,$\boldsymbol{w}_{\bar{y},\bar{k}}$ 同理,

$$S_{\bar{y},v}=\sum_{i=1}^{\lfloor\frac{|\bar{Y}|}{v}\rfloor}\frac{1}{i} \tag{5}$$

接着,如果分类器对非相关标记的预测值大于相关标记的预测值,即 $f_{\bar{y}}(\boldsymbol{X}_i)>f_y(\boldsymbol{X}_i)-1$,就按照式(6)更新 $\boldsymbol{W}_0,\boldsymbol{w}_{y,k},\boldsymbol{w}_{\bar{y},\bar{k}}$[21],其中 γ_t 为 SGD 算法更新的步长。接着进行下一轮迭代并在达到一定收敛准则后停止。最后,将模型应用于未见示例样本集 $\boldsymbol{T}=\{\boldsymbol{X}_t\}$,$\boldsymbol{X}_t=\{\boldsymbol{x}_{t_1},\boldsymbol{x}_{t_2},\cdots,\boldsymbol{x}_{t_{n_t}}\}$,输出预测标记。

$$\begin{cases}\boldsymbol{W}_0^{t+1}=\boldsymbol{W}_0^t-\gamma_t S_{\bar{Y},v}(\boldsymbol{w}_{\bar{Y},\bar{k}}^t\bar{\boldsymbol{x}}^{\mathrm{T}}-\boldsymbol{w}_{y,k}^t\boldsymbol{x}^{\mathrm{T}})\\ \boldsymbol{w}_{y,k}^{t+1}=\boldsymbol{w}_{y,k}^t+\gamma_t S_{\bar{Y},v}\boldsymbol{W}_0^t\boldsymbol{x}\\ \boldsymbol{w}_{\bar{y},\bar{k}}^{t+1}=\boldsymbol{w}_{\bar{y},\bar{k}}^t-\gamma_t S_{\bar{Y},v}\boldsymbol{W}_0^t\boldsymbol{x}\end{cases} \tag{6}$$

1.2　融入有向无环图结构

有向无环图结构层次性约束有两种情况[24]:情况 A 是如果某一节点的标记为正,那么它的所有父节点的标记也为正;情况 B 是如果某一节点的标记为正,那么它所有父节点中至少有一个节点的标记为正。其中情况 A 较为常用。对于给定测试样本包 $\boldsymbol{X}_T$ 的预测结果 $Y_T=\{y_1,y_2,\cdots,y_T\}\in\{0,1\}$,将每个预测标记看作一个节点,已知样本包对应 L 个标记,那么情况 A 优化问题表示为:

$$\begin{aligned}&\max_{\Psi}\sum_{i\in T}y_i\Psi_i\\ &\text{s.t. }\Psi_i\geqslant 0,\forall i\in T,\Psi_0=1,\sum_{i\in T}\Psi_i\leqslant L\end{aligned} \tag{7}$$

其中,若集合 $\boldsymbol{Y}_T$ 呈现情况 A 的层次结构,则称 $\Psi=\{\Psi_1,\Psi_2,\cdots,\Psi_T\}$ 为情况 A 的 nonincreasing 集合。

1.3　算法分析

使用算法 1 步骤 7 ~ 17 来逐步优化上述问题,其核心在于通过合并多个子节点为超节点来保证整体的非递增性。具体来讲,定义节点的超节点值(supernode value,SNV)为组成这个节点的所有标记的平均值,每一次迭代均会选取拥有最大 SNV 且未被分配的节点 S^*,如果 S^* 所有父节点的 Ψ 均为 1(即满足 1.2 节的情况),那么更新这个节点,否则,将 S^* 与未被分配且具有最小 SNV 的父节点合并,直到 $\sum_{i\in G}\Psi_i=L$。直观地讲,如果 S^* 不满足 AND-G 解释,那么自身的 SNV 一定比父节点的 SNV 大,将 S^* 与 SNV 最小的父节点合并能够让新的超节点拥有可能的最小 SNV。由于步骤 10 总会选取最大 SNV 的超节点,现存的其他超节点会比新合并的超节点更有可能被选中并进行更新。不断的迭代下,由步骤 6 得到的预测标记能够逐步优化为 AND-G 层次结构[22],最后输出预测结果。

2　算法仿真

2.1　实验数据

使用了三组蛋白质生物学功能预测数据集来对 MIMLDAG 算法进行实验仿真分析,其分别为

G蛋白偶联受体数据集、硫土杆菌(Geobacter sulfurreducens,GS)的蛋白质数据集和古细菌死海盐盒菌(Haloarcula marismortui, HM)的蛋白质数据集(见表1)。

表1　数据集的描述

Tab. 1　Descriptions of datasets

数据集		蛋白质数目	基因本体术语数目
GPCRs	MF	1 674	208
	BP	1 331	162
GS	MF	379	320
HM	MF	304	234

注:GPCRs表示G蛋白偶联受体, GS表示硫土杆菌, HM表示死海盐盒菌, MF表示分子功能, BP表示生物学过程。

2.1.1　G蛋白偶联受体数据集

G蛋白偶联受体数据集是从UniProt数据库[25]中下载得到共3 052个G蛋白偶联受体(G protein-coupled receptor, GPCR),再通过UniProt ID号从UniProt数据库中得到所有GPCR的FASTA格式序列。接着,将其输入NCBI的blastclust可执行程序,对GPCR序列进行去冗余处理,将得到的非冗余GPCR样本数据集提交到NCBI的Batch CD-Search服务器,得到GPCR的保守结构域。对于每一个结构域,从以下7个方面构建其特征:

1)三联氨基酸组成(conjoint triad)信息:把20种氨基酸依据其侧链体积与偶极矩分为6类。针对每个结构域,根据其氨基酸序列计算三联体出现频率,其特征维数为216[26]。

2)氨基酸关联(amino acid correlation, AAC)信息:依据上面的6类氨基酸信息来计算每个结构域中两两氨基酸间的AAC信息,对每个结构域,其ACC特征的维数为144[27]。

3)二级结构关联(secondary structure element correlation, SSC)信息:通过PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/ psiform.html)所提供的在线分析工具完成蛋白质二级结构的预测,然后计算3类二级结构类型在结构域中的关联信息 $SSC(\kappa)$, $\kappa \in \{2,4,8,16\}$。对每个结构域,其二级结构关联信息特征的维数为72。

4)进化信息:通过psiblast程序[28]产生位置特异性得分矩阵以表示结构域的进化信息。对于氨基酸长度为 n 的结构域,其位置特异性得分矩阵的维数为 $42n$。考虑因氨基酸长度不同导致矩阵大小不同的问题,把每个蛋白质的结构域当作一个示例,然后由算法miFV[29]统一为单一向量,对应到单个GPCR结构域的位置特异性得分矩阵特征维数为84。

5)信号肽特征信息(SignalP):通过一种基于神经网络的SignalP 4.0方法[30]从GPCR序列中提取信号肽特征信息。在SignalP 4.0中使用了两种类型的网络,首先使用跨膜数据序列作为负数据来训练得到SignalP-TM网络;然后在缺失这些负数据的情况下训练得到SignalP-noTM网络;最后使用简单的决策方案来选择使用哪个网络:如果SignalP-TM网络预测4个或者更多位置为跨膜位置,则SignalP-TM被用于最终的预测,否则使用SignalP-noTM网络预测。对于每个结构域,SignalP特征维数为84。

6)无序区域特征信息(Disorder):通过DISOPRED 2.43程序[31]预测得到蛋白质的无序区域特征信息。DISOPRED服务器允许用户提交一个蛋白质序列,然后返回每个无序区的无序概率估计值作为无序区域特征信息。对于每个结构域,Disorder特征维数为84。

7)SDK(scientific database maker)软件预测出的特征信息:通过SDK[17]预测蛋白质特征。它从Swiss-Port[32]数据库中提取蛋白质数据,同时对蛋白质进行数据分析,包括物理化学分布计算、同源序列搜索、多序列比对等,得到105维的数据特征。将非数据部分特征去掉,归一化后得到59维的蛋白质数据特征。

最后,对于样本空间中的每个结构域,共有特征维数743。

依据生物学过程与分子功能两个方面来描述蛋白质生物学功能。首先,根据GPCR蛋白质数据集的UniProt ID号从UniProt-GOA ftp站点(ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/)下载得到其对应基因本体学术语(GO terms)ID号及其对应的GO术语;然后,从基因本体学网站(http://geneontology.org/page/download-ontology)下载go.obo文件,分别得到BP与MF的GO术语对应的父节点;最后,基于样本中含有的GO terms及其所有父节点GO terms,构建标记的DAG层次性结构。对于BP,得到非冗余GPCRs样本1 331个,GO术语162个,GO术语的层次性结构深度为9;对于MF,得到非冗余GPCRs样本1 674个,GO术语208个,GO术语的层次性结构深度为12(见表1)。

2.1.2　GS 和 HM 蛋白质数据集

GS 的蛋白质数据集和 HM 的蛋白质数据集来自 http://www.lamda.nju.edu.cn/data_MIMLprotein.ashx[16],此处不再赘述。

2.2　性能比较

与多种经典的多示例多标记学习算法进行比较。这些算法分别为:基于径向基核函数神经网络的 MIMLRBF[33]、基于集成学习的 EnMIMLNN[16]、基于 K 邻近算法的 MIMLKNN[34]、基于支持向量机的 MIMLSVM[3] 和快速多示例多标记学习 MIMLfast[21]。

对于上述对比算法,本文选用了对应参考文献中的默认参数。其中, MIMLRBF 算法的缩放因子为 0.08,分数参数为 0.1;EnMIMLNN 算法中学习率设为 0.4;MIMLKNN 算法中聚类簇占据样本包的比例为 40%;MIMLSVM 算法中,高斯核半径 r 设置为 0.2;对 MIMLDAG 算法,低维共享空间的维度 m 设置为 50;对 MIMLfast 算法,共享空间维度设为 100。

此外,本文采用十倍交叉验证对 MIML 模型三种常用的评价指标 HL(hamming loss)[33]、MaF1(Macro-F1)[35]、MiF1(Micro-F1)[35] 进行评估。HL 表示样本的预测标记与真实标记之间的错误率。MaF1 先计算在每个标记上的 F1 值,然后求在所有标记上的平均值,MaF1 容易受到样本量少的标记的预测结果影响;MaF1 越大,表示模型性能越好。MiF1 计算在所有示例包和类别标记上预测结果的 F1 值;MiF1 越大,表示模型性能越好。

2.2.1　性能分析

表 2 展示了不同数据集下几种算法在不同指标上的性能变化,其中 ↑ 表明指标越大模型性能越好,↓ 反之。由表 2 可知,对于 GPCRs 的分子功能 MF,相比于其他五种方法,MIMLDAG 在 HL、MaF1、MiF1 上都获得了最好的性能。同样,对于 GPCRs 的生物学过程,MIMLDAG 在 HL、MaF1、MiF1 方面均取得了最好的性能。对于 GS 的分子功能,MIMLDAG 在 MaF1 获得了最优的性能。对于 HM 的分子功能,MIMLDAG 在 HL 上性能较弱,不如 EnMIMLNN,也略低于 MIMLRBF 和 MIMLSVM,在其他上都取得了最好的性能。由此可见,MIMLDAG 方法比其他多示例多标记学习方法取得了更好的性能。

2.2.2　时间效率分析

表 3 给出了所有算法在 4 种数据集上的时间开销。在算法训练及测试时间开销上,对于所有的数据集,MIMLDAG 方法的速度与最快的 MIMLfast 方法基本上接近,明显优于其他 4 种多示例多标记学习方法。

表 2　不同数据集上与多示例多标记学习方法的性能比较

Tab.2　Performance comparison with MIML methods on different datasets

数据集		评价指标	方法					
			MIMLDAG	MIMLRBF	MIMLKNN	MIMLSVM	EnMIMLNN	MIMLfast
GPCRs	MF	HL ↓	**0.147 0**	0.175 6	0.387 1	0.450 6	0.242 9	0.148 1
		MaF1 ↑	**0.032 4**	0.013 2	0.014 2	0.012 3	0.019 7	0.028 7
		MiF1 ↑	**0.231 4**	0.142 3	0.078 8	0.193 6	0.201 2	0.198 4
	BP	HL ↓	**0.143 7**	0.182 5	0.248 9	0.518 0	0.351 3	0.144 5
		MaF1 ↑	**0.029 2**	0.019 2	0.012 1	0.009 1	0.017 6	0.022 6
		MiF1 ↑	**0.215 9**	0.122 5	0.068 9	0.197 6	0.200 8	0.174 9
GS	MF	HL ↓	0.014 7	0.010 8	0.092 7	0.011 2	**0.010 0**	0.015 6
		MaF1 ↑	**0.035 6**	0.004 8	0.020 4	0.008 4	0.023 1	0.031 1
		MiF1 ↑	0.187 3	0.103 2	0.056 8	0.151 7	**0.235 2**	0.156 2
HM	MF	HL ↓	0.016 8	0.014 5	0.080 2	0.015 1	**0.011 8**	0.026 1
		MaF1 ↑	**0.042 3**	0.008 1	0.039 8	0.013 6	0.042 2	0.035 8
		MiF1 ↑	**0.345 3**	0.127 8	0.124 3	0.199 2	0.337 2	0.168 4

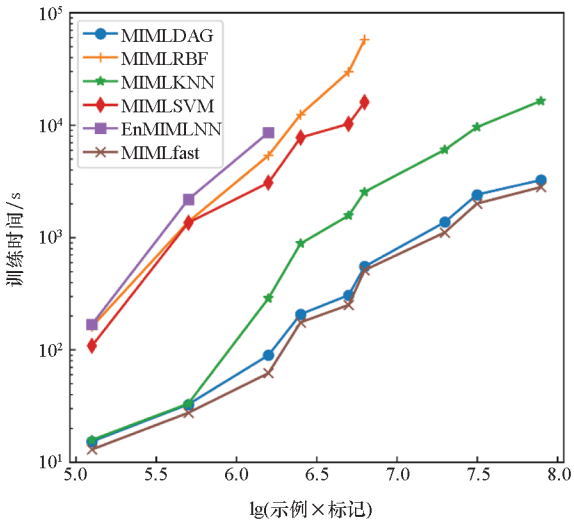
表 3　不同数据集上与各多示例多标记学习方法的时间开销比较

Tab. 3　Runtime comparison with MIML algorithms methods on different datasets

单位：s

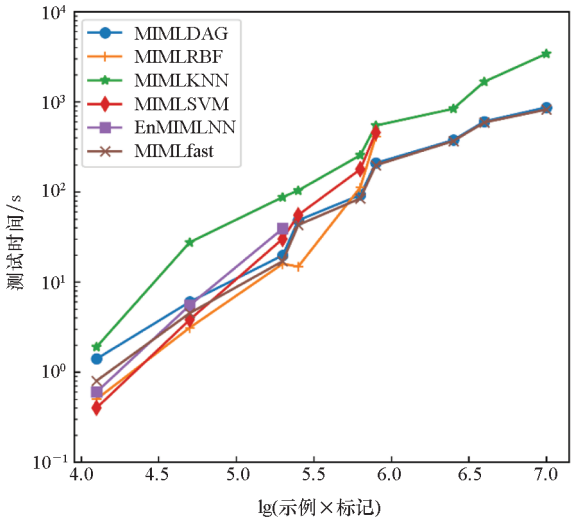
数据集		阶段	方法					
			MIMLDAG	MIMLRBF	MIMLKNN	MIMLSVM	EnMIMLNN	MIMLfast
GPCRs	MF	训练	36.7	3 025.7	165.8	29 731.4	9 162.4	**34.2**
		测试	22.2	177.7	55	112.0	35.3	**13.4**
	BP	训练	19.5	2 066.8	87.5	5 361.8	5 545.9	**18.3**
		测试	9.7	30.1	27.4	15.9	15.5	**8.9**
GS	MF	训练	21.3	1 464.3	68.9	3 341.7	2 534.8	**16.8**
		测试	12.5	33.2	21.4	15.3	12.4	**8.2**
HM	MF	训练	**15.5**	1 067.4	57.5	1 363.7	2 243.3	16.3
		测试	8.6	20.6	22.4	11.7	13.6	**7.9**

图 2 展示了各个算法在不同数据规模下的时间增长模式。由图可知，对于训练时间，EnMIMLNN 的增长速度最快，而 MIMLfast 和 MIMLDAG 增长速度较为缓慢；对于测试时间，MIMLKNN 算法的增长速度最快，当 lg(示例 × 标记) > 6.0 时，MIMLDAG 的增长速度最为缓慢。因此，MIMLDAG 算法随数据集的增长速率很低，基本与 MIMLfast 方法持平。MIMLDAG 算法共分三层：第一层，从原始数据集的特征空间训练出一个被所有标记共享的低维子空间；第二层，训练标记线性模型并通过随机梯度下降方法来优化排序损失；第三层，从层次性结构中找到一个子图，通过合并子节点来构建与有向无环图结构层次一致的多标记信息，得到未见示例样本的层次性标记集合。由式(1) ~ (6)可以推导出 MIMLDAG 算法前两层的时间复杂度为 $O[t \times L \times (d \times L) \times m]$，其中 t 为迭代轮数，L 为标记空间中的标记个数，d 为示例的特征维度，m 为共享低维子空间的维度($m \ll d$)。算法第三层(算法 1 步骤 7 ~ 17)的时间复杂度为 $O[L \times \log(L)]$，其中 L 为标记空间中的标记个数，也就是 DAG 中的节点数。因为前两层与第三层之间是串行连接，所以算法总的时间复杂度为 $O[t \times L \times (d \times L) \times m + L \times \log(L)]$。可以看出，MIMLDAG 算法与样本的数量无关，同时算法将样本原始特征空间映射到低



(a) 训练时间

(a) Training time



(b) 测试时间

(b) Testing time

图 2　6 种 MIML 算法在不同数量示例和标记上的训练和测试时间开销

Fig. 2　Runtime of training and testing on six MIML methods with various sizes of instances and labels

维共享空间,减少了内存消耗并降低了计算复杂度。

3　结论

本文提出面向蛋白质功能预测和有向无环图标记结构的多示例多标记学习算法。首先,算法训练一个共享矩阵将高维示例映射到低维子空间;然后,利用 SGD 优化排序方法初步训练出分类模型;最后,算法考虑了标记间的 DAG 结果,通过合并多个子节点为超节点,实现对标记的最终预测。本文运用 MIMLDAG 算法对多个蛋白质生物学功能预测数据集进行了学习。实验表明,相比于其他类似算法, MIMLDAG 拥有更好的预测性能与时间效率。在后续的研究中,拟考虑更多的数据集,进一步扩大算法的应用场景。

参考文献(References)

[1] ROUSU J, SAUNDERS C, SZEDMAK S, et al. Kernel-based learning of hierarchical multilabel classification models[J]. Journal of Machine Learning Research, 2006, 7: 1601 – 1626.

[2] SCHÖLKOPF B, PLATT J, HOFMANN T. Multi-instance multi-label learning with application to scene classification[C]// Proceedings of the 19th International Conference on Neural Information, 2006: 1609 – 1616.

[3] ZHOU Z H, ZHANG M L, HUANG S J, et al. MIML: a framework for learning with ambiguous objects [EB/OL]. (2008 – 08 – 24)[2021 – 02 – 18]. https://arxiv.org/abs/0808.3231v1.

[4] YANG H, ZHOU J T, CAI J F, et al. MIML-FCN + : multi-instance multi-label learning via fully convolutional networks with privileged information [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017: 5996 – 6004.

[5] NGUYEN C T, WANG X, LIU J, et al. Labeling complicated objects: multi-view multi-instance multi-label learning[C]//Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014: 2013 – 2019.

[6] ZHU Y, TING K M, ZHOU Z H. Discover multiple novel labels in multi-instance multi-label learning[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017: 2977 – 2983.

[7] HU H F, CUI Z K, WU J S, et al. Metric learning-based multi-instance multi-label classification with label correlation[J]. IEEE Access, 2019, 7: 109899 – 109909.

[8] SONG L Y, LIU J, QIAN B Y, et al. A deep multi-modal CNN for multi-instance multi-label image classification[J]. IEEE Transactions on Image Processing, 2018, 27(12): 6025 – 6038.

[9] XU K X, ZHAO Z Y, GU J P, et al. Multi-instance multi-label learning for gene mutation prediction in hepatocellular carcinoma[C]//Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society, 2020: 6095 – 6098.

[10] LI T G, YANG Y, SHEN H B. HMIML: hierarchical multi-instance multi-label learning of drosophila embryogenesis images using convolutional neural networks[C]//Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, 2018: 907 – 912.

[11] LI Y C, YIN C X, ZHONG S, et al. Multi-instance multi-label learning networks for aspect-category sentiment analysis[EB/OL]. (2020 – 10 – 06)[2021 – 02 – 18]. https://arxiv.org/abs/2010.02656v1.

[12] MERCAN C, AKSOY S, MERCAN E, et al. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images [J]. IEEE Transactions on Medical Imaging, 2018, 37(1): 316 – 325.

[13] ZHANG X Y, SHI H C, LI C S, et al. Multi-instance multi-label action recognition and localization based on spatio-temporal pre-trimming for untrimmed videos [C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12886 – 12893.

[14] LI Y C, YIN C X, ZHONG S H, et al. Multi-instance multi-label learning networks for aspect-category sentiment analysis[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 3550 – 3560.

[15] PAN Z S, WANG S F, ZHU M T, et al. Automatic waveform recognition of overlapping LPI radar signals based on multi-instance multi-label learning [J]. IEEE Signal Processing Letters, 2020, 27: 1275 – 1279.

[16] WU J S, HUANG S J, ZHOU Z H. Genome-wide protein function prediction through multi-instance multi-label learning[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2014, 11(5): 891 – 902.

[17] HAMMAMI R, ZOUHIR A, NAGHMOUCHI K, et al. SciDBMaker: new software for computer-aided design of specialized biological databases [J]. BMC Bioinformatics, 2008, 9: 121.

[18] LI Z J, LIAO B, LI Y, et al. Gene function prediction based on combining gene ontology hierarchy with multi-instance multi-label learning [J]. RSC Advances, 2018, 8(50): 28503 – 28509.

[19] ZHANG F H, SONG H, ZENG M, et al. A deep learning framework for gene ontology annotations with sequence- and network-based information [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(6): 2208 – 2217.

[20] ZHAO L L, XIE P J, HAO L F, et al. Gene ontology aided compound protein binding affinity prediction using BERT encoding[C]//Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, 2020: 1231 – 1236.

[21] HUANG S J, GAO W, ZHOU Z H. Fast multi-instance multi-label learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(11): 2614 – 2627.

[22] BI W, KWOK J T. Multi-label classification on tree- and DAG-structured hierarchies [C]//Proceedings of the 28th International Conference on International Conference on Machine Learning, 2011: 17 – 24.

[23] WESTON J, BENGIO S, USUNIER N. WSABIE: scaling up to large vocabulary image annotation[C]//Proceedings of the 22nd International Joint Conference on Artificial Intelligence, 2011: 2764 – 2770.

[24] VENS C, STRUYF J, SCHIETGAT L, et al. Decision trees

for hierarchical multi-label classification [J]. Machine Learning, 2008, 73(2): 185 – 214.

[25] The UniProt Consortium. The universal protein resource (UniProt) [J]. Nucleic Acids Research, 2009, 37 (Suppl1): D169 – D174.

[26] WANG Y C, WANG X B, YANG Z X, et al. Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature [J]. Protein and Peptide Letters, 2010, 17(11): 1441 – 1449.

[27] COULONDRE C, MILLER J H. Genetic studies of the *lac* repressor. Ⅲ. Additional correlation of mutational sites with specific amino acid residues [J]. Journal of Molecular Biology, 1977, 117(3): 525 – 567.

[28] AGRAWAL A, HUANG X Q. PSIBLAST_PairwiseStatSig: reordering PSI-BLAST hits using pairwise statistical significance[J]. Bioinformatics, 2009, 25(8): 1082 – 1083.

[29] WEI X S, WU J X, ZHOU Z H. Scalable multi-instance learning[C]//Proceedings of IEEE International Conference on Data Mining, 2014: 1037 – 1042.

[30] PETERSEN T N, BRUNAK S, VON HEIJNE G, et al. SignalP 4.0: discriminating signal peptides from transmembrane regions[J]. Nature Methods, 2011, 8(10): 785 – 786.

[31] WARD J J, SODHI J S, MCGUFFIN L J, et al. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life [J]. Journal of Molecular Biology, 2004, 337(3): 635 – 645.

[32] BOECKMANN B, BAIROCH A, APWEILER R, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003 [J]. Nucleic Acids Research, 2003, 31(1): 365 – 370.

[33] ZHANG M L, WANG Z J. MIMLRBF: RBF neural networks for multi-instance multi-label learning[J]. Neurocomputing, 2009, 72(16/17/18): 3951 – 3956.

[34] ZHOU Z H, ZHANG M L, HUANG S J, et al. Multi-instance multi-label learning [J]. Artificial Intelligence, 2012, 176(1): 2291 – 2320.

[35] WU J S, HU H F, YAN S C, et al. Multi-instance multilabel learning with weak-label for predicting protein function in electricigens [J]. BioMed Research International, 2015, 2015: 619438.