文章编号:1000-2367(2023)06-0021-09 DOI:10.16366/j.cnki.1000-2367.2023.06.003

基于 ReliefF 和最大相关最小冗余的多标记特征选择

孙林,徐枫,李硕,王振

(河南师范大学 计算机与信息工程学院,河南 新乡 453007)

摘 要:针对现有的特征选择模型未涉及特征和标记集之间的相关度,造成分类精度偏低等情况,提出了基于 ReliefF 和最大相关最小冗余(maximum Relevance and Minimum Redundancy,mRMR)的多标记特征选择.首先,运用互信息计算每个标记和标记集之间的相关度,使用每项相关度占其相关度之和的比例设计了标记权重,由此构建了特征和标记集间的相关度,初选与标记集相关度高的特征;其次,计算对象在特征上的距离,构建了新的特征权值更新公式,基于标记权重改进多标记 ReliefF 模型.然后,基于互信息和标记权重构建了最大相关性,设计了最小冗余性及其新的最大相关最小冗余评价准则,并将其应用于多标记特征选择,进一步剔除冗余特征;最后,设计了一种基于 ReliefF 和最大相关最小冗余的多标记特征选择算法,有效提高了多标记分类性能.在 8 个多标记数据集上测试所提算法的平均分类精度、覆盖率、汉明损失、1 错误率和排序损失,实验结果证明了该算法的有效性.

关键词:多标记学习;特征选择;标记权重;ReliefF;最大相关最小冗余

中图分类号:TP181 **文献标志码**:A

目前,多标记学习是数据挖掘、人工智能等领域的一个重要的研究方向[1].作为数据挖掘的预处理技术,多标记特征选择策略可分为过滤式、封装式和嵌入式[2].过滤式在效率上有一定的优势,而如何构建合适的评价指标来评价候选特征是过滤式的关键,并且不依赖特定分类器,由此选择出来的特征子集对于不同的分类器适用性更强[3].王晨曦等[4]通过融合标记权重与对象平均间隔构建了邻域信息熵,进而设计了信息粒化多标记特征选择模型.但是该方法未涉及标记之间以及特征与标记集之间的相关性.魏葆雅等[5]使用标记对对象的可分性赋予标记权重,并联合特征权重对特征进行排序,基于标记重要性提出了多标记特征选择模型.但是该模型未涉及特征与标记间的相关性.综合考虑上述缺陷,利用互信息计算标记权重,由此设计特征和标记集间的相关度,初选与标记集相关度高的特征,提高特征与标记集之间的相关性.ReliefF 是一种基于过滤式与特征加权的特征选择算法.陈平华等[6]结合 ReliefF 和多标记贡献值改进特征权值,基于互信息与 ReliefF 设计了多标记特征选择算法.但是这种方法未计算标记之间的相关性.REYES 等[7]提出了扩展的多标记 ReliefF 的特征选择模型.但是,该模型未涉及特征间的相关性和冗余性.刘海洋等[8]利用 ReliefF 算法度量标记间的依赖关系,提出了基于 ReliefF 剪枝的多标记分类算法,但是,该算法未考虑特征与标记的相关性.针对以上算法存在的问题,本文改进 ReliefF 方法,引入标记权重构建特征权值更新公式,提高特征与标记间的相关性.

最大相关最小冗余算法是一种有效的过滤式特征选择模型,其评价函数考虑了特征与类别、特征与特征之间的相关性,因此基于 mRMR 的多标记特征选择得到了越来越多的关注.LI 等[9]为了选择更相关和紧凑的特征子集以及探索标记相关性,基于互信息和 mRMR 构造了多标记特征选择模型.但是该算法计算复杂度较高.LIN 等[10]考虑了标记之间的相关性、特征依赖性与冗余性,基于 mRMR 设计了多标记特征选择模型.但是,它未涉及特征和标记之间的相关性而造成分类精度低且时间代价大.HUANG 等[11]将邻域逼近精度

收稿日期:2023-02-23;**修回日期**:2023-04-06.

基金项目:国家自然科学基金(62076089;61976082).

作者简介:孙林(1979—),男,河南南阳人,河南师范大学教授,博士,研究方向为粒计算、机器学习,E-mail:sunlin@htu.edu.cn.

通信作者:徐枫,E-mail:xufeng_2022@126.com.

与 mRMR 结合,基于邻域粗糙集构建了多标记特征选择方法.但是该方法的计算复杂度较大.SUN 等[12]提出了一种基于模糊邻域粗糙集和 mRMR 的缺失标记特征选择算法.然而,该算法没有考虑标记间相关性,造成去除冗余特征精度不高,不能完全剔除所有的冗余特征,降低多标记分类的预测结果.基于上述分析,使用互信息和标记权重改进最大相关算法,提出了新的 mRMR 评价准则,以此来衡量标记间的相关性以及特征之间的冗余度,有助于对多标记数据集去除冗余特征,获得最佳分类性能.

针对传统 ReliefF 算法仅能处理单标记数据并且未充分考虑特征和标记集之间的相关性,以及传统 mRMR 算法没有考虑标记间相关性而造成分类精度偏低等问题,对传统 ReliefF 算法和 mRMR 算法进行改进,提出了基于 ReliefF 和 mRMR 的多标记特征选择算法.首先,根据标记和标记集之间的互信息定义相关度,计算该相关度所占的比例来构建新的标记权重,构造特征和标记集之间的相关度,初选与标记集相关度较高的特征;然后,计算对象在特征上的距离,构建新的特征权值更新公式,并结合标记权重改进多标记 ReliefF 特征选择模型;最后,结合互信息和标记权重定义最大相关性,使用标记权重值与特征权值之和构建新的 mRMR 评价准则,有效提高模型的分类性能.

1 基础理论

1.1 互信息

给定 $A=\{a_1,a_2,\cdots,a_n\}$ 为一个随机变量,$p(a_i)$ 为变量 a_i 的先验概率,则 A 的信息熵[13] 表示为:

$$H(A)=\sum_{i=1}^{n}-p(a_i)\log_2 p(a_i). \tag{1}$$

给定 $A=\{a_1,a_2,\cdots,a_n\}$ 和 $B=\{b_1,b_2,\cdots,b_m\}$ 为随机变量,$p(a_i,b_j)$ 为 A 和 B 的联合概率,$i=1,2,\cdots,n$,$j=1,2,\cdots,m$,则 A 和 B 的联合信息熵[13] 表示为:

$$H(A,B)=-\sum_{i=1}^{n}\sum_{j=1}^{n}p(a_i,b_j)\log_2 p(a_i,b_j), \tag{2}$$

给定 $A=\{a_1,a_2,\cdots,a_n\}$ 和 $B=\{b_1,b_2,\cdots,b_m\}$ 为随机变量,$p(b_j\mid a_i)$ 为条件先验概率,$i=1,2,\cdots,n$,$j=1,2,\cdots,m$,则 B 在给定 A 下的条件熵[13] 表示为:

$$H(B\mid A)=-\sum_{i=1}^{n}\sum_{j=1}^{n}p(a_i,b_j)\log_2 p(b_j\mid a_i). \tag{3}$$

随机变量 A 和 B 的互信息[13]表示为:

$$MI(A;B)=\sum_{i=1}^{n}\sum_{j=1}^{n}p(a_i,b_j)\log_2\frac{p(a_i\mid b_j)}{p(a_i)}. \tag{4}$$

然后对互信息量进行归一化处理[13],归一化处理公式表示为:

$$NMI(A,B)=2\left[\frac{MI(A;B)}{H(A)+H(B)}\right]. \tag{5}$$

易证明 $NMI(A;B)\in[0,1]$.$NMI(A;B)=0$ 表示 A 和 B 相互独立,$NMI(A;B)=1$ 表示可通过 A 和 B 之一确定另一个.

1.2 ReliefF 算法

在 ReliefF 算法中,两个对象 R_1 和 R_2 在特征 f 上的距离[14] 为:

$$diff(f,R_1,R_2)=\frac{R_1(f)-R_2(f)}{\max(f)-\min(f)}, \tag{6}$$

其中,$R_1(f)$ 表示对象 R_1 在特征 f 上的值;$R_2(f)$ 表示对象 R_2 在特征 f 上的值;$\max(f)$ 表示在特征 f 上的最大值,$\min(f)$ 表示最小值.更新每个特征的权重的公式为:

$$W(f)=W(f)-\sum_{j=1}^{k}\frac{diff(f,R,H_j)}{mk}+\sum_{l\notin L_{R_i}}\frac{P(l)}{1-P(L_{R_i})}\sum_{i=1}^{m}\frac{diff(f,R,M_j)}{mk}, \tag{7}$$

其中,H_j 和 M_j 分别代表对象 R 的第 j 个近邻同类对象和异类对象,$diff(f,R,H_j)$ 和 $diff(f,R,M_j)$ 分别表示对象 R 与 H_j 和 M_j 分别在 f 上的距离,m 为算法的迭代次数,k 为近邻对象数,L_{R_i} 是对象 R_i 的标记,

$P(L_{R_i})$ 是对象 R_i 所属标记的概率,$P(l)$ 表示标记 l 的概率.

2　多标记特征选择方法

2.1　标记权重

假设一个多标记决策系统表示为 $MDS=\langle U,C,D,T\rangle$[15],其中 $U=\{x_1,x_2,\cdots,x_n\}$ 是对象集;C 是条件特征集和 D 是对象对应的标记空间;$T=\{(x_i,t_i)\mid i=1,2,\cdots,n\}$ 是在标记上的映射关系.若对象 x_i 有第 l 个类别标记,记为 $t_i(l)=1$,否则记为 $t_i(l)=0$;且 $\sum t_i\geq 1$,其中每个对象 x_i 由 f 维表示,即 $x_i\in R^f$,对应的标记集由 $t_i\in\{0,1\}^l$ 表示,这里 $l\in D$.

为了解决未考虑特征和标记集之间的相关度而造成分类精度偏低的问题,利用互信息和标记权重,计算特征和标记集之间的相关度,使其有效筛选出与标记集相关度较高的特征子集.

定义 1　在 $MDS=\langle U,C,D,T\rangle$ 中,$L\subseteq D$,$l_k\in L$,$k=1,2,\cdots,m$,基于互信息计算标记和标记集之间相关度公式为:

$$CBTL(l_i)=\frac{1}{m-1}\sum_{l_j\in L,j\neq i}NMI(l_i;l_j). \tag{8}$$

定义 2　在 $MDS=\langle U,C,D,T\rangle$ 中,$L\subseteq D$,$l_k\in L$,$k=1,2,\cdots,m$,计算每个标记和标记集相关度,并计算其相关度的和,用两者的比例来定义标记权重为:

$$WOL(l_k)=\frac{CBTL(l_k)}{\sum_{k=1}^{m}CBTL(l_k)}. \tag{9}$$

定义 3　在 $MDS=\langle U,C,D,T\rangle$ 中,$F\subseteq C$,$f_j\in F$,$j=1,2,\cdots,z$,$L\subseteq D$,$l_k\in L$,$k=1,2,\cdots,m$,根据互信息和标记权重计算特征 f 和标记集 L 之间的相关度为:

$$Corr(f,L)=\sum_{k=1}^{m}NMI(f;l_k)\cdot WOL(l_k). \tag{10}$$

2.2　改进的 ReliefF

为了解决传统 ReliefF 不适用于多标记特征选择的问题,根据标记权重设计新的 ReliefF 模型,由此构建特征权值更新公式,提高 ReliefF 算法的分类性能.

定义 4　在 $MDS=\langle U,C,D,T\rangle$ 中,$X\subseteq U$,$x_i,y_i\in X$,$i=1,2,\cdots,n$,$F\subseteq C$,$f_j\in F$,$j=1,2,\cdots,z$,任意两个对象 x_i 和 y_i 在特征 f_j 上的距离被定义为:

$$diff(x_i,y_i)=\frac{|x_i(f_j)-y_i(f_j)|}{\max(f_j)-\min(f_j)}, \tag{11}$$

其中,$x_i(f_j)$ 是 x_i 在 f_j 上的特征值;$y_i(f_j)$ 是 y_i 在 f_j 上的特征值;$\max(f_j)$ 和 $\min(f_j)$ 分别是 f_j 在对象空间上的最大值和最小值.

定义 5　在 $MDS=\langle U,C,D,T\rangle$ 中,$X\subseteq U$,$x_i\in X$,$i=1,2,\cdots,n$,$F\subseteq C$,$f\in F$,$L\subseteq D$,$l_k\in L$,$k=1,2,\cdots,m$,结合标记权重和距离定义特征权重更新公式为:

$$w_f=\sum_{k=1}^{m}WOL(l_k)\cdot\sum_{i=1}^{n}(diff(x_i,NM^l(x_i))-diff(x_i,NH^l(x_i))), \tag{12}$$

其中,$NM^l(x_i)$ 是在 l 中 x_i 的最近邻异类对象,$NH^l(x_i)$ 是在 l 中 x_i 的最近邻同类对象.$diff(x_i,NM^l(x_i))$ 和 $diff(x_i,NH^l(x_i))$ 分别是在 f 下 x_i 在 l 中与其最近异类对象的距离和最近同类对象的距离.

2.3　改进的 mRMR

为解决 mRMR 未涉及标记之间的相关性,导致删除冗余特征后的分类精度出现不理想的问题,运用互信息和标记权重更新最大相关性公式,并结合特征权值之和,提出新的 mRMR 评价准则,并将其应用于多标记特征选择.

定义 6　在 $MDS=\langle U,C,D,T\rangle$ 中,$L\subseteq D$,$l\in L$,$F\subseteq C$,$f_j\in F$,$j=1,2,\cdots,z$,结合互信息和标记

权重定义最大相关性的计算公式为:

$$\mathrm{MAX}(F,l)=\frac{1}{|F|}\sum_{f_j\in F}NMI(f_j;l)\cdot WOL(l). \tag{13}$$

定义 7 在 $MDS=\langle U,C,D,T\rangle$ 中,$F\subseteq C$,$f_i,f_j\in F$,$i,j=1,2,\cdots,z$,基于互信息定义最小冗余性的计算公式为:

$$\mathrm{MIN}(F)=\frac{1}{|F|^2}\sum_{f_i,f_j\in F}NMI(f_i,f_j). \tag{14}$$

定义 8 在 $MDS=\langle U,C,D,T\rangle$ 中,特征集合 $F\subseteq C$,$L\subseteq D$,$l\in L$,结合特征权重定义新的 mRMR 计算公式为:

$$MR(F)=\frac{1}{1+\mathrm{e}^{-\sum w(F)}}+\mathrm{MAX}(F,l)-\mathrm{MIN}(F). \tag{15}$$

其中,$\sum w(F)$ 为特征集合 F 中每个特征的特征权重之和.

2.4 多标记特征选择算法

由此,设计基于 ReliefF 和 mRMR 的多标记特征选择算法(Multilabel feature selection algorithm using ReliefF and mRMR,MFSRM).首先计算新的标记权重和每个特征和标记集之间的相关度,初次筛选特征子集;然后计算特征权重选择出中间特征子集;最后计算 MR 值得到最终特征排序.其伪代码为:

算法 1 MFSRM

输入 $MDS=\langle U,C,D,T\rangle$.

输出 最优特征子集 R.

步骤 1 For 每个标记 $l\in D$ 和每个特征 $f\in C$;

步骤 2 根据式(9)和式(10)分别计算标记权重 $WOL(l_k)$ 和特征和标记集之间的相关度 $Corr(f,D)$;

步骤 3 End For;

步骤 4 根据特征和标记集之间的 $Corr(f,D)$ 值初次筛选出特征子集 R_0(特征个数关系:$|R_0|=2|R_1|=4|R|$).

步骤 5 For $x_i\in U$;

步骤 6 计算 x_i 的 $NM^l(x_i)$ 和 $NM^l(x_i)$;

步骤 7 End For;

步骤 8 For 每个特征 $f\in C$;

步骤 9 根据式(12)计算特征权重 w_f;

步骤 10 End For;

步骤 11 根据特征权重选择出中间特征子集 R_1(特征个数关系:$|R_1|=2|R|$).

步骤 12 For 特征子集 R_1;

步骤 13 根据式(15)计算 $MR(R_1)$;

步骤 14 End For ;

步骤 15 对 MR 值进行排序并选择前 c 个特征作为最终特征子集 R.

3 实验分析

3.1 实验准备

为了测试 MFSRM 算法的分类性能,选取了 Mulan 数据库(http://mulan.sourceforge.net)中的 8 个数据集进行实验,表 1 描述了 8 个数据集的详细信息.依据文献[16]的平均分类精度(Average Precision,AP)、覆盖率(Coverage,CV)、汉明损失(Hamming Loss,HL)、1 错误率(One Error,OE)、排序损失(Ranking Loss,RL)指标作为分类性能和排序性能的 5 个评价指标.为了充分验证本文算法的有效性,采用上述的5 个基于对象的评价指标和选择特征比例来评价算法的性能,其中 AP 值越大表示算法性能越好(最优值为 1),其余 4 个指标值和特征选择比例越小则算法性能越好.采用多标记 K 最近邻(Multilabel k-nearest neighbor,ML-KNN)分类器,设置近邻个数为 10,平滑参数为 1.实验环境为 Windows 10、CPU Intel(R)Core(TM)i5-8500 3.00 GHz 和内存 8.00 GB,采用 MATLAB 2019a 工具箱进行编码.

3.2 ML-KNN 分类器上的实验分析

为了充分展示 MFSRM 算法在不同数据集上的有效性,选择 5 种对比算法:MLNB(Feature selection for multi-label naive Bayes classification)[17]、PMU(Pairwise Multivariate Mutual Information)[18]、MLRF(Relief for multi-label feature selection)[19]、MFSR(Multi-label feature selection algorithm based on improved ReliefF)[20]和 WFSNR(Weak label feature selection method based on neighborhood rough sets and

Relief)[21]，进一步呈现 MFSRM 算法在 ML-KNN 分类器上的 5 个指标(*AP*、*HL*、*CV*、*OE* 和 *RL*)的实验结果.各个数据集进行实验时所选的特征个数见文献[22].表 2 描述了 MFSRM 算法与 5 种多标记特征选择算法在5 个指标下的实验结果.

表 1　8 个多标记数据集的信息

Tab. 1　Information of eight multilabel datasets

序号	数据集	对象	特征	标记	训练集	测试集
1	Art	5 000	462	26	2 000	3 000
2	Reference	5 000	793	33	2 000	3 000
3	Computer	5 000	681	33	2 000	3 000
4	Education	5 000	550	33	2 000	3 000
5	Recreation	5 000	606	22	2 000	3 000
6	Health	5 000	612	32	2 000	3 000
7	Entertainment	5 000	640	21	2 000	3 000
8	Business	5 000	612	32	2 000	3 000

表 2　8 个数据集上 6 种算法在 5 个指标下的比较结果

Tab. 2　Comparison results of five metrics with six algorithms for eight datasets

指标	算法	Art	Reference	Computer	Education	Recreation	Health	Enter	Business
AP	MLNB	0.488 7	0.597 4	0.628 0	0.524 8	0.459 7	0.672 8	0.551 4	0.871 1
	PMU	0.479 5	0.614 8	0.628 5	0.543 4	0.439 8	0.671 4	0.558 9	0.875 2
	MLRF	0.453 7	0.584 3	0.615 8	0.484 3	0.411 0	0.644 1	0.498 5	0.865 1
	MFSR	0.488 1	0.617 7	0.623 3	0.672 8	0.445 9	0.672 8	0.558 0	0.872 4
	WFSNR	0.496 9	0.599 8	0.626 8	0.521 4	0.441 6	0.645 2	0.554 7	0.868 0
	MFSRM	0.515 1	0.638 8	0.636 8	0.556 6	0.525 7	0.678 8	0.594 0	0.875 6
HL	MLNB	0.061 7	0.031 8	0.040 3	0.042 9	0.062 3	0.044 0	0.062 0	0.028 2
	PMU	0.060 9	0.030 8	0.039 8	0.041 5	0.063 7	0.043 5	0.061 3	0.027 6
	MLRF	0.063 2	0.033 6	0.042 2	0.044 1	0.064 8	0.048 1	0.066 9	0.028 5
	MFSR	0.062 2	0.032 4	0.040 8	0.043 0	0.063 9	0.044 3	0.062 3	0.027 8
	WFSNR	0.062 1	0.030 5	0.040 7	0.042 6	0.063 6	0.046 6	0.061 6	0.028 6
	MFSRM	0.060 7	0.028 9	0.039 3	0.040 1	0.059 3	0.042 6	0.060 1	0.027 5
CV	MLNB	5.599 0	3.611 0	4.468 0	4.071 0	5.038 0	3.489 7	3.370 0	2.434 7
	PMU	5.416 7	3.379 0	4.476 0	3.973 7	5.126 7	3.403 0	3.361 7	2.318 3
	MLRF	5.929 7	3.760 3	4.698 3	4.493 3	5.503 7	3.737 3	3.734 0	2.480 0
	MFSR	5.515 3	3.521 3	4.533 2	4.149 1	5.175 4	3.498 5	3.331 3	2.367 5
	WFSNR	5.471 3	3.507 7	4.477 7	4.097 7	5.087 7	3.521 0	3.392 7	2.417 3
	MFSRM	5.398 0	3.286 0	4.377 7	3.899 7	4.794 0	3.475 0	3.127 7	2.375 7
OE	MLNB	0.663 3	0.502 3	0.451 3	0.621 7	0.702 3	0.417 0	0.607 7	0.126 7
	PMU	0.647 3	0.489 3	0.445 0	0.592 7	0.725 7	0.428 0	0.592 3	0.124 7
	MLRF	0.718 3	0.516 0	0.463 0	0.673 7	0.763 7	0.460 0	0.687 3	0.134 7
	MFSR	0.659 4	0.473 6	0.456 0	0.633 1	0.717 1	0.416 6	0.595 3	0.126 6
	WFSNR	0.656 0	0.504 7	0.453 3	0.627 7	0.722 7	0.470 3	0.596 3	0.132 3
	MFSRM	0.618 7	0.449 3	0.441 0	0.569 3	0.603 0	0.407 3	0.556 3	0.122 3
RL	MLNB	0.152 7	0.094 2	0.093 2	0.096 9	0.190 5	0.065 6	0.126 8	0.043 1
	PMU	0.151 2	0.087 2	0.093 5	0.093 6	0.194 2	0.063 7	0.126 1	0.041 2
	MLRF	0.170 6	0.098 4	0.098 9	0.108 6	0.210 2	0.072 7	0.143 8	0.045 2
	MFSR	0.154 1	0.091 2	0.095 6	0.099 6	0.196 6	0.066 3	0.124 6	0.043 0
	WFSNR	0.152 1	0.091 5	0.093 2	0.097 4	0.194 0	0.067 3	0.126 8	0.044 1
	MFSRM	0.149 0	0.083 9	0.091 2	0.091 5	0.175 6	0.065 6	0.115 7	0.042 8

由表 2 可知，在 *AP* 指标下，MFSRM 算法在除 Education 外的 7 个数据集上，MFSRM 算法的实验结

果均优于其他 5 种算法;尤其在 Recreation 数据集上,MFSRM 算法比次优的 MLNB 算法高 0.066;在 Education 数据集上,MFSRM 算法为次优,比最优的 MFSR 算法低 0.116 2,但 MFSRM 算法比其他 4 种算法高 0.013 2～0.072 3.在 *HL* 指标下,MFSRM 算法在 8 个数据集上的表现均为最优.在 *CV* 指标下,在 Business 数据集上,MFSRM 算法比最优的 PMU 算法和次优的 MFSR 算法分别高 0.057 4 和 0.008 2;在 Health 数据集上,MFSRM 算法比最优的 PMU 算法高 0.072;但在其他 6 个数据集上,MFSRM 算法均优于其他 5 种多标记特征选择算法.在 *OE* 指标下,在这 8 个数据集上,MFSRM 算法均为最优.在 *RL* 指标下,MFSRM 算法在除 Business 和 Health 以外的 6 个数据集上的实验结果均为最优;在 Business 数据集上,MFSRM 算法比最优的 PMU 算法高 0.001 6;在 Health 数据集上,MFSRM 算法为次优,比最优的 PMU 算法高 0.001 9,但总体上仍优于其他 4 种算法;在 Recreation 数据集上,MFSRM 算法明显优于其他 5 种算法,比其他算法低 0.014 9～0.034 6.从整体来看,MFSRM 算法在大部分数据集上的表现均为最优,在极个别数据集上的指标表现为次优,如在 *CV* 和 *RL* 指标下,Business 和 Health 这 2 个数据集上 MFSRM 算法的表现较差,原因是 Business 和 Health 这 2 个数据集是稀疏矩阵数据集,证明了 MFSRM 算法在这类稀疏矩阵数据集上的分类性能不佳.通过观察发现:这 6 种算法在 Art 和 Recreation 这 2 个数据集上的 *AP* 值均过低(低于或略高于 50%).究其原因可能是:在 Art 和 Recreation 这 2 个数据集上,收集数据和划分数据时可能对平均分类精度有所限制,从而导致在现存大多数流行算法的实验结果中呈现的平均分类精度均不高.综上所述,MFSRM 算法的性能得到了有效地提升.

在本节实验的第 2 部分是将 MFSLM 算法与第 1 部分的 4 种多标记特征选择算法:MLRF 算法、PMU 算法、MFSR 算法和 WFSNR 算法作对比分析,给出表 1 中 Art、Business、Computer、Entertainment、Reference 这 5 个数据集各特征选择算法在不同特征比例下的分类情况.图 1 展示了 5 种算法在不同特征比例下 *AP* 指标的变化趋势对比图.由于篇幅限制,其余 4 个指标的变化趋势图可以单独提供.

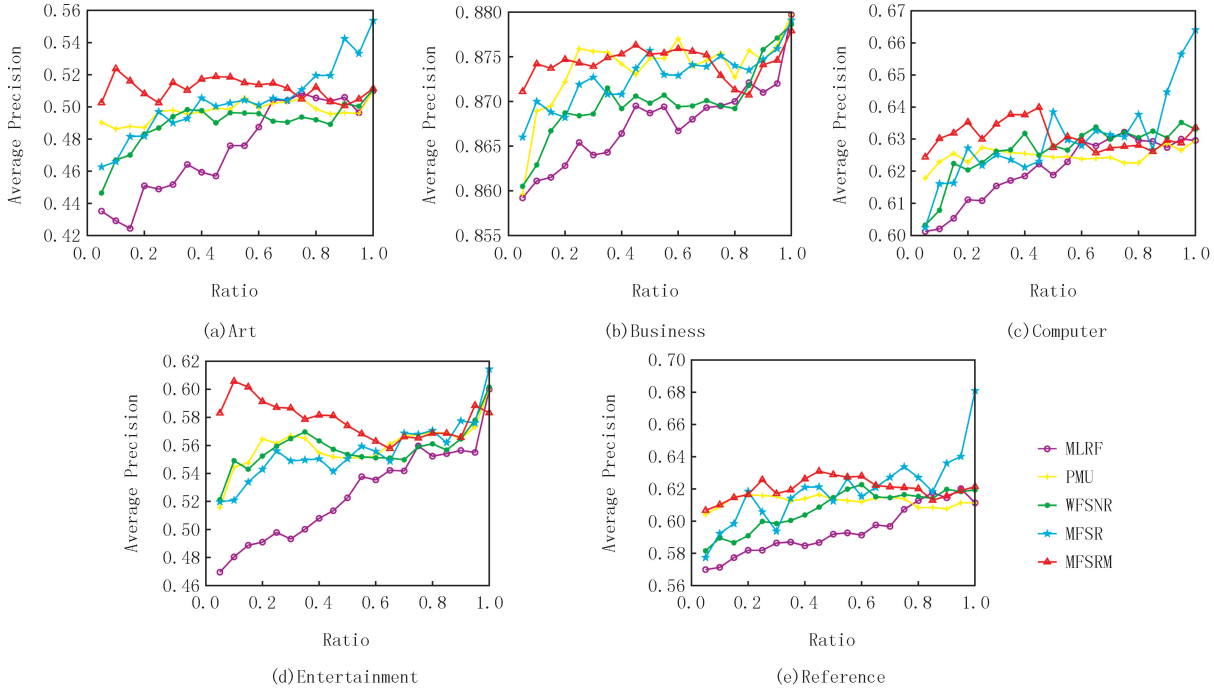


图1 5种对比算法在5个数据集上的*AP*指标

Fig.1 *AP* index of the five compared algorithms on the five datasets

由图 1 可知,在 *AP* 指标下,在 Art 和 Entertainment 这 2 个数据集上,MFSRM 算法在选择特征比例较小时,算法性能较好,在选择特征比例大于 0.7 时,MFSRM 算法的性能差于 MFSR 算法.在 Business 数据集上,在选择特征比例为 0.25～0.35 时,MFSRM 算法差于 PMU 算法;在选择特征比例为 0.40～0.45 和 0.65～0.70 时,MFSRM 算法为最优;在选择特征比例大于 0.5 时,MFSRM 算法性能较差.在 Computer 数据集上,在选

择特征比例小于 0.5 时，MFSRM 算法优于其他算法.在 Reference 数据集，MFSRM 算法在选择特征比例较小时优于其他算法，但选择特征比例大于 0.6 时，MFSRM 算法差于 MFSR 算法.

整体来看，在选择特征比例较小时，MFSRM 算法优于其他 4 种算法，其原因可能是：在选择特征比例较大时，冗余特征较多，分类性能下降.但综上所述，MFSRM 算法的改进是有效的，提升了算法的性能.

3.3　统计分析

本节使用 Friedman 测试[23]和 Nemenyi 测试[24]验证多标记特征选择算法对于不同数据集分类结果的统计重要性.Friedman 测试公式为：

$$\chi_F^2=\frac{12T}{s(s+1)}\left(\sum_{i=1}^{s}R_i^2-\frac{s(s+1)}{4}\right), \tag{16}$$

$$F_F=\frac{(T-1)\chi_F^2}{T(s-1)-\chi_F^2}, \tag{17}$$

其中，T 是数据集的数量，s 代表对比算法数量，R_i 是每个算法在所有数据集上的平均排序. 临界距离表示为：

$$CD_\alpha=q_\alpha\sqrt{\frac{s(s+1)}{6T}}, \tag{18}$$

其中，q_α 是测试临界值，α 是 Nemenyi 测试中的一个重要指标.

参考文献[23]的统计计算方法，检测本文 MFSRM 算法与其他 5 种对比算法在 AP、HL、CV、OE 和 RL 这 5 个指标上是否存在显著差异.表 2 中的实验结果对应的统计结果如表 3 所示，CD 图如图 2 所示.图 2 展示了表 2 中 5 个指标上的 6 种算法的对比结果.接下来，使用了 Friedman 检验判断这 6 种算法在分类性能上是否相同.由文献[25]中表 2.6 的 F 检验的常用临界值可知，在 $\alpha=0.1$，$s=6$ 且 $T=8$ 时，F 检验临界值是 2.019，而由表 3 可知，这 5 个评价指标的 F_F 值均大于临界值 2.019.因此拒绝“所有算法性能相同”这个假设，需要进行后续检验.于是，本文采用 Nemenyi 测试执行后续检验.而对于 Nemenyi 测试，当 $q_{0.1}=2.589$，$s=6$ 且 $T=8$ 时，CD 值为 2.421 8.文献[25]指出：若两个算法存在性能显著不同，则它们的平均序值之差会超出临界值 CD，否则它们之间有一条直线连接，表示它们没有显著差别.从图 2 中可以得到如下结论：本文 MFSRM 算法在 5 个评价指标上分类性能均排名第 1.详细来讲：在 AP 指标上，MFSRM 算法与 WFSNR 和 MLRF 这 2 种算法的性能显著不同，与其他 3 种算法的分类性能没有显著差别；在 HL、CV 和 RL 这 3 种指标上，MFSRM 算法与 PMU 和 MLNB 这 2 种算法之间不存在显著性差异，MFSRM 算法与 MFSR、WFSNR 和 MLRF 这 3 种算法之间存在显著性差异；在 OE 指标上，MFSRM 算法与 PMU 和 MFSR 这 2 种算法之间不存在显著性差异，MFSRM 算法显著优于与 MLNB、WFSNR 和 MLRF 这 3 种算法.总之，与其他 5 种对比算法相比，MFSRM 算法表现出了良好的分类性能.

表 3　5 个评价指标下 6 种算法的统计结果

Tab. 3　Statistical results of six algorithms under five metrics

指标	AP	HL	CV	OE	RL
χ_F^2	28.946 4	33.142 9	28.857 1	30.285 7	26.839 3
F_F	18.331 2	33.833 3	18.128 2	21.823 5	14.275 4

4　结　论

针对一些多标记特征选择算法未充分考虑特征和标记集之间的相关性导致算法分类性能偏低的问题，本文考虑标记之间的相关性以及特征与标记集之间的相关性，提出了基于 ReliefF 和 mRMR 的多标记特征选择算法.首先，基于互信息计算标记和标记集间的相关度来改进标记权重，并结合标记权重计算特征和标记集之间的相关度，提高了特征和标记集之间的相关性.然后，改进 ReliefF 模型，结合标记权重和基于对象在特征上的距离更新特征权值公式，提高算法分类精度.最后，结合互信息和标记权重定义最大相关性，使用特征权值之和构建了新的 mRMR 评价准则，得到最优特征子集.在 8 个多标记数据集上进行实验，结果表明所提算法性能有所提高.

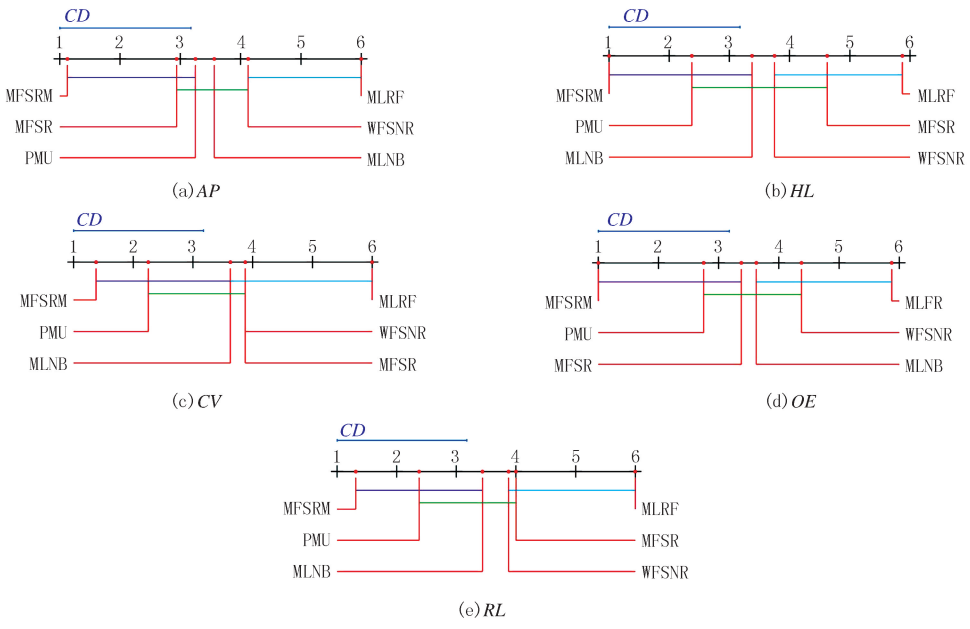


图2 MFSRM与其他5种算法在5个指标下的Nemenyi检验结果

Fig.2 Nemenyi test results of five metrics with MFSRM and the other five algorithms

参 考 文 献

[1] 张志浩,林耀进,卢舜,等.流缺失标记环境下的多标记特征选择[J].山东大学学报(理学版),2022,57(8):39-52.
ZHANG Z H,LIN Y J,LU S,et al.Multi-label feature selection with streaming and missing labels[J].Journal of Shandong University (Natural Science),2022,57(8):39-52.

[2] ZHANG P,SHENG J Y,GAO W F,et al.Multi-label feature selection method based on dynamic weight[J].Soft Computing,2022,26(6):2793-2805.

[3] SUN L,CHEN Y S,DING W P,et al.AMFSA:adaptive fuzzy neighborhood-based multilabel feature selection with ant colony optimization[J].Applied Soft Computing,2023,138:110211.

[4] 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 C X,LIN Y J,TANG L,et al.Multi-label feature selection based on information granulation[J].Pattern Recognition and Artificial Intelligence,2018,31(2):123-131.

[5] 魏葆雅,林梦雷,郑艺峰.基于标记重要性的多标记特征选择算法[J].湘潭大学自然科学学报,2017,39(4):1-5.
WEI B Y,LIN M L,ZHENG Y F.Multi-Label feature selection algorithm based on labeling-importance[J].Journal of Xiangtan University (Natural Science Edition),2017,39(4):1-5.

[6] 陈平华,黄辉,麦淼,等.结合 ReliefF 和互信息的多标签特征选择算法[J].广东工业大学学报,2018,35(5):20-25.
CHEN P H,HUANG H,MAI M,et al.Multi-label feature selection algorithm based on ReliefF and mutual information[J].Journal of Guangdong University of Technology,2018,35(5):20-25.

[7] REYES O,MORELL C,VENTURA S.Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context[J].Neurocomputing,2015,161:168-182.

[8] 刘海洋,王志海,张志东.基于 ReliefF 剪枝的多标记分类算法[J].计算机学报,2019,42(3):483-496.
LIU H Y,WANG Z H,ZHANG Z D.ReliefF based pruning model for multi-label classification[J].Chinese Journal of Computers,2019,42(3):483-496.

[9] LI F,MIAO D Q,PEDRYCZ W.Granular multi-label feature selection based on mutual information[J].Pattern Recognition,2017,67:410-423.

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

[11] HUANG M M,SUN L,XU J C,et al.Multilabel feature selection using Relief and minimum redundancy maximum relevance based on neighborhood rough sets[J].IEEE Access,2020,8:62011-62031.

[12] 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,2021,30(5):1197-1211.

[13] DOQUIRE G，VERLEYSEN M.Mutual information-based feature selection for multilabel classification[J].Neurocomputing，2013，122：148-155.

[14] LOU Q D，DENG Z H，CHOI K S，et al.Robust multi-label Relief feature selection based on fuzzy margin co-optimization[J].IEEE Transactions on Emerging Topics in Computational Intelligence，2021，6(2)：387-398.

[15] YIN T Y，CHEN H M，YUAN Z，et al.Noise-resistant multilabel fuzzy neighborhood rough sets for feature subset selection[J].Information Sciences，2023，621：200-226.

[16] SUN L，WANG T X，DING W P，et al.Feature selection using fisher score and multilabel neighborhood rough sets for multilabel classification[J].Information Sciences，2021，578：887-912.

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

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

[19] SPOLAOR N，CHERMAN E A，MONARD M C，et al.ReliefF for multi-label feature selection[C]//2013 Brazilian Conference on Intelligent Systems.[s.l.：s.n.]，2013：6-11.

[20] 孙林，陈雨生，徐久成.基于改进 ReliefF 的多标记特征选择算法[J].山东大学学报(理学版)，2022，57(4)：1-11.
SUN L，CHEN Y S，XU J C.Multilabel feature selection algorithm based on improved ReliefF[J].Journal of Shandong University(Natural Science)，2022，57(4)：1-11.

[21] 孙林，黄苗苗，徐久成.基于邻域粗糙集和 Relief 的弱标记特征选择方法[J].计算机科学，2022，49(4)：152-160.
SUN L，HUANG M M，XU J C.Weak label feature selection method based on neighborhood rough sets and Relief[J].Computer Science，2022，49(4)：152-160.

[22] 汪正凯，沈东升，王晨曦.基于文本分类的 Fisher Score 快速多标记特征选择算法[J].计算机工程，2022，48(2)：113-124.
WANG Z K，SHEN D S，WANG C X.Fisher score fast multi-label feature selection algorithm based on text classification[J].Computer Engineering，2022，48(2)：113-124.

[23] 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.

[24] DEMSAR J.Statistical comparisons of classifiers over multiple data sets[J].The Journal of Machine Learning Research，2006，7：1-30.

[25] 周志华.机器学习[M].北京：清华大学出版社，2016.

Multilabel feature selection algorithm using ReliefF and mRMR

Sun Lin，Xu Feng，Li Shuo，Wang Zhen

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

Abstract：The correlation between feature and label set is not deeply considered by existing multilabel feature selection models，which results in low classification accuracy. To address the issues，this paper proposed a multilabel feature selection method using ReliefF and maximum Relevance and Minimum Redundancy(mRMR). Firstly，based on the mutual information，the correlation degree between the label and the label-set was defined. A new label weighting was constructed by calculating the proportion of the correlation degree to the sum of the correlation degrees between all labels and the label set. Thus the relationship calculation between the feature and the label set was designed to select the feature subsets that are highly correlated with the label set. Secondly，by calculating the distance of the samples on the feature，a new feature weighting update formula was developed to improve the multilabel ReliefF model based on the label weighting. Thirdly，based on mutual information and the label weighting，the maximum correlation was constructed，the minimum redundancy and new maximum correlation and minimum redundancy evaluation criterion was constructed，which could be applied to multilabel feature selection to further eliminate redundancy features. Finally，a multilabel feature selection algorithm using ReliefF and mRMR was designed to effectively improve the performance of multilabel classification. The experiment was conducted on eight multilabel datasets to test the Average precision，Coverage rate，Hamming Loss，One Error rate and Ranking Loss of the proposed algorithm. The experimental results show that this presented algorithm is effective.

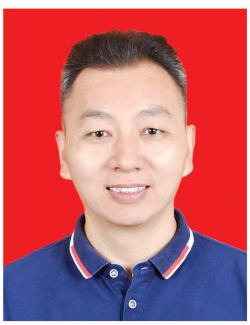
Keywords：multilabel learning；feature selection；label weighting；ReliefF；mRMR

[责任编校　陈留院　赵晓华]

本　期　专　家　介　绍

张明高院士，1999 年当选为中国工程院院士。部级有突出贡献专家，享受国务院政府特殊津贴.先后主持完成了十几项重大科研项目和国防重点工程.其中对信息与电子工程系统中关键的电波技术有独到深入的研究，取得了一系列具有国际先进水平的创造性成果，先后获国家级和部级科技进步奖 5 项，光华科技基金奖一等奖.特别是改进和发展了 5 项国际电联建议书中的关键技术模式，得到了世界各国电波传播领域权威专家的公认；在对流层散射通信、卫星通信、航天飞船通信、陆地移动通信、固定通信等诸多方面发挥了重大作用，从多方面推动了高新技术的发展，为我国赢得了国际声誉.

张安学，西安交通大学教授，博士生导师，电磁与信息技术研究所所长，现为中国天线学会专业委员会委员，中国电磁环境效应产业技术创新战略联盟理事，超高速电路设计与电磁兼容教育部重点实验室学术委员会委员，陕西省天线与控制技术重点实验室学术委员会委员，多功能材料与结构教育部重点实验室和陕西省深空探测智能信号处理重点实验室学科带头人，科技部重点研发计划评审专家，探月工程有效载荷评审专家，中国电子学会优秀博士论文指导教师.主持国家自然科学基金 2 项、863 项目 1 项、总装探索和预研项目 8 项、研究所及企业合作项目 40 余项；授权中国发明专利 25 件、美国发明专利 1 件，发表论文 300 余篇，其中 SCI 收录 200 余篇.获省部级科技进步二等奖 2 项，省高校科学技术研究优秀成果特等奖和一等奖各 1 项，多项研究成果得到推广应用.目前的研究方向涉及新型天线与微波器件设计、雷达信号处理、多天线通信系统与阵列信号处理、微波测试理论与系统设计等.



孙林，天津科技大学特聘教授、天科人才，博士后，博士生导师，获得“河南省科技创新杰出青年”(省杰青)、“河南省高层次人才”、“河南省教育厅学术技术带头人”、“河南省高等学校青年骨干教师”等称号.2003 年和 2007 年毕业于河南师范大学，分别获计算机科学与技术专业学士和硕士学位.现为中国人工智能学会粒计算与知识发现专业委员会委员、知识工程与分布式智能专业委员会青年委员.主要研究方向为大数据挖掘技术与应用、粒计算理论与应用、机器学习、生物信息学.近 5 年以第一作者在国际国内顶级科技期刊 *Information Fusion*、*IEEE Transactions on Fuzzy Systems*、《软件学报》等期刊上发表高水平学术论文 30 余篇(SCI 影响因子≥4)，包括4 篇ESI、16 篇Top、7 篇SCI 一区和 18 篇 SCI 二区，以及 1 篇一级顶尖中文期刊；授权发明专利 30 余件；在科学出版社出版学术专著 4 部.主持国家自然科学基金项目(面上基金 2 项、青年基金 1 项)、中国博士后科学基金面上项目、河南省科技创新人才计划、河南省重点科技攻关计划等 10 余项.荣获河南省自然科学学术奖一等奖 5 项和二等奖3 项、河南省高等教育省级教学成果一等奖 3 项.于 2019－2023 年连续荣获“河南师范大学优秀硕士学位论文指导教师”，并于 2020 年、2022 年和 2023 年分别荣获“河南省优秀硕士学位论文指导教师”.