ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 3,May 2020,pp. 420－430
DOI:10. 16337/j. 1004-9037. 2020. 03. 005
© 2020 by Journal of Data Acquisition and Processing
http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

边缘标记弱化的多标记特征选择算法

王一宾[1,2]，吴　陈[1]，程玉胜[1,2]，江健生[1,2]

(1. 安庆师范大学计算机与信息院，安庆，246133;2. 安徽省高校智能感知与计算重点实验室，安庆，246133)

摘　要：在多标记学习中,特征选择是处理数据高维问题和提升分类性能的一种有效手段,然而现有特征选择算法大多是基于标记分布大致平衡这一假设,鲜有考虑标记分布不平衡的问题。针对这一问题,本文提出了一种边缘标记弱化的多标记特征选择算法(Multi-label feature selection algorithm with weakening marginal labels,WML),计算不同标记下正负标记的频数比率作为该标记的权值,然后通过赋权方式弱化边缘标记,将标记空间信息融入到特征选择的过程中,得到一组更为高效的特征序列,提升标记对样本描述的精确性。在多个数据集上的实验结果表明,本文算法具有一定优势,通过稳定性分析和统计假设检验进一步证明本文算法的有效性和合理性。

关键词：多标记学习;特征选择;标记分布;边缘标记

中图分类号：TP391　　**文献标志码：**A

Multi-label Feature Selection Algorithm with Weakening Marginal Labels

WANG Yibin[1,2], WU Chen[1], CHENG Yusheng[1,2], JIANG Jiansheng[1,2]

(1. School of Computer and Information, Anqing Normal University, Anqing, 246133, China;2. The University Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing, 246133, China)

Abstract: In multi-label learning, feature selection is an effective method to deal with high-dimensional data problems and improve classification performance. However, most of the existing feature selection algorithms are based on the assumption that the label distribution is roughly balanced, and rarely consider the problem of unbalanced label distribution. To solve this problem, this paper proposes a multi-label feature selection algorithm with weakening marginal labels (WML). The algorithm calculates the frequency ratio of positive and negative labels under different labels as the weight of the label, weakens the marginal label by weighting method, and integrates the label space information into the process of feature selection to obtain a more efficient feature sequence, thus improving the accuracy of label description of samples. The experimental results on several datasets show that the proposed algorithm has certain advantages. The effectiveness and rationality of the proposed algorithm are further proved by stability analysis and statistical hypothesis test.

Key words: multi-label learning; feature selection; label distribution; marginal label

---

基金项目：安徽省高校重点自然科学基金(KJ2017A352)资助项目;安徽省高校重点实验室基金(ACAIM160102)资助项目;安徽省自然科学基金(1908085MF194)资助项目。
收稿日期：2019-10-30;**修订日期：**2019-12-10

引　言

特征选择[1-6]作为一种降维手段被广泛运用在多标记学习中[7],许多学者对此进行了探究并取得了卓越的成果。例如,文献[8]结合标记权重和分类间隔构造出邻域信息度量方法;文献[9]提出一种基于互信息的过滤型特征选择方法;文献[10]通过对子空间学习的研究,提出了基于非负稀疏表示的多标记特征选择方法等。然而现有的特征选择算法多数是基于标记分布平衡这一假设,相对于标记不平衡问题却鲜有考虑。如图1是数据集Arts的标记分布图,可以看出第5,8,11,13,16,21和26个标记分布较为均衡,其他标记则表现为不平衡分布,相似的情况在其他数据集中也有体现。而通过对标记分布问题的研究,可以充分挖掘标记空间内的信息,加深各类标记对样本的描述程度。

众所周知,一个实例中是否存在该标记往往取决于实例的特征属性。如有人出现"鼻塞""流鼻涕"等症状时,或许是因为"流感"导致的,又或者是"鼻炎"所引发的,二者皆有可能,但如果还伴随着"全身酸痛"和"发烧"等症状时,则是"流感"的缘故更大一些,称此类现象为标记的不平衡性。现实世界中普遍存在着标记的不平衡性现象,而对于此类现象的研究较为罕见,传统的处理方式多半是对不平衡数据进行重采样或抽样处理,将其转变成平衡数据再进行深究,造成的结果可能是原有数据集的属性因此而改变且分类精度有所折损。如果在分类过程中加入不同标记信息,这样不仅能保留原有数据集中特征空间的原始属性,同时对分类器的分类精度也有大幅提高。目前对其的研究主要为:文献[11]提出了一种类不平衡下异方差线性判别分析(Linear discriminant analysis, LDA)的动态线性模型;文献[12]提出了一种解决二元分类中的类不平衡问题的启发式方法;文献[13]通过使用单类支持向量机(Support vector machine, SVM)和欠采样技术来研究类不平衡和类重叠问题;文献[14]通过二元混淆矩阵的分类性能来度量类不平衡的影响等。可见,现有的研究多数是针对单标记下的不平衡性问题,而对多标记下的不平衡性却鲜有研究。

同时不难察觉,不同于只有单一语义的单标记学习,在多标记学习中,当特征空间的变化对标记空间中某类标记影响甚微或者无影响时,意味着该类标记与特征之间的关联性较为微弱或无关,该类标记被称之为边缘标记。例如在标记空间中时常会出现某类单一标记全为0或全为1的情况,该类标记所含有用信息较少且冗余,有时甚至包含错误的信息,那么需对此类标记进行弱化处理,减少冗余无关信息,提高算法预测准确率。而对于标记空间中0或1标记约占1/2概率的情况,即当标记分布大致均衡时,该类标记能提供更为丰富的有效信息,同时对其进行相应的强化处理可以高效突出有用信息,进一步提升分类的精度。

针对多标记下的标记分布不平衡问题,本文提出了边缘标记弱化的多标记特征选择算法(Weakening marginal labels, WML):首先对标记空间进行处理,统计每个标记下正负类标记样本出现的频数,然后对正负类样本频数进行比值处理,构造一个权值矩阵将比值结果存储其中;然后运用赋权方法弱化标记空间中的边缘标记,同时利用信息熵等相关知识来衡量标记与特征的关联性;最后依据所构模型提出本文算法。在多个数据集上的实验结果表明本文算法在分类精度和稳定性方面具有一定优越性。

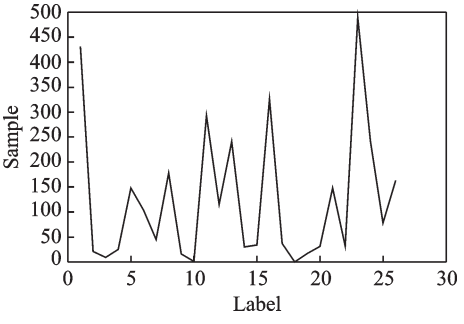

图1　Arts数据集标记分布图

Fig.1　Label distribution of Arts dataset

1　相关知识

1.1　多标记学习框架

在多标记学习框架中,针对一个样本由多个特征和标记所构成的多义性现象问题,将正确标记[15]

对应上更多未知实例是学习的目的。

假定特征集合 $T=\{t_1,t_2,t_3,\cdots,t_n\}$ 由 n 个特征所构成，标记集合 $L=\{l_1,l_2,l_3,\cdots,l_m\}$ 由 m 个标记所构成，在集合 L 中：0代表无该标记，1代表有该标记。多标记数据集为

$$\text{DataSet}=\{(T_i,L_i)|1\leqslant i\leqslant z,T_i\in T,L_i\in L\} \tag{1}$$

式中：z 表示共有 z 个样本；i 表示第 i 个样本。

1.2 评价指标

本文通过对实验评价来验证算法的高效性，使用了5类评价指标[16]：Average precision (AP)，Ranking loss(RL)，Coverage(CV)，One error(OE)和Hamming loss(HL)。

测试集用 $\{(\boldsymbol{x}_i,Y_i)\}_{i=1}^{m}\subset\mathbf{R}^d\times\{+1,-1\}^L$ 表示，标记集合 $h(x)$ 是通过算法预测所得，而排序函数 $\text{rank}_f(x,l)\in\{1,2,\cdots,L\}$ 依据预测函数 $f_l(x)$ 所定义。

Average precision(AP)：衡量排序正确的平均分数，即

$$\text{AP}(f)=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|R_i|}\sum_{l\in R_i}\left(\frac{\{k|\text{rank}_f(\boldsymbol{x}_i,k)\leqslant\text{rank}_f(\boldsymbol{x}_i,l),k\in R_i\}}{\text{rank}_f(\boldsymbol{x}_i,l)}\right) \tag{2}$$

式中 R_i 为隶属于 x_i 的相关标记集合。

Ranking loss(RL)：度量不相关标记低于相关标记的情况，即

$$\text{RL}(f)=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|R_i||\overline{R_i}|}\left|\{(l,k)|\text{rank}_f(\boldsymbol{x}_i,l)\geqslant\text{rank}_f(\boldsymbol{x}_i,k),(l,k)\in R_i\times\overline{R_i}\}\right| \tag{3}$$

式中：$R_i=\{l|Y_{il}=+1\}$ 和 $R_i=\{l|Y_{il}=-1\}$ 所构成的集合分别对应和样本 x_i 相关和无关的类别集合[17]；$\overline{R_i}$ 为集合 R_i 在标记空间中的补集。

Coverage(CV)：考察需要多少步方可遍历所有的相关标记，即

$$\text{CV}(f)=\frac{1}{p}\sum_{i=1}^{p}\max_{y\in Y_i}\text{rank}_f(x_i,y)-1 \tag{4}$$

One error(OE)：度量最高排序标记并不正确的的次数情况，有

$$\text{OE}(f)=\frac{1}{m}\sum_{i=1}^{m}[\,Y_{i,l_i}=-1\,]\qquad l_i=\arg\max_{k=1,2,\cdots,L}f_k(\boldsymbol{x}_i) \tag{5}$$

Hamming loss(HL)：评估预测标记和真实标记在单一标记下的非正确匹配情况，有

$$\text{HL}(h)=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{L}\sum_{l=1}^{L}[\,h_l(\boldsymbol{x}_i)\neq Y_{il}] \tag{6}$$

式中：在5个评价指标中，除了AP值，其他评价指标越小越好。

1.3 相关信息熵与互信息

定义 1[5] 在信息论中，信息熵越大，代表其不确定性越大，定义为

$$H(X)=-\sum_{X}p(x)\log p(x) \tag{7}$$

式中：$H(\cdot)$ 为分类器，可得 x_i 的预测标记向量。

定义 2[5] 当服从 $p(x,y)$ 分布时，$H(X,Y)$ 定义为

$$H(X,Y)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(x,y)=H(X)+H(Y|X)=H(Y)+H(X|Y) \tag{8}$$

定义 3[5] 当给定随机变量 X 时，则 Y 的条件熵为

$$H(Y|X)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(y|x) \tag{9}$$

定义4[5] 通过上述定义可知,互信息表示

$$I(X;Y)=H(X)+H(Y)-H(X,Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (10)$$

定义1—4公式中的各变量含义参见文献[5]。由此可知,若$I(X;Y)$越大,则X与Y之间关联越紧密,若X与Y无关,则$I(X;Y)=0$。通过文氏图(Venn diagram)可以更直观地展示信息熵与互信息的关联性(见图2)。

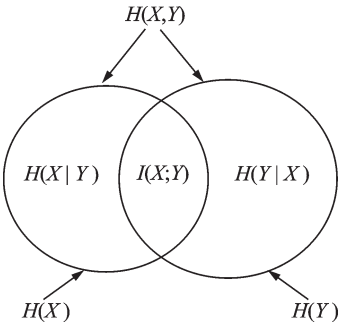

图2 信息熵与互信息的关系图

Fig.2 Relationships between information entropy and mutual information

1.4 特征与标记集合的相关性

多标记学习中,经过对特征与标记集合之间关联性的分析,互信息可定义为

定义5[5] 当给定特征f和标记集合$L=\{l_1,l_2,l_3,\cdots,l_m\}$,$\forall l\in L$,$i=1,2,\cdots,m$时,$I(f;l_i)$表现为$f$与$l_i$的互信息,则$f$与$L$的互信息为

$$\mathrm{IML}(f;L)=\sum_{i=1}^{m}I(f;l_i) \quad (11)$$

当$I(f;l_i)\geqslant 0$时,$\mathrm{IML}(f;L)\geqslant 0$,其中当$f$和$L$完全独立时,两者互信息最小;反之,$\mathrm{IML}(f;L)$越大,表示$f$和$L$关联性越强。

同理,有如下公式

$$\mathrm{IML}(f;L)=\mathrm{IML}(L;f) \quad (12)$$

$$\mathrm{IML}(f;L)=\sum_{i=1}^{m}H(l_i)-\sum_{i=1}^{m}H(l_i|f) \quad (13)$$

推理可得

$$\mathrm{IML}(f;L)=\sum_{i=1}^{m}H(l_i) \quad (14)$$

即L完全由f所决定,$\mathrm{IML}(f;L)$表示为l_i的不确定度之和,此时互信息最大。

2 边缘标记弱化的多标记特征选择算法

2.1 WML算法模型

在多标记数据集中,标记对样本描述程度常由于标记分布不平衡问题而产生偏差,目前算法大多对其鲜有考虑。针对这一问题,先计算各标记下的正负样本频数比率,再对相应标记进行赋权处理,最后构造标记权值模型,具体见如下定义。

定义6 给定多标记下的一个样本空间$U=\{x_1,x_2,\cdots,x_z\}$,则其中的特征空间为$T=\{t_1,t_2,t_3,\cdots,t_n\}$,相对应的标记空间为$L=\{l_1,l_2,l_3,\cdots,l_m\}$。对于$\forall l\in L$,$i=1,2,\cdots,m$,在第$i$个标记下正类样本数和负类样本数分别表示为$|l_i^+|$和$|l_i^-|$,其权值可如下定义

$$\omega_i^+=\frac{l_i^+}{l_i^-} \qquad |l_i^+|<|l_i^-| \quad (15)$$

$$\omega_i^-=\frac{l_i^-}{l_i^+} \qquad |l_i^-|<|l_i^+| \quad (16)$$

在传统的单标记特征选择过程中,给定了样本的特征f和类别标记l,联合互信息等相关知识用

$I(f;l)$表示两者的关联性。而在多标记学习中,由于考虑到标记权值模型下的各样本所含信息量的不同,所以通过赋权的方式对f和L关联性进行深入研究,挖掘出更多有用信息。而定义7给出了具体加权方式。

定义7 假定在多标记框架下,利用U表示为一个样本空间,而与此相对应的$L=\{l_1,l_2,l_3,\cdots,l_m\}$表示为标记空间,$\forall l\in L, i=1,2,\cdots,m$,针对标记和特征之间的关联性量化问题,利用赋权互信息方式对此进行处理,同时结合定义6和式(11)得

$$\mathrm{IML}(f;L^*)=\frac{1}{m}\sum_{i=1}^{m}\omega_i^* I(f;l_i) \tag{17}$$

式中:"$*$"为"$+$"表示正标记数少于负标记数,"$*$"为"$-$"表示负标记数少于正标记数,即

$$\begin{cases}\mathrm{IML}(f;L^+)=\dfrac{1}{m}\sum\limits_{i=1}^{m}\omega_i^+ I(f;l_i)=\dfrac{1}{m}\sum\limits_{i=1}^{m}\dfrac{l_i^+}{l_i^-}I(f;l_i)\\ \mathrm{IML}(f;L^-)=\dfrac{1}{m}\sum\limits_{i=1}^{m}\omega_i^- I(f;l_i)=\dfrac{1}{m}\sum\limits_{i=1}^{m}\dfrac{l_i^-}{l_i^+}I(f;l_i)\end{cases} \tag{18}$$

2.2 WML算法描述

考虑到边缘标记与实例的特征属性之间的联系,以及标记空间中的不同标记所含信息多少各不相同,为更加准确地描述样本实例且更高效地提取有效信息,本文主要通过弱化边缘标记来过滤冗余标记信息,而相对应的分布较为均衡的有用标记信息则得到了强化,从而校正了标记对样本描述的精确度。首先对标记空间进行处理,统计每个标记下正负类标记样本出现的频数,然后对正负标记进行对比处理,结果作为相应标记的权值进行加权处理,得到一组不同重要度的特征序列。本文提出WML算法,算法流程图如图3所示。

算法步骤:

输入:多标记训练集D

输出:排序后的特征序列R

(1) $\mathrm{IML}=\varnothing$;

(2) for each $f_j\in F$

(3) $R=\varnothing$

(4) for each $l_i\in L$

(5) 根据定义6和计算每个标记下的权值ω_i^*

(6) end

(7) 根据式(17)计算每个特征f_j在不同类别标记下的互信息$\mathrm{IML}(f_j;L_i^*)$

(8) end

(9) 依据第7步计算结果将其按照降序排序,输出重排序所得的一组新的特征序列R。

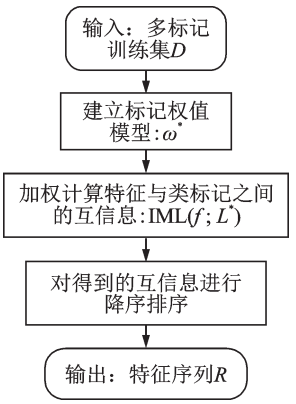

图3 算法流程图

Fig.3 Algorithm flowchart

3 实验数据及其结果分析

3.1 实验数据

为验证算法WML的有效性与合理性,本文采用了9个常用公开数据集,相关信息见表1,数据来自http://mulan.sourceforge.net/datasets.html。

表1 多标记数据集

Table 1 Multi-label datasets

数据集	样本数	特征数	标记数	训练样本	测试样本
Arts	5 000	462	26	2 000	3 000
Business	5 000	438	30	2 000	3 000
Computer	5 000	681	33	2 000	3 000
Education	5 000	550	33	2 000	3 000
Health	5 000	612	32	2 000	3 000
Recreation	5 000	606	22	2 000	3 000
Reference	5 000	793	33	2 000	3 000
Society	5 000	636	27	2 000	3 000
Yeast	2 417	103	14	1 499	918

3.2 实验结果

本文实验均在Matlab2016a中运行,具体硬件环境为Inter(R) Core(TM) i7-7700HQ CPU @ 2.80 GHz,8 GB内存,实验使用的电脑系统为现今普遍的Windows 10操作系统。同时也运用具有代表性且有说服力的ML-kNN[18]作为本文的基础实验分类器。

实验所用算法包括:基于最大相关性的属性约简算法(Multi-label dimensionality reduction via dependence maximization,MDDM[19]),该算法利用两种投影策略,根据原始特征与标记空间最大相关性将原始数据投影到较低维的特征空间;基于多变量互信息的多标记特征选择算法(Pairwise multivariate mutual information, PMU[20]),该算法是通过选择与标记空间互信息最大的特征生成特征子集;多标签朴素贝叶斯分类的特征选择算法(Feature selection for multi-label naive Bayes classification, MLNB[21]),此算法主要是以遗传算法和主成分分析为基础,进而采用贝叶斯分类法实现特征提取;基于标记相关性的多标记特征选择算法(Multi-label feature selection with label correlation,MUCO[22])。而MDDMspc与MDDMproj两种方法由于MDDM在实验过程中采取的参数并不相同而进行区分。

通过所做实验可知,本文算法与MDDM,PMU和MUCO等方法得到一组用特征序列表示的结果,所以在实验过程中将所选的特征子集个数按照MLNB算法实验进行相同的设置,本文在运行实验过程中的基础实验分类器ML-kNN,而在实验过程中设置的平滑系数为$s=1$以及$k=10$。实验结果如表2—6所示。表中数值依据"↑"表示为其值若越大则越优,"↓"为越小越优,最优结果已黑色加粗。与此同时,为了凸显各算法间的性能差异,本文采用显著性水平5%的成对T检验[23]进行算法对比,并在表格中用•/◦表示本文算法优于/差于对比算法,底行括号中数字为最优个数。

3.3 实验结果分析

(1) 表2中实验结果表明:对于评价指标AP,通过对比MDDMspc,MDDMproj,PMU,MLNB和MUCO这5个算法,在8个数据集上能看出本文算法均取得最优值,仅在Education上次于MUCO算法0.008 6,表明本文算法性能较为突出,在9个数据集上的平均结果显示本文算法排在第一,MUCO算法排在第二。

(2) 表3中实验结果表明:对于评价指标RL,本文算法在9个数据集中有6个数据集结果优于其他算法,MUCO算法在数据集Education和Society上比本文算法分别提升0.003 4和0.001 6,PMU算法在Yeast上仅比本文算法少0.004 4,依据平均结果得出,本文算法排列第一,性能也是最好。

(3) 表4中实验结果表明:对于评价指标CV,其中有6个数据集的值都是最小的,本文算法在Education和Society两个数据集上比MUCO算法增加0.124 7和0.040 6,而在Yeast数据集上比PMU算法

表 2 各算法在平均精度上的结果

Table 2 Results of each algorithm in average precision (↑)

Dataset	MDDMspc	MDDMproj	PMU	MLNB	MUCO	WML
Arts	0.507 2 •	0.494 3 •	0.494 4 •	0.499 1 •	0.519 2 •	**0.534 8**
Business	0.873 6 •	0.873 2 •	0.875 4 •	0.871 3 •	0.877 0 •	**0.879 2**
Computer	0.634 5 •	0.628 4 •	0.627 6 •	0.639 1 •	0.640 3 •	**0.645 7**
Education	0.538 9 •	0.542 5 •	0.546 5 •	0.547 8 •	**0.575 3 °**	0.566 7
Health	0.665 4 •	0.650 2 •	0.680 2 •	0.688 0 •	0.685 7 •	**0.701 8**
Recreation	0.471 7 •	0.470 3 •	0.436 5 •	0.479 0 •	0.477 5 •	**0.532 8**
Reference	0.612 6 •	0.610 6 •	0.616 9 •	0.623 4 •	0.630 1 •	**0.634 4**
Society	0.561 5 •	0.568 1 •	0.588 1 •	0.589 4 •	0.593 9 •	**0.595 1**
Yeast	0.721 3 •	0.721 0 •	0.747 3 •	0.735 5 •	0.735 0 •	**0.747 8**
Average	0.620 7(0)	0.617 6(0)	0.623 7(0)	0.630 3(0)	0.637 1(1)	**0.648 7(8)**

表 3 各算法在排位损失上的结果

Table 3 Results of each algorithm in ranking loss (↓)

Dataset	MDDMspc	MDDMproj	PMU	MLNB	MUCO	WML
Arts	0.152 1 •	0.155 5 •	0.152 7 •	0.154 2 •	0.150 4 •	**0.143 1**
Business	0.042 2 •	0.042 2 •	0.041 3 •	0.041 9 •	0.040 3 •	**0.040 2**
Computer	0.091 6 •	0.093 4 •	0.094 1 •	0.091 0 •	0.089 6 •	**0.089 5**
Education	0.091 4 •	0.092 4 •	0.091 1 •	0.092 2 •	**0.085 6 °**	0.089 0
Health	0.066 3 •	0.069 8 •	0.063 8 •	0.064 1 •	0.060 8 •	**0.059 2**
Recreation	0.183 8 •	0.185 9 •	0.195 5 •	0.187 9 •	0.185 7 •	**0.167 4**
Reference	0.088 8 •	0.088 9 •	0.086 8 •	0.088 9 •	0.086 5 •	**0.084 0**
Society	0.150 0 •	0.148 4 •	0.144 2 •	0.145 6 •	**0.141 6 °**	0.143 2
Yeast	0.199 0 •	0.204 1 •	**0.178 6 °**	0.187 1 •	0.190 9 •	0.183 0
Average	0.118 4(0)	0.120 1(0)	0.116 5(1)	0.117 0(0)	0.114 6(2)	**0.111 0(6)**

表 4 各算法在覆盖率上的结果

Table 4 Results of each algorithm in coverage (↓)

Dataset	MDDMspc	MDDMproj	PMU	MLNB	MUCO	WML
Arts	5.474 0 •	5.555 3 •	5.491 7 •	5.504 0 •	5.418 7 •	**5.213 7**
Business	2.346 0 •	2.330 3 •	2.318 7 •	2.348 3 •	2.268 7 •	**2.260 0**
Computer	4.398 7 •	4.443 7 •	4.501 3 •	4.374 0 •	4.348 7 •	**4.306 0**
Education	3.898 7 •	3.920 3 •	3.899 0 •	3.918 3 •	**3.693 0 °**	3.817 7
Health	3.505 7 •	3.621 7 •	3.407 0 •	3.416 3 •	3.315 3 •	**3.242 3**
Recreation	4.940 3 •	4.947 0 •	5.136 7 •	4.995 3 •	5.003 3 •	**4.584 7**
Reference	3.439 0 •	3.446 0 •	3.366 0 •	3.431 3 •	3.358 0 •	**3.287 7**
Society	5.842 3 •	5.800 0 •	5.660 3 •	5.739 0 •	**5.610 7 °**	5.651 3
Yeast	6.813 7 •	6.818 1 •	**6.491 3 °**	6.692 8 •	6.605 7 •	6.631 8
Average	4.517 6(0)	4.542 5(0)	4.474 7(1)	4.491 0(0)	4.402 5(2)	**4.332 8(6)**

表5 各算法在1-错误上的结果

Table 5 Results of each algorithm in One error (↓)

Dataset	MDDMspc	MDDMproj	PMU	MLNB	MUCO	WML
Arts	0.634 0 •	0.648 7 •	0.653 7 •	0.643 3 •	0.613 7 •	**0.588 0**
Business	0.128 7 •	0.128 0 •	0.122 7 •	0.131 7 •	0.120 7 •	**0.119 0**
Computer	0.440 3 •	0.449 0 •	0.446 7 •	0.432 0 •	0.433 0 •	**0.427 0**
Education	0.610 0 •	0.597 3 •	0.592 0 •	0.582 7 •	**0.552 0 °**	0.565 0
Health	0.427 0 •	0.440 3 •	0.408 0 •	0.394 7 •	0.401 3 •	**0.380 0**
Recreation	0.679 3 •	0.682 7 •	0.721 0 •	0.664 3 •	0.667 0 •	**0.597 7**
Reference	0.484 3 •	0.488 7 •	0.486 7 •	0.470 3 •	0.463 7 •	**0.456 0**
Society	0.495 3 •	0.481 3 •	0.459 3 •	0.454 0 •	0.453 3 •	**0.448 0**
Yeast	0.259 3 •	0.252 7 •	**0.233 1 °**	0.256 0 •	0.252 7 •	0.242 9
Average	0.462 0(0)	0.463 2(0)	0.458 1(1)	0.447 7(0)	0.439 7(1)	**0.424 8(7)**

表6 各算法在海明损失上的结果

Table 6 Results of each algorithm in Hamming loss (↓)

Dataset	MDDMspc	MDDMproj	PMU	MLNB	MUCO	WML
Arts	0.060 7 •	0.061 2 •	0.061 5 •	0.061 2 •	0.060 5 •	**0.058 8**
Business	0.027 7 •	0.027 7 •	0.027 3 •	0.028 3 •	0.027 3 •	**0.027 1**
Computer	0.040 6 •	0.040 6 •	0.041 3 •	0.040 1 •	0.040 7 •	**0.038 7**
Education	0.042 6 •	0.042 2 •	0.040 9 •	0.040 5 •	**0.040 1 °**	0.040 7
Health	0.044 1 •	0.045 6 •	0.044 6 •	0.041 5 •	0.044 4 •	**0.040 2**
Recreation	0.062 0 •	0.061 6 •	0.063 3 •	0.061 1 •	0.060 8 •	**0.058 6**
Reference	0.032 2 •	0.031 1 •	0.030 6 •	0.029 6 •	0.030 9 •	**0.028 9**
Society	0.058 0 •	0.057 7 •	0.056 1 •	0.055 9 •	**0.054 5 °**	0.055 7
Yeast	0.220 9 •	0.224 6 •	0.208 9 •	0.208 0 •	0.209 0 •	**0.202 8**
Average	0.065 4(0)	0.065 8(0)	0.063 8(0)	0.062 9(0)	0.063 1(2)	**0.061 3(7)**

增加0.140 5,本文算法在和其他所有算法进行对比时,9个数据集上有6个占优,在其他数据集上名列前三,而在CV的平均结果名列第一。

(4) 表5中实验结果表明:对于评价指标OE,本文算法在7个数据集上其结果都是最小,MUCO算法在Education数据集上只比本文算法减少了0.013 0,PMU算法也只比本文算法在Yeast数据集上减少0.009 8,充分表明了本文算法的优越性,而且在各数据集的平均结果上同样排名第一。

(5) 表6中实验结果表明:在HL上,本文算法在9个数据集中有7个占优,这表明了本文算法的性能效果最好,而本文算法也仅在Education和Society数据集上稍逊色于MUCO算法,同样在平均结果上本文算法位居第一。

综上所述,本文在9个标记数据集和5个评价指标上进行了大量实验对算法WML的有效性和合理性进行了验证。在5个评价指标上,本文提出的WML算法在大多数数据集上都排列第一,在其他数据集也都位居前列,表明本文算法优于当前多数的多标记特征选择算法。而其原因在于本文算法充分考虑了标记分布不平衡问题,通过计算不同标记下正负标记的频数作为该标记的权值,运用赋权方法弱化标记空间中的边缘标记,同时利用信息熵等相关知识来衡量标记与特征的关联性;即在保留原有数据集中特征空间的原始属性的情况下,将标记空间的信息加入到了特征选择过程中,从而选出了信息更丰富的特征。

3.4 统计假设检验及算法稳定性分析

为了更好体现本文算法WML在9个数据集下和其他算法所对比的合理性，本文通过结合统计学等相关知识进行显著性水平为5%的Nemenyi检验[24]来验证实验结果。若对比算法之间在所有多标记数据集下进行对比的平均排序的差值低于临界差(Critical difference，CD)，则认为这两个对比算法之间不存在显著性的差异，否则认为是有显著性的差异。图4显示了各算法的对比结果，依据式(17)计算CD＝2.513 5($k=6, N=9$)，在坐标轴上数字越小，算法性能在此方面则表示越好。图中不同彩色实线相连接的算法表示两者之间并不存在显著性差异，反之则有显著性差异。由CD图结果显示，本文算法在各指标上排名均占优，即有

$$\mathrm{CD}_{\alpha}=q_{\alpha}\sqrt{\frac{k(k+1)}{6N}} \tag{19}$$

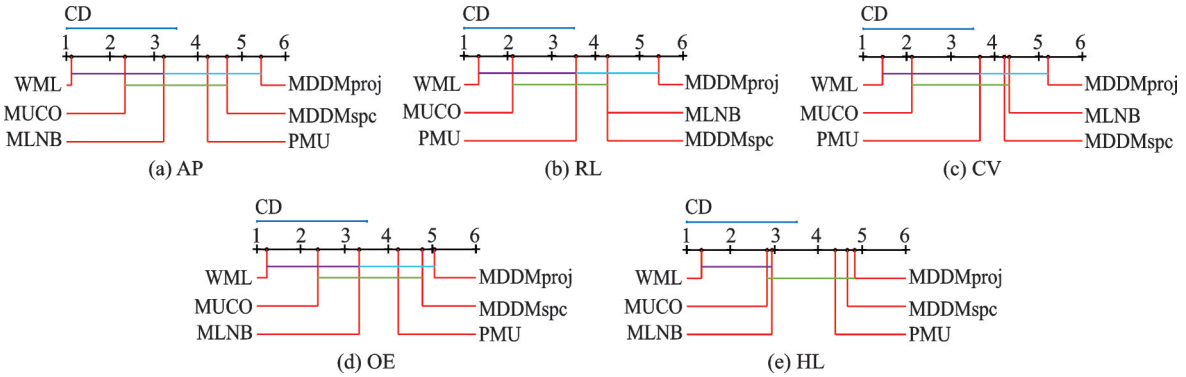

图4 各算法Nemenyi检验的性能对比

Fig.4 Performance comparison of Nemenyi test by different algorithms

为了对本文算法进行稳定性分析[25]，本文采用雷达图的形式来表示。由于在各数据集上预测分类时的实验结果会有所偏差，出于对此的考虑，本文将实验结果进行标准归一化在[0.1，0.5]区间内，然后利用归一化处理后的实验数值度量算法稳定指数。图5给出了各算法的稳定性。

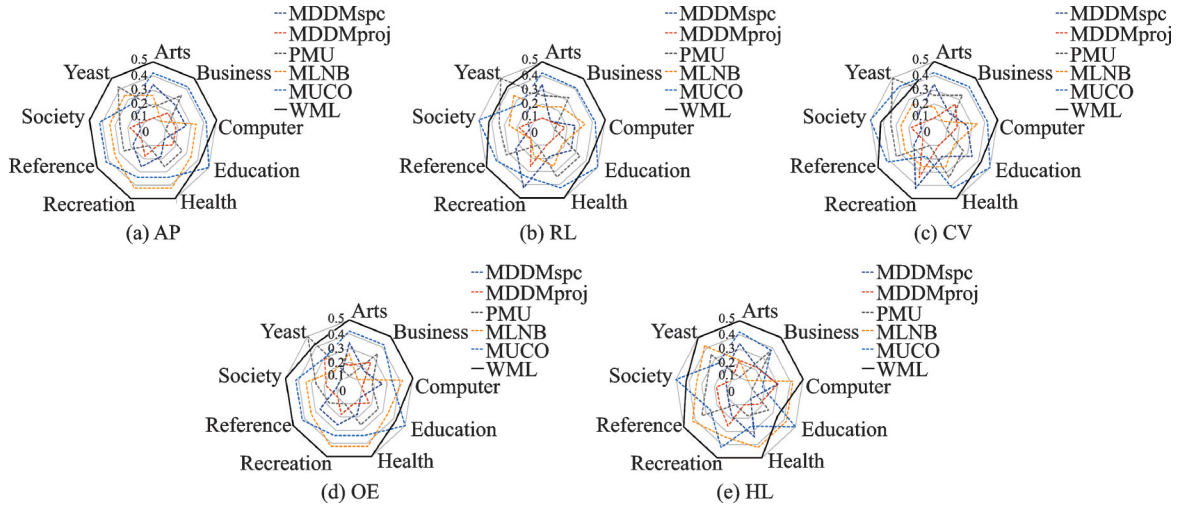

图5 各算法在9个数据集和不同评价指标中的稳定性

Fig.5 Stability of each algorithm in nine datasets with different evaluation indexes

通过观察图5可知:在评价指标AP中,本文算法在稳定指数值为[0.4,0.5]范围内均展现出十分稳定的效果;在评价指标RL中,本文算法同样是相当稳定的解决方案,各值均在[0.4,0.5]内;在评价指标CV中,本文算法在8个数据集上对比其他算法得出了格外稳定的结果;相似的结果在其他评价指标中也有体现。

4 结束语

由于不同标记所含信息量的不同,对样本空间的描述程度有所偏差,对于该标记分布不平衡问题,本文通过联合信息熵和互信息等相关知识,提出了一种边缘标记弱化的多标记特征选择算法。在保留原有数据集中特征空间的原始属性的情况下,利用不同标记下的正负标记比率权值来提升标记对样本描述程度的精确性,进而运用该赋权方法弱化标记空间中的边缘标记,因为在特征选择过程中加入了标记空间的信息,所以选出了含有更加丰富信息的特征。实验结果表明,本文算法在现有的特征选择算法中具有一定的优越性。但是本文在进行特征选择时仅考虑了标记空间信息,并未考虑到特征空间的信息以及高维特征间的相关性问题,因此对其有待进一步的研究。

参考文献:

[1] 李炜,巢秀琴.改进的粒子群算法优化的特征选择方法[J].计算机科学与探索,2019,13(6):990-1004.
LI Wei, CHAO Xiuqin. Improved particle swarm optimization method for feature selection[J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(6): 990-1004.

[2] LIN Yaojin, HU Qinghua, LIU Jinghua, et al. Multi-label feature selection based on neighborhood mutual information[J]. Applied Soft Computing, 2016, 38: 244-256.

[3] 程玉胜,李雨,王一宾,等.动态滑动窗口加权互信息流特征选择[J].南京大学学报(自然科学版),2018,54(5):974-985.
CHENG Yusheng,LI Yu,WANG Yibin,et al. Streaming feature selection with weighted fuzzy mutual information based on dynamic sliding window[J].Journal of Nanjing University(Natural Sciences),2018,54(5):974-985.

[4] 徐少成,李东喜.基于随机森林的加权特征选择算法[J].统计与决策,2018,34(18):25-28.
XU Shaocheng, LI Dongxi. Weighted feature selection algorithm based on random forest[J]. Statistics and Decision, 2018, 34(18): 25-28.

[5] 刘景华, 林梦雷, 王晨曦, 等. 基于局部子空间的多标记特征选择算法[J]. 模式识别与人工智能, 2016, 29(3): 240-251.
LIU Jinghua,LIN Menglei,WANG Chenxi,et al.Multi-label feature selection algorithm based on local subspace[J].Pattern Recognition and Artificial Intelligence,2016,29(3):240-251.

[6] 黄琴,钱文彬,王映龙,等.面向代价敏感的多标记不完备数据特征选择算法[J].小型微型计算机系统,2018,39(12):2617-2624.
HUANG Qin, QIAN Wenbin, WANG Yinglong, et al. Multi-label feature selection algorithm with incomplete data based on cost sensitivity[J]. Journal of Chinese Computer Systems, 2018, 39(12): 2617-2624.

[7] CHENG Y S, ZHAO D W, ZHAN W F, et al. Multi-label learning of non-equilibrium labels completion with mean shift[J]. Neurocomputing, 2018, 321: 92-102.

[8] 王晨曦, 林耀进, 唐莉, 等. 基于信息粒化的多标记特征选择算法[J]. 模式识别与人工智能, 2018, 31(2): 123-131.
WANG Chenxi,LIN Yaojin,TANG Li,et al.Multi-label feature selection based on information granulation[J].Pattern Recognition and Artificial Intelligence,2018,31(2):123-131.

[9] 许行,张凯,王文剑.一种小样本数据的特征选择方法[J].计算机研究与发展,2018,55(10):2321-2330.
XU Hang, ZHANG Kai, WANG Wenjian. A feature selection method for small samples[J]. Journal of Computer Research and Development, 2018, 55(10): 2321-2330.

[10] 蔡志铃,祝峰.非负稀疏表示的多标签特征选择[J].计算机科学与探索,2017,11(7):1175-1182.
CAI Zhiling, ZHU William. Multi-label feature selection via non-negative sparse representation[J]. Journal of Frontiers of Computer Science and Technology, 2017, 11(7): 1175-1182.

[11] GYAMFI K S, BRUSEY J, HUNT A, et al. A dynamic linear model for heteroscedastic LDA under class imbalance[J]. Neurocomputing, 2019, 343: 65-75.

[12] KAUR P, GOSAIN A. FF-SMOTE: A metaheuristic approach to combat class imbalance in binary classification[J]. Applied Artificial Intelligence, 2019, 33(5): 420-439.

[13] DEVI D, BISWAS S K, PURKAYASTHA B. Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique[J]. Connection Science, 2019, 31(2): 105-142.

[14] LUQUE A, CARRASCO A, MARTIN A, et al. The impact of class imbalance in classification performance metrics based on the binary confusion matrix[J]. Pattern Recognition, 2019, 91: 216-231.

[15] 李志欣, 卓亚琦, 张灿龙, 等. 多标记学习研究综述[J]. 计算机应用研究, 2014, 31(6): 1601-1605.
LI Zhixin, ZHUO Yaqi, ZHANG Canlong, et al. Survey on multi-label learning[J]. Application Research of Computers, 2014, 31(6): 1601-1605.

[16] ZHANG M L, ZHOU Z H. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819-1837.

[17] KIRA K, RENDELL L A. The feature selection problem: Traditional methods and a new algorithm[C]//Proceedings of the Tenth National Conference on Artifical Intelligence. Combridge, MA: MIT Press, 1992, 2: 129-134.

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

[19] ZHANG Y, ZHOU Z H. Multilabel dimensionality reduction via dependence maximization[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2010, 4(3): 14.

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

[21] ZHANG M L, PEÑA J M, ROBLES V. Feature selection for multi-label naive Bayes classification[J]. Information Sciences, 2009, 179(19): 3218-3229.

[22] LIN Y, HU Q, LIU J, et al. Streaming feature selection for multilabel learning based on fuzzy mutual information[J]. IEEE Transactions on Fuzzy Systems, 2017, 25(6): 1491-1507.

[23] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, P A: Association for Computational Linguistics, 2002: 311-318.

[24] ZHANG M L, WU L. Lift: Multi-label learning with label-specific features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(1): 107-120.

[25] 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.

作者简介:

王一宾(1970-),男,教授,硕士生导师,研究方向:多标记学习、机器学习和软件安全等,E-mail: wangyb07@mail.ustc.edu.cn。

吴陈(1993-),男,硕士研究生,研究方向:机器学习、数据挖掘和统计等,E-mail: 1024031783@qq.com。

程玉胜(1969-),男,教授,硕士生导师,研究方向:数据挖掘、粗糙集和机器学习等,E-mail:chengyshaq@163.com。

江健生(1982-),男,讲师,研究方向:机器学习、数字图像处理等,E-mail: jiangjsh2009@163.com。

(编辑:刘彦东)