doi:10.3969/j.issn.1001-4616.2022.03.014

基于 AP 聚类和互信息的弱标记特征选择方法

孙　林,施恩惠,司珊珊,徐久成

(河南师范大学计算机与信息工程学院,河南 新乡 453007)

[摘要]　特征选择是多标记学习中重要的预处理过程. 针对现有多标记分类方法没有考虑标记占比对特征和标记相关性的影响,以及不能有效处理弱标记数据等问题,提出一种基于仿射传播(affinity propagation,AP)聚类和互信息的弱标记特征选择方法. 首先,在 AP 聚类的基础上,结合剩余标记信息和样本相似性,构建概率填补公式,预测缺失标记值,有效补齐缺失标记;然后,使用先验概率定义标记占比,结合互信息构建相关性度量,评估特征与标记集之间的相关程度;最后,设计一种弱标记特征选择算法,有效提高弱标记数据的分类性能. 在 6 个多标记数据集上进行仿真实验,结果表明,该算法在多个指标上获得了良好的分类性能,优于当前多种相关的多标记特征选择算法,有效验证了所提算法的有效性.

[关键词]　多标记学习,特征选择,AP 聚类,互信息,缺失标记
[中图分类号]TP399　[文献标志码]A　[文章编号]1001-4616(2022)03-0108-08

Weak Label Feature Selection Method Based on AP Clustering and Mutual Information

Sun Lin,Shi Enhui,Si Shanshan,Xu Jiucheng

(College of Computer and Information Engineering,Henan Normal University,Xinxiang 453007,China)

Abstract:Feature selection is an important preprocessing process in multi-label learning. To address the issues that some multi-label classification methods do not consider the influence of the proportion of label on the correlation between features and label sets and cannot efficiently deal with weak label data,a weak label feature selection method based on affinity propagation(AP)clustering and mutual information was proposed. Firstly,to effectively fill in all missing labels,the combination of the remaining label information with the similarity of samples was performed based on AP clustering,and then a probability filling formula was constructed to predict the values of missing labels. Secondly,the prior probability was used to define the proportion of label,which was combined with mutual information to develop the correlation metric for evaluating the correlation degree between features and label sets. Finally,a weak label feature selection algorithm was designed to effectively improve the classification performance of the weak label data. The simulation experimental results and analysis under six multi-label datasets show that the algorithm achieves better classification performance on multiple metrics and is superior to many related multi-label feature selection algorithms at present. All these can verify the effectiveness of the proposed algorithm.

Key words:multi-label learning,feature selection,AP clustering,mutual information,missing labels

目前,多标记学习在神经网络、机器学习等领域引起了广泛关注,且被应用于各种现实任务中,作为多标记学习的重要内容,特征选择旨在消除冗余特征、降维获取有用的信息以提升分类性能[1]. 传统的多标记特征选择算法假设多标记数据集的标记空间是完整且不缺失的[2]. 但在实际应用中,标记会因为人为或者设备的原因缺失或者无法获取,由此产生了大量的弱标记数据,即数据存在标记缺失或无标记的情况[3]. 然而,不完整的标记空间将导致特征和标记集相关性度量不准确,并且在特征选择过程中会丢失一些有价值的特征[4-7]. 因此,如何处理带缺失标记的弱标记数据问题显得尤为重要. 目前,缺失标记的处理方法有两种较为普遍,分别是填补法、粗糙集模型扩展法[3,8]. 例如,Zhu 等[9]利用线性回归模型来填补缺

收稿日期:2021-04-25.
基金项目:国家自然科学基金项目(62076089、61772176、61976082)、河南省科技攻关项目(212102210136).
通讯作者:孙林,博士,副教授,研究方向:粒计算、数据挖掘、生物信息学. E-mail:sunlin@htu.edu.cn

失标记,并将正则化作用于特征选择矩阵,选择最优特征子集;Jiang 等[10]基于标记压缩和局部特征相关性,补全缺失标记. 由于粗糙集模型[11]处理缺失标记的方法不多,而填补法简单方便、直接有效[8],因而本文采用填补法处理缺失标记. 目前,回归填补需要花费大量时间[8];K 最近邻填补需要设定 K 值且处理大规模数据集的效果不佳[12-13];而基于聚类的填补方法不受缺失数据的影响,时间复杂度相对较小,具有普遍的适用性[8,13]. 另外,基于 AP 聚类的填补方法与其他聚类算法相比,不需要预先设定聚类个数,可以将所有的样本都看作聚类中心,按照样本间的信息传递实现聚类,根据类中相似对象进行填补,且多次运行后的聚类结果比较稳定[8,13]. 因此,本文借鉴 AP 聚类算法优势来处理缺失标记,结合原有完整标记信息和样本相似性,有效补齐所有的缺失标记.

由于互信息能够检测变量之间的非线性关系,实现有效的特征选择[2],因此很多基于互信息的多标记特征选择算法被提出. 例如,Lee 等[14]提出通过最大化特征和标记之间的相关性选择特征子集;Lin 等[15]结合互信息,按照特征和标记的最大相关、特征间的最小冗余准则,筛选特征;Sun 等[16]将最大相关最小冗余转化为约束凸优化函数,构造应用于多标记学习的特征选择算法. 但是,上述这些算法都假设标记空间中所有标记具有相同的占比,而忽略了标记空间中标记占比可能会对特征和标记集相关性的影响,进而导致相关性计算的不准确. 为了解决这个问题,Shi 等[17]提出标记占比并应用于多标记特征选择. 基于此,引入标记占比改进互信息公式,度量特征和标记集之间的相关性,选择最优特征子集.

针对以上问题,本文运用 AP 聚类算法将样本编号,按照缺失标记数排序,结合样本相似度和概率填补预测缺失标记值,补全缺失标记;将标记先验概率作为标记的占比,结合互信息构建相关性度量评估特征和标记集之间的相关程度,降序排列选择最优特征子集. 实验表明该算法的性能比目前相关算法更好.

1　AP 聚类

AP 聚类[18]是 Frey 和 Dueck 提出的高效聚类算法. 该聚类算法是通过数据点间的消息传递,确定聚类中心,以及给每个聚类中心分配数据点[13]. 相似度矩阵的计算是消息传递的基础,在 AP 聚类中用欧式距离衡量数据点之间的相似度. 假设 $X=\{x_1,x_2,\cdots,x_n\}\in R^{n\times f}$ 表示样本空间,对于任意样本 $x_i,x_k\in X$,相似度公式[8,13]表示为:

$$s(x_i,x_k)=-\parallel x_i-x_k\parallel^2. \tag{1}$$

AP 聚类有吸引度和依赖度两类消息传递,用来确定聚类中心,以及给每个聚类中心分配数据点. 假设 $\boldsymbol{R}$ 表示吸引度矩阵,$\boldsymbol{A}$ 表示依赖度矩阵. 对于任意样本 $x_i,x_k\in X$,吸引度矩阵更新公式[13]表示为:

$$r(x_i,x_k)\leftarrow s(x_i,x_k)-\max\{a(x_i,x_{k'})+s(x_i,x_{k'})\},k'\neq k. \tag{2}$$

对于任意样本 $x_i,x_k\in X$,依赖度矩阵更新公式[13]表示为:

$$a(x_i,x_k)\leftarrow\begin{cases}\min\left\{0,r(x_k,x_k)+\sum\limits_{i'\notin\{i,k\}}\max\{0,r(x_{i'},x_k)\}\right\}, & i\neq k,\\ \sum\limits_{i'\neq k}\max\{0,r(x_{i'},x_k)\}, & i=k.\end{cases} \tag{3}$$

2　弱标记特征选择方法

2.1　基于 AP 聚类的缺失标记填补

定义 1　假设 $X=\{x_1,x_2,\cdots,x_n\}\in R^{n\times f}$ 表示样本空间,$Y=\{y_1,y_2,\cdots,y_n\}\in R^{n\times L}$ 表示标记空间,数据集 X 经过聚类划分为 k 个不相交的簇 $C=\{c_1,c_2,\cdots,c_k\}$,根据样本是否含有缺失标记将数据集划分为完备数据集 D_1 和不完备数据集 D_2. 如果样本 x_i 有第 j 个标记,则 $x_i(l_j)=1$,否则 $x_i(l_j)=-1$. 如果样本 x_i 的第 j 个标记值缺失,则设 $x_i(l_j)=0$.

定义 2　假设 $X=\{x_1,x_2,\cdots,x_n\}\in R^{n\times f}$ 表示样本空间,$Y=\{y_1,y_2,\cdots,y_n\}\in R^{n\times L}$ 表示标记空间,$x_i\in D_2$,$x_i(l_j)=0$,$x_m(l_j)=1$ 或 -1,且 $x_i,x_m\in C_q(q=1,2,\cdots,k)$. x_i 与同一簇 C_q 中所有 x_m 的相似度累加和表示为:

$$s=\sum s(x_i,x_m). \tag{4}$$

定义 3　假设 $X=\{x_1,x_2,\cdots,x_n\}\in R^{n\times f}$ 表示样本空间,$Y=\{y_1,y_2,\cdots,y_n\}\in R^{n\times L}$ 表示标记空间,$x_i\in D_2$,

$x_i(l_j)=0$,$x_m(l_j)=1$ 或-1,且 $x_i,x_m\in C_q(q=1,2,\cdots,k)$,缺失标记预测值可表示为:

$$\text{pre\_label}=\sum\frac{s(x_i,x_m)}{s}\cdot x_m(l_j). \tag{5}$$

由于相似样本之间具有相似标记,利用式(4)和式(5)预测缺失标记的值. pre_label 即为 $x_i(l_j)$ 的预测值,且取值范围为$[-1,1]$.

定义 4　假设 $X=\{x_1,x_2,\cdots,x_n\}\in R^{n\times f}$表示样本空间,$Y=\{y_1,y_2,\cdots,y_n\}\in R^{n\times L}$表示标记空间,$x_i(l_j)=0$,根据计算得到 pre_label 值,$x_i(l_j)$可表示为:

$$x_i(l_j)=\begin{cases}1, & \text{pre\_label}\geqslant 0,\\ -1, & \text{pre\_label}<0.\end{cases} \tag{6}$$

2.2　基于互信息的特征选择

定义 5　假设 $X=\{x_1,x_2,\cdots,x_n\}\in R^{n\times f}$表示样本空间,$Y=\{y_1,y_2,\cdots,y_n\}\in R^{n\times L}$表示标记空间,$F=\{p_1,p_2,\cdots,p_f\}\in R^f$ 表示特征空间. 对于任意特征 $p_i\in F$ 和标记 $l_j\in L$,p_i 和 l_j 的互信息公式定义为:

$$I(p_i;l_j)=H(p_i)-H(l_j|p_i), \tag{7}$$

式中,$H(p_i)$表示特征 p_i 的信息熵,$H(l_j|\ p_i)$表示标记 l_j 的条件熵. 式(7)计算特征和标记之间的相关性.

定义 6　假设 $X=\{x_1,x_2,\cdots,x_n\}\in R^{n\times f}$表示样本空间,$Y=\{y_1,y_2,\cdots,y_n\}\in R^{n\times L}$表示标记空间,对于任意标记 $l_j\in L$,l_j 的占比公式定义为:

$$W(l_j)=\frac{n(l_j)}{n}, \tag{8}$$

式中,n 表示样本的个数,$n(l_j)$表示包含标记 l_j 的样本个数. 如果标记 l_j 的占比很高,则表明很多样本包含标记 l_j,因此可认为标记 l_j 相对重要. 另外,根据式(8)可知,一般情况下每个标记的 $W(l_j)$值相差不大.

因为在大多数多标记数据集中,特征值通常是连续或者高度离散的,所以当使用互信息衡量特征之间的冗余度时,结果几乎为 0. 文献[19]指出,如果使用互信息度量特征和标记之间的相关度,则可以根据相关度值选择特征子集,忽略特征冗余的影响. 因此,本文将不考虑特征间的冗余.

定义 7　假设 $X=\{x_1,x_2,\cdots,x_n\}\in R^{n\times f}$表示样本空间,$Y=\{y_1,y_2,\cdots,y_n\}\in R^{n\times L}$表示标记空间,$F=\{p_1,p_2,\cdots,p_f\}\in R^f$ 表示 f 维的特征空间. 对于任意特征 $p_i\in F$ 和标记 $l_j\in L$,p_i 和 L 的相关度被定义为:

$$R_{p_i,L}=\sum_{j=1}^{L}I(p_i;l_j)\cdot W(l_j). \tag{9}$$

2.3　算法描述

算法 1　基于 AP 聚类和互信息的弱标记特征选择算法
(Weak Label Feature Selection Method Based on AP Clustering and Mutual Information,WFSAM)

输入:多标记数据集 X
输出:最优特征子集 S
Step 1:进行 AP 聚类,将样本划分为 k 个簇 $C=\{c_1,c_2,\cdots,c_k\}$;
Step 2:For $i=1:N\cdot L\cdot 0\%(20\%、40\%、60\%)$
　　随机产生缺失标记,将标记值置 0;
　End For
Step 3:按照标记是否缺失,将样本划分为 D_1 和 D_2. 将样本编号按照缺失标记个数从小到大排序;
Step 4:对于含有缺失标记的样本 x_i,按照公式(4)-(6)确定缺失标记值;
Step 5:检查 x_i 的标记是否补全,若已补全,则将其加入 D_1,继续补全其他样本的缺失标记,直至 D_2 为空;若无,返回步骤 4;
Step 6:For each $l_j\in L$
　　根据公式(8)计算 $W(l_j)$;
　End For
Step 7:For $p_i\in F$
　　根据公式(9)计算特征和标记集的相关性;
　End For
Step 8:将相关度值降序排序,输出最优特征子集 S.

假设多标记数据集有 N 个样本、L 个标记和 f 个特征,聚类产生 k 个簇,标记随机缺失后,共有 p 个样本

具有缺失标记. WFSAM 的时间复杂度计算如下:步骤 1 聚类的时间复杂度为 $O(N^2\log N)$;步骤 2-步骤 5 补全标记的时间复杂度为 $O(kpL)$;步骤 6-步骤 8 特征和标记之间相关度计算的时间复杂度为 $O(Nf)$,由此可计算出 WFSAM 算法总的时间复杂度为 $O(N^2\log N+Nf)$.

3　实验结果与讨论

3.1　实验数据与准备

本文选用 6 个多标记数据集,数据集来自 http://mulan.sourceforge.net,如表 1 所示. 将基于 k 最近邻的多标记方法(multi-label k-nearest neighbors,MLKNN)[20]作为特征选择的分类器(近邻数为 10,平滑参数为 1). 使用平均精度(average precision,AP)、排序损失(ranking loss,RL)、覆盖率(coverage,CV)和 1-错误率(one error,OE)作为评价指标[3,7]. 其中,AP 值越大,表示算法性能越好,用"↑"表示,且最优值为 1;其他指标值越小,则表示算法性能越好,用"↓"表示,且最优值为 0.

表 1　多标记数据集信息

Table 1　Multi-label datasets information

数据集	实例	特征	标记	训练集	测试集	领域
Arts	5 000	462	26	2 000	3 000	Text
Computers	5 000	681	33	2 000	3 000	Text
Entertainment	5 000	640	21	2 000	3 000	Text
Enron	1 702	1 001	53	1 123	579	Text
Recreation	5 000	606	22	2 000	3 000	Text
Science	5 000	743	40	2 000	3 000	Text

3.2　不同缺失比率下多标记特征选择算法的分类结果

为了验证算法的有效性,将所提的 WFSAM 算法与 MFML[7]、PMU[14]、MDMR[15]、MDDMspc[21]、MDDMproj[21]、MLNB[22]和 MLFRS[23]算法进行比较. 这些算法均使用对应文献中的最佳实验参数. 实验选取特征排序的前 30%作为特征子集,对比的实验数据与结果选自文献[7]. 表 2-表 5 展示了 8 个多标记特征选择算法分别在 0%、20%、40%和 60%的缺失比率下在 6 个多标记数据集上的实验结果. 最优结果为粗体表示.

表 2　0%缺失标记下 4 个指标的实验结果对比

Table 2　Comparison results of four metrics under 0% missing labels

指标	数据集	MDDMspc	MDDMproj	MLNB	PMU	MDMR	MLFRS	MFML	WFSAM
AP(↑)	Arts	0.507 2	0.494 3	0.499 1	0.494 4	0.498 4	0.518 3	0.530 5	**0.534 4**
	Computers	0.634 5	0.628 4	0.639 1	0.627 6	0.626 3	0.636 3	0.636 0	**0.642 2**
	Enron	0.633 5	0.617 9	0.624 2	0.634 4	0.637 2	0.565 4	**0.637 4**	0.627 2
	Entertainment	0.554 5	0.556 8	0.557 5	0.555 4	0.554 1	0.499 6	0.570 1	**0.592 8**
	Recreation	0.471 7	0.470 3	0.479 0	0.436 5	0.479 6	0.495 0	0.491 2	**0.518 7**
	Science	0.454 7	0.443 0	0.461 3	0.441 6	0.449 7	0.483 2	0.470 9	**0.500 6**
	Average	0.542 7	0.535 1	0.543 4	0.531 7	0.540 9	0.533 0	0.556 0	**0.569 3**
RL(↓)	Arts	0.152 1	0.155 5	0.154 2	0.152 7	0.149 8	0.146 6	0.145 8	**0.144 9**
	Computers	0.091 6	0.093 4	0.091 0	0.094 1	0.094 9	0.092 9	0.090 3	**0.089 6**
	Enron	0.096 9	0.097 6	0.093 7	0.094 2	**0.093 6**	0.108 0	0.093 7	0.099 6
	Entertainment	0.124 9	0.126 9	0.125 4	0.123 9	0.124 7	0.149 1	0.120 2	**0.117 4**
	Recreation	0.183 8	0.185 9	0.187 9	0.195 5	0.182 9	0.179 1	0.179 5	**0.176 8**
	Science	0.138 8	0.141 7	0.136 4	0.139 4	0.138 3	0.129 2	0.135 5	**0.128 1**
	Average	0.131 4	0.133 5	0.131 4	0.133 3	0.130 7	0.134 2	0.127 5	**0.126 1**
CV(↓)	Arts	5.474 0	5.555 3	5.504 0	5.491 7	5.401 7	5.333 3	5.321 7	**5.312 0**
	Computers	4.398 7	4.443 7	4.374 0	4.501 3	4.513 7	4.473 3	**4.347 0**	4.350 3
	Enron	13.556 1	13.514 7	13.183 1	13.247 0	13.207 3	14.664 9	**13.153 7**	13.626 9
	Entertainment	3.342 7	3.393 7	3.354 0	3.312 0	3.327 3	3.827 0	3.224 7	**3.180 7**
	Recreation	4.940 3	4.947 0	4.995 3	5.136 7	4.905 7	4.818 7	4.827 7	**4.810 0**
	Science	6.948 3	7.084 0	6.836 7	6.998 7	6.962 7	6.554 0	6.813 3	**6.535 0**
	Average	6.443 4	6.489 7	6.374 5	6.447 9	6.386 4	6.611 9	**6.281 4**	6.302 5
OE(↓)	Arts	0.634 0	0.648 7	0.643 3	0.653 7	0.645 7	0.615 7	0.598 3	**0.581 7**
	Computers	0.440 3	0.449 0	**0.432 0**	0.446 7	0.449 7	0.439 7	0.440 3	0.434 0
	Enron	0.283 2	0.317 8	0.316 1	0.279 8	**0.271 2**	0.381 7	**0.271 2**	0.307 4
	Entertainment	0.609 7	0.596 0	0.593 0	0.605 0	0.606 0	0.668 7	0.592 7	**0.545 0**
	Recreation	0.679 3	0.682 7	0.664 3	0.721 0	0.669 0	0.641 7	0.657 7	**0.614 7**
	Science	0.682 3	0.694 3	0.671 3	0.701 0	0.689 7	0.641 3	0.658 7	**0.611 3**
	Average	0.554 8	0.564 8	0.553 3	0.567 9	0.555 2	0.564 8	0.536 5	**0.515 7**

表 3　20%缺失标记下 4 个指标的实验结果对比

Table 3　Comparison results of four metrics under 20% missing labels

指标	数据集	MDDMspc	MDDMproj	MLNB	PMU	MDMR	MLFRS	MFML	WFSAM
AP(↑)	Arts	0.498 0	0.494 6	0.507 6	0.495 6	0.491 3	0.432 7	0.518 0	**0.537 2**
	Computers	0.625 6	0.625 6	0.618 9	0.620 1	0.621 3	0.598 9	**0.631 4**	0.628 5
	Enron	0.598 1	0.616 3	0.611 0	0.612 1	0.615 1	0.560 9	**0.616 4**	0.589 2
	Entertainment	0.550 3	0.565 1	0.556 4	0.552 1	0.544 8	0.500 1	0.568 5	**0.598 4**
	Recreation	0.449 2	0.459 9	0.462 1	0.442 7	0.471 9	0.388 3	0.468 8	**0.508 9**
	Science	0.449 5	0.439 7	0.451 9	0.435 8	0.445 7	0.378 3	0.462 1	**0.500 3**
	Average	0.528 5	0.533 5	0.534 7	0.526 4	0.531 7	0.476 5	0.544 2	**0.560 4**
RL(↓)	Arts	0.151 5	0.154 3	0.151 1	0.152 5	0.152 9	0.178 7	0.150 1	**0.145 3**
	Computers	0.095 0	0.094 1	0.097 7	0.096 6	0.096 7	0.102 2	**0.091 9**	**0.091 9**
	Enron	0.101 9	0.100 1	0.101 2	0.097 5	**0.096 5**	0.112 1	0.097 8	0.103 3
	Entertainment	0.129 0	0.124 1	0.129 5	0.125 2	0.127 8	0.150 4	0.123 4	**0.115 7**
	Recreation	0.196 4	0.188 6	0.192 9	0.198 3	0.187 9	0.215 4	0.187 5	**0.178 6**
	Science	0.142 1	0.139 3	0.140 7	0.141 0	0.140 2	0.157 7	0.137 8	**0.131 8**
	Average	0.136 0	0.133 4	0.135 5	0.135 2	0.133 7	0.152 8	0.131 4	**0.127 7**
CV(↓)	Arts	5.448 3	5.535 7	5.463 7	5.494 0	5.500 7	6.129 0	5.424 7	**5.292 3**
	Computers	4.550 7	4.526 8	4.653 0	4.630 7	4.608 7	4.821 0	4.445 7	**4.402 3**
	Enron	13.944 7	13.844 6	13.930 9	13.578 6	**13.481 9**	15.074 3	13.573 4	14.095 0
	Entertainment	3.430 3	3.314 3	3.453 0	3.332 3	3.407 3	3.876 7	3.312 7	**3.142 7**
	Recreation	5.194 7	5.024 0	5.100 3	5.202 7	5.019 3	5.603 7	5.014 3	**4.830 0**
	Science	7.082 3	6.985 7	7.042 0	7.053 3	7.058 0	7.800 7	6.928 7	**6.719 0**
	Average	6.608 5	6.538 5	6.607 2	6.548 6	6.512 7	7.217 6	6.449 9	**6.413 6**
OE(↓)	Arts	0.653 7	0.653 7	0.632 7	0.649 7	0.655 0	0.753 0	0.621 0	**0.583 7**
	Computers	0.448 7	0.449 3	0.457 7	0.450 7	0.452 0	0.480 7	**0.445 3**	0.454 0
	Enron	0.321 2	0.317 8	0.321 2	0.314 3	0.309 2	0.361 0	**0.297 1**	0.338 5
	Entertainment	0.604 3	0.587 0	0.591 7	0.614 7	0.613 0	0.671 3	0.586 0	**0.538 7**
	Recreation	0.707 3	0.699 0	0.686 3	0.714 0	0.675 3	0.789 0	0.687 0	**0.625 0**
	Science	0.688 3	0.700 7	0.677 3	0.712 3	0.695 3	0.783 3	0.669 3	**0.610 3**
	Average	0.570 6	0.567 9	0.561 2	0.576 0	0.566 6	0.639 7	0.551 0	**0.525 0**

表 4　40%缺失标记下 4 个指标的实验结果对比

Table 4　Comparison results of four metrics under 40% missing labels

指标	数据集	MDDMspc	MDDMproj	MLNB	PMU	MDMR	MLFRS	MFML	WFSAM
AP(↑)	Arts	0.484 0	0.483 9	0.493 3	0.487 4	0.487 2	0.431 8	0.501 6	**0.512 4**
	Computers	0.627 4	0.615 2	**0.629 6**	0.625 7	0.625 0	0.596 8	0.627 4	0.623 5
	Enron	0.609 8	0.596 4	0.599 0	0.615 5	**0.618 7**	0.568 9	0.617 4	0.566 2
	Entertainment	0.476 2	0.476 2	0.467 1	0.508 3	0.539 1	0.493 4	0.523 3	**0.592 8**
	Recreation	0.471 0	0.466 3	0.464 4	0.459 0	0.470 4	0.386 1	0.475 1	**0.511 3**
	Science	0.430 4	0.435 3	0.450 2	0.421 9	0.433 4	0.393 6	0.460 3	**0.486 5**
	Average	0.516 5	0.512 2	0.517 3	0.519 6	0.529 0	0.478 4	0.534 2	**0.548 8**
RL(↓)	Arts	0.156 3	0.157 1	0.154 1	0.156 3	0.155 8	0.178 1	0.151 9	**0.148 8**
	Computers	**0.091 6**	0.097 4	0.097 3	0.094 2	0.094 6	0.102 5	**0.091 6**	0.091 8
	Enron	0.100 2	0.103 1	0.098 5	0.098 2	**0.097 6**	0.108 8	0.098 4	0.103 8
	Entertainment	0.155 5	0.155 5	0.269 5	0.140 5	0.131 9	0.150 1	0.133 3	**0.118 7**
	Recreation	0.191 4	0.191 4	0.193 3	0.192 4	0.189 2	0.223 7	0.187 0	**0.178 7**
	Science	0.146 7	0.145 9	0.139 7	0.144 9	0.143 6	0.156 1	0.139 9	**0.135 1**
	Average	0.140 3	0.141 7	0.158 7	0.137 8	0.135 5	0.153 2	0.133 7	**0.129 5**
CV(↓)	Arts	5.577 7	5.564 7	5.507 7	5.559 7	5.551 7	6.087 3	5.459 7	**5.388 0**
	Computers	4.450 7	4.666 3	4.632 0	4.531 7	4.539 3	4.859 3	**4.437 3**	4.458 3
	Enron	13.737 5	13.932 6	13.573 4	13.580 3	13.521 6	14.544 0	**13.516 6**	14.207 3
	Entertainment	4.002 7	4.002 7	4.030 8	3.674 0	3.512 0	3.876 0	3.546 0	**3.230 3**
	Recreation	5.089 3	5.091 3	5.146 7	5.119 0	5.061 7	5.790 3	4.996 3	**4.805 0**
	Science	7.319 3	7.282 3	7.057 3	7.272 7	7.209 3	7.782 7	7.068 0	**6.864 7**
	Average	6.696 2	6.756 7	6.658 0	6.622 9	6.565 9	7.156 6	6.504 0	**6.492 3**
OE(↓)	Arts	0.674 7	0.675 3	0.651 3	0.665 3	0.665 7	0.754 3	0.648 0	**0.623 7**
	Computers	0.447 7	0.462 3	**0.441 7**	0.450 0	0.446 0	0.483 7	0.443 0	0.460 7
	Enron	**0.314 3**	0.354 3	0.352 3	0.329 9	0.319 5	0.371 3	0.316 1	0.383 4
	Entertainment	0.709 7	0.709 7	0.623 1	0.669 3	0.628 3	0.689 7	0.655 7	**0.542 3**
	Recreation	0.683 7	0.686 0	0.685 3	0.691 0	0.682 7	0.795 0	0.675 3	**0.619 0**
	Science	0.705 0	0.706 0	0.674 3	0.721 7	0.705 3	0.765 3	0.664 3	**0.624 0**
	Average	0.589 2	0.598 9	0.571 3	0.587 9	0.574 6	0.643 2	0.567 1	**0.542 2**

表 5　60%缺失标记下 4 个指标的实验结果对比

Table 5　Comparison results of four metrics under 60% missing labels

指标	数据集	MDDMspc	MDDMproj	MLNB	PMU	MDMR	MLFRS	MFML	WFSAM
AP(↑)	Arts	0.478 6	0.463 6	0.473 8	0.475 9	0.476 6	0.430 4	0.491 1	**0.501 9**
	Computers	0.618 0	0.615 7	0.615 7	0.619 2	0.613 3	0.597 1	0.622 5	**0.623 7**
	Enron	0.591 2	0.596 4	0.614 6	0.603 3	**0.622 5**	0.568 9	0.591 4	0.547 3
	Entertainment	0.563 7	0.549 9	0.556 3	0.546 0	0.543 5	0.462 1	0.541 8	**0.585 0**
	Recreation	0.432 3	0.443 4	0.459 2	0.419 2	0.457 1	0.374 9	0.447 5	**0.497 9**
	Science	0.420 1	0.415 9	0.442 2	0.416 9	0.435 8	0.393 0	0.443 3	**0.466 5**
	Average	0.517 3	0.514 2	0.527 0	0.513 4	0.524 8	0.471 1	0.522 9	**0.537 1**
RL(↓)	Arts	0.160 7	0.164 4	0.158 0	0.160 2	0.160 3	0.177 1	0.156 4	**0.151 1**
	Computers	0.101 3	0.099 4	0.100 2	0.100 4	0.101 0	0.106 2	0.098 2	**0.092 7**
	Enron	0.101 9	0.102 1	0.099 4	0.097 7	**0.097 2**	0.108 8	0.098 6	0.109 1
	Entertainment	0.124 6	0.127 2	0.128 9	0.129 7	0.130 3	0.157 0	0.129 6	**0.117 6**
	Recreation	0.198 0	0.196 3	0.193 1	0.198 7	0.191 3	0.221 6	0.195 8	**0.180 1**
	Science	0.151 1	0.150 2	0.145 2	0.148 1	0.146 9	0.153 5	0.143 1	**0.139 8**
	Average	0.139 6	0.139 9	0.137 5	0.139 1	0.137 8	0.154 0	0.137 0	**0.131 7**
CV(↓)	Arts	5.711 3	5.811 7	5.643 0	5.706 7	5.700 0	6.111 3	5.637 0	**5.465 0**
	Computers	4.782 7	4.778 0	4.812 7	4.768 0	4.799 7	4.988 7	4.738 0	**4.539 3**
	Enron	13.827 3	13.965 5	13.649 4	13.490 5	**13.459 4**	14.544 0	13.521 6	14.702 9
	Entertainment	3.321 3	3.369 0	3.434 0	3.435 3	3.444 0	4.021 0	3.436 7	**3.212 3**
	Recreation	5.214 0	5.226 7	5.097 3	5.205 7	5.083 3	5.722 3	5.231 7	**4.828 7**
	Science	7.504 7	7.437 7	7.289 0	7.408 0	7.367 0	7.664 0	7.192 3	**7.111 0**
	Average	6.726 9	6.764 8	6.654 2	6.669 0	6.642 2	7.175 2	**6.626 2**	6.643 2
OE(↓)	Arts	0.664 7	0.694 0	0.685 0	0.677 3	0.674 3	0.758 0	0.654 7	**0.627 7**
	Computers	0.459 3	0.459 7	0.454 7	0.454 0	0.466 0	0.478 7	**0.452 0**	0.459 3
	Enron	0.361 0	0.350 6	0.335 1	0.340 2	**0.309 2**	0.371 3	0.354 1	0.473 2
	Entertainment	0.593 7	0.611 3	0.604 3	0.620 0	0.622 3	0.704 7	0.628 3	**0.552 7**
	Recreation	0.740 7	0.715 0	0.694 0	0.750 3	0.696 3	0.806 3	0.705 7	**0.641 3**
	Science	0.717 3	0.719 0	0.682 7	0.716 7	0.695 0	0.766 7	0.698 0	**0.650 7**
	Average	0.589 5	0.591 6	0.576 0	0.593 1	0.577 2	0.647 6	0.582 1	**0.567 5**

对表 2 分析可知:(1)在 AP 指标对比结果中,WFSAM 在 Enron 数据集上仅优于 MDDMproj、MLNB 和 MLFRS,比其余 4 个算法低 0.006 3~0.010 2,但 WFSAM 的平均值最优. (2)在 RL 指标对比结果中,WFSAM 在 Enron 数据集上仅优于 MLFRS,但与分类最优的 MDMR 仅差 0.006,且 WFSAM 在其余 5 个数据集上表现最优. (3)在 CV 指标对比结果中,WFSAM 在 Computers、Enron 数据集上比 MFML 分别略相差 0.003 3、0.473 2,但是 WFSAM 在其余 4 个数据集上表现最优. (4)在 OE 指标对比结果中,WFSAM 在 Computers 数据集上与 MLNB 仅差 0.002;在 Enron 数据集上,WFSAM 与 MDMR、MFML 相差 0.036 2;在其余 4 个数据集上,WFSAM 表现最优. 总体来说,在标记缺失比率为 0%时,WFSAM 明显优于其他 7 种算法.

对表 3 分析可知:(1)在 AP 指标对比结果中,WFSAM 在 Computers 数据集上仅次于 MFML;在 Enron 数据集上,WFSAM 仅优于 MLFRS,略差于其余 6 种算法. 但是 WFSAM 的平均值最优. (2)在 RL 对比结果中,WFSAM 在 Enron 数据集上仅优于 MLFRS,但在其余 5 个数据集上 WFSAM 表现最优,同时 WFSAM 的平均值也最优. (3)在 CV 指标对比结果中,WFSAM 在 Enron 数据集上仅优于 MLFRS,与其余 6 个算法相差 0.150 3~0.613 1;在其余 5 个数据集上 WFSAM 表现最优. (4)在 OE 指标对比结果中,WFSAM 在 Computers、Enron 数据集上与 MFML 分别仅相差 0.008 7、0.041 4;在其余 4 个数据集上 WFSAM 表现最优. 同时 WFSAM 的平均值也最优. 总体来说,在标记缺失比率为 20%时,WFSAM 分类性能最优.

对表 4 分析可知:(1)在 AP 指标对比结果中,WFSAM 在 Computers 数据集上仅优于 MDDMproj 和 MLFRS;在 Enron 数据集上,WFSAM 略差于其他 7 种算法,与最优算法 MDMR 仅相差 0.052 5;在其余 4 个数据集上 WFSAM 性能最优. (2)在 RL 和 CV 指标对比结果中,WFSAM 在 Computers 数据集上的 RL 结果与 MDDMspc 和 MFML 仅相差 0.000 2,CV 结果与 MFML 仅相差 0.021 0;在 Enron 数据集上,WFSAM 的 RL 和 CV 结果仅优于 MLFRS;在其余 4 个数据集上 WFSAM 表现最优. 同时 WFSAM 的平均值也最优. (3)在 OE 指标对比结果中,WFSAM 在 Computers、Enron 数据集上与最优算法 MLNB、MDDMspc 分别仅相差 0.019 0、0.069 1;在其余 4 个数据集上 WFSAM 表现最优. 同时 WFSAM 的平均值也最优. 总体来说,在标记缺失比率为 40%时,WFSAM 分类效果最佳.

对表 5 分析可知:(1)WFSAM 在 Enron 数据集上的 AP、RL 和 CV 结果略次于 MDMR,但在其余 5 个数据集上均取得最优的分类效果. (2)在 OE 指标对比结果中,WFSAM 在 Computers 数据集上与 MFML 仅相差 0.007 3;在 Enron 数据集上,与 MDMR 仅相差 0.164 0;在其余 4 个数据集上 WFSAM 分类性能最佳. 同时 WFSAM 的平均值最优. 总体来说,在标记缺失比率为 60%时,WFSAM 优于其他对比算法.

综上所述,根据表 2-表 5 的结果,随着标记缺失比率的增加,所有算法分类效果均越来越差,标记缺失时的分类效果明显低于标记完整时的分类效果,这充分说明标记的有效利用有益于弱标记特征选择.

3.3　统计分析

本文使用 Friedman 和 Bonferroni-Dunn 测试[24-26]来分析所有实验结果的统计意义,计算公式为:

$$\chi_F^2=\frac{12T}{s(s+1)}\left(\sum_{i=1}^{s}R_i^2-\frac{s(s+1)^2}{4}\right), \tag{10}$$

$$F_F=\frac{(T-1)\chi_F^2}{T(s-1)-\chi_F^2}, \tag{11}$$

式中,T 表示不同评价指标下总的数据集个数,s 表示算法的个数,R_i 表示某一算法在所有数据集上的平均排序. 临界距离[26]可以表示为:

$$CD_\alpha=q_\alpha\sqrt{\frac{s(s+1)}{6T}}, \tag{12}$$

式中,q_α 是测试的临界列表值,α 是 Bonferroni-Dunn 测试的重要度.

表 6 展示了在不同的标记缺失比率下统计计算的 F_F 值. 由表 6 分析可知,当 $\alpha=0.1$ 时,Friedman 检验下零假设被拒绝. 因而,可以利用 Bonferroni-Dunn 测试进一步比较 8 种算法在统计分析上的不同,进而讨论算法之间的相对性能[7]. 为了直观地显示 WFSAM 与其他比较算法的相对性能,图 1 呈现了 8 种算法在不同标记缺失比率下的 CD 值,其中每个算法的平均排序沿数轴绘制,轴上的最小值位于左侧,因此,左侧排序的算法更好[2-3,7]. 在图 1 中,当所有算法两两比较时,将没有显著差异的算法用粗线连接起来,如图 1(a)所示,WFSAM 与 MDDMspc、MLNB 与 MDDMproj 之间有粗线相连,表明每对算法之间无显著差异. 根据图 1 可得:(1)标记缺失比率为 0%和 60%时,WFSAM 明显优于其他算法;(2)标记缺失比率为 20%和 40%时,WFSAM 略低于 MFML,但与其余算法相比,优势仍然明显. 综上所述,在不同标记缺失比率下,WFSAM 与 MFML 效果大致相同,但优于其他 6 种比较算法. 根据 Bonferroni-Dunn 测试,当 $\alpha=0.1$,$q_\alpha=2.45$ 时,$CD=1.732$,其中 $s=8$,$T=24$.

表 6　Friedman 测试的 F_F 值($s=8$ 且 $T=24$)

Table 6　F_F of the Friedman test($s=8$ and $T=24$)

缺失比率	0%	20%	40%	60%
F_F	12.462 6	14.268 5	8.609 9	14.168 1

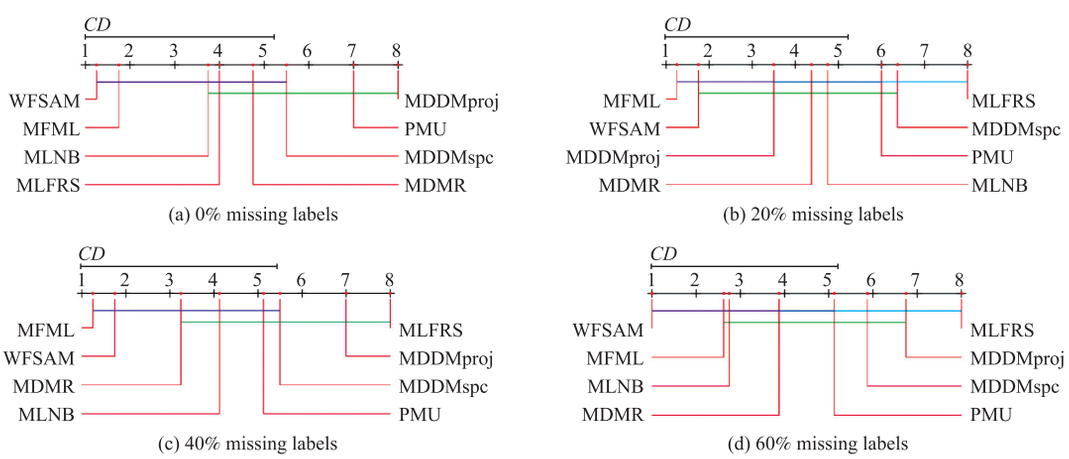

图 1　WFSAM 与其他比较算法的 Bonferroni-Dunn 检验

Fig. 1　The Bonferroni-Dunn test of WFSAM against the other compared algorithms

4　结论

本文提出一种基于 AP 聚类和互信息的弱标记特征选择方法. 该方法在 AP 聚类的基础上,结合剩余标记信息和样本相似性,对缺失标记进行填补,使用标记占比改进互信息,度量特征和标记集的相关性,设

计弱标记特征选择算法,搜索最优特征子集. 在 6 个多标记数据集上,与 7 种算法对比,实验结果显示,缺失标记的填补技术和互信息的改进均是有效的. 但是,本文仅依靠特征和标记集的相关性选择特征子集,忽略了标记间的依赖关系,因此,在未来的研究工作中,针对大规模复杂的弱标记数据集,需要考虑标记相关性,结合线性回归、稀疏正则化等理论,进一步研究多标记分类的弱监督学习问题.

[参考文献]

[1] 袁京洲,高昊,周家特,等. 基于树结构的层次性多示例多标记学习[J]. 南京师大学报(自然科学版),2019,42(3):80-87.

[2] SUN L,YIN T Y,DING W P,et al. Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems[J]. Information sciences,2020,537:401-424.

[3] SUN L,YIN T Y,DING W P,et al. Feature selection with missing labels using multilabel fuzzy neighborhood rough sets and maximum relevance minimum redundancy[J]. IEEE transactions on fuzzy systems,2021,DOI:10.1109/TFUZZ.2021.3053844.

[4] 徐海峰,张雁,刘江,等. 基于变异系数和最大特征树的特征选择方法[J]. 南京师大学报(自然科学版),2021,44(1):111-118.

[5] 刘艳,程璐,孙林. 基于 K-S 检验和邻域粗糙集的特征选择方法[J]. 河南师范大学学报(自然科学版),2019,47(2):21-28.

[6] 邓威,郭钇秀,李勇,等. 基于特征选择和 Stacking 集成学习的配电网网损预测[J]. 电力系统保护与控制,2020,48(15):108-115.

[7] WANG C X,LIN Y J,LIU J H. Feature selection for multi-label learning with missing labels[J]. Applied intelligence,2019,49(8):3027-3042.

[8] 应臻奕. 基于 AP 聚类的不完备数据处理方法的研究与实现[D]. 北京:北京邮电大学,2018.

[9] ZHU P F,XU Q,HU Q H,et al. Multi-label feature selection with missing labels[J]. Pattern recognition,2018,74:488-502.

[10] JIANG L,YU G X,GUO M Z,et al. Feature selection with missing labels based on label compression and local feature correlation[J]. Neurocomputing,2020,395:95-106.

[11] 薛占熬,庞文莉,姚守倩,等. 基于前景理论的直觉模糊三支决策模型[J]. 河南师范大学学报(自然科学版),2020,48(5):31-36.

[12] 李征,李斌. 一种基于关联规则与 K-means 的领域本体构建方法[J]. 河南师范大学学报(自然科学版),2020,48(1):24-32.

[13] 韦修喜,黄华娟,周永权. 基于 AP 聚类的约简孪生支持向量机快速分类算法[J]. 计算机工程与科学,2019,41(10):1899-1904.

[14] LEE J,KIM D W. Feature selection for multi-label classification using multivariate mutual information[J]. Pattern recognition letters,2013,34(3):349-357.

[15] LIN Y J,HU Q H,LIU J H,et al. Multi-label feature selection based on max-dependency and min-redundancy[J]. Neurocomputing,2015,168:92-103.

[16] SUN Z Q,ZHANG J,DAI L,et al. Mutual information based multi-label feature selection via constrained convex optimization[J]. Neurocomputing,2019,329:447-456.

[17] SHI E H,SUN L,XU J C,et al. Multilabel feature selection using mutual information and ML-ReliefF for multilabel classification[J]. IEEE access,2020,8:145381-145400.

[18] FREY B J,DUECK D. Clustering by passing messages between data points[J]. Science,2007,315(5814):972-976.

[19] 徐洪峰,孙振强. 多标签学习中基于互信息的快速特征选择方法[J]. 计算机应用,2019,39(10):2815-2821.

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

[21] ZHANG Y,ZHOU Z Z. Multilabel dimensionality reduction via dependence maximization[J]. ACM transactions on knowledge discovery from data,2010,4(3):1-21.

[22] ZHANG M L,PENA J M,ROBLES V. Feature selection for multilabel naive Bayes classification[J]. Information sciences,2009,179:3218-3229.

[23] LIN Y J,LI Y W,WANG C X,et al. Attribute reduction for multi-label learning with fuzzy rough set[J]. Knowledge-based systems,2018,152:51-61.

[24] FRIEDMAN M. A comparison of alternative tests of significance for the problem of m rankings[J]. Annals of mathematical statistics,1940,11(1):86-92.

[25] 孙林,赵婧,徐久成,等. 基于改进帝王蝶优化算法的特征选择方法[J]. 模式识别与人工智能,2020,33(11):981-994.

[26] DEMIAR J,SCHUURMANS D. Statistical comparisons of classifiers over multiple data sets[J]. Journal of machine learning research,2006,7(1):1-30.

[责任编辑:丁　蓉]