

基于分解策略的多标签在线特征选择算法*

张永伟^{1,2}, 朱 祁^{1,2}, 吴永城^{1,2}

(1.南瑞集团(国网电力科学研究院)有限公司, 江苏 南京 210003;

2.南京南瑞智慧交通科技有限公司, 江苏 南京 210032)

摘要: 在线学习方法是用于大规模数据集的、高效且可扩展的机器学习算法。然而,在对多标签数据集进行特征选择时,传统的在线多标签学习方法需要访问数据集的所有特征,当数据集具有较高维度时,这种在线学习方式并不能适用于实际情景。针对多标签数据集的特征选择,在现有研究的基础上,使用二类分解策略,提出基于分解策略的多标签在线特征选择算法。该算法利用稀疏正则化和截取方法进行在线特征选择,降低计算复杂度。实验表明,算法的特征选择性能优于其他多标签在线特征选择算法。

关键词: 特征选择; 在线学习; 多标签分类; 二类分解策略

中图分类号: TP305; TP181

文献标识码: A

DOI: 10.19358/j.issn.2097-1788.2022.04.010

引用格式: 张永伟,朱祁,吴永城. 基于分解策略的多标签在线特征选择算法[J]. 网络安全与数据治理, 2022, 41(4): 65-71, 77.

Online multi-label feature selection algorithm based on binary relevance strategy

Zhang Yongwei^{1,2}, Zhu Qi^{1,2}, Wu Yongcheng^{1,2}

(1.NARI Group Corporation(State Grid Electric Power Research Institute), Nanjing 210003, China;

2.Nanjing NARI Intelligent Transport Technology Co., Ltd., Nanjing 210032, China)

Abstract: Online learning method is an efficient and extensible machine learning algorithm for large-scale applications. However, when selecting features for multi-label datasets, the traditional online multi-label learning algorithms need to access all the features of the dataset. When the dataset has a higher dimension, this online learning method cannot be applied to the actual situation. Based on the existing research, a multi-label online feature selection algorithm based on decomposition strategy is proposed by using binary relevance strategy. This algorithm uses sparse regularization and interception methods for online feature selection to reduce computational complexity. The experimental results show that the feature selection performance of the algorithm is better than that of other multi-label online feature selection algorithms.

Key words: feature selection; online learning; multi-label classification; binary relevance strategy

0 引言

近年来,随着多标签分类问题的深入研究,出现了大量的多标签分类算法^[1-2]。目前,在多标签分类中,存在四种主要的处理策略:数据分解法、算法扩展法、混合法和集成法^[3-4]。特征选择是多标签分类问题中的一个重要课题,并且已经进行了广泛研究。对于分类,特征选择的目标是通过相关特征的一个子集来构建有效的预测模型,通过消除不相关和冗余特征,可以减轻维度灾难的影响,提高

泛化性能,加快学习过程,提高模型预测的性能。特征选择已在许多领域得到应用,特别是在涉及高维数据的问题中。

虽然已经进行了广泛研究,但大多数现有的特征选择研究都局限于批量学习,假定特征选择任务是以离线/批量学习的方式进行的,而且训练实例的特征是先验的。这样的假设并不总是适用于训练样本以顺序方式到达的实际应用。与批量学习方式相比,在线学习方式^[5]则采用增量的方式处理数据集,相对而言,计算代价要小于批量学习算法。在现有的多标签在线分类算法中计算数

* 基金项目:国家电网科技项目(52460D220001)

据的全部特征信息是需要代价的。尤其是存在高维数据和数据冗余时,传统的多标签在线分类算法,需大量计算且分类性能较差。本文利用在线学习的优势,研究了多标签在线特征选择问题,旨在通过有效地探索在线学习方法来解决多标签特征选择问题。具体而言,多标签在线特征选择的目标是研究在线分类器,其仅涉及用于分类的少量和固定数量的特征。当处理高维度的连续训练数据时,如在线垃圾邮件分类任务(其中传统的批量特征选择方法不能直接应用),在线特征选择尤为重要和必要。

本文提出一种基于分解策略的多标签在线特征选择算法——MLFS_{CD}算法。该算法采用二类分解策略,将多标签特征选择问题分解为二类特征选择问题,然后应用二类在线特征选择算法,完成多标签在线分类任务。通过多标签特征选择得到更好的特征子集,能高效地处理高维稀疏数据集。通过实验比较,结果显示本文提出的 MLFS_{CD} 算法能够有效地降低数据维度,同时算法的分类性能也优于其他多标签在线特征选择算法。

1 相关工作

在机器学习算法中,在线学习是一类高效且扩展性较好的方法,这种增量的方式使在线学习能够有效地应对数据规模较大的分类任务。其中,感知器算法是一种简单、经典的在线学习方法^[6],它通过预测样本标签与真实的样本标签是否相等,然后根据损失函数的大小,以梯度下降的方式求解并更新分类器模型。随着在线学习方法的深入研究,大量的单标签在线分类算法被提出^[7-8]。其中,OGD 算法是以在线梯度下降算法优化不同的损失函数定义的目标函数。CW 算法则通过最小化新的权值向量分布与之前的权值向量分布之间的 KL 散度(Kullback-Leibler Divergence),来保证正确分类的概率大于设定的阈值,该算法采用一个比较主动的方式更新权值向量的分布,可以实现快速更新模型。SCW 算法则克服了 CW 算法线性不可分、数据和噪声不敏感的缺点,通过在 CW 算法模型中加入了惩罚项,提高算法分类性能。

相比单标签在线分类问题,多标签在线分类问题的研究则更为复杂,也更符合现实应用。文献[9]用二类相关分解策略,结合已有的二类在线“被动-进攻”主动学习算法,提出基于分解策略的多标签

在线“被动-进攻”主动学习算法。文献[10]同样利用二类分解策略,提出一种基于判别采样和镜像梯度下降规则的多标签在线主动学习算法,从而更有助于分类器收敛到最佳状态。

尽管相关文献对多标签在线分类算法进行了研究,但大多数多标签在线分类算法都需要访问数据样本的所有特征^[11-12]。因此,为了降低模型学习成本和计算代价,在不影响模型的预测能力的前提下,有学者提出使用特征选择方法^[13-14]以降低学习成本和计算代价。依照不同的选择标准,可以将多标签特征选择方法分为三种:过滤(filter)法,包裹(wrapper)法和嵌入式(embedded)法^[15-16]。过滤法是在分类算法开始之前通过测量样本特征与类标签之间的相关性产生特征子集。包裹法依赖于预定的分类算法的性能来确定最优特征子集。包裹法针对特定模型,易于产生较高的分类性能,但在计算上通常比过滤法花费更大代价。嵌入法旨在将特征选择过程嵌入到分类器训练过程中,将特征选择和分类同时进行。它们通常比包裹法更快并且能够为学习算法提供合适的特征子集。

目前,有关多标签在线特征选择算法的文献很少。本文利用在线学习的方式,只允许分类模型访问少量或固定数量的特征,其目标是从高维的多标签稀疏数据集中学习线性分类器,进而解决特征选择问题,并对分类模型中的非零元素的数量进行严格的约束。本文提出的在线特征选择算法采用二类分解策略,将多标签分解成多个独立的二分类问题。该策略简单高效、计算代价低。在实验中采用 5 个评价指标进行比较,结果显示 MLFS_{CD} 相比其他算法都有比较好的性能,尤其是在特征数较多的数据集上,优势更加明显。

2 多标签在线特征选择算法

本文在处理多标签特征选择问题时,设样本标签的集合为 $Q=\{1, 2, \dots, q\}$, q 表示样本标签个数;假设样本 X 所对应的相关标签集合 $L \subseteq 2^Q$ 。给定一个大小为 N 且独立同分布的训练样本集合 D :

$$D=[(X_1, L_1), \dots, (X_i, L_i), \dots, (X_n, L_n)] \quad (1)$$

另外,标签集合也可以表示成二进制的形式:

$$Y=[y_1, \dots, y_i, \dots, y_N] \in \{0, 1\}^{q \times N} \quad (2)$$

其中,第 i 个样本对应的标签也可以表示为一个二进制标签: $y_i \in \{-1, 1\}^q$, 1 表示样本 X_i 包含该标签, -1 表示不包含该标签。

2.1 二类分解策略

多标签分类算法中,二类分解策略(Binary Relevance, BR)是一种数据分解的方法,将多标签分类问题分解成 q 个相互独立的二类分类问题。对于一个标签集合为 $Q=\{1, 2, \dots, q\}$ 的多标签分类问题,在二类分解方法中,被分解为 q 个二类分类的问题,每一个二类分类问题对应标签集合 Q 中一个可能的类别标签。

二类分解的过程是首先为标签集合中的每一个标签建立对应的二类的训练集合,对于第 j 个类别标签,可以将训练样本集合分解为如下形式:

$$D=\{(X_i, y_i) | 1 \leq i \leq n, 1 \leq j \leq q\} \quad (3)$$

其中 y_j 为 -1 或 1, -1 表示不是样本的相关标签, 1 表示是样本的相关标签。接着, BR 方法利用单标签分类的二类分类算法来训练一个二类分类器 h_j 。这样,就会产生 q 个二类分类器,每一个多标签训练集 D 中的样本都将参与这 q 个分类器的分类过程。在二类分类器 h_j 中,如果样本 X_i 与标签 j 相关联,则被作为一个正类的样本,否则,被作为一个负类的样本,在预测阶段,将未知标签的样本在每一个二类分类器中的预测结果结合起来,就可以得到预测出的未知样本的标签集合 L^q ,并且满足:

$$L^q=\{j | h_j(X) > 0, 1 \leq j \leq q\} \quad (4)$$

与其他的多标签数据分解策略方法相比, BR 方法不仅简单有效,而且计算代价较低。不仅如此,二类相关分解方法假设标签之间是相互独立的,对标签的添加或删除都不会对剩余的其他标签产生影响,对某一个标签的训练和预测也不会对其他标签的分类过程产生影响,这样有利于在线实现该方法。因此, BR 方法适合于处理大规模的多标签分类问题。

2.2 二类在线特征选择(OFS)算法

二类在线特征选择 OFS 算法^[17]是基于向量的分类器模型 $W \in \mathbb{R}^n$, 其中最多包含 B 个非零元素,采用形式 $\text{sign}(W \cdot X)$ 表示,其中向量 W 中的每个元素表示分配给每个样本 X_i 的特征权重。在 t 时刻, OFS 算法的权重为 W_t , $y_t(W_t \cdot X_t)$ 表示预测函数。

当预测值大于 0 时,分类器预测正确。为了更好地预测样本标签, OFS 算法使用在线梯度下降(OGD)算法来优化模型 W 。

$$\min_w f(W) = \frac{1}{2} \|W - W_t\|^2 + Cl(W_t; (X_t, y_t)) \quad (5)$$

其中, $C > 0$ 是惩罚参数,损失函数为铰链损失(hinge loss):

$$l(W; (X, y)) = \max(0, 1 - y_t(W_t \cdot X_t)) \quad (6)$$

t 时刻,分类器更新模型,其中 η 表示学习率:

$$W_{t+1} = W_t + \eta y_t X_t \quad (7)$$

为了更好地进行特征选择,在更新模型中使用截取函数,确保更新分类器模型 W 中的元素是最大的 B 个元素。算法 1 给出了截取函数的具体步骤。

算法 1 截取函数

Truncate($\hat{W}, B$)

If $\|W\|_0 > B$ then

$W = \hat{W}^B$; 其中 $\hat{W}^B$ 为 W 中除了 B 个最大值之外,其他元素为零的向量。

else

$W = \hat{W}$;

end if

return W

2.3 基于分解策略的多标签在线特征选择算法

结合二类分解策略,本文提出了基于分解策略的多标签在线特征选择算法——MLFS_{CD} 算法。在多标签在线分类算法中,分类器以在线学习的方式预测样本 X_t 的标签集合,最后根据样本标签的真实集合计算损失来判断分类器是否更新。然而,对于大数据量的数据集而言,这些数据具有复杂化与高维化的特点,尤其对于稀疏数据集而言更是存在着大量的冗余性和无关性的特征。而这些特征增加了机器学习算法的复杂度和运行时间,同时降低了模型预测的准确性。为了解决以上问题,本文使用在线特征选择的处理方式。对于多标签数据集,每个样本都包含多个标签,本文在不考虑标签相关性的条件下,使用分解策略将多标签特征选择问题分解成二类特征选择问题。具体地, MLFS_{CD} 算法首先初始化权重矩阵:

$$W = [W_1, \dots, W_t]_{q \times n} \quad (8)$$

在 t 时刻,对于样本 X_t , MLFS_{CD} 算法首先计算其预测函数:

$$\hat{y}_{t,j} = \text{sign}(W_{t,j}^T X_t) \quad (9)$$

其中, $\text{sign}(\cdot)$ 表示符号函数。利用分解策略方式,分类器将样本 X_t 分解成 q 个独立的二类在线 OFS 算法。每个样本 X_t 含有 q 个标签,在 t 时刻,如果分类器预测样本 X_t 错误,则更新分类器权重 $W_{t,j}$:

$$W_{t+1,j} = W_{t,j} + \eta y_{t,j} X_i \quad (10)$$

然后,根据投影公式对分类器做二次操作:

$$W_{t+1,j} = \min \left\{ 1, \frac{1/\sqrt{\lambda}}{\|W_{t+1,j}\|_2} \right\} * W_{t+1,j} \quad (11)$$

最后,利用截取方法选择 B 个特征。算法 2 概括了 MLFS_{GD} 算法的具体步骤。

算法 2 MLFS_{GD} 算法

输入:正则化参数 λ ; 步长 η ; 选择特征数 B 。

初始化: $W_i = 0$;

for $t = 1, 2, \dots$

for $j = 1$ to q

分类器预测样本 X_i 标签:

$$\hat{y}_{t,j} = \text{sign}(W_{t,j}^T X_i);$$

给出样本 X_i 的真实标签: $y_{t,j} \in Y$

if $l < 0$ then

根据式(10)更新 W ;

根据式(11)对分类器 W 做二次操作;

根据截取函数 $\text{Truncate}(W, B)$ 选择特征数;

end if

end for

end for

3 实验结果与分析

3.1 实验设置

实验中采用了 MuLan 提供的多标签分类数据集,数据主要集中在文本、视频和图像等领域。表 1 列出了六个多标签基准数据集 (Mediamill、Scene、Delicious、Bibtex、Rcv1v2 和 Tmc2007) 的详细信息。本文采用了 5 个多标签评价指标 (汉明损失 (Hamming loss)、覆盖率 (coverage)、排序损失 (ranking loss)、平均精度 (average precision) 和 1-错误率 (one error)) 来进行算法性能的比较。其中,为了方便表示,对于覆盖率指标,将该指标结果做了归一化处理,即将覆盖率数值除以数据集的标签数。

本文在操作系统 Windows 10 环境中进行实验,使用 MATLAB2018b 开发编码。实验中,将所选特征的数量设置为每个数据集的 10% ($0.1 \times \text{维度}$),正则化参数 λ 设为 0.01,学习率 η 设为 0.2。所有多标签在线算法都使用相同的参数。对所有实验数据集进行 10 次随机排列,然后对结果进行平均得到最终实验结果。

本文使用三个基于特征选择的多标签特征选择算法: MOFS^[18]、MLPE_{mm} 和 MLRAND 与本文提出的 MLFS_{GD} 算法进行比较。MOFS 算法是将多标签数据集划分为多个单标签数据集的一种在线特征选择算法,解决不平衡数据集的分类性能问题。MLPE_{mm} 算法是基于感知器的多标签算法,利用分解策略,将基于感知器的特征选择算法推广至多标签分类。MLRAND 算法采用随机查询的方式,随机选择固定数量的特征值,并利用分解策略来解决多标签分类问题。

3.2 实验结果与分析

本文首先在两个特征数较多的 Rcv1v2 和 Tmc2007 数据集上,给出汉明损失、平均精度、1-错误率和覆盖率四种指标的比较结果,如图 1 和图 2 所示。其中图 1 表示四种算法在数据集 Rcv1v2 上四种不同评价指标的比对结果。可以看出本文提出的 MLFS_{GD} 算法优于其他算法。图 2 表示四种算法在数据集 Tmc2007 上不同评价指标的对比结果,可以看到在评价指标 1-错误率和平均精度上,MLFS_{GD} 算法明显优于其他算法,在另外两种指标中,MLFS_{GD} 算法优于 MOFS 算法和 MLRAND 算法,仅次于 MLPE_{mm} 算法。同时可以看到随着选择特征比例的不断增加,所有算法的分类性能没有特别明显的提升,这也证明特征选择的优势。

为了更好地评估本文提出的 MLFS_{GD} 算法,表 2~表 6 给出了在六个数据集上四种算法的实验结果。

表 1 多标签基准数据集信息

数据集	领域	训练样本	测试样本	特征数	标签数	标签基数	标签密度
Mediamill	视频	30 993	19 914	120	101	4.38	0.043
Scene	图像	1 211	1 196	294	6	1.07	0.178
Delicious	文本	12 920	3 185	500	983	19.02	0.019
Bibtex	文本	4 480	2 515	1 836	159	2.40	0.015
Rcv1v2	文本	3 000	3 000	47 236	101	2.88	0.029
Tmc2007	文本	21 519	7 077	49 060	22	2.16	0.098

表 2 是四种算法在六个不同数据集上的汉明损失值。可以看出, $MLFS_{GD}$ 算法取得比较低值, 尤其在数据特征数比较大的 $Tmc2007$ 和 $Rev1v2$ 数据集上, $MLFS_{GD}$ 算法的汉明损失值小于 $MOFS$ 和 $MLRAND$ 算法。表 3 所示为四种算法在不同数据集上的 1-错误率值, 在六个数据集上除了 $Delicious$ 数据集以外, $MLFS_{GD}$ 算法都取得最小的值, 特别是数据集 $Tmc2007$ 和 $Rev1v2$ 。表 4 是四种算法的排序损失值, $MLFS_{GD}$ 算法在大部分数据集上为最优, 其中, 在 $Tmc2007$ 数据集上, $MLPE_{lmm}$ 算法要优于文本提出的 $MLFS_{GD}$ 算法。表 5 所示为四种算法的覆盖率值, 从表中可以观察到, $MLFS_{GD}$ 算法除了在 $Tmc2007$ 数据集次优之外, 其他数据集上结果最优。表 6 所示为四种算法的平均精度值, $MLFS_{GD}$ 算法在所有数据集上均为最优。通过以上实验证明, 本文提出的 $MLFS_{GD}$ 算法在做多标签在线特征选择时能够取得比较好的结果。

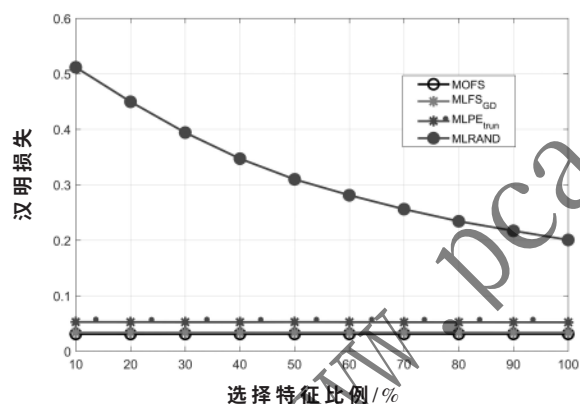
4 结论

本文提出了基于分解策略的多标签在线特征

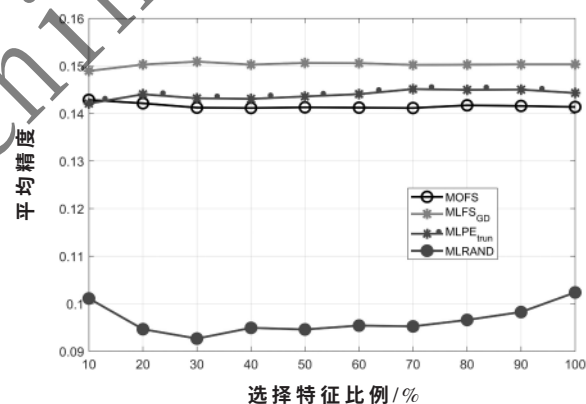
选择算法, 将多标签在线特征选择问题分解成多个二类在线特征选择问题, 进而对样本进行特征选择。实验基于六个数据集, 使用不同的多标签评价指标, 比较 $MLFS_{GD}$ 算法和其他三种在线特征选择算法在特征子集个数从 10% 增加到 100% 时的性能。实验表明, $MLFS_{GD}$ 算法在处理多标签在线特征选择时, 性能优于其他在线算法。由于该算法假设样本之间相互独立, 并没有考虑样本标签相关性, 未来工作中, 将就标签的相关性进行进一步研究。

参考文献

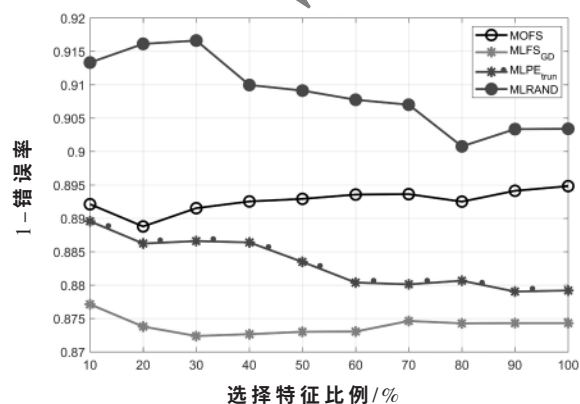
- [1] XU J H. Multi-label lagrangian support vector machine with random block coordinate descent method[J]. Information Sciences, 2016, 329: 184-205.
- [2] 牟甲鹏, 蔡剑, 余孟, 等. 基于标签相关性的类属属性多标签分类算法[J]. 计算机应用研究, 2020, 37(9): 2656-2658, 2673.
- [3] 朱旭东, 熊赞云. 基于多层次注意力与图模型的图像多标签分类算法[J]. 计算机工程, 2022, 48(4): 173-



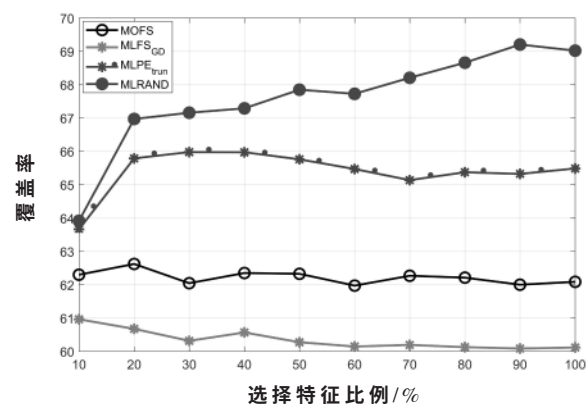
(a) 四种算法在 $Rev1v2$ 上的汉明损失



(b) 四种算法在 $Rev1v2$ 上的平均精度

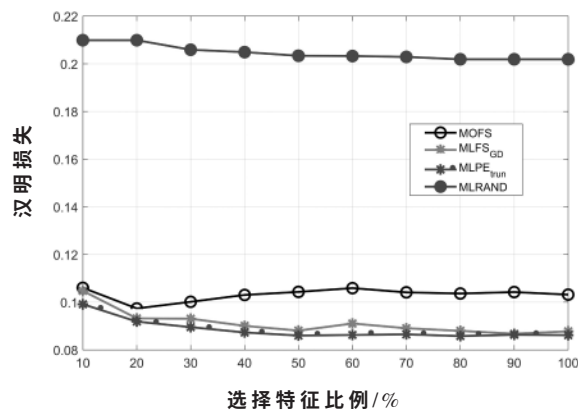


(c) 四种算法在 $Rev1v2$ 上的 1-错误率

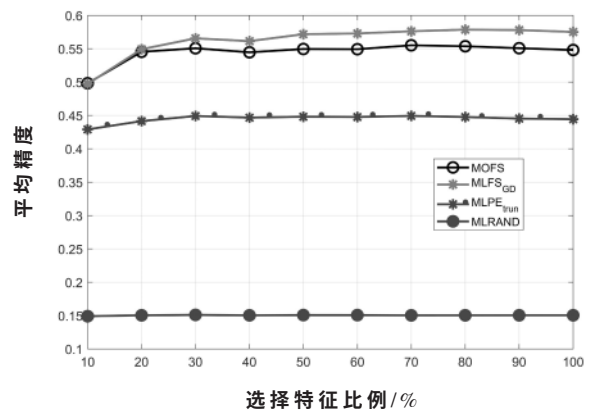


(d) 四种算法在 $Rev1v2$ 上的覆盖率

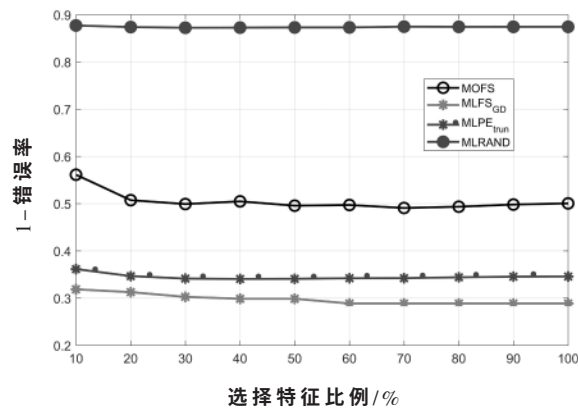
图 1 四种算法在 $Rev1v2$ 上的比较结果



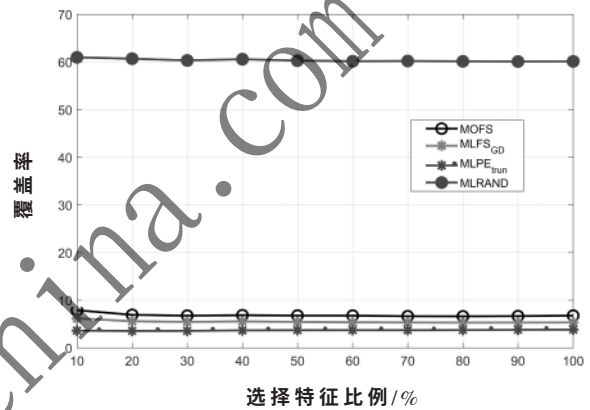
(a) 四种算法在 Tmc2007 上的汉明损失



(b) 四种算法在 Tmc2007 上的平均精度



(c) 四种算法在 Tmc2007 上的 1-错误率



(d) 四种算法在 Tmc2007 上的覆盖率

图 2 四种算法在 Tmc2007 上的比较结果

表 2 四种算法在六个数据集上的汉明损失

数据集	MLFS _{GD}	MOFS	MLPE _{train}	MLRAND
Mediamill	0.038 $\pm$ 0.004	0.552 $\pm$ 0.048	0.055 $\pm$ 0.010	0.054 $\pm$ 0.015
Scene	0.162 $\pm$ 0.015	0.176 $\pm$ 0.034	0.214 $\pm$ 0.028	0.289 $\pm$ 0.005
Delicious	0.281 $\pm$ 0.062	0.785 $\pm$ 0.008	0.451 $\pm$ 0.003	0.915 $\pm$ 0.064
Bibtex	0.017 $\pm$ 0.001	0.018 $\pm$ 0.001	0.048 $\pm$ 0.001	0.169 $\pm$ 0.110
Rcv1v2	0.034 $\pm$ 0.004	0.036 $\pm$ 0.001	0.052 $\pm$ 0.001	0.570 $\pm$ 0.121
Tmc2007	0.104 $\pm$ 0.016	0.107 $\pm$ 0.020	0.095 $\pm$ 0.001	0.209 $\pm$ 0.024

表 3 四种算法在六个数据集上的 1-错误率

数据集	MLFS _{GD}	MOFS	MLPE _{train}	MLRAND
Mediamill	0.215 $\pm$ 0.037	0.512 $\pm$ 0.254	0.367 $\pm$ 0.193	0.750 $\pm$ 0.226
Scene	0.449 $\pm$ 0.064	0.534 $\pm$ 0.061	0.553 $\pm$ 0.088	0.823 $\pm$ 0.015
Delicious	0.332 $\pm$ 0.014	0.545 $\pm$ 0.021	0.392 $\pm$ 0.036	0.226 $\pm$ 0.162
Bibtex	0.482 $\pm$ 0.018	0.639 $\pm$ 0.033	0.516 $\pm$ 0.013	0.857 $\pm$ 0.117
Rcv1v2	0.876 $\pm$ 0.005	0.895 $\pm$ 0.010	0.882 $\pm$ 0.020	0.913 $\pm$ 0.199
Tmc2007	0.303 $\pm$ 0.041	0.576 $\pm$ 0.065	0.353 $\pm$ 0.011	0.872 $\pm$ 0.156

表 4 四种算法在六个数据集上的排序损失

数据集	MLFS _{GD}	MOFS	MLPE _{min}	MLRAND
Mediamill	0.126+/-0.020	0.597+/-0.103	0.230+/-0.035	0.308+/-0.042
Scene	0.449+/-0.064	0.534+/-0.061	0.553+/-0.088	0.498+/-0.028
Delicious	0.473+/-0.064	0.663+/-0.011	0.695+/-0.012	0.942+/-0.049
Bibtex	0.164+/-0.007	0.246+/-0.010	0.208+/-0.005	0.618+/-0.019
Rev1v2	0.395+/-0.007	0.399+/-0.026	0.436+/-0.006	0.684+/-0.078
Tmc2007	0.182+/-0.016	0.236+/-0.024	0.095+/-0.002	0.390+/-0.101

表 5 四种算法在六个数据集上的覆盖率

数据集	MLFS _{GD}	MOFS	MLPE _{min}	MLRAND
Mediamill	0.379+/-0.038	0.734+/-0.029	0.472+/-0.050	0.537+/-0.042
Scene	0.019+/-0.026	0.204+/-0.042	0.218+/-0.028	0.430+/-0.023
Delicious	0.806+/-0.029	0.951+/-0.004	0.894+/-0.004	0.919+/-0.003
Bibtex	0.271+/-0.008	0.398+/-0.039	0.306+/-0.007	0.636+/-0.021
Rev1v2	0.593+/-0.010	0.598+/-0.037	0.661+/-0.009	0.665+/-0.035
Tmc2007	0.239+/-0.021	0.294+/-0.026	0.163+/-0.002	0.535+/-0.085

表 6 四种算法在六个数据集上的平均精度

数据集	MLFS _{GD}	MOFS	MLPE _{min}	MLRAND
Mediamill	0.578+/-0.032	0.270+/-0.113	0.447+/-0.079	0.291+/-0.096
Scene	0.722+/-0.040	0.662+/-0.048	0.646+/-0.049	0.416+/-0.016
Delicious	0.181+/-0.013	0.138+/-0.008	0.110+/-0.005	0.033+/-0.004
Bibtex	0.425+/-0.014	0.304+/-0.017	0.376+/-0.009	0.054+/-0.010
Rev1v2	0.151+/-0.002	0.141+/-0.005	0.143+/-0.007	0.112+/-0.005
Tmc2007	0.512+/-0.030	0.508+/-0.050	0.428+/-0.005	0.145+/-0.089

178, 190.

- [4] 杨岚雁, 靳敏, 张迎春, 等. 一种基于关联规则的 MLKNN 多标签分类算法[J]. 计算机工程与科学, 2020, 42(7): 1309-1317.
- [5] 翟婷婷, 高阳, 朱俊武. 面向流数据分类的在线学习综述[J]. 软件学报, 2020, 31(4): 912-931.
- [6] ROSENBLATT F. The perceptron: a probabilistic model for information storage and organization in the brain[J]. Psychological Review, 1958(65): 386-408.
- [7] SCHHA B, DS A, JING L C, et al. Online learning: a comprehensive survey[J]. Neurocomputing, 2021, 459(12): 249-289.
- [8] ACHARYA S, DVORKIN Y, KARRI R. Causative cyber-attacks on online learning-based automated demand response systems[J]. IEEE Transactions on Smart Grid, 2021, 12(4): 3548-3559.
- [9] GUO X, ZHANG Y, XU J. Online multi-label passive aggressive active learning algorithm based on binary relevance[C]//International Conference on Neural Information Processing, 2017(10): 256-266.
- [10] 龚楷伦, 翟婷婷, 唐鸿成. 一种面向多标签分类的在线主动学习算法[J]. 山东大学学报(工学版), 2022, 52(2): 80-88.
- [11] SAHOO D, PHAM Q, LU J, et al. Online deep learning: learning deep neural networks on the fly[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018: 2660-2666.
- [12] BENG S V, BUSA-FEKETE R, MESAOUDI-PAULA E, et al. Preference-based online learning with dueling bandits: a survey[J]. Journal of Machine Learning Research, 2021, 22(7): 101-108.
- [13] 张要, 马盈仓, 朱恒东, 等. 结合流形学习与逻辑回归的多标签特征选择[J]. 计算机工程, 2022, 48(3): 90-99, 106.
- [14] 蔡剑, 牟甲鹏, 余孟池, 等. 基于特征选择和标签相

(下转第 77 页)

- [7] FANG S H, LIN T. Principal component localization in indoor WLAN environments[J]. IEEE Transactions on Mobile Computing, 2012, 11(1): 100–110.
- [8] DENIS B. Joint distributed synchronization and positioning in UWB Ad hoc networks using TOA[J]. IEEE Transactions on Microwave Theory and Techniques, 2012, 15(4): 1896–1911.
- [9] GANERIWAL S, KUMAR R, SRIVASTAVA M B. Timing-sync protocol for sensor network[C]//Proceedings of the 1st International Conference on Embedded Networked Sensor Systems, 2003: 138–149.
- [10] YANG W, HUANG Y. Wireless sensor network based coal mine wireless and integrated security monitoring information system[C]//Proceedings of the 6th International Conference on Networking, 2007: 1896–1911.
- [11] HE Y, WEI G, CHEN X. Taxonomy-based multiple attribute group decision making method with probabilistic uncertain linguistic information and its application in supplier selection[J]. Journal of Intelligent & Fuzzy Systems, 2021, 11(2): 3237–3250.
- [12] LU J, ZHANG S, WU J, et al. COPRAS method for multiple attribute group decision making under picture fuzzy environment and their application to green supplier selection[J]. Technological and Economic Development of Economy, 2021(27): 369–385.
- [13] LI D F. The GOWA operator based approach to multiattribute decision making using intuitionistic fuzzy sets[J]. Mathematical and Computer Modelling, 2021(53): 1182–1196.
- [14] YE J. Hesitant interval neutrosophic linguistic set and its application in multiple attribute decision making[J]. International Journal of Machine Learning and Cybernetics, 2019, 10(4): 667–678.

(收稿日期: 2022-06-29)

作者简介:

李锋(1981-), 男, 硕士, 副教授, 主要研究方向: 软件算法与图像识别。

(上接第 71 页)

- 关性的多标签分类算法[J]. 计算机与数字工程, 2021, 49(10): 1967–1972, 1997.
- [15] 李郅琴, 杜建强, 聂斌, 等. 特征选择方法综述[J]. 计算机工程与应用, 2019, 55(24): 10–19.
- [16] 施后军, 潘峰, 龙福海, 等. 特征选择方法研究综述[J]. 微电子学与计算机, 2022, 39(3): 1–8.
- [17] WANG J L, ZHAO P L, HOI S C H, et al. Online feature selection and its applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(3): 698–710.

- [18] LIU J, GUO Z W, SUN Z W, et al. Online multi-label feature selection on imbalanced data sets[J]. Communications in Computer and Information Science, 2018(812): 165–174.

(收稿日期: 2022-07-08)

作者简介:

张永伟(1988-), 通信作者, 男, 硕士, 工程师, 主要研究方向: 机器学习。E-mail: zywei_1988@163.com。

朱祁(1987-), 男, 硕士, 高级工程师, 主要研究方向: 数据挖掘。

吴永城(1987-), 男, 硕士, 高级工程师, 主要研究方向: 云计算、数据挖掘。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部

www.pcachina.com