【电子与信息科学 / Electronics and Information Science】

不平衡标记差异性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, 安徽安庆 246133; 2) 安徽省高校智能感知与计算重点实验室, 安徽安庆 246133

摘 要: 针对现有的特征选择算法大多未考虑不同标记对样本的描述程度可能存在差异的问题, 提出一种不平衡标记差异性多标记特征选择算法 (multi-label feature selection algorithm with imbalance label otherness, MSIO), 将不同标记下正负标记的频率分布作为该标记的权值加入到特征选择的过程中, 并修正传统的信息熵计算方法, 从而得到一组更高效特征序列. 以多标记 k 近邻 (multi-label k -nearest neighbor, ML- k NN) 为基础分类器, 在 Mulan 数据库的 11 个多标记基准数据集上, 对基于最大相关性的多标记维数约简 (multi-label dimensionality reduction via dependence maximization, MDDM) 算法、基于多变量互信息的多标记特征选择算法 PMU (pairwise multivariate mutual information)、多标记朴素贝叶斯分类的特征选择 (feature selection for multi-label naive Bayes classification, MLNB) 算法、基于标记相关性的多标记特征选择 (multi-label feature selection with label correlation, MUCO) 算法和 MSIO 算法进行评价, 实验结果和统计假设检验说明, MSIO 算法稳定性佳且分类精度高, 具有一定的有效性和优越性.

关键词: 人工智能; 多标记学习; 特征选择; 不平衡数据; 标记相关性; 信息熵; 标记差异性

中图分类号: TP311; TP181 **文献标志码:** A **doi:** 10.3724/SP.J.1249.2020.03234

Multi-label feature selection algorithm with imbalance label otherness

WANG Yibin^{1, 2}, WU Chen¹, CHENG Yusheng^{1, 2}, and JIANG Jiansheng^{1, 2}

1) School of Computer and Information, Anqing Normal University, Anqing 246133, Anhui Province, P. R. China

2) The University Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing 246133, Anhui Province, P. R. China

Abstract: In view of the fact that most of the existing feature selection algorithms do not consider the possible differences existing in the sample description by different labels, a multi-label feature selection algorithm with imbalance label otherness (MSIO) is proposed. The frequency distributions of positive and negative labels under different labels are added to the process of feature selection as the label weight, the traditional method of calculating information entropy is modified to get a more efficient feature sequence. Based on ML- k NN (multi-label k -nearest neighbor), the features are classified on 11 multi-label benchmark datasets of Mulan database, and the algorithms of multi-label dimensionality reduction via dependency maximization (MDDM), pairwise multivariate mutual information (PMU), feature selection for multi-label naive Bayes classification (MLNB), multi-label feature selection with label correlation (MUCO) and MSIO algorithm are evaluated. Experimental results and statistical hypothesis tests show that MSIO algorithm has good stability, high classification accuracy, and certain effectiveness and superiority.

Received: 2019-04-13; **Accepted:** 2019-06-03

Foundation: Natural Science Research Funds of Education Department of Anhui Province (KJ2017A352); Key Laboratory of Data Science and Intelligence Application, Fujian Province University (D1801); Anhui Province Key Laboratory of Affective Computing & Advanced Intelligent Machine (ACAIM160102)

Corresponding author: Professor WANG Yibin. E-mail: wangyb07@mail.ustc.edu.cn

Citation: WANG Yibin, WU Chen, CHENG Yusheng, et al. Multi-label feature selection algorithm with imbalance label otherness [J]. Journal of Shenzhen University Science and Engineering, 2020, 37(3): 234-242. (in Chinese)

Key words: artificial intelligence; multi-label learning; feature selection; imbalanced data; label correlation; information entropy; label otherness

现实世界广泛存在着多标记学习对象, 多标记学习已日渐成为机器学习、数据挖掘和深度学习^[1-2]等领域的研究热点之一. 这些标记之间并非相互独立的, 而是有着一定的关联, 如一篇新闻报道可能同时属于“经济”、“体育”和“国家”; 人的面部表情可能同时被“开心”、“激动”和“兴奋”等标记. 因此, 如何利用标记的相关性构造出泛化性较强的分类器是多标记学习的关键之一^[3]. 刘军煜等^[4]通过挖掘标记之间的关联性, 提出一种关联规则挖掘的多标记分类算法. 何志芬等^[5]提出基于多标记分类和标记相关性的联合学习. 蔡亚萍等^[6]利用标记的局部相关性进行多标记学习和分类. 吴磊等^[7]通过构建标记的类属属性提出基于类属属性的多标记学习算法. 王一宾等^[8]通过关联规则分析标记空间并提出一种基于回归核极限学习机的多标记分类算法. 例如, 在某个实例中, 是否有该标记与实例的特征属性密切相关, 如“感冒”和“肺炎”都有“发烧”和“咳嗽”症状, 但若还有“流鼻涕”症状, 则“感冒”的可能性要大于“肺炎”, 此现象被称为标记的不平衡性. 标记的不平衡性广泛存在于现实世界中, 也正是由于标记的不平衡性造成了不同标记对样本实例的描述程度存在一定的差异性, 有些标记出现的频率较多, 能描述大部分的样本; 而有些标记仅仅存在于少量样本中, 但往往这一小部分的标记却包含了很多信息. 可见, 实例的特征会直接影响到标记的结果, 因此, 研究特征与标记的相关性十分重要.

传统处理不平衡性标记的方法大多是先通过抽样或重采样将不平衡数据处理为平衡数据再进行研究, 但是这种方式常会改变原数据集属性, 丢失部分信息, 降低分类器的分类精度. 若能将不同标记包含的信息加入到分类过程中, 则不仅能保留特征空间的原始属性, 还能提高分类器的精度. 在多标记学习中, 为更准确的描述样本实例, 往往需要大量特征, 且特征越多描述越准确. 但随着特征数据的增加, 弱相关特征和冗余特征也增多, 严重影响到分类器的分类精度, 甚至造成误分类. 因此, 需先对特征数据进行降维. 特征选择是一种广泛使用且有效的降维方法, 经过解析样本特征与标记之间的相关性, 选择出相关性高且冗余性小的特征作为特征子集进行分类训练与预测^[9]. 张振海等^[10]提

出一类基于信息熵的多标记特征选择算法. 刘景华等^[11]基于互信息提出基于局部子空间的特征选择算法. LIN 等^[12-13]通过扩展互信息提出一种基于邻域互信息的多标记特征选择算法, 进而又提出一种基于模糊互信息的特征选择算法.

研究发现, 在多标记学习中, 由于标记对样本的描述存在差异性, 即在每个标记下正类与负类出现的频率不一样, 这种标记频率分布可为多标记学习的研究提供一定的辅助信息, 从而提高分类的精度^[14]. 本研究提出一种不平衡标记差异性多标记特征选择 (multi-label feature selection algorithm with imbalance label otherness, MSIO) 算法, 首先计算标记空间中, 记录每个标记下正标记 (正类) 样本和负标记 (负类) 样本出现的频率分布, 并作为相应标记的权值保存在权值矩阵中; 其次, 考虑到标记空间中的标记包含一些辅助信息结合信息熵设计相应的度量方法以度量特征与标记之间的相关性; 最后, 根据所构造出的模型提出不平衡标记差异性多标记特征选择算法. 在 11 个常用公开数据集上与 5 个常用的多标记特征选择算法^[15]对比, 证明 MSIO 算法可提高分类器的分类精度.

1 相关知识

1.1 多标记学习框架

多标记学习是一种针对实际生活中普遍存在的多义性现象的学习框架, 在此框架下, 样本由多个特征和多个标记构成, 学习目的是将未知的实例对应上更多正确的标记^[16].

假设 T 是由 n 个特征组成的特征集合 $T = \{t_1, t_2, \dots, t_n\}$, L 是由 m 个标记组成的标记集合 $L = \{l_1, l_2, \dots, l_m\}$, 在标记集合中, 有该标记为 1, 否则为 0, 则含有 z 个样本的多标记数据集表示为 $\text{DataSet} = \{(T_i, L_i) \mid 1 \leq i \leq z, T_i \in T, L_i \in L\}$ (1)

1.2 信息熵与互信息

定义 1^[11] 若 $X = \{x_1, x_2, \dots, x_m\}$ 为随机变量, x_i 的概率为 $p(x_i)$, 则 X 的不确定性期望为

$$H(X) = - \sum_{i=1}^m p(x_i) \lg p(x_i) \quad (2)$$

$H(X)$ 亦被称为随机变量 X 的熵, 其值越小表示 X 的期望和不确定性程度越小.

定义 2^[11] 设随机变量 $X = \{x_1, x_2, \dots, x_m\}$ 和 $Y = \{y_1, y_2, \dots, y_n\}$, 则 X 和 Y 的联合熵为

$$H(X, Y) = - \sum_{i=1}^m \sum_{j=1}^n p(x_i, y_j) \lg p(x_i, y_j) \quad (3)$$

定义 3^[11] 设随机变量 $X = \{x_1, x_2, \dots, x_m\}$, $Y = \{y_1, y_2, \dots, y_n\}$, 则 Y 在 X 条件下的条件熵为

$$H(Y|X) = - \sum_{i=1}^m \sum_{j=1}^n p(x_i, y_j) \lg p(y_j|x_i) \quad (4)$$

$H(Y|X)$ 可用来度量 Y 在给定 X 时的不确定性程度.

定义 4^[11] 若 X 和 Y 为已给定的随机变量, 则定义 X 与 Y 之间的互信息为

$$I(X; Y) = - \sum_{i=1}^m \sum_{j=1}^n p(x_i, y_j) \lg \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \quad (5)$$

$I(X; Y)$ 用于衡量随机变量 X 和 Y 的相关性, $I(X; Y)$ 越大, 表明 X 与 Y 之间的相关性越大. 同时, 互信息还满足以下关系

$$\begin{aligned} I(X; Y) &= H(X) + H(Y) - H(X, Y) = \\ &= H(X) - H(X|Y) = \\ &= H(Y) - H(Y|X) \end{aligned} \quad (6)$$

2 不平衡标记差异性多标记特征选择

2.1 特征与标记集合的互信息

在多标记学习中, 一个样本由多个特征和标记描述, 则特征与标记集合之间的互信息可定义为:

定义 5^[11] 对于给定描述样本的特征 f 和标记集合 $L = \{l_1, l_2, \dots, l_m\}$, $\forall l_i \in L, i = 1, 2, \dots, m$, 若特征 f 与标记 l_i 之间的互信息为 $I(f; l_i)$, 则特征 f 与标记集合 L 之间的互信息为

$$\text{IML}(f; L) = \sum_{i=1}^m I(f; l_i) \quad (7)$$

由 $I(f; l_i) \geq 0$ 可知, $\text{IML}(f; L) \geq 0$. 当 f 和 L 相互独立时, 等号成立, 此时特征与标记之间不提供任何信息, 即该特征与标记空间互信息为最小值. 互信息值越大, 特征与标记之间的关系越密切.

同理, 式(8)和式(9)成立.

$$\text{IML}(f; L) = \text{IML}(L; f) \quad (8)$$

$$\text{IML}(f; L) = \sum_{i=1}^m H(l_i) - \sum_{i=1}^m H(l_i|f) \quad (9)$$

根据式(9)可得, 若

$$\text{IML}(f; L) = \sum_{i=1}^m H(l_i) \quad (10)$$

则表明标记集合 L 可完全由特征 f 确定, 两者的不确定性之和取得最大值.

2.2 不平衡标记差异性多标记特征选择模型

在多标记数据集中, 常因标记的不平衡性导致不同标记对样本描述程度有所不同, 而目前多数算法并未考虑这种情况. 为此, 本研究根据标记空间中的每个标记下的正负样本个数对标记赋予权值, 建立标记权值模型, 具体定义为:

定义 6 给定样本空间 $U = \{x_1, x_2, \dots, x_z\}$, 其特征空间 $T = \{t_1, t_2, \dots, t_n\}$, 标记空间 $L = \{l_1, l_2, \dots, l_m\}$. 对于 $\forall l_i \in L, i = 1, 2, \dots, m$, 样本总数为 z , 在第 i 个标记下的正标记和负标记数目分别为 $|l_i^+|$ 和 $|l_i^-|$, 权值为 $\omega_i^+ = |l_i^+|/z$ 和 $\omega_i^- = |l_i^-|/z$.

对于传统的单标记运用互信息进行特征选择过程中, 假使用两个特征 f_1 和 f_2 对样本进行刻画, 则对于特征与特征之间的冗余性用 $I(f_1; f_2)$ 进行有效表示; 倘若样本的特征用给定的 f 表示, 样本的类别标记用 l 表示, 则对于特征与标记类别之间的相关程度用 $I(f; l)$ 进行有效表示. 而在现有的多标记学习中, 各个样本可用 1 个特征向量来表示, 也可能隶属于多个类别标记, 同时, 充分考虑到标记权值模型下的各样本所含有的有用信息, 因此需对加权特征和标记集合之间的相关性进行更深层次的探究. 定义 7 给出了加权特征与标记集合之间的互信息定义.

定义 7 对于给定样本空间 $U = \{x_1, x_2, \dots, x_z\}$, 标记空间 $L = \{l_1, l_2, \dots, l_m\}$, $\forall l_i \in L, i = 1, 2, \dots, m$, 由定义 6 和式(7)可知, 加权特征 f 与标记之间的互信息为

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

其中, “*” 可为 “+” (正标记) 或 “-” (负标记), 则 ω_i^+ 和 ω_i^- 分别为定义 6 中的正负标记的权值.

2.3 MSIO 算法描述

考虑到标记对样本的描述存在一定的差异性, 为使描述的样本实例更准确, 同时也为了更充分地挖掘存在于少量样本中但包含众多有用信息的一些标记, 首先计算标记空间中同一标记下正负标记的权值; 其次, 利用加权信息熵度量特征与标记空间之间的相关性, 并由此得出两组不同的特征重要度排序; 最后, 将两组重要度不同的特征进行融合排序, 得出最终特征序列. MSIO 算法伪代码如图 1.

输入: 多标记训练集 D
 输出: 排序后的特征序列 rank

- 1) $IML = \emptyset$;
- 2) for each $f_i \in F$
- 3) rank = $\emptyset$
- 4) for each $l_j \in L$
- 5) 根据定义 5 计算和 $IML(f_i; L_j^+)$ 和 $IML(f_i; L_j^-)$
- 6) end
- 7) 根据式 (11) 计算每个特征 f_i 在同一标记的不同情况下的互信息
- 8) end
- 9) 根据 7) 得到的互信息降序排序, 得到正标记的特征序列 rank⁺ 和负标记的特征序列 rank⁻
- 10) 对 rank⁺ 和 rank⁻ 融合排序, 得到最终序列 rank, 并输出

图1 MSIO 算法伪代码

Fig. 1 Pseudocode of MSIO algorithm

分析图 1 的算法可见, MSIO 算法先通过统计标记空间中类标记下的正负标记密度, 然后计算标记空间与特征空间的相关性, 并按照标记密度对特征进行赋权得分输出两组序列, 再按照一定关系对两组序列加权得出最终序列. 此方法简单易行且运行快速, 既考虑了多种因素, 又能有效提高分类器的分类精度. MSIO 算法流程图如图 2.

图2 MSIO 算法流程图

Fig. 2 Flowchart of MSIO algorithm

3 实验数据及结果分析

3.1 实验数据

采用 Mulan 数据库^[16] 中 11 个常用的公开实验数据集(表 1), 来验证算法 MSIO 的有效性.

表1 多标记数据集^[17]Table 1 Multi-label datasets^[17]

数据集	样本数	特征数	标记数	训练样本数	测试样本数
Arts	5 000	462	26	2 000	3 000
Business	5 000	438	30	2 000	3 000
Cal500	502	68	174	251	251
Computer	5 000	681	33	2 000	3 000
Education	5 000	550	33	2 000	3 000
Emotions	593	72	6	391	202
Health	5 000	612	32	2 000	3 000
Rec	5 000	606	22	2 000	3 000
Ref	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 评价指标

利用平均准确率(average precision, AP)、排位损失(ranking loss, RL)、1-错误(one-error, OE)和海明损失(Hamming loss, HL) 4 个评价指标^[18] 对多标记实验实验结果进行验证和评价度量.

测试集为 $\{(x_i, Y_i)\}_{i=1}^m \subset \mathbf{R}^d \otimes \{+1, -1\}^L$. 其中, Y_i 为隶属于样本 x_i 的相关标记集合, 经算法预测得到的标记集合记为 $h(x)$; 符号 $\otimes$ 表示维度相乘. 根据预测函数 $f_i(x)$ 定义排序函数为 $\text{rank}_f(x, l) \in \{1, 2, \dots, L\}$.

AP 为评估预测标记排在前列且正确存在于相关样本标记的平均概率, 如式 (12). 该值越大表示分类效果越好, 最优值为 1.

$$AP(f) = \frac{1}{m} \sum_{i=1}^m \left(\frac{1}{|Y_i|} \sum_{l \in Y_i} \frac{(k | \text{rank}_f(x_i, k) \leq \text{rank}_f(x_i, l), k \in Y_i)}{\text{rank}_f(x_i, l)} \right) \quad (12)$$

其中, Y_i 为隶属于样本 x_i 的相关标记集合.

RL 用于评估无关标记在相关标记的样本前列的多少, 如式 (13). 该值越小表示分类效果越好, 最优值为 0.

$$RL(f) = \frac{1}{m} \sum_{i=1}^m \left(\frac{1}{|Y_i| |\bar{Y}_i|} \left| \{ (l, k) \mid \text{rank}_f(x_i, l) \geq \text{rank}_f(x_i, k), (l, k) \in Y_i \bar{Y}_i \} \right| \right) \quad (13)$$

其中, $\bar{Y}_i$ 为集合 Y_i 在标记空间中的补集.

HL 用于衡量样本在单一标记上的非正确匹配情况, 如式 (14). 该值越小, 分类效果越好, 最优值为 0.

$$HL(h) = \frac{1}{m} \sum_{i=1}^m \frac{1}{L} |h(x_i) \neq Y_i| \quad (14)$$

其中, Y_i 为隶属于 x_i 的相关标记集合; $h(\cdot)$ 为分类器, 即可得 x_i 的预测标记向量.

OE 用于衡量最高排序中样本的标记不存在于相关标记集合中的情况, 如式 (15). 该值越小, 表示分类效果越好, 最优值为 0.

$$OE(h) = \frac{1}{m} \sum_{i=1}^m \{ [\arg \max_{l \in L} f(x_i, l)] \notin Y_i \} \quad (15)$$

在 4 个评价指标中, 除去 AP 值越大越优, 其余的越小越优.

3.3 实验结果及分析

本研究实验代码均在 Matlab2016a 中运行, 硬件环境 Inter® Core™ i7-7700HQ CPU @ 2.80 GHz, 8

Gbyte 内存; 操作系统为 Windows 10. 以多标记 k 近邻 (multi-label k -nearest neighbor, ML- k NN)^[19] 作为基础分类器, 对基于最大相关性的多标记维数约简 (multi-label dimensionality reduction via dependence maximization, MDDM) 算法^[20-21]、基于多变量互信息的多标记特征选择算法 PMU (pairwise multivariate mutual information)^[22]、多标记朴素贝叶斯分类的特征选择 (feature selection for multi-label naive Bayes classification, MLNB) 算法^[23]、基于标记相关性的多标记特征选择 (multi-label feature selection with label correlation, MUCO) 算法^[13] 和 MSIO 算法的 AP、RL、OE 和 HL 值进行排序. 其中, MDDM 算法按照参数所选择的不同分为 MDDM_{spc} 与 MDDM_{proj} 算法. 由于 MDDM、PMU、MUCO 和 MSIO 算法得到的是一组特征序列, 于是设置特征子集的个数与 MLNB 算法一致, 并设 ML- k NN 中的平滑系数 $s = 1$, 近邻个数 $k = 10$. 表 2 至表 5 列举了 6 种算法在数据集集中的 AP、RL、OE 和 HL 值.

表 2 六种算法在 11 个数据集集中的平均准确率排序¹⁾

Table 2 Average precision ranking of 6 algorithms in 11 datasets

数据集	MDDM _{spc}	MDDM _{proj}	PMU	MLNB	MUCO	MSIO
Arts	0.507 2	0.494 3	0.494 4	0.499 1	0.519 2	0.542 1
Business	0.873 6	0.873 2	0.875 4	0.871 3	0.877 0	0.876 7
Cal500	0.482 3	0.482 3	0.479 6	0.477 6	0.482 8	0.472 4
Computer	0.634 5	0.628 4	0.627 6	0.639 1	0.640 3	0.644 4
Education	0.538 9	0.542 5	0.546 5	0.547 8	0.575 3	0.565 4
Emotions	0.773 4	0.762 7	0.708 1	0.752 9	0.775 5	0.776 7
Health	0.665 4	0.650 2	0.680 2	0.688 0	0.685 7	0.697 0
Rec	0.471 7	0.470 3	0.436 5	0.479 0	0.477 5	0.514 4
Ref	0.612 6	0.610 6	0.616 9	0.623 4	0.630 1	0.638 6
Society	0.561 5	0.568 1	0.588 1	0.589 4	0.593 9	0.593 8
Yeast	0.721 3	0.721 0	0.747 3	0.735 5	0.735 0	0.748 0
平均 ²⁾	0.622 0(0)	0.618 5(0)	0.618 2(0)	0.627 6(0)	0.635 7(4)	0.642 7(7)

¹⁾ 平均准确率指标越大越好, 灰底数值表示在该指标上取得的最优结果; ²⁾ 括号内数字表示在 11 个数据集中算法获得最优值的个数

表3 六种算法在11个数据集中的排位损失排序¹⁾

Table 3 Ranking loss of 6 algorithms in 11 datasets

数据集	MDDM _{spc}	MDDM _{proj}	PMU	MLNB	MUCO	MSIO
Arts	0.152 1	0.155 5	0.152 7	0.154 2	0.150 4	0.140 6
Business	0.042 2	0.042 2	0.041 3	0.041 9	0.0403	0.041 5
Cal500	0.190 8	0.190 8	0.189 1	0.191 3	0.190 5	0.194 3
Computer	0.091 6	0.093 4	0.094 1	0.091 0	0.089 6	0.088 1
Education	0.091 4	0.092 4	0.091 1	0.092 2	0.085 6	0.088 2
Emotions	0.176 2	0.194 0	0.265 6	0.205 5	0.185 0	0.195 8
Health	0.066 3	0.069 8	0.063 8	0.064 1	0.060 8	0.059 8
Rec	0.183 8	0.185 9	0.195 5	0.187 9	0.185 7	0.174 6
Ref	0.088 8	0.088 9	0.086 8	0.088 9	0.086 5	0.082 4
Society	0.150 0	0.148 4	0.144 2	0.145 6	0.141 6	0.142 8
Yeast	0.199 0	0.204 1	0.178 6	0.187 1	0.190 9	0.180 5
平均 ²⁾	0.130 2(1)	0.133 2(0)	0.136 6(2)	0.131 8(0)	0.127 9(3)	0.126 2(5)

¹⁾ 排位损失指标越小越好, 灰底数值表示在该指标上取得的最优结果; ²⁾ 括号内的数字表示在11个数据集中算法获得最优值的个数

表4 六种算法在11个数据集中的1-错误上排序¹⁾

Table 4 One-error ranking of 6 algorithms in 11 datasets

数据集	MDDM _{spc}	MDDM _{proj}	PMU	MLNB	MUCO	MSIO
Arts	0.634 0	0.648 7	0.653 7	0.643 3	0.613 7	0.574 3
Business	0.128 7	0.128 0	0.122 7	0.131 7	0.120 7	0.121 0
Cal500	0.147 4	0.147 4	0.119 5	0.143 4	0.107 6	0.107 6
Computer	0.440 3	0.449 0	0.446 7	0.432 0	0.433 0	0.429 0
Education	0.610 0	0.597 3	0.592 0	0.582 7	0.552 0	0.564 0
Emotions	0.321 8	0.3515	0.391 1	0.376 2	0.316 8	0.302 0
Health	0.427 0	0.440 3	0.408 0	0.394 7	0.401 3	0.382 7
Rec	0.679 3	0.682 7	0.721 0	0.664 3	0.667 0	0.623 7
Ref	0.484 3	0.488 7	0.486 7	0.470 3	0.463 7	0.454 7
Society	0.495 3	0.481 3	0.459 3	0.454 0	0.453 3	0.4547
Yeast	0.259 3	0.252 7	0.233 1	0.256 0	0.252 7	0.2505
平均 ²⁾	0.420 7(0)	0.424 3(0)	0.421 3(1)	0.413 5(0)	0.398 3(4)	0.387 7(7)

¹⁾ 1-错误指标越小越好, 灰底数值表示在该指标上取得的最优结果; ²⁾ 括号内数字表示在11个数据集中算法获得最优值的个数

表5 六种算法在11个数据集中的海明损失排序¹⁾

Table 5 Hamming loss ranking of 6 algorithms in 11 datasets

数据集	MDDM _{spc}	MDDM _{proj}	PMU	MLNB	MUCO	MSIO
Arts	0.060 7	0.061 2	0.061 5	0.061 2	0.060 5	0.058 9
Business	0.027 7	0.027 7	0.027 3	0.028 3	0.027 3	0.027 1
Cal500	0.139 6	0.139 6	0.139 9	0.142 6	0.139 5	0.144 5
Computer	0.040 6	0.040 6	0.041 3	0.040 1	0.040 7	0.038 0
Education	0.042 6	0.042 2	0.040 9	0.040 5	0.040 1	0.039 6

(续表 5)

数据集	MDDM _{spc}	MDDM _{proj}	PMU	MLNB	MUCO	MSIO
Emotions	0.240 9	0.245 0	0.250 8	0.245 0	0.231 8	0.225 2
Health	0.044 1	0.045 6	0.044 6	0.041 5	0.044 4	0.042 1
Rec	0.062 0	0.061 6	0.063 3	0.061 1	0.060 8	0.059 4
Ref	0.032 2	0.031 1	0.030 6	0.029 6	0.030 9	0.029 6
Society	0.058 0	0.057 7	0.056 1	0.055 9	0.054 5	0.055 4
Yeast	0.220 9	0.224 6	0.208 9	0.208 0	0.209 0	0.203 9
平均 ³⁾	0.088 1(0)	0.088 8(0)	0.087 7(0)	0.086 7(2)	0.085 4(2)	0.084 0(8)

¹⁾ 海明损失指标越小越好, 灰底数值表示在该指标上取得的最优结果; ²⁾ 括号内数字表示在 11 个数据集中算法获得最优值的个数

由表 2 至表 5 可见:

1) MSIO 算法在 4 种评价指标中的最优数目和平均值均位列第 1, 且在 Art、Computer、Rec 和 Ref 数据集上的所有评价指标均优于其他算法。

2) 在 AP 指标上, 仅 MUCO 算法在部分数据集上占优, 但最优的 Cal500 数据集也仅提高了 2.15%; 在 RL 指标上, 6 种算法各有千秋, 其中, 在 Emotions 数据集中 MDDM_{spc} 算法占优最大, 为 11.12%;

在 OE 和 HL 指标上, 最优的分别高 7.46% 和 3.58%。

图 3 为 6 种算法在不同评价指标上的箱型图。由图 3 可见, 6 种算法在 4 种评价指标上展现的分类性能对比中, MSIO 算法在箱形图中的中位数表现明显占优。可见, MSIO 算法稳定性更好的, 且分类精度更高, 性能优于其他特征选择算法。

图 3 六种算法在不同评价指标上的箱型图

Fig. 3 Box-plot of six algorithms on different evaluation indexes

3.4 算法性能统计分析

运用统计学知识, 对在 11 个数据集上的实验结果进行显著性水平为 5% 的 Nemenyi 统计假设检验, 若在所有数据集上两个对比算法平均排序的差低于临界差 (critical difference, CD), 则认为它们无显著性差异, 否则, 认为这两个对比算法有显著性差异. 本研究设显著性水平 $\alpha = 0.05$, $q_\alpha = 2.850$ (第 k 个对应数), 算法个数 $k = 6$, 数据集个数 $N = 11$, 则

$$CD = q_\alpha \sqrt{k(k+1)/6N} = 2.274 \quad (16)$$

图 4 给出了各个算法在不同评价指标下的对比. 坐标轴上的刻度描述了各种算法在不同指标中的平均排序, 轴上数字越小, 表明算法性能越优. 不同线型相接的算法表示在性能之间无显著性差异. 由图 4 可见, MSIO 算法排名除在 RL 指标上比 MUCO 稍微逊色, 在其余指标中均明显较优.

图 4 各算法 Nemenyi 检验的性能对比

Fig. 4 Performance comparison of Nemenyi test of each algorithm

结 语

通过对不同标记在样本空间的描述程度存在一定的差异性的思考, 结合信息熵的相关知识, 提出一种不平衡标记差异性特征选择算法 MSIO, 通过不同标记下的正负标记权值修正传统的信息熵, 由于加入了标记空间的信息, 使选出的特征具有更加丰富的信息. 在多组数据集上的多个评价指标中, MSIO 算法性能都优于目前多数的特征选择算法. 但是, MSIO 算法在进行特征选择时仅考虑了特征与标记的相关性, 而未对特征空间本身进行冗余性约简, 这也是下一步的研究方向.

基金项目: 安徽省高校重点科研资助项目 (KJ2017A352); 福建省高校重点实验室开放课题资助项目 (D1801); 安徽省高校重点实验室基金资助项目 (ACAIM160102)

作者简介: 王一宾 (1970—), 安庆师范大学教授. 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

引 文: 王一宾, 吴 陈, 程玉胜, 等. 不平衡标记差异性多标记特征选择算法[J]. 深圳大学学报理工版, 2020, 37(3): 234-242.

参考文献 / References:

- [1] PAN Xiaoyong, FAN Yongxian, JIA Jue, et al. Identifying RNA-binding proteins using multi-label deep learning [J]. Science China Information Sciences, 2019, 62(1): 19103.
- [2] ROMAN-RANGEL E, MARCHAND-MAILLET S. Inductive t-SNE via deep learning to visualize multi-label images [J]. Engineering Applications of Artificial Intelligence, 2019, 81: 336-345.
- [3] CHENG Yusheng, ZHAO Dawei, ZHAN Wenfa, et al. Multi-label learning of non-equilibrium labels completion with mean shift [J]. Neurocomputing, 2018, 321: 92-102.
- [4] 刘军煜, 贾修一. 一种利用关联规则挖掘的多标记分类算法[J]. 软件学报, 2017, 28(11): 2865-2878.
LIU Junyu, JIA Xiuyi. Multi-label classification algorithm based on association rule mining [J]. Journal of Software, 2017, 28(11): 2865-2878. (in Chinese)
- [5] 何志芬, 杨 明, 刘会东. 多标记分类和标记相关性的联合学习[J]. 软件学报, 2014, 25(9): 1967-1981.
HE Zhifen, YANG Ming, LIU Huidong. Joint learning of multi-label classification and label correlations [J]. Journal of Software, 2014, 25(9): 1967-1981. (in Chinese)
- [6] 蔡亚萍, 杨 明. 一种利用局部标记相关性的多标记特征选择算法[J]. 南京大学学报自然科学版, 2016, 52(4): 693-704.

- CAI Yaping, YANG Ming. A multi-label feature selection algorithm by exploiting label correlations locally [J]. Journal of Nanjing University Natural Sciences, 2016, 52(4): 693-704. (in Chinese)
- [7] 吴磊, 张敏灵. 基于类属属性的多标记学习算法[J]. 软件学报, 2014, 25(9): 1992-2001.
- WU Lei, ZHANG Minling. Label-specific features on multi-label learning algorithm [J]. Journal of Software, 2014, 25(9): 1992-2001. (in Chinese)
- [8] 王一宾, 程玉胜, 何月, 等. 回归核极限学习机的多标记学习算法[J]. 模式识别与人工智能, 2018, 31(5): 419-430.
- WANG Yibin, CHENG Yusheng, HE Yue, et al. Multi-label learning algorithm of regression kernel extreme learning machine [J]. Pattern Recognition and Artificial Intelligence, 2018, 31(5): 419-430. (in Chinese)
- [9] 黄莉莉, 汤进, 孙登第, 等. 基于多标签 ReliefF 的特征选择算法[J]. 计算机应用, 2012, 32(10): 2888-2890.
- HUANG Lili, TANG Jin, SUN Dengdi, et al. Feature selection algorithm based on multi-label ReliefF [J]. Journal of Computer Applications, 2012, 32(10): 2888-2890. (in Chinese)
- [10] 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. (in Chinese)
- [11] 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. (in Chinese)
- [12] 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.
- [13] LIN Yaojin, HU Qinghua, LIU Jinghua, et al. Streaming feature selection for multilabel learning based on fuzzy mutual information [J]. IEEE Transactions on Fuzzy Systems, 2017, 25(6): 1491-1507.
- [14] 程玉胜, 赵大卫, 钱坤. 近邻标签空间非平衡化标签补全的多标签学习[J]. 模式识别与人工智能, 2018, 31(8): 740-749.
- CHENG Yusheng, ZHAO Dawei, QIAN Kun. Multi-label learning for non-equilibrium labels completion in neighborhood labels space [J]. Pattern Recognition and Artificial Intelligence, 2018, 31(8): 740-749. (in Chinese)
- [15] 程玉胜, 陈飞, 王一宾. 基于粗糙集的数据流多标记分布特征选择[J]. 计算机应用, 2018, 38(11): 3105-3111.
- CHENG Yusheng, CHEN Fei, WANG Yibin. Feature selection for multi-label distribution learning with streaming data based on rough set [J]. Journal of Computer Applications, 2018, 38(11): 3105-3111. (in Chinese)
- [16] 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. (in Chinese)
- [17] TSOU MAKAS G, SPYROMITROS-XIOUFIS E, VILCEK J, et al. Mulan: a java library for multi-label learning [DB/OL]. (2011-07-12). <http://mulan.sourceforge.net/datasets.html>.
- [18] LIN Yaojin, HU Xuegang, WU Xindong. Quality of information-based source assessment and selection [J]. Neurocomputing, 2014, 133: 95-102.
- [19] ZHANG Minling, ZHOU Zhihua. ML-kNN: a lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40(7): 2038-2048.
- [20] 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, 2017, 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. (in Chinese)
- [21] ZHANG Yin, ZHOU Zhihua. Multilabel dimensionality reduction via dependence maximization [J]. ACM Transactions on Knowledge Discovery from Data, 2010, 4(3): 14.
- [22] LEE J, KIM D W. Feature selection for multi-label classification using multivariate mutual information [J]. Pattern Recognition Letters, 2013, 34(3): 349-357.
- [23] ZHANG Minling, PEÑA J M, ROBLES V. Feature selection for multi-label naive Bayes classification [J]. Information Sciences, 2009, 179(19): 3218-3229.

【中文责编: 英子; 英文责编: 木柯】