doi:10.3969/j.issn.1001-4616.2019.03.011

基于树结构的层次性多示例多标记学习

袁京洲[1],高　昊[1],周家特[1],冯巧遇[2],吴建盛[1]

(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 210023)

[**摘要**]　针对多示例多标记学习中标记间树结构的问题,将多示例学习、多标记学习和树结构标记优化方法有机融合,提出了基于树结构标记的层次性多示例多标记学习方法 TreeMIML. TreeMIML 先将样本中的多个示例转化为单示例,然后通过多标记学习得到新样本的标记,最后通过树结构标记优化方法学习样本的最终标记. 实验结果证明,TreeMIML 方法在 G 蛋白偶联受体的生物学功能预测上获得了很好的分类性能,优于目前最好的多示例多标记学习和多标记学习方法.

[**关键词**]　层次性多示例多标记学习,树结构,G 蛋白偶联受体,生物学功能,多示例学习
[**中图分类号**]TP399　[**文献标志码**]A　[**文章编号**]1001-4616(2019)03-0080-08

A Hierarchical Multi-Instance Multi-Label Learning for Tree Structure Among Labels

Yuan Jingzhou[1], Gao Hao[1], Zhou Jiate[1], Feng Qiaoyu[2], Wu Jiansheng[1]

(1.School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)
(2.College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: This paper proposed a novel hierarchical multi-instance multi-label learning algorithm named TreeMIML to solve the challenge of tree structure among labels in multi-instance multi-label learning (MIML), by integrating multi-instance learning, multi-label learning and tree-structure optimization scheme. TreeMIML first converts multiple instances in each sample into single instance, then obtains sample outputs by multi-label learning, and finally optimizes the outputs to obtain the labels of unseen samples by a tree-structure optimization method. The experimental results show that our TreeMIML algorithm achieves good classification performance in predicting biological functions of G protein-coupled receptors, which is superior to state-of-the-art multi-instance multi-label learning and multi-label learning methods.

Key words: hierarchical multi-instance multi-label learning, tree structure, G protein-coupled receptors, biological functions, multi-instance leanring

传统的监督学习在很多研究领域得到了非常广泛的应用,但对于真实世界很多的学习问题,由于一个对象经常同时对应于多个示例及多个概念标记,它往往不能较好地利用传统的监督学习框架进行建模.

为了解决此问题,Zhou & Zhang 提出了多示例多标记学习(multi-instance multi-label learning, MIML)的学习框架[1]. 在 MIML 学习框架下,每个训练样本用多个示例进行表示且同时拥有多个概念标记. 由于多示例多标记学习模型具备强大的表示能力,因此得到了研究者们广泛的关注,并提出了各种多示例多标记学习的算法. 典型的有基于退化框架下的算法:MIMLSVM[1]、MIMLBOOST[2]、MIMLNN[3] 和 SISL-MIML[4];基于正则化机制的算法:D-MIMLSVM 算法[3] 和 M3MIML 算法[5];把单示例数据恢复为 MIML 形式数据的方法:INSDIF[6];生成式方法:DBA[7];MIML 距离度量学习[8];MIML 示例预测算法:Rankloss-Sim[9];在每个包中找到相应类别的关键示例的方法:KISAR[10];快速 MIML 预测算法:MIMLfast[11] 等. 同

收稿日期:2019-07-05.
基金项目:国家自然科学基金(61872198、81771478、61571233)、江苏省高校自然科学基金(18KJB416005)、江苏省高等学校自然科学研究项目(17KJA510003)、南京邮电大学科研基金(NY218092).
通讯联系人:吴建盛,副教授,研究方向:机器学习和生物信息学. E-mail:jansen@njupt.edu.cn

样,多示例多标记学习也在图像分类[12]、文本分类、web 网页分类[5]、视频标注[13-14]、基因表达模式分析[15]及蛋白质功能预测[16]等诸多方面得到了很好的应用. 我们研究多示例多标记学习的核心内容是能够有效利用标记间的关系. 在很多的多示例多标记学习场景中,标记间表现出树型的结构关系. 如果独立地分类学习每个标记,伴随着标记个数增加,输出空间呈指数式增长,标记间的区分难度和每个标记所需的训练样本也将急剧增加,这将会导致模型构建时巨大的存储和时间开销,而且在一些样本量少的标记上模型难以得到很好的泛化性能. 因此,急需开发一种基于树结构标记关系的多示例多标记学习方法.

本文有机整合了多种方法:单示例化方法、多标记学习与树结构优化方法,提出了层次性多示例多标记学习算法 TreeMIML,有效利用了标记间的树型依赖关系. 我们利用该算法进行了 G 蛋白偶联受体的生物学功能预测,实验表明我们算法的性能比目前最好的多示例多标记学习和多标记学习方法的性能还要好.

1 TreeMIML 算法

1.1 算法框架图

本文提出了一种基于标记间树结构的层次性多示例多标记学习算法 TreeMIML.

该算法首先把多示例转化为单示例,再通过多标记学习得到测试样本的预测值,最后通过树结构标记优化方法得到测试样本的多个标记. TreeMIML 算法的整体框架如图 1 所示.

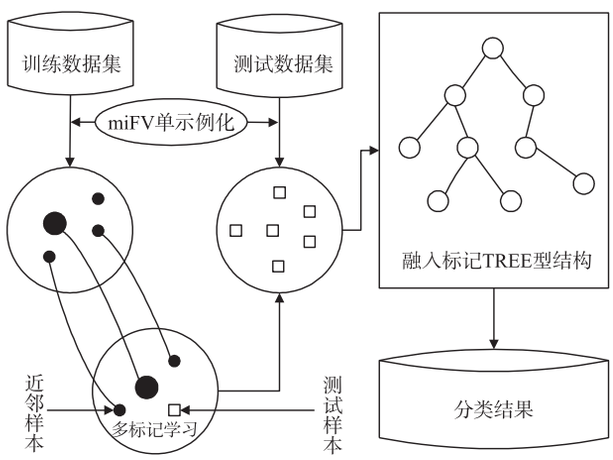

图 1 TreeMIML 算法框架图

Fig. 1 TreeMIML algorithm framework

1.2 样本单示例化

本文采用了一种高效、可扩展的单示例化算法 miFV[17],示例包通过其映射函数得到新的特征向量,它可将更多的信息编码到新的特征向量中,而且该算法计算效率高,即便使用线性分类器也能获得很好的预测结果.

miFV 单示例化方法中的包表示:假设 $R=\{R_q, q=1,\cdots,Q\}$ 是 Q 观测值的一个样本. 设概率密度函数 p,它用参数 λ 模拟 R 中元素的生成过程. 那么样本 R 可以用以下梯度向量来描述:

$$\boldsymbol{G}_\lambda^R = \nabla_\lambda \log_p(R|\lambda). \tag{1}$$

直观地说,梯度描述了如何调整参数 p 以更好地拟合数据 R. 注意,$\boldsymbol{G}_\lambda^R$ 的维数由参数 p 的数量决定,与样本 Q 的大小无关. 换句话说,它将可变长度集合 R 转换成固定长度向量 $\boldsymbol{G}_\lambda^R$. 以下单示例化方法中的映射函数 M_f,可以将具有不同示例数目的包映射到固定长度的特征向量中.

Sánchez J,et al.[18]提出了 Fisher Kernel(FK)来度量两个样本 R_1 和 R_2 之间的相似度:

$$K_{\mathrm{FK}}(R_1,R_2)=\boldsymbol{G}_\lambda^{R_1}\boldsymbol{F}_\lambda^{-1}\boldsymbol{G}_\lambda^{R_2}, \tag{2}$$

其中 $\boldsymbol{F}_\lambda$ 是 p 的 Fisher 信息矩阵:

$$\boldsymbol{F}_\lambda = E_{R\sim p}[\nabla_\lambda \log_p(R|\lambda)\,\nabla_\lambda \log_p(R|\lambda)'], \tag{3}$$

由于 $\boldsymbol{F}_\lambda$ 是对称且正定的,因此它具有 Cholesky 分解 $\boldsymbol{F}_\lambda = L_\lambda' L_\lambda$. 那么式(2)中的 FK 可以被明确地重写为一个点积形式:

$$K_{\mathrm{FK}}(R_1,R_2)=\boldsymbol{f}_\lambda^{R_1}{}'\boldsymbol{f}_\lambda^{R_2}, \tag{4}$$

$$\boldsymbol{f}_\lambda^R = L_\lambda \boldsymbol{G}_\lambda^R = L_\lambda \nabla_\lambda \log_f(R|\lambda). \tag{5}$$

上述式(5)给出的归一化梯度向量就是 Fisher 向量,也即 miFV 单示例化方法中包层面的表示形式.

我们把一个包 X_i 当作上面提到的一个样本 R. 假设包中示例独立同分布,R 中的 R_q 是由 p 独立生成的,因此,概率密度函数 p 选择高斯混合模型(GMM),从而可以使用最大似然估计(MLE)来估计训练包上的 GMM 中 p 的参数.

我们用 $\Omega=\{\omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, k=1,\cdots,K\}$ 表示第 K 个 GMM 的参数,其中 ω_k 为第 k 个高斯混合权重,$\boldsymbol{\mu}_k$ 为

均值向量,$\boldsymbol{\Sigma}_k$ 为协方差矩阵. 那么给定任意样本包 $X_i=\{x_1^i,x_2^i,\cdots,x_{n_i}^i\}$,有 $L(X_i|\Omega)=\log p(X_i|\Omega)$. 使用高斯混合模型(GMM)描述包中示例如下:

$$L(X_i|\Omega)=\sum_{j=1}^{n_i}\log p(x_j^i|\Omega)=\sum_{j=1}^{n_i}\log\sum_{k=1}^{K}\omega_k p_k(x_j^i|\Omega), \tag{6}$$

其中 p_k 是第 k 个高斯过程:

$$p_k(x_j^i|\Omega)=\frac{\exp\{-0.5(x_j^i-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(x_j^i-\boldsymbol{\mu}_k)\}}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_k|^{1/2}}. \tag{7}$$

以下我们将描述多示例单示例化方法中的映射函数 M_f 如何将一个包映射为 Fisher 向量. 在式(4)中分别对 GMM 模型中的参数 $\Omega=\{\omega_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\}$ 求梯度并归一化:

$$(\nabla_{\omega_k}L(X_i|\Omega))_{\text{Normalization}}=f_{\omega_k}^{X_i}=\frac{1}{\sqrt{\omega_k}}\sum_{j=1}^{n_i}(\gamma_j(k)-\omega_k), \tag{8}$$

$$(\nabla_{\mu_k}L(X_i|\Omega))_{\text{Normalization}}=\boldsymbol{f}_{\mu_k}^{X_i}=\frac{1}{\sqrt{\omega_k}}\sum_{j=1}^{n_i}\gamma_j(k)\left(\frac{x_j^i-\boldsymbol{\mu}_k}{\sigma_k}\right), \tag{9}$$

$$(\nabla_{\sigma_k}L(X_i|\Omega))_{\text{Normalization}}=\boldsymbol{f}_{\sigma_k^2}^{X_i}=\frac{1}{\sqrt{\omega_k}}\sum_{j=1}^{n_i}\gamma_j(k)\frac{1}{\sqrt{2}}\left[\frac{(x_j^i-\boldsymbol{\mu}_k)^2}{\sigma_k^2}-1\right], \tag{10}$$

其中 $\gamma_j(k)$ 是由第 k 个高斯混合模型生成 x_j^i 的概率. 在获得梯度并归一化之后,剩下的步骤就是计算 L_Ω. 在论文[18-19]中提出的方法为我们提供了一个 L_Ω 的封闭形式,同时它也可以被有效地解决.

注意到,样本包 X_i 中示例的维度是 d,并且我们可以看出 $f_{\omega_k}^{X_i}$ 是标量,$\boldsymbol{f}_{\mu_k}^{X_i}$ 和 $\boldsymbol{f}_{\sigma_k^2}^{X_i}$ 都是 d 维向量. 那么映射函数 M_f 以高斯混合模型 p 将样本包 X_i 映射为 $(f_{\omega_k}^{X_i},\boldsymbol{f}_{\mu_k}^{X_i},\boldsymbol{f}_{\sigma_k^2}^{X_i})$ 组成的 $(2d+1)K$-维的 Fisher 向量.

1.3　多标记学习

本文采用了一种惰性多标记学习方法 ML-KNN[20]. 给定多标记训练集 $D=\{(X_1,Y_1),\cdots,(X_m,Y_m)\}$ 中包含 q 个标记. 将未知样本 x 的标记向量记为 $\boldsymbol{y}_x$,其中 $\boldsymbol{y}_x(j)$ 表示第 j 个标记,如果 $j\in Y$,$\boldsymbol{y}_x(j)=\mathbf{1}$,反之为 $\mathbf{0}$. 设 $N(x)$ 表示 x 在训练集中的 k 个近邻样本构成的集合,多标记学习 ML-KNN 构建模型步骤如下:

Step 1　计算 $N(x)$ 中属于标记 $y(j)$ 的样本个数 C_j:

$$C_j=\sum_{a\in N(x)}\boldsymbol{y}_{x_a}(j),j\in Y. \tag{11}$$

Step 2　令 E_j 表示样本 x 属于标记 $y(j)$ 的事件(Event),$P(E_j|C_j)$ 表示当 x 的 k 近邻集合 $N(x)$ 中有 C_j 个样本属于标记 $y(j)$ 的条件下,事件 E_j 成立的后验概率,相反的,$P(-E_j|C_j)$ 表示 $N(x)$ 中有 C_j 个样本属于标记 $y(j)$ 的条件下,事件 E_j 不成立的后验概率. 多标记函数表达式如下:

$$Y=\{y(j)|f(x,y(j))=P(E_j|C_j)/P(-E_j|C_j)>1,1\leqslant j\leqslant q\}. \tag{12}$$

基于贝叶斯定理,可将求解后验概率转换为求先验概率和条件概率:

$$f(x,y(j))=\frac{P(E_j|C_j)}{P(-E_j|C_j)}=\frac{P(E_j)\times P(C_j|E_j)}{P(-E_j)\times P(C_j|-E_j)}, \tag{13}$$

则多标记分类器可重写为:

$$Y=\{y(j)|P(E_j)\times P(C_j|E_j)/P(-E_j)\times P(C_j|-E_j)>1,1\leqslant j\leqslant q\}, \tag{14}$$

式中,$P(E_j)$ 表示事件 E_j 成立时的先验概率,$P(-E_j)$ 表示事件 E_j 不成立时的先验概率,$P(C_j|E_j)$ 表示事件 E_j 成立时,近邻集合 $N(x)$ 中有 C_j 个样本属于标记 $y(j)$ 的条件概率.

Step 3　根据训练集的频率估计,先验概率可利用统计训练集中属于每个标记的样本数估计得到,根据先验概率从而可得到其条件概率.

$$P(E_j)=s+\sum_{i=1}^{N}y_{x_i}(j)/\frac{s}{2}+N, \tag{15}$$

$$P(-E_j)=1-P(E_j),\quad 1\leqslant j\leqslant q, \tag{16}$$

其中,s 是平滑参数,一般设为 1.

通过多标记学习方法,可以得到给定测试样本的标记集合 $\hat{y}$.

1.4 标记树结构优化

目前,很多多示例多标记学习中,标记之间存在关联,且在许多情况下,标记呈现层次性树(TREE)结构. 但是现在的研究中总会忽略标记之间的层次依赖性,标记树结构优化算法充分考虑了标记之间的层次性结构,提高了算法的效率.

树型结构假设根节点为正标记,除根节点之外,只有当其父节点被标记为正时,节点 i 才可能被标记为正. 具体来说,基于树型层次性结构构建中,给定测试样本 X_t,由 1.3 节可以得到测试样本 X_t 的标记集合 $\hat{y}$. 然后在本节中,融入标记之间的树型层次性结构,重新优化调参,最终得到测试样本的层次性标记集合. 其中一个关键的步骤就是找到满足以下问题的多标记集合 $\hat{y}^*$.

(1)与 1.3 节中粗略估计的标记集合 $\hat{y}$ 最大程度相似;

(2)满足标记的树型层次性结构;

(3)预先设定的正标记个数为 L.

以下将详细介绍如何在大数据下融合标记之间的树型层次性结构有效地训练模型.

CSSA 方法[21]假设已知 X_t 具有 L 个标记,如果树标记之间不存在结构化问题,那么就可以简单地选择 $\hat{y}$ 中的 L 个最大的标记作为预测标记结果. $\hat{y}$ 是测试样本在每个标记上的输出值,$\boldsymbol{\psi}=[\psi_1,\cdots,\psi_d]^{\mathrm{T}}$,$\psi_i\in\{0,1\}$,$\psi_i=0$ 表示 i 节点为正标记,否则为负标记,表述为优化问题则为:

$$\begin{aligned}&\max_{\psi}\sum_{i=1}^{d}\hat{y}_i\psi_i\\&\text{s.t. }\sum_{i=1}^{d}\psi_i=L,\end{aligned}\tag{17}$$

那么使优化式(17)最大化,就能得到测试样本的 L 个正标记集合.

由此可知,当标记的层次结构为树型时(记为 T),自然地就是扩展上述式(17)中的约束 ψ. 在后文中,这样的 ψ 被称为 T-nonincreasing,再考虑到从根节点到叶子结点,ψ 值是不会增加的. 那么优化问题可以表述为以下方式:

$$\begin{aligned}&\max_{\psi}\sum_{i\in T}\hat{y}_i\psi_i\\&\text{s.t. }\psi_i\in\{0,1\},\ \forall i\in T,\\&\sum_{i\in T}\psi_i=L,\\&\psi\text{ is }T\text{-nonincreasing.}\end{aligned}\tag{18}$$

在层次性结构中,根节点必然是正标记,不然模型预测毫无意义,加上 T-nonincreasing 的非增长性质,我们可以重新考虑以上问题,即用边界约束 $0\leqslant\psi_i\leqslant1$ 代替约束(20),(19)简化为分数背包问题,并且通过贪心算法很容易求解. 重新考虑之后的优化模型如下表示:

$$\begin{aligned}&\max_{\psi}\sum_{i\in T}\omega_i\psi_i\\&\text{s.t. }\psi_i\geqslant0,\ \forall i\in T,\\&\psi_0=1,\ \sum_{i\in T}\psi_i\leqslant L,\end{aligned}\tag{19}$$

$$\psi\text{ is }T\text{-nonincreasing},\tag{20}$$

其中 T 的根索引为 0,$\omega_i\in R$. 它的主要思想是通过将非单调树节点缩合成超节点来保证 ψ_i 是 nonincreasing 的. 超节点 S 是通过合并节点(或者超节点)与其父节点形成的,然后被分配一个超节点值(SNV),它是所有组成节点上的 ψ_i 值的平均值,由此产生的凝聚分类和选择算法都是迭代的. 并且,在每次迭代中,选择具有最大 SNV 的并且未分配的超节点 S^*,如果将 $\psi(S^*)$ 赋值为 1,不违背 T-nonincreasing 属性($\psi(pa(S^*))=1$),则分配将永久化,否则 S^* 与其父节点结合形成超节点. 重复该过程直到 $\sum_{i=1}^{d}\psi_i=L$.

TreeMIML 算法的伪代码如表 1 所示. 其中 N 是样本个数,输入 $D=\{(X_i,Y_i)\mid1\leqslant i\leqslant N\}$ 是训练样本数据集. $T=\{X_t\}$,$X_t=\{x_{t1},x_{t2}\cdots x_{tn_t}\}$ 是未知标记的测试样本,$\Omega=\{\omega_p,\boldsymbol{\mu}_p,\boldsymbol{\Sigma}_p,p=1,\cdots,P\}$ 是第 P 个 GMM(高

斯混合模型)的参数. k 是样本的近邻个数,L 是层次性结构中预设正标记个数,s 是平滑参数.

2　实验仿真和分析

2.1　数据集

G 蛋白偶联受体(G protein-coupled receptors, GPCR)是细胞信号传导过程中的重要蛋白质,它与很多人体疾病有关,预测 GPCR 的功能信息具有重要的价值[21]. 大部分 GPCR 蛋白质都有不止一个保守结构域,每个结构域可以独立执行生物学功能,也可以与相邻结构域共同作用执行功能,同时生物学功能也不止一种. 研究发现,GPCR 蛋白质功能预测这一科学问题从此角度可抽象为多示例多标记学习问题[16-17],每个蛋白质对应于多示例多标记学习中的一个样本对象,每个结构域和功能分别对应于一个示例和一个标记;描述蛋白质的生物学功能的方式有很多种,其中使用最广泛的是基因本体学(gene ontology, GO)[22]. 基因本体学根据分子功能、生物学过程、细胞组分 3 个 ontologies 来描述蛋白质信息,并且这 3 个 ontologies 下面又可以独立出不同的亚层次,不同的亚层次层层向下构成树型分支结构. 因为 GPCR 蛋白质的细胞组分基本都位于细胞膜上,本文只研究分子功能、生物学过程两个 ontologies.

表 1　TreeMIML 算法伪代码

Table 1　TreeMIML algorithm pseudocode

$\hat{y}^* = \text{TreeMIML}(D, X_t, N, k, s, L)$

for $i = 1, \cdots, N$

　将包 X_i 映射为 Fisher 向量:$\boldsymbol{f}_{\Omega}^{X_i} \leftarrow M_f(X_i, p)$

　$[\boldsymbol{f}_{\Omega}^{X_i}]_j \leftarrow \text{sign}([\boldsymbol{f}_{\Omega}^{X_i}]_j)\sqrt{|[\boldsymbol{f}_{\Omega}^{X_i}]_j|}$

　$\boldsymbol{f}_{\Omega}^{X_i} \leftarrow \boldsymbol{f}_{\Omega}^{X_i} / \|\boldsymbol{f}_{\Omega}^{X_i}\|_2$

for $j \subset q$ do

　$P(E_j) = s + \sum_{i=1}^{m} y(j) / \frac{s}{2} + m$

　$P(-E_j) = 1 - P(E_j)$

for $j \in \{0, \cdots, k\}$ do

　$P(C_j | E_j) = \dfrac{s + c[\gamma]}{s \times (k+1) + \sum_{r=0}^{k} c[\gamma]}$

　$P(C_j | -E_j) = \dfrac{s + c'[\gamma]}{s \times (k+1) + \sum_{r=0}^{k} c'[\gamma]}$

for $j \subset q$ do

　$\hat{y} = f(x, y(j)) = \dfrac{P(E_j | C_j)}{P(-E_j | C_j)} = \dfrac{P(E_j) \times P(C_j | E_j)}{P(-E_j) \times P(C_j | -E_j)}$

　$C_j = \sum_{a \in N(x)} \boldsymbol{y}_{x_a}(j)$

通过式(19)对树结构标记进行优化,得到 $\hat{y}^*$

2.2　实验结果

2.2.1　与多示例多标记学习的比较

我们与 4 种多示例多标记学习算法进行了比较,分别是:MIMLRBF[23]、ENMIMLNNBP[24]、MIMLNN[3]、MIMLSVM[1]. 这些算法均采用参考文献中的最佳参数,对于 MIMLRBF 算法,缩放因子设为 0.08,分数参数设为 0.1;MIMLKNN 算法中聚类簇的数目设置为样本包的 40%;MIMLSVM 算法中,高斯核半径 r 设置为 0.2;ENMIMLNNBP 算法中学习率设为 0.4.

表 2 显示了我们的模型与多示例多标记学习方法的比较结果. 其中 ↑ 表示评价指标值越大,性能越好,↓ 则相反;最优的结果用粗体标注. 通过表 2 可以看到,在预测 GPCRs 的 GO 分子功能和生物学过程上,我们算法的性能比其他 4 种基于多示例多标记学习的预测方法好. 例如,对 GO 分子功能,我们模型比次优的 MIMLRBF 算法高 0.028 6(Hamming Loss 值);对 GO 生物学过程,我们模型比次优的 MIMLRBF 算法高 0.038 8(Hamming Loss 值). 由此可以看出,通过树结构标记优化方法学习样本的最终标记性能优异,树结构有效提高了算法性能.

表 2　与多示例多标记学习的比较

Table 2　Comparison with multi-instance multi-label learning

type	indicators	TreeMIML	MIMLRBF	MIMLNN	MIMLSVM	ENMIMLNNBP
MF	Hamming Loss ↓	**0.147 0**	0.175 6	0.387 1	0.450 6	0.242 9
	Ranking Loss ↓	**0.117 0**	0.190 8	0.186 3	0.273 6	0.281 4
	OneError ↓	**0.289 7**	0.415 0	0.418 0	0.407 5	0.442 1
	Average Precision ↑	**0.886 8**	0.714 5	0.678 1	0.147 5	0.814 2
BP	Hamming Loss ↓	**0.143 7**	0.182 5	0.248 9	0.518 0	0.351 3
	Ranking Loss ↓	**0.191 6**	0.278 2	0.250 3	0.263 8	0.452 3
	OneError ↓	**0.229 2**	0.456 7	0.381 7	0.413 5	0.684 5
	Average Precision ↑	**0.871 8**	0.744 9	0.689 3	0.060 7	0.513 7

注:↑ 表示值越大,其模型性能越优;↓ 相反. MF:Molecular function;BP:Biological process. 最优结果用黑体进行标注.

2.2.2 与多标记学习的比较

我们与5种多标记学习算法进行了比较,分别是RAKEL[25]、MLKNN[20]、ECC[26]、IBLR-ML[27]、Meta-Labeler[28]. 5种对比算法均采用参考文献中使用的默认参数. 表3列出了与多标记学习算法在HLoss、AP上的性能比较. 通过表3可知,我们算法的性能比其他的多标记学习算法优越. 例如,在模型上,我们的比最好的算法ECC在GO分子功能上高0.008 1(Hamming Loss值),在GO生物学过程上比最优的多标记学习MLKNN高0.035 2(Hamming Loss值). 多标记学习部分使算法在GO分子功能和GO生物学过程上都得到了有效提升.

表3 与多标记学习算法的比较

Table 3 Comparison with multi-label learning algorithm

type	indicators	TreeMIML	RAKEL	ECC	MLKNN	IBLR-ML	MetaLabeler
MF	Hamming Loss ↓	**0.167 5**	0.247 0	0.175 6	0.187 1	0.450 6	0.642 9
	Ranking Loss ↓	**0.153 6**	0.217 0	0.190 8	0.186 3	0.273 6	0.281 4
	OneError ↓	**0.185 0**	0.429 0	0.415 0	0.218 0	0.407 5	0.442 1
	Average Precision ↑	**0.860 2**	0.786 8	0.714 5	0.868 1	0.147 5	0.794 2
BP	Hamming Loss ↓	**0.143 7**	0.193 5	0.182 5	0.178 9	0.418 0	0.451 3
	Ranking Loss ↓	**0.191 6**	0.293 9	0.278 2	0.250 3	0.263 8	0.452 3
	OneError ↓	**0.229 2**	0.426 0	0.456 7	0.281 7	0.413 5	0.684 5
	Average Precision ↑	**0.831 8**	0.773 6	0.744 9	0.835 0	0.060 7	0.513 7

注:↑表示值越大,其模型性能越优;↓相反. MF:Molecular function;BP:Biological process. 最优结果用黑体进行标注.

2.2.3 单示例化方法的影响

我们与另外两种单示例化方法进行了比较,即miVLAD[29]和mean(均值). miVLAD首先将所有示例聚类形成质心,然后由映射函数基于质心将每个包映射为单一向量;直接平均的方法就是将每个样本包的所有示例求平均,从而形成由单一示例描述每个样本. 实验结果如表4. 由表4可知,在GPCR的GO功能的预测中,无论是分子功能,还是生物学过程,3种单示例方法的模型性能排名如下:miFV>miVLAD>mean,因此在我们算法中采用了miFV单示例化.

表4 TreeMIML中单示例化方法对性能的影响

Table 4 The performance impact of the single instance approach in TreeMIML

type	single instance	Hamming Loss ↓	Ranking Loss ↓	OneError ↓	Average Precision ↑
MF	miFV	**0.167 5**	**0.153 6**	**0.215 0**	**0.860 2**
	miVLAD	0.239 4	0.226 7	0.423 7	0.658 9
	mean	0.252 6	0.265 0	0.563 8	0.462 8
BP	miFV	**0.143 7**	**0.191 6**	**0.221 6**	**0.831 8**
	miVLAD	0.285 0	0.185 1	0.632 0	0.655 6
	mean	0.202 7	0.236 2	0.583 7	0.422 0

注:↑表示值越大,其模型性能越优;↓相反. MF:Molecular function;BP:Biological process. 最优结果用黑体进行标注.

2.2.4 不同参数对算法性能的影响

表5显示了多标记学习ML-KNN方法中样本不同近邻个数k对算法性能的影响. 由表5可知,当近邻个数$k=20$时,模型性能最佳.

表6和表7显示了标记数量L对算法性能的影响. 对GPCR的分子功能(MF)进行预测时,$L=35$模型性能最佳(表6);对GPCR的生物学过程(BP)进行预测时,$L=20$模型性能最佳(表7).

表5 TreeMIML中近邻个数k对性能的影响

Table 5 The effect of neighbor number k in TreeMIML on performance

evaluation index	number of neighbors				
	10	15	20	25	30
Hamming Loss ↓	0.167 5	0.167 5	0.167 5	0.167 5	0.167 5
Ranking Loss ↓	0.155 7	0.156 8	**0.153 6**	0.160 2	0.164 3
OneError ↓	0.215 3	0.224 3	**0. 215 0**	0.227 8	0.234 6
Average Precision ↑	0.857 8	0.857 0	**0. 860 2**	0.853 1	0.848 9

注:↑表示值越大,其模型性能越优;↓相反. 最优结果用黑体进行标注.

表 6 标记数量对 GO 分子功能预测性能的影响

Table 6 The effect of the label number on GO molecular functional pridiction

evaluation index	Label number						
	15	20	25	30	35	40	45
Hamming Loss ↓	0.406 1	0.397 7	0.294 2	0.219 6	**0.171 1**	0.193 2	0.246 3
Ranking Loss ↓	0.287 4	0.255 8	0.213 6	0.200 7	**0.157 8**	0.181 2	0.245 5
OneError ↓	0.389 5	0.362 2	0.338 7	0.284 3	**0.248 1**	0.253 3	0.260 4
Average Precision ↑	0.584 3	0.692 6	0.778 2	0.796 6	**0.823 5**	0.811 4	0.785 6

注:↑表示值越大,其模型性能越优;↓相反. 最优结果用黑体进行标注.

表 7 标记数量对 GO 生物学过程预测性能的影响

Table 7 The effect of the label number on GO biological process pridiction

evaluation index	Label number					
	10	15	20	25	30	35
Hamming Loss ↓	0.181 7	0.157 3	**0.146 5**	0.152 2	0.170 3	0.223 1
Ranking Loss ↓	0.243 8	0.243 0	**0.197 7**	0.214 1	0.224 6	0.239 3
OneError ↓	0.237 9	0.239 3	**0.234 6**	0.240 8	0.238 5	0.243 9
Average Precision ↑	0.726 4	0.781 1	**0.815 4**	0.808 6	0.775 3	0.750 7

注:↑表示值越大,其模型性能越优;↓相反. 最优结果用黑体进行标注.

3 结语

本文将样本单示例化、多标记学习与树结构标记优化方法进行了有机融合,提出了一种新的算法:TreeMIML(基于树结构标记的层次性多示例多标记学习),并将 TreeMIML 学习算法应用到 G 蛋白偶联受体的 GO 生物学功能预测上. 实验结果证明,本文的方法优于多种多示例多标记学习和多标记学习方法.

[参考文献]

[1] ZHOU Z H,ZHANG M L,HUANG S J,et al. MIML:a framework for learning with ambiguous objects[EB/OL]. [2018-01-29]. https://arxiv.org/abs/0808.3231v1#.

[2] 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:MIT Press,2006.

[3] ZHOU Z H,ZHANG M L,HUANG S J,et al. Multi-instance multi-label learning[J]. Artificial intelligence,2008,176(1):2291-2320.

[4] NGUYEN N. A new SVM approach to multi-instance multi-label learning[C]//IEEE International Conference on Data Mining. Sydney:IEEE,2011.

[5] ZHANG M L,ZHOU Z H. M3MIML:a maximum margin method for multi-instance multi-label learning[C]//Proceedings of the 8th IEEE International Conference on Data Mining(ICDM 2008). Pisa:IEEE,2008.

[6] ZHANG M L, ZHOU Z H. Multi-label learning by instance differentiation[C]//Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence. Vancouver:AAAI,2007.

[7] YANG S,ZHA H,HU B. Dirichlet-bernoulli alignment:a generative model for multi-class multi-label multi-instance corpora[J]. Advances in neural information processing systems,2009,22:2143-2150.

[8] JIN R, WANG S, ZHOU Z H. Learning a distance metric from multi-instance multi-label data[C]//IEEE Conference on Computer Vision & Pattern Recognition. Miami:IEEE,2010.

[9] BRIGGS F,FERN X Z,RAICH R. Rank-loss support instance machines for MIML instance annotation[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2012:534-542.

[10] LI Y F,HU J H,JIANG Y,et al. Towards discovering what patterns trigger what labels[C]//Twenty-Sixth AAAI Conference on Artificial Intelligence. Toronto,2012.

[11] SHENG J H, WEI G, ZHI H Z. Fast multi-instance multi-label learning[J]. IEEE transactions on pattern analysis and machine intelligence,2014,3:1868-1874.

[12] ZHA Z J,HUA X S,MEI T,et al. Joint multi-label multi-instance learning for image classification[C]//IEEE Conference on Computer Vision and Pattern Recognition. Anchorage:IEEE,2008.

[13] XU X S, JIANG Y, XUE X, et al. Semi-supervised multi-instance multi-label learning for video annotation task[C]// Proceedings of the 20th ACM International Conference on Multimedia. Nara:ACM,2012.

[14] XU X S. Ensemble multi-instance multi-label learning approach for video annotation task[C]//Proceedings of the 19th International Conference on Multimedea. Scottsdale:DBLP,2011.

[15] LI Y X,JI S,KUMAR S,et al. Drosophila gene expression pattern annotation through multi-instance multi-label learning[C]// International Jont Conference on Artifical Intelligence. San Francisco:Morgan Kaufmann Publishers Inc,2009.

[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] WU J S,HU H F,YAN S C,et al. Multi-instance multilabel learning with weak-label for predicting protein function in electricigens[EB/OL]. [2018-01-29]. http://www.hindawi.com/journals/bmri/2015/619438/.

[18] SÁNCHEZ J,PERRONNIN F,MENSINK T,et al. Image classification with the fisher vector:theory and practice[J]. International journal of computer vision,2013,105(3):222-245.

[19] PERRONNIN F,DANCE C. Fisher kernels on visual vocabularies for image categorization[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis:IEEE Computer Society,2007.

[20] ZHANG M L,ZHOU Z H. ML-KNN:a lazy learning approach to multi-label learning[J]. Pattern recognition,2007,40(7): 2038-2048.

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

[22] CAMON E,MAGRANE M,BARRELL D,et al. The gene ontology annotation(goa) database:sharing knowledge in uniprot with gene ontology[J]. Nucleic acids research,2004,32(Suppl 1):D262-D266.

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

[24] 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.

[25] TSOUMAKAS G, VLAHAVAS I. Random *k*-labelsets:an ensemble method for multilabel classification[C]//European Conference on Machine Learning. Berlin,Heidelberg:Springer,2007.

[26] READ J,PFAHRINGER B,HOLMES G,et al. Classifier chains for multi-label classification[J]. Machine learning,2011, 85(3):333.

[27] CHENG W,HÜLLERMEIER E. Combining instance-based learning and logistic regression for multilabel classification[J]. Machine learning,2009,76(2/3):211-225.

[28] TANG L,RAJAN S,NARAYANAN V K. Large scale multi-label classification via metalabeler[C]//Proceedings of the 18th International Conference on World Wide Web. New York:ACM,2009.

[29] WEI X S,WU J,ZHOU Z H. Scalable algorithms for multi-instance learning[J]. IEEE transactions on neural networks and learning systems,2016,28(99):1-13.

[责任编辑:丁　蓉]