doi:10.3969/j.issn.1672-1292.2023.01.009

基于邻域近似误差率的多标记特征选择

潘思远[1,2],刘园奎[1,2],毛　煜[1,2],林耀进[1,2]

(1.闽南师范大学计算机学院,福建 漳州 363000)
(2.闽南师范大学计算机学院数据科学与智能应用福建省高等学校重点实验室,福建 漳州 363000)

[**摘要**]　多标记学习可以同时处理与一组标记相关的数据,多标记学习的研究对于多义性对象的学习建模具有十分重要的意义. 与传统的单标记学习一样,数据的高维性是多标记学习的阻碍,因此数据降维是一项十分重要的工作,而特征选择是一种有效的数据降维技术. 提出了基于邻域近似误差率的多标记特征选择算法. 首先,在邻域粗糙集理论的基础上,引入实例的边界来对所有实例进行粒度化. 其次,基于邻域决策误差率提出了邻域近似误差率的策略来评价特征. 最后,在公开的数据集上进行了大量的实验,结果表明所提算法的有效性.

[**关键词**]　多标记学习,特征选择,邻域近似误差率
[**中图分类号**]O643/X703　[**文献标志码**]A　[**文章编号**]1672-1292(2023)01-0066-09

Multi-Label Feature Selection Based on Neighborhood Approximation Error Rate

Pan Siyuan[1,2],Liu Yuankui[1,2],Mao Yu[1,2],Lin Yaojin[1,2]

(1.School of Computer Science,Minnan Normal University,Zhangzhou 363000,China)
(2.Fujian Key Laboratory of Granular Computing and Application,School of Computer Science,Minnan Normal University,Zhangzhou 363000,China)

Abstract:Multi-label learning can process data associated with a set of labels simultaneously,and the study of multi-label learning is very important for learning modeling of polysemous objects. As with traditional single-label learning,the high dimensionality of data is an obstacle to multi-label learning,so data dimensionality reduction is a very important task,and feature selection is an effective data dimensionality reduction technique. A multi-label feature selection algorithm is proposed on the basis of the neighborhood approximation error rate. Firstly,based on the neighborhood rough set theory,the boundaries of instances are introduced to granularize all instances. Secondly,a neighborhood approximation error rate strategy is proposed to evaluate features based on the neighborhood decision error rate. Finally,extensive experiments are conducted on publicly available datasets.The results show the effectiveness of the proposed algorithm.

Key words:multi-label learning,feature selection,neighborhood approximation error rate

在经典的监督学习中,每个实例相对于多个候选标签只属于一个标签. 然而,在许多实际应用中,一个实例通常同时与多个概念关联[1]. 对于图 1(a)所示的关于南非世界杯的新闻报道,既可以认为它属于体育类,也可以认为它属于非洲类,该报道可能还谈及了本次世界杯对南非在经济层面的影响从而属于经济类. 对于图 1(b)所示的图像而言,既可以认为它属于日落类,也可以认为它属于云、树木甚至乡村类. 这样的例子有很多. 一个对象对应一个标签无法准确描述这种场景,因此,多标记分类任务的研究吸引了越来越多的兴趣[2].

在许多现实应用中,多标记数据通常具有成千上万个特征[3]. 例如,从一组文档或网页中提取出数百万个信息性词汇来表示其相关主题. 同时,从图像中提取成千上万的特征来反映其各种语义. 一般来说,对于给定的学习任务,许多特征是冗余或无关的,高维数据可能会给学习算法带来许多挑战,如计算负担、

收稿日期:2022-09-15.
基金项目:国家自然科学基金项目(62076116)、福建省自然科学基金项目(2020J01811、2020J01792、2021J02049).
通讯作者:林耀进,博士,教授,研究方向:数据挖掘,粒计算. E-mail:zzlinyaojin@163.com

(a) 一篇报道　　(b) 一幅景色图画

图 1　多标记实例

Fig. 1　Examples of multi-label

过拟合和性能不佳[4]. 为了解决这一问题,提出了多种基于降维的方法. 这些方法可分为两大类:多标记特征映射和多标记特征选择. 多标记特征映射是一种通过变换或映射将原有高维特征空间转化为新的低维特征空间的方法,新构造的特征通常是原有特征的组合[5]. 然而,特征映射的结果存在缺乏解释,特征映射后生成的新特征失去物理解释.

与多标记特征映射不同,多标记特征选择直接从原始特征空间中选取特征子集,并保留所选特征的物理意义,更具有可解释性,本文将讨论多标记特征选择方法. 多标记特征选择方法通常分为 3 大类[6-8]:过滤式、包装式和嵌入式. 过滤式方法将特征选择与分类器学习分离开来[9]. 包装式方法使用预定学习算法的预测精度来确定所选特征的质量[10-12]. 嵌入式方法同时实现了模型拟合和特征选择. 2015 年,段等[11]提出了基于邻域粗糙集的多标记分类特征选择算法,但其评价特征条件过于苛刻. 由于邻域决策误差率(neighborhood decision error rate,NDER)最小化准则[12]在处理特征选择时具有较强的鲁棒性,但该方法还未应用到多标记特征选择中,且需要大量使用 KNN 分类器辅助操作,从而导致时间复杂度高、降维效率低. 本文使用邻域近似误差率代替邻域决策误差率,并且将其扩展到多标记特征选择中,提出基于邻域近似误差率的多标记特征选择算法(multi-label feature selection based on neighborhood approximation error rate,NAER). NAER 利用标记之间的邻域近似误差率最小化,从而选择邻域近似误差率(neighborhood approximation error rate,NAER)最小的特征子集.

1　基础知识

本文所提方法基于多标记学习和邻域决策误差率,下面就相关概念和基本知识予以介绍.

1.1　多标记学习下的邻域模型

设 U 是非空集合,若 $\forall x_i,x_j,x_k\in U$,存在唯一确定的实函数 Δ 与之对应,且 Δ 满足:

(1)$\Delta(x_i,x_j)\geqslant 0$,当且仅当 $x_i=x_j$,$\Delta(x_i,x_j)=0$;

(2)$\Delta(x_i,x_j)=\Delta(x_j,x_i)$;

(3)$\Delta(x_i,x_k)\leqslant\Delta(x_i,x_j)+\Delta(x_j,x_k)$;

则称 Δ 是 U 上的距离函数,$\langle U,\Delta\rangle$ 是度量空间. 其中 P-范数距离函数定义为

$$\Delta_P(x_i,x_j)=\left[\sum_{l=1}^{N}(x_{li},x_{lj})^P\right]^{\frac{1}{P}}. \tag{1}$$

当 $P=1$ 时,Δ 表示曼哈顿距离. 当 $P=2$ 时,Δ 表示为欧氏距离.

给定多标记决策系统 $\langle U,F,L\rangle$,$U=\{x_1,x_2,\cdots,x_n\}$ 表示样本的集合,$F=\{f_1,f_2,\cdots,f_t\}$ 是描述样本特征的集合,$L=\{l_1,l_2,\cdots,l_k\}$ 是样本的标记集合. 那么,样本 x 在标记 l_i 下的间隔为

$$m_{l_i}(x)=\Delta_{l_i}(x,NM_{l_i}(x))-\Delta_{l_i}(x,NT_{l_i}(x)),\ \forall l_i\in L. \tag{2}$$

式中,$NM_{l_i}(x)$ 代表样本 x 根据标记 l_i 得到的最近异类样本,$NT_{l_i}(x)$ 代表样本 x 根据标记 l_i 得到的最近同类样本.

样本 x 在标记空间 L 下的平均间隔为

$$m^{\mathrm{neu}}(x)=\frac{1}{k}\sum_{i=1}^{k}m_{l_i}(x). \tag{3}$$

给定特征集合 B,若 $m^{\mathrm{neu}}(x)\geqslant 0$,$x$ 的邻域为

$$\delta_B^{\mathrm{neu}}(x)=\{y\mid \Delta(x,y)\leqslant m^{\mathrm{neu}}(x),y\in U\}. \tag{4}$$

若 $m^{\mathrm{neu}}(x)\leqslant 0$,则令 $m^{\mathrm{neu}}(x)=0$.[10]

1.2　邻域决策误差率

定义 1　给定邻域决策系统 $NAT=\langle U,C,D\rangle$,U 是样本的集合,C 是特征空间的集合,D 是标记,$x_i\in U$,给定特征集合 B. $P(\omega_j\mid\delta(x_i))$,$j=1,2,\cdots,c$,是该邻域内属于 ω_j 类的概率,那么定义 x_i 的邻域决策函数为

$$ND(x_i)=\omega_i,P(\omega_l\mid\delta(x_i))=\max_j P(\omega_j\mid\delta(x_i)). \tag{5}$$

式中,$P(\omega_j\mid\delta(x_i))=n_j/N$,$N$ 是邻域内样本的数量,n_j 是第 j 类样本的数量.

引入 0-1 误分类损失函数

$$\lambda(\omega(x_i)\mid ND(x_i))=\begin{cases}0, & \omega(x_i)=ND(x_i),\\ 1, & \omega(x_i)\neq ND(x_i).\end{cases} \tag{6}$$

式中,$\omega(x_i)$ 为 x_i 的真实类别.

定义 2[12]　邻域决策误差率($NDER$)定义为

$$NDER=\frac{1}{n}\sum_{i=1}^{n}\lambda(\omega(x_i)\mid ND(x_i)). \tag{7}$$

式中,n 为样本的总量.

邻域决策误差率本质上就是根据样本邻域内类的分布信息,按照多数决策原则重新给各样本分配决策类,进而统计实际类别与重新分配的类别的差异率.

给定邻域决策系统 $NDT=\langle U,C,D\rangle$,Δ 是 U 上的度量,$\delta\geqslant 0$. 以下表达式成立:

(1)$\gamma_A(D)\leqslant 1-NDER$;

(2)如果邻域决策系统是一致的,$\gamma_A(D)=1-NDER$,并且 $\gamma_c(D)=1$,$NDER=0$. 称 $1-NDER$ 为邻域识别率(neighborhood recognition rate,NRR),即

$$NRR=1-NDER. \tag{8}$$

2　基于邻域近似误差率的多标记特征选择模型

本文将构建基于邻域近似误差率的多标记特征选择模型.

2.1　基于邻域近似误差率的多标记特征选择算法模型

邻域近似误差率是用来评估子空间分类能力的,找到邻域近似误差率最小的子空间,便可进行特征选择,从而提出一个新的特征选择准则——邻域近似误差率最小化准则.

给定多标记邻域近似系统 $MNAT=\langle U,C,L\rangle$,$U$ 是样本的集合,C 是特征空间的集合,L 是标记的集合,给定特征集合 B,由式(3)得到样本 x_i 在标记空间 L 下的平均间隔 $m^{\mathrm{neu}}(x_i)$,$\delta_B^{\mathrm{neu}}(x_i)$ 是样本 x_i 在邻域近似空间的邻域.

引入余弦相似度函数[13]作为获取近似度的方法,余弦相似度函数为

$$S(x_i,y_j)=\frac{\sum_{z=1}^{k}x_{i_z}\times y_{j_z}}{\sqrt{\sum_{z=1}^{k}(x_{i_z})^2}\times\sqrt{\sum_{z=1}^{k}(y_{j_z})^2}},\ y_j\in\delta_B^{\mathrm{neu}}(x_i). \tag{9}$$

式中,x_i 为当前样本,y_j 为 x_i 邻域内的样本,k 是标记个数.

近似度为

$$\alpha_B^{\mathrm{neu}}(x_i)=\frac{\sum_{j=1}^{|\delta_B^{\mathrm{neu}}(x_i)|}S(x_i,y_j)}{|\delta_B^{\mathrm{neu}}(x_i)|},\ y_j\in\delta_B^{\mathrm{neu}}(x_i). \tag{10}$$

式中,S 函数由式(9)得出,$|\delta_B^{neu}(x_i)|$是 x_i 邻域内样本的数量. x_i 为当前样本,y_j 为 x_i 邻域内的样本.

引入 0-1 误分类损失函数:

$$\lambda^{neu}(x_i|\delta_B^{neu}(x_i))=\begin{cases}0, & \alpha_B^{neu}(x_i)\geqslant\theta,\\ 1, & \alpha_B^{neu}(x_i)<\theta.\end{cases} \tag{11}$$

式中,θ 为误差阈值,取值范围是[0.1,0.9]. 邻域近似误差率($NAER$)定义为:

$$NAER=\frac{1}{n}\sum_{i=1}^{n}\lambda^{neu}(x_i|\delta_B^{neu}(x_i)). \tag{12}$$

式中,n 为样本的总量.

邻域识别率(γ_B^{neu})为:

$$\gamma_B^{neu}=1-NAER. \tag{13}$$

给定定义属性 $B,a,B\subset C,a\notin B,a$ 相对于 D 的重要度($Msig$)为:

$$Msig^{neu}(a,B,D)=\gamma_{B\cup a}^{neu}-\gamma_B^{neu}. \tag{14}$$

2.2 基于邻域近似误差率的多标记特征选择算法

邻域近似误差率最小化准则可以用于估计特征的分类能力,那么将其与某一搜索策略相结合,即可实现基于邻域近似误差率最小化准则的特征选择. 根据以上定义,考虑到计算的效率,基于重要度指标,设计了一个前向贪心特征选择算法,提出基于邻域近似误差率的多标记特征选择算法,算法步骤如下.

算法 1 基于邻域近似误差率的多标记特征选择算法

输入:多标记邻域决策系统$\langle U,C,D\rangle$和平均间隔 $m^{neu}(x)$

输出:特征子集 red

1. $\phi\rightarrow red$
2. for each $a_i\in C-red$
3. if $C-red=\phi$
4. return red
5. else
6. Compute $Msig^{neu}(a_i,B,D)=\gamma_{B\cup a_i}^{neu}-\gamma_B^{neu}$
7. 选择 a_k,满足:
8. $Msig^{neu}(a_k,B,D)=\max_i(Msig^{neu}(a_i,B,D))$
9. if $Msig^{neu}(a_k,B,D)>0$
10. $red\cup a_k\rightarrow red$
11. else
12. return red
13. end if
14. end if
15. end for

算法 1 主要包含的工作是不断从候选特征中获取$\max_i(Msig^{neu}(a_i,B,D))$及使$\max_i(Msig^{neu}(a_i,B,D))>0$的特征 a_k,直到$\max_i(Msig^{neu}(a_i,B,D))\leqslant 0$或全部计算完毕. 邻域近似误差率结合前项贪心特征选择算法可使每次计算获得最优特征. 该 NAER 算法不仅可提高计算速度还可保证算法的有效性.

3 实验分析

3.1 实验数据

在公开的 12 个多标记数据集上进行实验,包括 Arts(A)、Birds(B)、Business(C)、CAL500(D)、CHD49(E)、Emotions(F)、Enron(G)、Flags(H)、Sample(I)、Scene(J)、Water-quality(K)和 Yeast(L). 表 1 给出了数据集所对应的详细信息.

其中 Emotions 和 CAL500 是音乐数据集,音频数据集 Birds 用于预测鸟类物种,生物数据集 Yeast 包

含 2 417 个酵母基因的微阵列表达和系统发育概况. 化学数据集 Water-quality 用于预测斯洛文尼亚河流的水质,医学数据集 CHD49 包含中医冠心病的信息. Scene 是包含 2 407 张自然场景图像的图像数据集,Flags 也是图像数据集. Arts、Business 数据集被广泛应用于分类,包括 Enron 数据集. 其为安然公司电子邮件语料库的一个子集,用一组类别进行标记,它是基于收集的电子邮件信息,这些信息被分为 53 个主题类别,如公司战略、幽默和法律建议. 另外还有 Sample 数据集.

表 1　实验数据集

Table 1　Datasets for the experiment

数据集	样本数	特征数	类别数	训练样本数	测试样本数
Arts(A)	5 000	462	26	2 000	3 000
Birds(B)	645	260	20	322	323
Business(C)	5 000	438	30	2 000	3 000
CAL500(D)	502	68	174	251	251
CHD49(E)	555	49	6	372	183
Emotions(F)	593	72	6	391	202
Enron(G)	1 702	1 001	53	1 123	579
Flags(H)	194	19	7	129	65
Sample(I)	600	294	5	400	200
Scene(J)	2 407	294	6	1 211	1 196
Water-quality(K)	1 059	16	14	710	349
Yeast(L)	2 417	103	14	1 499	918

3.2　评价指标

假设有 d 维的实例空间 $\boldsymbol{X}=\boldsymbol{R}^{m*d}$ 和拥有 M 个标记的标记空间 $\boldsymbol{L}=\{-1,+1\}^{M}$. 给定多标记训练集 $\boldsymbol{D}=\{(\boldsymbol{x}_i,\boldsymbol{Y}_i)\mid 1\leqslant i\leqslant n\}$ 和多标记测试集 $\boldsymbol{Z}=\{(\boldsymbol{x}_i,\boldsymbol{Y}_i)\mid 1\leqslant i\leqslant m\}$,其中 $\boldsymbol{x}_i\in\boldsymbol{X}$ 是 d 维的特征向量,$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{id})$. $\boldsymbol{Y}_i\in\boldsymbol{L}$ 是正确的标记集合. 多标记的学习任务从训练集中学到一个预测函数 $f:\boldsymbol{x}\rightarrow\boldsymbol{y}$,根据预测函数得到 $\boldsymbol{Y}'\in\boldsymbol{L}$. 而 $rank_f(\cdot,\cdot)$ 是对应的排序函数[14]. 为衡量特征选择效果选取的 4 个评价指标为:

(1) Average Precision(AP):AP 统计了在样本的类标记的排序序列中,排在相关标记之前的标记依然是相关标记的情况. 该指标越大则系统性能越好.

$$AP=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|\boldsymbol{Y}_i|}\sum_{y\in\boldsymbol{Y}_i}\frac{|\{\boldsymbol{y}'\in\boldsymbol{Y}_i:rank_f(\boldsymbol{x}_i,\boldsymbol{y}')\leqslant rank_f(\boldsymbol{x}_i,\boldsymbol{y})\}|}{rank_f(\boldsymbol{x}_i,\boldsymbol{y})}. \tag{15}$$

(2) Hamming Loss(HL):HL 评估了误分类的情况的比例. 其取值越小则系统性能越好,

$$HL=\frac{1}{m}\sum_{i=1}^{m}\frac{|\boldsymbol{Y}'_i\oplus\boldsymbol{Y}_i|}{M}. \tag{16}$$

式中,$\oplus$ 是异或运算.

(3) Ranking Loss(RL):在该指标统计了样本的类标记的排序序列中,出现错误排序的情况,该指标越小则系统性能越好,

$$RL=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|\boldsymbol{Y}_i||\bar{\boldsymbol{Y}}_i|}|\{(\boldsymbol{y}',\boldsymbol{y}'')\mid f(\boldsymbol{x}_i,\boldsymbol{y}')\leqslant f(\boldsymbol{x}_i,\boldsymbol{y}''),(\boldsymbol{y}',\boldsymbol{y}'')\in\boldsymbol{Y}_i\times\bar{\boldsymbol{Y}}_i\}|. \tag{17}$$

式中,$\bar{\boldsymbol{Y}}_i$ 的是集合 $\boldsymbol{Y}_i$ 的补集.

(4) One-error(OE):该评价指标用于考察在样本的类别标记排序序列中,序列最前端的标记不属于相关标记集合的情况,该指标取值越小系统性能越好,

$$OE=\frac{1}{m}\sum_{i=1}^{m}[[\arg\max_{\boldsymbol{y}\in\boldsymbol{Y}}f(\boldsymbol{x}_i,\boldsymbol{y})]]\notin\boldsymbol{Y}_i]. \tag{18}$$

3.3　实验方法

本实验采用多标记分类器 ML-KNN($K=10$)对已选择的特征子集进行分类精度的评估. 实验平台统一采用 Matlab R2016a,并且所有的实验都在同一台 Inter E5-2620,2.10 GHz,112 GB 内存的计算机上运行. 为了验证 NAER 算法在多标记数据集上进行特征选择的有效性,同时为了能够对比算法分类的性能差异,本文采用 5 种算法和 4 个评价指标作为对比算法进行实验,算法包括 MLNB[15]、MDDMspc[16]、

MDDMproj[16]、PMU[17]以及 MFSNC[18]. 在 NAER 算法中,$0.1 \leqslant \theta \leqslant 0.9$. 由于 MDDMspc、MDDMproj、PMU 以及 MFSNC 将得到特征排序,因此选取前 k(NAER 算法得到的特征子集个数)特征作为特征子集.

3.4 实验结果与分析

为了验证所提 NAER 算法的有效性,首先通过表格记录实验结果,直观比较各种算法在 4 种多标记分类评价指标 Average Precision、Hamming Loss、Ranking Loss 和 One-error 上的效果,然后进一步使用雷达图直观分析各种算法在以上 4 个分类评价指标上的分类性能,最后通过使用折线图突出 NAER 算法的有效性.

表 2~表 5 分别描述了 6 种不同的多标记特征选择算法在 Average Precision、Hamming Loss、Ranking Loss 和 One-error 评价指标上的 ML-KNN 分类结果. 表 2~表 5 中"↑"表示取值越大越好,"↓"表示取值越小越好. 表 6 为各算法在 4 个评价指标下的平均排序. 其中,最后一行为算法的平均精度(Average),加粗数据为不同评价指标下的最优结果.

根据表 2~表 6 中数据可得如下结论:

(1)NAER 算法在 12 个数据集上的 4 个评价指标的平均性能均排名第一. 在 4 个评价指标上,至少在 7 个及以上(超过一半)的数据集上的性能最优,由表 6 可见平均排名均为第一. 故从总体性能上看,NAER 算法在各评价指标上性能都是最优的.

(2)NAER 算法在 ML-KNN 分类器中,在 Birds、Emotions、Flags、Scene 和 Water-quality 数据集上的分类性能在所有指标上都是最优的. 并且在其他数据集上的分类性能和最优值较为接近,故 NAER 算法在 12 个数据集上的分类性能比较稳定.

表 2 基于 ML-KNN 分类器的 *AP* 值(↑)

Table 2 *AP* values based on ML-KNN classifier(↑)

数据集	MLNB 算法	MDDMspc 算法	MDDMproj 算法	PMU 算法	MFSNC 算法	NAER 算法
Arts	0.214 6	0.481 7	0.478 2	0.509 6	**0.531 0**	0.520 0
Birds	0.611 0	0.634 0	0.633 9	0.668 3	0.655 6	**0.693 1**
Business	0.255 9	0.868 0	0.864 7	0.872 0	**0.879 7**	0.873 4
CAL500	0.374 1	0.481 2	0.481 2	0.480 4	0.476 9	**0.484 6**
CHD49	0.779 6	0.779 0	0.773 8	0.777 3	0.753 7	**0.782 8**
Emotions	0.754 8	0.756 0	0.762 2	0.697 7	0.707 6	**0.778 6**
Enron	0.251 2	0.554 8	0.554 8	**0.641 5**	0.611 4	0.537 9
Flags	0.800 7	0.805 6	0.805 6	0.748 4	0.786 8	**0.835 4**
Sample	0.667 2	0.656 1	0.684 9	0.704 1	0.655 7	**0.710 8**
Scene	0.631 5	0.743 9	0.781 5	0.740 1	0.797 4	**0.826 1**
Water-quality	0.641 3	0.636 4	0.636 4	0.640 9	0.636 2	**0.657 1**
Yeast	0.706 8	0.769 7	**0.780 9**	0.743 9	0.744 2	0.750 9
Average	0.557 4	0.680 5	0.680 5	0.685 4	0.686 4	**0.704 2**

表 3 基于 ML-KNN 分类器的 *HL* 值(↓)

Table 3 *HL* values based on ML-KNN classifier(↓)

数据集	MLNB 算法	MDDMspc 算法	MDDMproj 算法	PMU 算法	MFSNC 算法	NAER 算法
Arts	0.065 3	0.061 7	0.061 3	0.059 8	**0.058 9**	0.060 4
Birds	0.073 4	0.059 9	0.068 3	0.062 5	0.056 8	**0.056 2**
Business	0.028 9	0.028 3	0.028 7	0.027 8	**0.026 7**	0.027 4
CAL500	0.187 0	0.140 5	0.140 5	0.139 8	0.140 4	**0.138 6**
CHD49	0.739 5	0.758 7	**0.722 2**	0.739 5	0.743 2	0.725 0
Emotions	0.261 6	0.248 3	0.248 3	0.277 2	0.277 2	**0.241 7**
Enron	0.071 1	0.057 3	0.057 3	**0.051 3**	0.056 4	0.058 4
Flags	0.301 1	0.327 5	0.327 5	0.336 3	0.338 5	**0.283 5**
Sample	0.235 0	0.247 0	0.230 0	**0.227 0**	0.239 0	0.233 0
Scene	0.178 4	0.141 0	0.124 7	0.146 5	0.115 9	**0.109 4**
Water-quality	0.867 2	0.857 1	0.857 1	0.885 8	0.869 4	**0.843 8**
Yeast	0.238 2	0.226 1	0.223 6	0.207 0	0.206 5	**0.203 2**
Average	0.270 6	0.262 8	0.257 5	0.263 4	0.260 7	**0.248 4**

表 4　基于 ML-KNN 分类器的 *RL* 值(↓)

Table 4　*RL* values based on ML-KNN classifier(↓)

数据集	MLNB 算法	MDDMspc 算法	MDDMproj 算法	PMU 算法	MFSNC 算法	NAER 算法
Arts	0.179 0	0.158 6	0.158 7	0.149 4	**0.146 2**	0.147 4
Birds	0.168 3	0.151 1	0.138 1	0.124 9	0.156 2	**0.122 1**
Business	0.049 4	0.043 8	0.044 5	0.043 4	**0.039 4**	0.042 5
CAL500	0.245 8	0.190 9	0.190 9	0.188 2	0.193 2	**0.188 1**
CHD49	0.228 9	**0.222 0**	0.237 6	0.228 3	0.253 5	0.224 9
Emotions	0.219 7	0.205 6	0.203 0	0.265 9	0.249 0	**0.184 6**
Enron	0.116 1	0.109 0	0.109 0	**0.095 4**	0.099 2	0.110 0
Flags	0.223 8	0.226 2	0.226 2	0.276 7	0.248 7	**0.210 3**
Sample	0.289 2	0.288 8	0.252 5	0.249 6	0.297 1	**0.237 9**
Scene	0.237 7	0.172 1	0.136 5	0.154 0	0.125 4	**0.108 0**
Water-quality	0.303 2	0.323 6	0.323 6	0.303 3	0.317 8	**0.295 6**
Yeast	0.211 6	0.194 2	0.184 7	0.183 9	0.182 8	**0.180 0**
Average	0.206 1	0.190 5	0.183 8	0.188 6	0.192 4	**0.171 0**

表 5　基于 ML-KNN 分类器的 *OE* 值(↓)

Table 5　*OE* values based on ML-KNN classifier(↓)

数据集	MLNB 算法	MDDMspc 算法	MDDMproj 算法	PMU 算法	MFSNC 算法	NAER 算法
Arts	0.751 0	0.673 0	0.678 7	0.622 3	**0.590 3**	0.616 0
Birds	0.532 5	0.489 2	0.520 1	0.448 9	0.427 8	**0.393 2**
Business	0.136 7	0.132 3	0.135 7	0.126 7	**0.119 7**	0.125 0
CAL500	0.243 0	0.167 3	0.167 3	0.135 5	**0.107 6**	0.123 5
CHD49	**0.240 4**	0.245 9	0.251 4	0.251 4	0.300 5	**0.240 4**
Emotions	0.326 7	0.356 4	0.331 7	0.455 4	0.415 8	**0.306 9**
Enron	0.419 7	0.400 7	0.400 7	**0.286 7**	0.340 2	0.412 8
Flags	0.246 2	0.184 6	0.184 6	0.261 5	0.246 2	**0.138 5**
Sample	0.505 0	0.545 0	0.510 0	0.460 0	0.530 0	**0.455 0**
Scene	0.581 1	0.406 4	0.357 9	0.431 4	0.330 3	**0.278 4**
Water-quality	0.335 2	0.329 5	0.329 5	0.372 5	0.361 0	**0.320 9**
Yeast	0.260 3	0.326 7	0.306 9	0.246 2	0.248 4	**0.230 9**
Average	0.381 5	0.354 8	0.347 9	0.341 5	0.334 8	**0.303 5**

表 6　各算法在 4 个评价指标下的平均值

Table 6　Average value of each algorithm under 4 evaluation metrics

评价指标	MLNB 算法	MDDMspc 算法	MDDMproj 算法	PMU 算法	MFSNC 算法	NAER 算法
AP	4.833 3	3.333 3	3.250 0	3.750 0	3.833 3	**1.666 7**
HL	5.083 3	3.916 7	3.166 7	3.250 0	3.166 7	**1.833 3**
RL	4.916 7	3.666 7	3.666 7	3.166 7	3.666 7	**1.583 3**
OE	4.500	3.833 3	3.666 7	3.666 7	3.083 3	**1.583 3**
Average	4.833 325	3.687 5	3.437 525	3.458 35	3.437 5	**1.666 65**

由图 2 可知,NAER 算法的性能较其他算法更优秀. 为了更直观地对比 NAER 算法和 5 个对比算法之间分类性能的稳定性,采用雷达图进行实验结果分析. 图 3 展现了在各个数据集下,不同算法的稳定性. 由图 3 可得以下结论:

(1)在 4 个评价指标下,NAER 算法的覆盖面积皆更为饱满. 说明 NAER 算法得到的解更加优异、分类性能更好.

(2)在 4 个评价指标下,NAER 算法至少在 7 个数据集上拥有最优的性能.

(3)在 4 个评价指标下,NAER 算法的覆盖面积远大于其他算法,说明 NAER 算法能够获得更稳定的解.

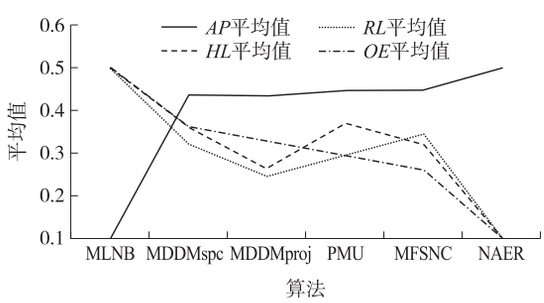
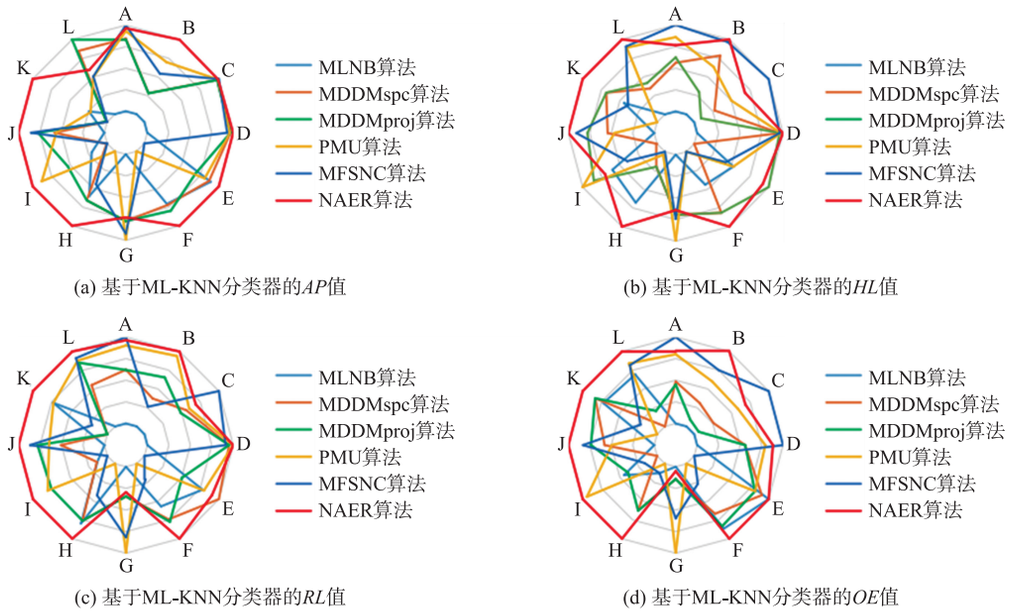


图 2　12 个数据集中不同指标下平均值的对比

Fig. 2　Line plot showing the comparison of the mean values of algorithm with different metrics and 12 datasets



(a) 基于ML-KNN分类器的*AP*值　(b) 基于ML-KNN分类器的*HL*值

(c) 基于ML-KNN分类器的*RL*值　(d) 基于ML-KNN分类器的*OE*值

图 3　蜘蛛图显示算法在不同指标和 12 个数据集下的稳定性

Fig. 3　Spider plot showing the stability of the algorithm with different metrics and 12 datasets

使用 Friedman[19] 检验和 Nemenyi[20] 检验进一步探讨各算法的性能是否有显著差异. Friedman 检验统计定义为 F_F,不同评价指标的 F_F 的值如表 7 所示,在显著性水平 $\alpha=0.05$ 下,临界值 CD 均为 2.176 7. F_F 值大于临界值,因此"所有算法的性能相同"的假设错误. q_a 是 Nemenyi 检验的临界值,根据文献[20]可知 q_a 值的大小. 对于 Nemenyi 测试,当 $\alpha=0.05$ 时,有 $q_a=2.85$,$CD=2.176\ 7$,其中 $k=6$,$N=12$. 根据算法的平均排序值绘制图 4,图中的坐标轴上画出了各对比算法的平均排序,最左边的平均排序最高,若两种算法在所有数据集上的平均排序的差高于临界值 CD,则认为它们有显著性差异. 用一根加粗的线段将性能没有显著差异的算法组连接起来. 根据图 4 可以得出以下结论:

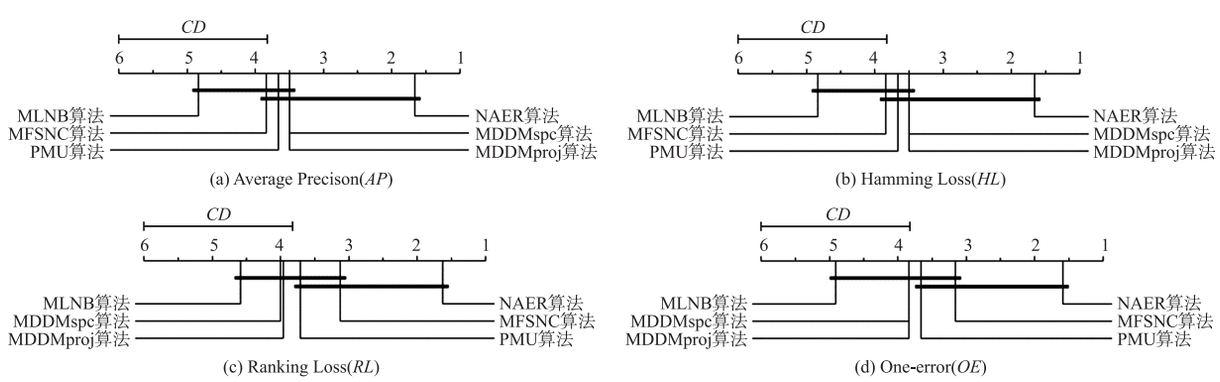
表 7　不同指标下的 Friedman 统计

Table 7　Friedman statistics on different evaluation measures

评价指标	F_F 值	临界值
AP	4.750 0	2.176 7
HL	4.858 1	2.176 7
RL	5.800 0	2.176 7
OE	4.821 9	2.176 7

(1) NAER 算法在 4 个指标上与 MLNB 算法都有显著性差异.

(2) 在 NAER 算法与 MDDM 算法在 *RL* 和 *OE* 指标有显著性差异.



(a) Average Precison(*AP*)　(b) Hamming Loss(*HL*)

(c) Ranking Loss(*RL*)　(d) One-error(*OE*)

图 4　通过 Nemenyi 测试比较 NAER 算法与其他算法的性能差异性

Fig. 4　Comparing the performance variability of NAER with other algorithms by Nemenyi test

根据上述实验的结果表明,NAER 算法的稳定性远高于其他算法,比其他算法具有更强的稳定性和更好的分类性能.

4　结论

提出了一种基于邻域近似误差率的多标记特征选择算法. 首先,不断从候选特征中获取使得邻域识

别率达到最大且邻域识别率大于 0 的特征子集,直到最大邻域识别率为 0 或全部计算完毕. 邻域近似误差率结合前向贪心特征选择算法可以使每次计算获得最优特征. 本文所提的 NAER 算法可以在提高计算速度的同时,保证算法的有效性. 实验结果显示,在 12 个层次化结构数据集上,NAER 算法能够选择出较优的特征子集.

[参考文献](References)

[1] FAKHARI A, MOGHADAM A. Combination of classification and regression in decision tree for multi-labeling image annotation and retrieval[J]. Applied Soft Computing,2013,13(2):1292-1302.

[2] GAO W,ZHOU Z H. On the consistency of multi-label learning[C]//Proceedings of the 24th Annual Conference on Learning Theory. PMLR 19:341-358,2011.

[3] GU Q,LI Z,HAN J. Correlated multi-label feature selection[C]//Proceedings of the 20th ACM International Conference on Information and Knowledge Management. Glasgow,Scotland:Association for Computing Machinery,2011.

[4] DAI J H, XU Q. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification[J]. Applied Soft Computing,2013,13(1):211-221.

[5] LIN Y J,HU Q H,LIU J H,et al. Multi-label feature selection based on neighborhood mutual information[J]. Applied Soft Computing,2016,38:244-256.

[6] SECHIDIS K,NIKOLAOU N,BROWN G. Information theoretic feature selection in multi-label data through composite likelihood [C]//Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition(SPR)and Structural and Syntactic Pattern Recognition(SSPR). Joensuu,Finland,2014.

[7] SPOLAOR N,CHERMAN E A,MONARD M C,et al. A comparison of multi-label feature selection methods using the problem transformation approach[J]. Electronic Notes in Theoretical Computer Science,2013,292:135-151.

[8] SPOLAOR N,MONARD M C,TSOUMAKAS G,et al. Label construction for multi-label feature selection[C]//2014 Brazilian Conference on Intelligent Systems. San Carlos,Venezuela,2014.

[9] SLAVKOV I,KARCHESKA J,KOCEV D,et al. ReliefF for hierarchical multi-label classification[J]. International Workshop on New Frontiers in Mining Complex Patterns. Springer,Cham,2013:148-161.

[10] GHARROUDI,ELGHAZEL,AUSSEM. A comparison of multi-label feature selection methods using the random forest paradigm [C]//Canadian Conference on Artificial Intelligence. Montreal,QC,Canada,2014.

[11] 段洁,胡清华,张灵均,等. 基于邻域粗糙集的多标记分类特征选择算法[J]. 计算机研究与发展,2015,52(1):56-65.

[12] HU Q H,PEDRYCZ W,YU D R,et al. Selecting discrete and continuous features based on neighborhood error minimization[J]. IEEE Transactions on Systems,Man and Cybernetics,Part B,2009,40(1):137-150.

[13] GAO T L,JIA X H,JIANG R,et al. SaaS service combinatorial trustworthiness measurement method based on Markov Theory and cosine similarity[J]. Security and Communication Networks,2022:7080367.

[14] 陈超逸,林耀进,唐莉,等. 基于邻域交互增益信息的多标记流特征选择算法[J]. 南京大学学报(自然科学),2020,56(1):30-40.

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

[16] ZHANG Y,ZHOU Z H. Multi-label dimensionality reduction via dependence maximization[J]. ACM Transactions on Knowledge Discovery from Data,2010,4(3):1-21.

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

[18] 卢舜,林耀进,吴镒潾,等. 基于多粒度一致性邻域的多标记特征选择[J]. 南京大学学报(自然科学),2022,58(1):60-70.

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

[20] NEMENYI P B. Distribution-free multiple comparisons[M]. Princeton,State of New Jersey:Princeton University,1963.

[责任编辑:陈 庆]