CCML 2021 第十八届中国机器学习会议
The 18th China Conference on Machine Learning
2021年8月6-8日 中国·长沙
AUG 6-8, 2021 CHANGSHA, CHINA

一种面向多标签分类的在线主动学习算法

龚楷伦　翟婷婷　唐鸿成
(扬州大学信息工程学院，江苏扬州，225127)

摘要：

针对现有算法多标签分类器收敛效率低和标签查询策略未考虑特征辨别能力的弊端，提出一种新的多标签在线主动学习算法， 该算法能在较小的数据标注代价下取得较好的多标签分类性能。算法使用镜像梯度下降规则更新其子分类器，在识别需要查询真实标签的实例时，算法采用了基于特征判别信息的采样策略。实验结果表明：该算法的多标签分类性能优于其对比算法。

简介：

多标签分类因在文本分类、图像识别、信息检索等多个领域的广泛应用，如今已经成为了机器学习领域的研究热点。多标签分类的**目的**是构建一个有效预测未知样本所属的标签集合的多标签分类器，供下游任务使用。

在线主动学习可以有效降低数据标注的**成本**，并且在处理流式数据和大规模数据方面优势明显，能很好的应对当今“**大数据时代**”的社会现状。

House	Tree	Beach	Cloud	Mountain	Animal
Yes	Yes	no	Yes	no	no

多标签分类问题示例

现存问题：

现有的面向多标签分类的在线主动学习算法，大都适用于**单一**领域；且采用**一阶**的在线更新方式，这不利于多标签分类器收敛到最佳状态。

同时，现有算法的**采样策略**仅考虑预测的不确定程度，忽略了实例中包含的特征的**辨别信息**，有效利用这些信息有助于识别出需要查询的关键实例。

研究内容：

1、本文提出的算法使用**镜像梯度下降**更新规则实现在线更新，同时引入对角矩阵$\boldsymbol{H}_t^{(i)}$用于记录实例中特征的更新信息，该规则使得子分类器的各个维度在更新过程中都有其**自适应的**学习速率，有助于分类器更有效的收敛到最佳状态。

2、算法的**采样策略**不仅利用了预测的不确定程度：$p_t^{(i)} = |(\boldsymbol{w}_t^{(i)})^{\top}\boldsymbol{x}_t|$，还利用了实例中包含的特征的辨别信息：$v_t^{(i)} = \mathbf{x}_t^{\top}(\mathbf{H}_{t-1}^{(i)})^{-1}\mathbf{x}_t$。实现该算法的**伪代码**如右图所示。

伪代码：

基于判别采样和镜像梯度下降规则的MLAMD-_D算法

输　入：超参数：$\delta>0$、$\eta>0$和$b>0$；

初始化：$\forall\, i \in [C]$, 设置 $\boldsymbol{w}_1^{(i)} = 0$ 且 $\boldsymbol{H}_0^{(i)} = \delta\boldsymbol{I}$;

for t = 1, 2, ... do

for i = 1 to C

接收实例$\mathbf{x}_t$，得到预测结果$\hat{y}_t^{(i)} = \mathrm{sign}[(\mathbf{w}_t^{(i)})^{\top}\mathbf{x}_t]$；

从参数为$b/(b+q_t^{(i)})\P[q_t^{(i)}>0])$的努利分布中，提取随机变量$Z_t^{(i)} \in \{1, 0\}$，其中$q_t^{(i)} = p_t^{(i)} - \frac{\eta}{2}v_t^{(i)}$；

if $Z_t^{(i)} = 1$, then 查询$y_t^{(i)} \in \{\pm 1\}$,

计算$\boldsymbol{g}_t^{(i)} = \begin{cases} -y_t{}^{(i)}\,\boldsymbol{x}_t, & \textbf{if } y_t{}^{(i)}\,(\boldsymbol{w}_t^{(i)})^{\top}\boldsymbol{x}_t < 1; \\ \mathbf{0}, & \text{其他情况。} \end{cases}$；

else 设置 $\boldsymbol{g}_t^{(i)} = \mathbf{0}$;

end if

设置 $\mathbf{G}_{1:t}^{(i)} = [\boldsymbol{g}_1^{(i)}, \cdots, \boldsymbol{g}_t^{(i)}]$；

计算 $\boldsymbol{H}_t^{(i)} = \delta\boldsymbol{I} + \mathrm{diag}\left(\boldsymbol{s}_t^{(i)}\right)$, 其中 $s_{t,j}^{(i)} = \|\mathbf{G}_{1:t,j}^{(i)}\|_2$；

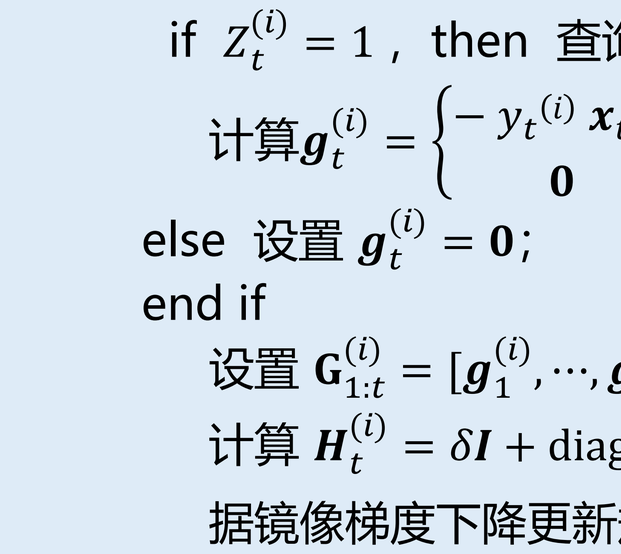
据镜像梯度下降更新规则计算$\mathbf{w}_{t+1}^{(i)}$；

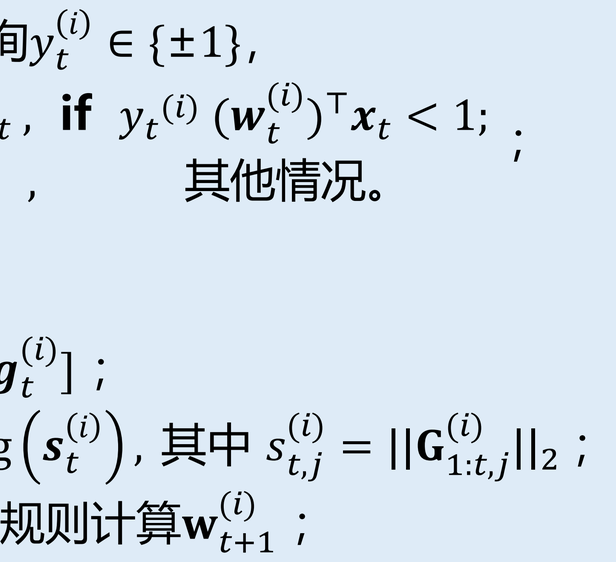
end for

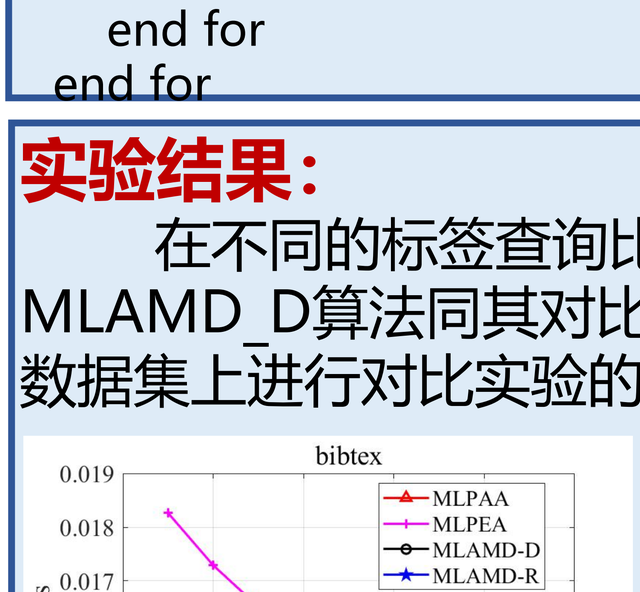
end for

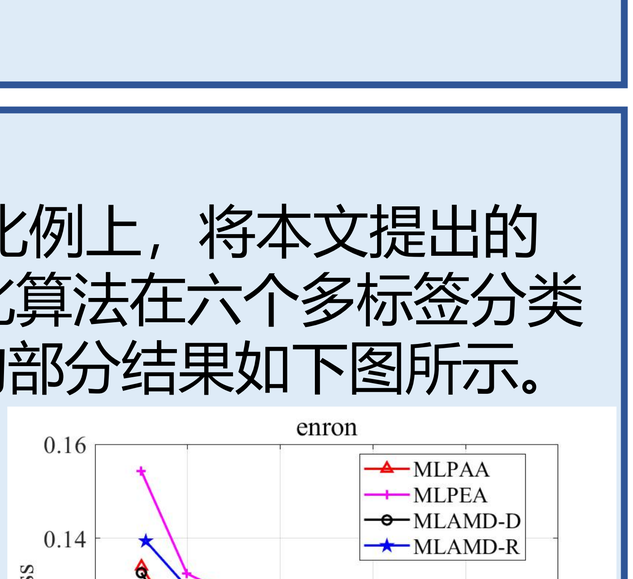
实验结果：

在不同的标签查询比例上，将本文提出的MLAMD_D算法同其对比算法在六个多标签分类数据集上进行对比实验的部分结果如下图所示。









总结：

本文提出了一种新的面向多标签分类的在线主动学习算法，该算法能有效降低数据标注的成本，同时在六个多标签分类数据集上执行的对比实验证明了本文所提算法的可行性与有效性。

投稿人 姓名：龚楷伦　翟婷婷　唐鸿成
单位：　　扬州大学信息工程学院