ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 3,May 2023,pp. 539－548
DOI:10. 16337/j. 1004-9037. 2023. 03. 004

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于标记补充的多标记特征选择算法

余　鹰，张志强，钱　进，万　明

（华东交通大学软件学院，南昌 330013）

摘　要：已有的多标记特征选择方法主要根据特征与标记之间的依赖度以及特征与特征之间的冗余度确定每个特征的重要度，然后根据重要度进行特征选择，常常忽略标记关系对特征选择的影响。针对上述问题，引入邻域互信息设计了基于标记补充的多标记特征选择算法（Multi-label feature selection algorithm based on label complementarity，MLLC），该算法将依赖度、冗余度以及标记关系作为特征重要度的评价要素，然后基于这3个要素重新设计特征重要度评估函数，使得选取的特征能够获得更佳的分类性能。最后，在6个多标记数据集上验证了MLLC算法的有效性和鲁棒性。

关键词：多标记学习；特征选择；邻域互信息；标记补充

中图分类号：TP391　　**文献标志码：**A

Multi-label Feature Selection Based on Label Complementarity

YU Ying，ZHANG Zhiqiang，QIAN Jin，WAN Ming

（School of Software，East China Jiaotong University，Nanchang 330013，China）

Abstract：Multi-label feature selection is an important research component in the field of multi-label learning. Existing multi-label feature selection methods mainly measure the importance of each feature based on the dependency between features and labels，and the redundancy among features. Then，feature ranking is performed based on feature importance，often ignoring the influence of label relationships on feature importance. To solve this problem，a multi-label feature selection algorithm based on label complementarity（MLLC）is designed，which introduces neighbourhood mutual information. The algorithm takes dependency，redundancy and label relationships as the evaluation elements of feature importance. And then it redesigns the feature importance evaluation function based on these three elements，so as to select features with stronger discriminative power and achieve better classification performance. Finally，the effectiveness and robustness of the algorithm are verified on six classical multi-label datasets.

Key words: multi-label learning; feature selection; neighborhood mutual information; label complementarity

引　　言

在传统的机器学习中，常常假设样本仅由一个类别标记属性描述，即单标记学习，但是这种假设通常不符合现实世界的真实情况。在现实世界中，一个样本可能同时包含多种语义信息。例如在文本分

收稿日期：2022-10-09；**修订日期：**2022-11-17

类中,一篇报道可以同时属于经济类和政治类;在生物学领域中,一个基因可以同时具有蛋白质合成、转录和翻译等功能;在场景分类中,一张照片可以同时被标注上沙滩、游客和房屋等标记。此时,传统的单标记学习框架已经无法有效地处理多标记样本,为此多标记学习框架[1-4]应运而生。在此框架下,每个样本由一组特征向量描述,而该样本可能同时隶属于多个类别,学习的目标是对待分类对象进行标记属性标注。多标记数据的分析与挖掘是机器学习和模式识别领域的重点研究内容之一[5],近年来受到国内外众多学者的广泛关注,在理论研究方面取得了重要进展,同时多标记学习方法在情感分类[6-7]、图像视频自动标注[8-10]和生物信息学[11]等领域也得到了良好的应用。随着大数据技术的发展,数据规模不断增长,描述数据对象的特征空间维度也随之增长。高维数据容易引发维数灾难[12],并且普遍存在的噪声特征会对分类产生不利影响。因此,对高维多标记数据进行特征选择[13]已经成为多标记学习不可或缺的部分。

目前,关于多标记特征选择研究已经取得了许多有意义的成果,其中基于信息论设计多标记特征选择算法是研究人员主要采用的方法之一。例如,Hu等[14]将传统的信息熵与邻域粗糙集相结合,提出邻域互信息的概念[15],然后Lin等[16]基于邻域互信息设计了一种多标记特征选择方法,从悲观、中立、乐观3种认知角度,定义3种邻域阈值,在不同阈值下讨论特征的重要性;Lee等[17]设计了一种称为D2F的多标记特征选择算法,该方法通过互信息和交互信息估计多个变量之间的依赖关系来选择特征;张振海等[18]采用特征与标记集合之间的信息增益来度量特征与标记集合之间的关系,并针对信息增益的不同阈值设计了不同的特征选择策略,这些阈值可以根据应用场合自动调整,增强了算法的自适应性;Li等[19]介绍了一种关于互信息的粒度多标记特征选择方法,它通过引入局部标记相关性来获取标记信息粒,然后为每个标记信息粒选择最大相关和最小冗余的特征子集;Sun等[20]提出了一种新颖的多标记特征选择方法,该方法采用虚拟信息和约束凸优化来考虑特征相关性和标记相关性信息;Liu等[21]提出了一种用于流标记的多标记特征选择方法,该方法首先基于互信息和区分指数为每个标记选择特定于该标记的特征,然后融合所有特定标记的特征以产生最终的特征子集。

现有的基于信息论的多标记特征选择方法通常将标记和特征之间的相关性作为特征重要度的评价指标,同时尽量减少特征冗余,进而选择出更有利于判别的有效特征。在这个过程中,标记之间的关系大多被忽略,但事实上标记之间的关系会影响特征选择,这种影响可能是积极的,也可能是消极的。因此,需要识别其他标记在特征选择中的作用,从而可以充分发挥其它标记的积极作用,尽量避免不利影响。为了充分利用标记之间的关系,本文引入了邻域信息熵作为相关性度量标准,提出了基于标记补充的多标记特征选择算法(Multi-label feature selection based on label complementarity, MLLC)。

1 相关知识

1.1 多标记问题描述

多标记数据可表示为一个决策系统MDT $=<U,F,D>$,其中论域 $U=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}\in\mathbf{R}^{n\times m}$ 表示包含 n 个样本的非空有限集合,$\boldsymbol{x}_i\in U$ 是论域 U 中的第 i 个样本,它是一个 m 维向量。$F=\{f_1,f_2,\cdots,f_m\}$ 为描述样本的一组条件特征集合,f_m 表示条件特征集合中的第 m 个特征。$D=\{l_1,l_2,\cdots,l_k\}$ 为 k 个类别标记。

1.2 邻域互信息

定义1[22] $U=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}\in\mathbf{R}^{n\times m}$ 是实数空间上的一个非空有限集合,$\boldsymbol{x}_i$ 是给定的任意样本,若有特征子集 $B\subseteq F$,则样本 x_i 在特征子集 B 上的邻域定义为:$\delta_B(\boldsymbol{x}_i)=\{\boldsymbol{x}_j|\boldsymbol{x}_j\in U,\Delta(\boldsymbol{x}_i,\boldsymbol{x}_j)\leqslant\delta,\delta\geqslant$

0}。其中,函数$\Delta(\boldsymbol{x}_i,\boldsymbol{x}_j)$为距离公式,本文采用欧式距离(即$p=2$)来计算邻域。

$$\Delta(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left(\sum_{j=1}^{m}\left\|f(\boldsymbol{x}_i,c_j)-f(\boldsymbol{x}_j,c_j)\right\|^p\right)^{\frac{1}{p}} \tag{1}$$

δ为邻域阈值,它定义了邻域的大小,其计算方法为

$$\delta=\frac{1}{\lambda\cdot n}\sum_{i=1}^{m}\frac{\sigma(f_i)}{\lambda} \tag{2}$$

式中:$\sigma(f_i)$表示特征f_i的标准差,λ为超参数。$\delta_B(\boldsymbol{x}_i)$为$\boldsymbol{x}_i$在阈值$\delta$上生成的邻域,亦称为$\boldsymbol{x}_i$的邻域粒子。$\delta$越大,$\boldsymbol{x}_i$邻域内的实例越多,即$\left\|\delta_B(\boldsymbol{x}_i)\right\|$越大。

定义2[16]　对于一个多标记邻域决策系统,其中特征子集$B\subseteq F$,对于任意实例$\boldsymbol{x}_i\in U$在特征子集B上的邻域记为$\delta_B(x_i)$,那么实例$\boldsymbol{x}_i$的邻域不确定性被定义为

$$NH_\delta^{x_i}(B)=-\ln\frac{\left\|\delta_B(\boldsymbol{x}_i)\right\|}{n} \tag{3}$$

即整个多标记数据集的平均邻域不确定性可以定义为

$$NH_\delta(B)=-\frac{1}{n}\sum_{i=1}^{n}\ln\frac{\left\|\delta_B(\boldsymbol{x}_i)\right\|}{n} \tag{4}$$

定义3[16]　给定特征子集$B,C\subseteq F$,实例$\boldsymbol{x}_i$在$B\cup C$特征子集下的邻域为$\delta_{B\cup C}(\boldsymbol{x}_i)$,那么联合邻域熵为

$$NH_\delta(B,C)=-\frac{1}{n}\sum_{i=1}^{n}\ln\frac{\left\|\delta_{B\cup C}(\boldsymbol{x}_i)\right\|}{n} \tag{5}$$

根据邻域粗糙集性质$\delta_{B\cup C}(\boldsymbol{x}_i)=\delta_B(\boldsymbol{x}_i)\cap\delta_C(\boldsymbol{x}_i)$,可推导出

$$NH_\delta(B,C)=-\frac{1}{n}\sum_{i=1}^{n}\ln\frac{\left\|\delta_B(\boldsymbol{x}_i)\cap\delta_C(\boldsymbol{x}_i)\right\|}{n} \tag{6}$$

定义4[16]　给定两个特征子集$B,C\subseteq F$,那么B在C下的条件邻域熵可以定义为

$$NH_\delta(B|C)=-\frac{1}{n}\sum_{i=1}^{n}\ln\frac{\left\|\delta_{B\cup C}(\boldsymbol{x}_i)\right\|}{\left\|\delta_C(\boldsymbol{x}_i)\right\|} \tag{7}$$

定义5[16]　给定两个特征子集$B,C\subseteq F$,它们之间的邻域互信息可以定义为

$$\mathrm{NMI}_\delta(B;C)=-\frac{1}{n}\sum_{i=1}^{n}\ln\frac{\left\|\delta_B(\boldsymbol{x}_i)\right\|.\left\|\delta_C(\boldsymbol{x}_i)\right\|}{n\left\|\delta_{B\cup C}(\boldsymbol{x}_i)\right\|} \tag{8}$$

由上述定义可知,$\mathrm{NMI}_\delta(B;C)=NH_\delta(B)-NH_\delta(B|C)$。如果特征子集$B$和$C$相互独立,那么$NH_\delta(B)=NH_\delta(B|C)$,则$\mathrm{NMI}_\delta(B;C)=0$,此时$B$和$C$之间的邻域互信息最小;如果$B$完全依赖于$C$,则$NH_\delta(B|C)=0$,即$\mathrm{NMI}_\delta(B;C)=NH_\delta(B)$,此时$B$和$C$之间的邻域互信息最大。

定义6[16]　给定特征子集$B\subseteq F$,标记子集$l\subseteq D$,则它们之间的邻域互信息可以定义为

$$\mathrm{NMI}_\delta(B;l)=-\frac{1}{n}\sum_{i=1}^{n}\ln\frac{\left\|\delta_B(\boldsymbol{x}_i)\right\|.\left\|l_{x_i}\right\|}{n\left\|\delta_{B\cup l}(\boldsymbol{x}_i)\right\|} \tag{9}$$

1.3　基于邻域互信息的多标记特征选择算法

在基于邻域互信息进行多标记特征选择时,传统的做法是根据最大相关-最小冗余原则进行特征选

择,此时相关性和冗余度均用邻域互信息度量。根据前面的定义,对于给定的集合B和C,如果它们之间的邻域互信息越大,那么相关性就越大。因此,可根据定义计算特征与标记之间的邻域互信息,如定义7所示,然后选择相关性大的特征,因为相关性越大的特征对于该标记越具判别力;同时,计算候选特征与已选特征之间的邻域互信息,如定义8所示,然后选择相关性小的特征,因为特征之间的相关性越小,说明特征越不同,即不存在特征冗余。最终选择的特征是对相关性和冗余度进行平衡后得到的结果。

定义7[23] 给定实例集合$U=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,候选特征$f$以及类别标记集合$D=\{l_1,l_2,\cdots,l_k\}$,其中类别标记$l_i\in D$,那么候选特征$f$和标记集合$D$之间的依赖性可以表示为

$$\mathrm{Dep}(f)=\sum_{l_i\in D}\mathrm{NMI}_\delta(f;l_i) \tag{10}$$

定义8[23] 设$S=\{f_1,f_2,\cdots,f_q\}$为已选特征子集,f为候选特征,则f和已选特征子集S间的冗余可以定义为

$$\mathrm{Red}(f)=\frac{1}{q}\sum_{f_s\in S}\mathrm{NMI}_\delta(f;f_s) \tag{11}$$

此时,根据最大相关-最小冗余原则,选择与标记集合D高度相关并且与已选特征子集S弱相关的候选特征f的目标函数可以表示为式(12),然后可以通过优化方法求得最优解。

$$J(\mathrm{Dep},\mathrm{Red})=\sum_{l_i\in L}\mathrm{NMI}_\delta(f;l_i)-\frac{1}{q}\sum_{f_s\in S}\mathrm{NMI}_\delta(f;f_s) \tag{12}$$

2 基于标记补充的多标记特征选择算法模型

2.1 目标函数

由上述内容可知,基于邻域互信息进行多标记特征选择时,传统的方法一般是依据最大相关-最小冗余原则。此时,仅考虑了特征与标记之间的相关性以及特征与特征之间的冗余。但事实上,在计算某个标记与候选特征邻域互信息时,其他标记对其有不同的作用,可能是正面的,也可能是负面的,也可能是不相关的。为了正确评估所选特征的好坏,有必要考虑标记关系对候选特征的影响。假设F是特征集,$f\in F$是候选特征,$l_i,l_j\in D$是2个标记,根据文献[24]的定义,将这两个标记之间的关系划分为3种类型:

(1)标记独立。如果给定标记l_j,候选特征f为标记l_i提供的信息量不变,即$\mathrm{NMI}(f;l_i|l_j)=\mathrm{NMI}(f;l_i)$,那么标记$l_i$和$l_j$之间相互独立。

(2)标记补充。如果给定标记l_j,候选特征f为标记l_i提供的信息量增大,即$\mathrm{NMI}(f;l_i|l_j)\geqslant\mathrm{NMI}(f;l_i)$,那么标记$l_i$和$l_j$之间存在标记补充关系。

(3)标记冗余。如果给定标记l_j,候选特征f为标记l_i提供的信息量减小,即$\mathrm{NMI}(f;l_i|l_j)\leqslant\mathrm{NMI}(f;l_i)$,那么标记$l_i$和$l_j$之间存在标记冗余关系。

显然,满足标记补充关系的标记能够筛选出为标记提供更多信息的候选特征。相反,具有标记冗余关系的标记具有一定的干扰作用,它会导致候选特征提供的信息量减少,那么一个好的候选特征就不仅应能为标记提供大量的判别信息,同时它还应该可以从其他标记获取更多关于该标记的补充信息。所以,在进行特征选择时,除了关注每个标记与候选特征之间的邻域互信息外,还应该考虑如何充分利用存在标记补充关系的其它标记提供的信息,同时减少存在标记冗余关系的标记的负面影响。

定义9　基于上述讨论,标记之间的关系可定义为

$$\mathrm{NMI}(f;\ l_i;\ l_j)=\mathrm{NMI}(f;\ l_i|l_j)-\mathrm{NMI}(f;\ l_i) \tag{13}$$

如果$\mathrm{NMI}(f;\ l_i;\ l_j)=0$,说明$l_i$和$l_j$的关系是标记独立的;如果$\mathrm{NMI}(f;\ l_i;\ l_j)<0$,那么$l_i$和$l_j$的关系就是标记冗余;如果$\mathrm{NMI}(f;\ l_i;\ l_j)>0$,说明$l_i$和$l_j$的关系是标记补充。

定义10　基于上述分析,标记补充信息的定义为

$$\mathrm{Cmp}(f)=\sum_{l_j\in D-\{l_i\}}\max\left\{0,\mathrm{NMI}\left(f;\ l_i;\ l_j\right)\right\} \tag{14}$$

因此本文提出的基于标记补充的多标记特征选择方法的特征重要度评估函数为

$$J(\mathrm{Dep},\mathrm{Red},\mathrm{Cmp})=\sum_{l_i\in D}\left\{\mathrm{NMI}_\delta\left(f;\ l_i\right)+\sum_{l_j\in D-\{l_i\}}\max\left\{0,\mathrm{NMI}_\delta\left(f_k;\ l_i;\ l_j\right)\right\}-\frac{1}{k}\sum_{f_s\in S}\mathrm{NMI}_\delta\left(f;f_s\right)\right. \tag{15}$$

2.2　算法描述

由上述分析可知,基于标记补充的多标记特征选择算法的具体流程如算法1所述。首先,输入参数λ,求得邻域的阈值,并根据阈值计算出每一个样本的邻域粒子。然后,计算候选特征与标记集合之间的邻域互信息,即候选特征与标记集合之间的依赖度$\mathrm{Dep}(f)$,特征与特征之间的邻域互信息,即候选特征和已选特征集合之间的冗余度$\mathrm{Red}(f)$,以及候选特征与标记之间的补充关系$\mathrm{Cmp}(f)$。最后根据评估函数$J(\mathrm{Dep},\mathrm{Red},\mathrm{Cmp})$计算候选特征$f$的重要度,然后将重要度最大的特征加入当前的特征子集S中。如此不断往复,直到已选特征的个数等于特征总数m。

算法1　基于标记补充的多标记特征选择算法MLLC

输入:多标记决策系统$\mathrm{MDT}=<U,F,D>$和超参数λ

输出:排序好的特征集合S

步骤1　初始化所选特征子集$S\leftarrow\varnothing$

步骤2　对于$\forall f\in F-S$,根据定义7计算候选特征和标记集合之间的依赖度$\mathrm{Dep}(f)$

步骤3　根据定义8计算候选特征f与已选特征集合S之间的冗余$\mathrm{Red}(f)$

步骤4　根据定义10计算候选特征f与标记之间的补充关系$\mathrm{Cmp}(f)$

步骤5　最后根据评估函数$J(\mathrm{Dep},\mathrm{Red},\mathrm{Cmp})$选择$f\in F$

步骤6　$S\leftarrow S\cup\{f\}$,$F\leftarrow F-\{f\}$

步骤7　重复步骤3~6,直到$|S|=m$

步骤8　输出根据特征重要度排好序的特征集合S,算法结束。

3　实验及结果分析

3.1　实验准备

为了验证MLLC的有效性,从MuLan数据库中选择了6个多标记数据集进行对比实验,相关信息如表1所示。其中,样本集、训练集、测试集、特征数以及标记数分别表示样本总数,训练样本个数、测试样本个数、特征数量以及标记数量,基数表示每个样本的平均标记数量,训练集和测试集直接采用了MuLan数据库中的样本数据

表1　多标记数据集

Table 1　Multi-label datasets

数据集	样本集	训练集	测试集	特征数	标记数	基数
Birds	645	322	323	260	20	1.014
Emotion	593	391	202	72	6	1.478
Cal500	502	251	251	68	174	20.578
Yeast	800	500	300	103	14	7.197
Scene	800	500	300	294	6	1.254
Flags	194	129	65	19	7	0.422

集原有的划分。基础分类器采用了ML*k*NN算法,最近邻个数设为10,平滑因子设为1。此外,采用5个常用的性能评价指标度量多标记算法的性能,包括汉明损失(Hamming loss)、1-错误率(One-error)、排序损失(Ranking ss)、覆盖率(Coverage)和平均分类精度(Average precision),其中汉明损失、1-错误率、排序损失以及覆盖率取值越小,算法性能越优。

3.2 实验结果与分析

为了验证MLLC算法的有效性,本文选择了3种多标记特征选择算法进行比较,它们分别是MLACO[25]、PMU[26]、MLRF[27],这些方法都会根据特征重要度获得所选特征的排序集合。其中,MLACO是基于蚁群优化的多标记特征选择方法,它通过同时引入计算标记相关性的有监督启发式函数与计算特征空间冗余性的无监督启发式函数在特征空间迭代搜索最优子集,是一种性能优异的多标记特征选择方法;PMU是一种基于多变量互信息的多标记特征选择方法,它通过三元互信息度量特征子集与标记集合间的相关性;MLRF是基于ReliefF的多标记特征选择算法,它通过引入汉明距离在标记集合上寻找样本的同类与异类近邻。

首先,通过各种多标记特征选择算法计算得到一组特征子集,然后基于这些特征子集进行分类性能的比较。为了保证实验比较的合理性,需要选取的特征数量按照多标记数据集原有特征总数量的一定比率设置,即文献[28]推荐的方法。如果$m<100$,将选取原始特征按重要度排序后的前40%作为已选特征;如果$100\leqslant m<500$,原始特征按重要度排序后的前30%将作为已选特征;如果$m\geqslant 500$,将选取原始特征按重要度排序后的前20%作为已选特征。上述对比算法的参数设置与原论文保持一致,本文邻域阈值参数的选择是本文最关注问题。为了确定超参数λ的取值,在各个多标记数据集上测试了λ在0.2到0.9之间取不同值时,分类器的性能表现,其中步长为0.02,然后选择使分类器性能最佳的超参数用于对比实验。表2~6列出了4种算法在不同数据集上的不同性能表现,符号"↓"表示该性能指标取值越小,分类性能越优;符号"↑"表示该性能指标取值越大,分类性能越优。黑体数据表示在该项评价指标上该取值为最优结果。

表2 各个算法的平均分类精度对比(↑)

Table 2 Comparison of average precision of each algorithm (↑)

数据集	MLACO	PMU	MLRF	MLLC
Birds	0.696 9	0.614 2	0.668 6	**0.701 7**
Emotion	0.781 1	0.757 4	0.763 4	**0.804 6**
Cal500	0.477 77	0.479 3	0.475 4	**0.482 3**
Yeast	0.721 3	0.724 3	0.716 3	**0.729 3**
Scene	0.936 6	0.939 3	0.931 0	**0.951 4**
Flags	0.802 4	0.787 1	**0.808 4**	0.800 1
平均值	0.736 01	0.716 93	0.727 18	**0.744 90**

表3 各个算法的覆盖率对比(↓)

Table 3 Comparison of coverage of each algorithm (↓)

数据集	MLACO	PMU	MLRF	MLLC
Birds	4.520 1	4.699 7	4.486 1	**4.470 6**
Emotion	3.005 0	3.138 6	3.069 3	**2.891 1**
Cal500	130.593 6	**130.334 7**	130.513 9	130.761 0
Yeast	7.900 0	8.006 7	8.006 7	**7.896 7**
Scene	1.330 0	1.313 3	1.330 0	**1.240 0**
Flags	4.846 2	4.952 8	4.784 6	**4.692 3**
平均值	25.365 82	25.407 63	25.365 10	**25.325 28**

表4 各个算法的1-错误率对比(↓)

Table 4 Comparison of one error of each algorithm (↓)

数据集	MLACO	PMU	MLRF	MLLC
Birds	0.374 6	0.526 3	0.424 1	**0.352 9**
Emotion	0.306 9	0.331 7	0.326 7	**0.272 3**
Cal500	0.147 4	0.115 5	0.147 4	**0.107 6**
Yeast	0.256 7	**0.250 0**	0.263 3	**0.250 0**
Scene	0.090 0	0.086 7	0.090 0	**0.076 7**
Flags	**0.184 6**	0.276 9	0.230 8	0.246 2
平均值	0.226 7	0.264 52	0.247 05	**0.217 62**

表5　各个算法的汉明损失对比(↓)

Table 5　Comparison of Hamming loss of each algorithm (↓)

数据集	MLACO	PMU	MLRF	MLLC
Birds	0.051 1	0.067 0	0.057 7	**0.050 9**
Emotion	0.232 6	0.234 3	0.236 0	**0.201 3**
Cal500	0.142 3	**0.139 7**	0.141 1	0.141 8
Yeast	0.218 1	0.220 7	0.222 1	**0.216 7**
Scene	0.045 0	0.050	**0.044 4**	0.046 1
Flags	**0.288 0**	0.325 3	0.290 1	0.309 9
平均值	0.162 85	0.172 833	0.165 233	**0.161 12**

表6　各个算法的排序损失对比(↓)

Table 6　Comparison of ranking loss of each algorithm (↓)

数据集	MLACO	PMU	MLRF	MLLC
Birds	0.130 0	0.143 4	0.132 0	**0.125 7**
Emotion	0.184 0	0.210 1	0.200 0	**0.163 1**
Cal500	0.189 9	**0.188 3**	0.190 5	0.189 5
Yeast	0.203 0	0.200 5	0.209 6	**0.198 9**
Scene	0.038 8	0.039 2	0.043 5	**0.028 3**
Flags	0.229 5	0.247 9	**0.216 2**	0.219 9
平均值	0.162 53	0.171 57	0.165 30	**0.154 23**

从表中可知,相比于其他多标记特征选择算法而言,MLLC算法在6个实验数据集上的整体表现优于其它算法,大部分性能指标都达到了最优值:

(1)在平均分类精度评价指标上,MLLC算法在除Flags数据集以外的其他数据集上都取得了最佳性能。虽然在Flags数据集上的表现略逊于MLACO和MLRF算法,但是平均性能依然最优。

(2)在覆盖率评价指标上,MLLC算法仅在Cal500数据集上稍逊于其他算法,但是平均排序结果达到最优。

(3)在1-错误率评价指标上,MLLC算法仅在Flags数据集上性能稍逊于MLACO和MLRF算法,在其他数据集上都表现最优性能,且平均排序结果达到最优。

(4)在汉明损失评价指标上,MLLC算法在Cal500数据集上稍逊于PMU和MLRF算法,在Scene和Flags数据集上稍逊于MLACO和MLRF算法,在其他数据集上都表现最优性能,且平均排序结果达到最优。

(5)在排序损失评价指标上,MLLC算法在Cal500数据集上稍逊于PMU算法,在Flags数据集稍逊于MLRF算法,在其他数据集上都表现最优性能,且平均排序结果达到最优。

上述实验分析表明,相较于其他3种多标记特征选择算法,由MLLC算法选取的特征子集的分类效果具有突出的优势,同时在特征选择数量上,MLLC算法选取的特征与原始特征空间相比较,特征数量削减了60%以上,实现了特征空间有效降维的目标,这些表明MLLC算法具有良好的有效性和鲁棒性。

其次,为了避免出现局部优势带来的误差影响,本文在部分数据集上分析了算法分类性能随已选特征数量改变而变化的情况。限于篇幅,本文只选择了各种多标记特征选择算法在3个数据集上的不同性能表现进行展示,性能变化趋势分别如图1~5所示。

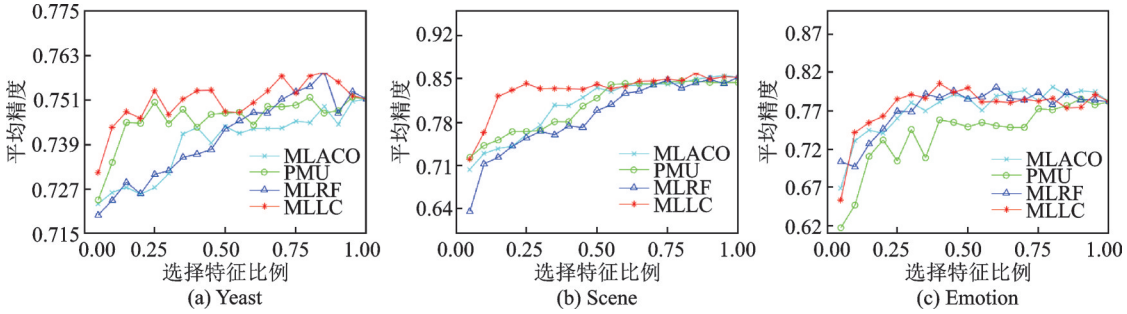

图1　不同算法的平均分类精度变化情况

Fig.1　Variation of average precision of different algorithms

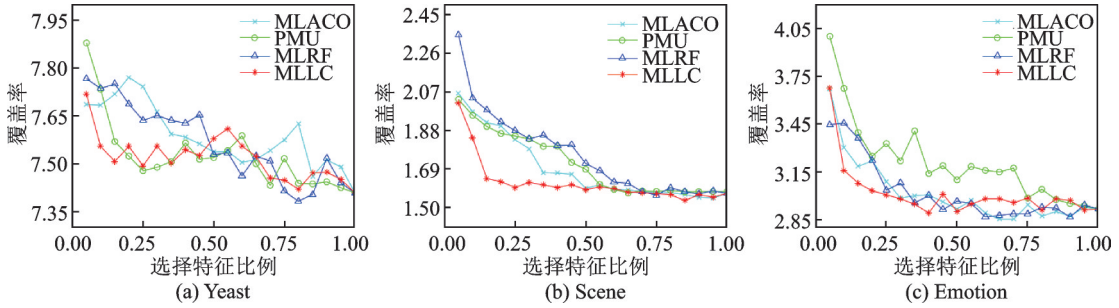

图2 不同算法的覆盖率变化情况

Fig.2 Variation of the coverage of different algorithm

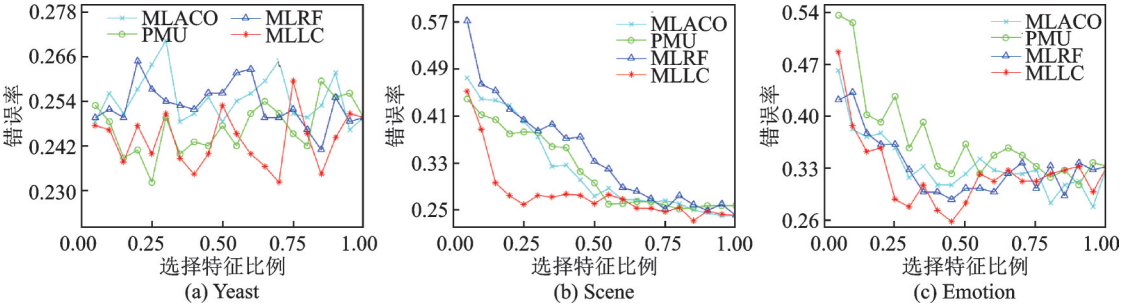

图3 不同算法的1-错误率变化情况

Fig.3 Variation of the one-error of different algorithm

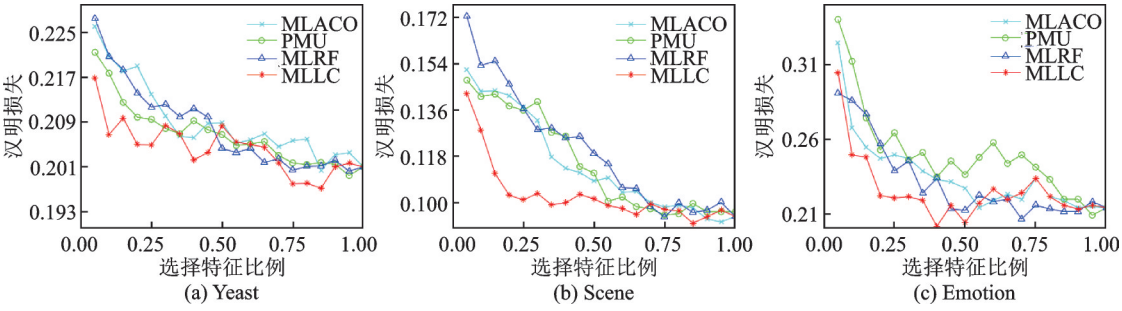

图4 不同算法的汉明损失变化情况

Fig.4 Variation of the Hamming loss of different algorithm

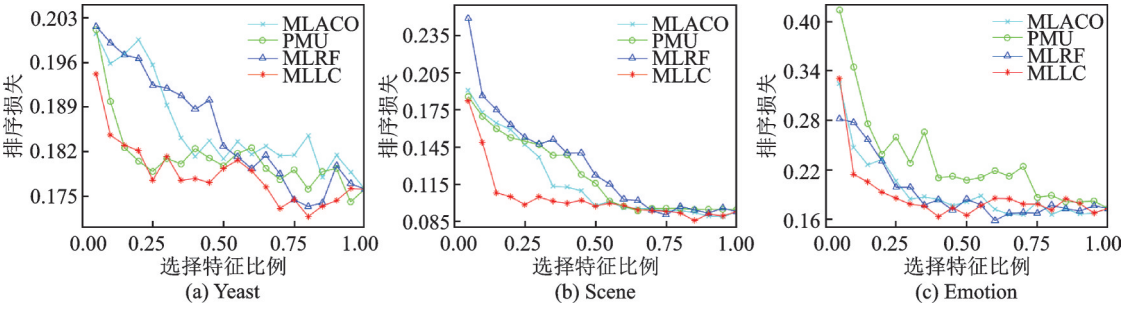

图5 不同算法的排序损失变化情况

Fig.5 Variation of the ranking loss of different algorithm

由图可见,随着选择特征的比例逐渐增大,各个算法在3个数据集上的平均分类精度呈现逐步上升的趋势,而其他4个性能指标则呈现逐步下降的趋势。本文所提出的算法MLLC在各个性能指标上的表现整体优于其他3个算法,特别是在平均分类精度、汉明损失和排序损失这3个性能指标上的优势最为明显。从图中也可以发现,本文所提出的MLLC算法在选择特征比例较小时也能达到较好的分类性能,表明所提出的算法能够实现特征选择的目的。

4　结束语

本文针对多标记特征选择中常常忽略标记关系的问题,重新设计了多标记特征重要度度量函数,不仅引入邻域互信息来度量特征与标记之间的依赖性以及特征与特征之间的冗余,同时也考虑了标记关系对候选特征的影响。采用邻域互信息进行特征选择时,无需对连续型数据进行离散化,从而避免了信息的丢失。最后,通过与多种经典多标记特征选择算法进行对比分析,证明了本文所提出的MLLC算法具有一定的有效性和鲁棒性。由于所提算法没有考虑标记的分布情况,这也是在未来工作中需要继续完善的地方。

参考文献:

[1] ZHANG Jia, LI Candong, SUN Zhenqiang, et al. Towards a unified multi-source-based optimization framework for multi-label learning[J]. Applied Soft Computing Journal, 2019, 76: 425-435.

[2] ZHANG Changqing, YU Ziwei, FU Huazhu, et al. Hybrid noise-oriented multilabel learning[J]. IEEE Transactions on Cybernetics, 2019, 50(6): 1-14.

[3] 余鹰.多标记学习研究综述[J].计算机工程与应用, 2015, 51(17): 20-27.
YU Ying. Summarize of multi-label learning research[J]. Computer Engineering and Application, 2015, 51(17): 20-27.

[4] KASHEF S, NEZAMABADI-POUR H. A label-specific multi-label feature selection algorithm based on the Pareto dominance concept[J]. Pattern Recognition, 2019, 88: 654-667.

[5] TAN Anhui, JI Xiaowan, LIANG Jiye, et al. Weak multi-label learning with missing labels via instance granular discrimination [J]. Information Sciences, 2022, 594: 200-216.

[6] LIU S M, CHEN J H. A multi-label classification based approach for sentiment classification[J]. Expert Systems with Applications, 2015, 42(3): 1083-1093.

[7] HUANG Shu, PENG Wei, LI Jinghuan, et al. Sentiment and topic analysis on social media: A multi-task multi-label classification approach[C]//Proceedings of the 5th Annual ACM Web Science Conference.[S.l.]: ACM, 2013: 172-181.

[8] WU Baoyuan, LYU S, HU Baogang, et al. Multi-label learning with missing labels for image annotation and facial action unit recognition[J]. Pattern Recognition, 2015, 48(7): 2279-2289.

[9] YU Ying, PEDRYCZ W, MIAO Duoqian. Neighborhood rough sets based multi-label classification for automatic image annotation[J]. International Journal of Approximate Reasoning, 2013, 54(9): 1373-1387.

[10] 李永豪,胡亮,高万夫.基于稀疏系数矩阵重构的多标记特征选择[J].计算机学报, 2022, 45(9): 1827-1841.
LI Yonghao, HU Liang, GAO Wangfu. Multi-label feature selection based on sparse coefficient matrix reconstruction[J]. Journal of Computer, 2022, 45(9): 1827-1841.

[11] ZHANG Mingling, ZHOU Zhihua. Multilabel neural networks with applications to functional genomics and text categorization [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1338-1351.

[12] HUANG Miaomiao, SUN Lin, XU Jiucheng, et al. Multilabel feature selection using relief and minimum redundancy maximum relevance based on neighborhood rough sets[J]. IEEE Access, 2020, 8: 1-1.

[13] CHENG Sibao, ZHANG Yumei, DING Q C H, et al. Extended adaptive Lasso for multi-class and multi-label feature selection [J]. Knowledge-Based Systems, 2019, 173: 28-36.

[14] HU Qinghua, ZHANG Lei, ZHANG D, et al. Measuring relevance between discrete and continuous features based on neighborhood mutual information[J]. Expert Systems with Applications, 2011, 38: 10737-10750.

[15] 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.

[16] LIN Yaojin, HU Qinghua, LIU jinghua, et al. Multi-label feature selection based on max-dependency and min-redundancy[J]. Neurocomputing , 2015, 168: 92-103.

[17] LEE J, KIM D W. Mutual information-based multi-label feature selection using interaction information[J]. Expert Systems With Applications, 2015, 42(4): 2013-2025.

[18] 张振海,李士宁,李志刚,等.一类基于信息熵的多标签特征选择算法[J].计算机研究与发展,2013,50(6): 1177-1184.
ZHANG Zhenhai, LI Shining, LI Zhigang, et al. Multi-label feature selection algorithm based on information entropy [J]. Journal of Computer Research and Development, 2013, 50(6): 1177-1184.

[19] LI Feng, MIAO Duoqian, PEDRYCZ W. Granular multi-label feature selection based on mutual information[J]. Pattern Recognition, 2017, 67: 410-423.

[20] SUN Zhengqian, ZHANG Jia, DAI Liang, et al. Mutual information based multi-label feature selection via constrained convex optimization[J]. Neurocomputing, 2019, 329: 447-456.

[21] LIU Jinghua, LI Yuwen, ZHANG jia, et al. Feature selection for multi-label learning with streaming label[J]. Neurocomputing, 2020, 387: 268-278.

[22] LIU Jinghua, LIN Yaojin, LI Yuwen, et al. Online multi-label streaming feature selection based on neighborhood rough set[J]. Pattern Recognition, 2018, 84: 273-287.

[23] QIAN Wenbin, LONG Xuandong, WANG Yinglong, et al. Multi-label feature selection based on label distribution and feature complementarity [J]. Applied Soft Computing, 2020, 90: 106-167.

[24] ZHANG Ping, LIU Guixia, GAO Wanfu, et al. Multi-label feature selection considering label supplementation[J]. Pattern Recognition, 2021, 120: 108137.

[25] MOHSEN P, MOHAMMAD B D, HOSSEIN N. MLACO: A multi-label feature selection algorithm based on ant colony optimization[J]. Knowledge-Based Systems, 2020, 192(C): 105285.

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

[27] SPOLAOR N, CHERMAN E A, MONARD M C, et al. Relief for multi-label feature selection [C]//Proceedings of the 2013 Brazilian Conference on Intelligent Systems. Fortaleza, Brazil:[s.n.], 2014: 19-24.

[28] KASHEF S, NEZAMABADI-POUR H, NIKPOUR B. Multilabel feature selection: A comprehensive review and guiding experiments [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(10):e1240.

作者简介:

余鹰(1979-),**通信作者**,女,副教授,博士,研究方向:多标记学习、计算机视觉、粒计算,E-mail: yuyingjx@163.com。

张志强(1995-),男,硕士研究生,研究方向:多标记学习、粒计算。

钱进(1975-),男,博士,教授,硕士生导师,研究方向:粒计算、大数据挖掘和机器学习等。

万明(1997-),男,硕士研究生,研究方向:多标记学习、粒计算。

(编辑:刘彦东)