

基于邻域交互增益信息的多标记流特征选择算法

陈超逸¹, 林耀进^{1,2*}, 唐 莉^{1,2}, 王晨曦^{1,2}

(1. 闽南师范大学计算机学院, 漳州, 363000; 2. 数据科学与智能应用福建省教育厅重点实验室, 漳州, 363000)

摘要: 现有的多标记特征选择一般假设特征空间是固定已知的, 然而实际应用中很多特征是需要提取过程中实时地进行筛选. 为此, 提出基于邻域交互增益信息的多标记在线流特征选择算法. 首先, 基于多标记邻域互信息和邻域交互增益信息提出在线相关性分析与在线冗余性分析两种策略来评价特征; 其次, 基于邻域交互增益信息构建了在线流多标记特征选择的目标优化函数; 最后, 在六个多标记数据集和四个评价指标上, 实验结果证明了该算法的有效性和稳定性.

关键词: 在线流特征, 多标记学习, 邻域熵, 邻域交互增益信息

中图分类号: TP391

文献标识码: A

Streaming multi-label feature selection based on neighborhood interaction gain information

Chen Chaoyi¹, Lin Yaojin^{1,2*}, Tang Li^{1,2}, Wang Chenxi^{1,2}

(1. School of Computer Science, Minnan Normal University, Zhangzhou, 363000, China; 2. Key Laboratory of Data Science and Intelligence Application, Department of Education of Fujian Province, Zhangzhou, 363000, China)

Abstract: The existing multi-label feature selection methods generally assume that the feature space is fixed and known. However, a lot of features need to be filtered in real-time during the extraction in practical application. Therefore, a streaming multi-label feature selection based on neighborhood interaction gain information is proposed. Firstly, we propose online correlation analysis and online redundancy analysis to evaluate features based on multi-label neighborhood mutual information and neighborhood interaction gain information. Secondly, based on neighborhood interaction gain information, we construct an objective optimization function for streaming multi-label feature selection. Finally, experimental results on six multi-label datasets and four criteria demonstrate the effectiveness and stability of the algorithm.

Key words: online stream features, multi-label learning, neighborhood entropy, neighborhood interaction gain

多标记学习旨在构建一个分类模型, 将相应的示例映射到多个类标记上. 多标记学习被广泛应用于现实生活中的许多领域, 如图像分类^[1]、文本分类^[2]、基因功能分类^[3]和音乐情感识别^[4]. 如一张图片的标记可能有“飞机”“山丘”“蓝天”等不同语义信息(见图1); 一篇新闻报道中有不同的

关键字, 如“C罗”“转会”“尤文图斯”; 疾病的诊断^[5]中, 每个病人可能同时有“心脏病”“高血压”“慢性胃炎”等多种疾病.

在多标记学习^[5]中, 数据特征空间的高维性易导致维数灾难, 造成分类性能的降低, 因此降低多标记数据的特征维数有显著的意义. 目前, 多



图1 多标记图片示例

Fig. 1 A picture with multi-label

标记特征降维技术主要分特征映射和特征选择两类. 多标记特征映射是将原高维特征向量映射到低维特征空间中的过程, 主要有线性判别分析 (Linear Discriminant Analysis, LDA)^[6]、依赖度最大化的多标记维数约简 (Multi-label Dimensionality reduction via Dependence Maximization, MD-DM)^[7] 和多标记语义搜索 (Multi-label informed latent semantic indexing, MLSI)^[8] 等特征映射方法. 虽然多标记特征映射降低了特征的维度, 但使特征失去了原有的物理意义, 导致映射后的特征与标记之间的因果关系难以解释. 而多标记特征选择^[9] 利用某种度量指标对原始特征进行排序或选择最优的特征子集. 多标记特征选择一般分过滤、包装、嵌入三种方法, 其中过滤式多标记特征选择方法因与分类器无关、因果解释清楚而受到广泛的关注. 目前, 许多基于不同度量指标的过滤式多标记特征选择算法被提出, 如依赖性分析^[10]、线性判别分析^[6] 及互信息^[11-12].

然而, 上述方法只能处理静态的多标记特征选择. 大量实际应用场景中多标记数据的特征通常无法一次性全部获得, 而要根据实际需求或时间顺序逐步提取相应特征. 如在无人车的自动驾驶过程中, 无人车根据实际需求自动切换传感器及传输时间顺序提取目标样本特征, 然后进行实时特征处理. 为了使到达的特征被及时处理, 许多在线流多标记特征选择算法被提出. 程玉胜等^[13] 提出动态滑动窗口加权互信息流特征选择, Lin et al 提出^[14] 基于模糊互信息的多标记流特征选择算法, Liu et al^[15] 提出基于邻域依赖度在线分析的多标记流特征选择算法 (Online Multi-label Streaming Feature Selection Based on Neighbor-

hood Rough Set, OM-NRS). 这些多标记流特征选择算法虽能有效地在流环境下选择一组较强差异能力的特征, 但也存在高计算代价、选择的特征数量多等缺点.

为了处理流环境下的多标记特征选择问题, 本文基于邻域交互增益信息提出一种多标记流特征选择算法 (Streaming Multi-label Feature Selection, SMFS). 首先利用平均间隔粒化不同标记下样本, 并定义了多标记学习下的邻域交互增益信息; 其次, 对新到特征进行在线相关分析与在线冗余分析, 基于邻域交互增益信息构建在线多标记特征选择的优化目标函数; 最后, 大量实验验证了所提算法的有效性.

本文的主要贡献: (1) 根据不同的标记, 计算样本的平均间隔, 以此进行邻域粒化. (2) 在流特征的环境下, 考虑特征与标记的相关性以及特征间的条件冗余性, 提出度量特征的有效性指标. (3) 定义了邻域交互增益信息和特征有效性指标, 并设计特征选择方法, 对已选特征子集进行约简. 实验结果证明, 本文的算法能够选出一个高质量特征子集, 有效地提高分类器的预测性能.

1 多标记学习下的邻域熵与邻域互信息

本节主要介绍多标记学习环境下的邻域熵与邻域互信息.

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:

$$(1) \Delta(x_i, x_j) \geq 0, \text{ 当且仅当 } x_i = x_j, \Delta(x_i, x_j) = 0;$$

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

$$(3) \Delta(x_i, x_k) \leq \Delta(x_i, x_j) + \Delta(x_j, x_k)$$

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:

$$\Delta_p(x_i, x_j) = \left[\sum_{l=1}^N (x_{il} - x_{jl})^p \right]^{\frac{1}{p}} \quad (1)$$

当 $P=1$ 时, Δ 表示为曼哈顿距离; 当 $P=2$ 时, Δ 表示为欧式距离.

定义 1^[16] 给定多标记决策系统 $\langle U, F, L \rangle$, $U = \{x_1, x_2, \dots, x_n\}$ 表示样本集合, $F = \{f_1, f_2, \dots, f_l\}$

是描述样本特征集, $L = \{l_1, l_2, \dots, l_m\}$ 是样本的标记集合. 则样本 x 在标记 l_i 下的间隔为:

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

其中, $NM_{l_i}(x)$ 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.

定义 2^[16] 给定多标记决策系统 $\langle U, F, L \rangle$, $U = \{x_1, x_2, \dots, x_n\}$ 表示样本集合, $F = \{f_1, f_2, \dots, f_t\}$ 是描述样本特征的集合, $L = \{l_1, l_2, \dots, l_m\}$ 是样本的标记集合. 那么, 样本 x 在标记空间 L 下的平均间隔为:

$$m^{neu}(x) = \frac{1}{m} \sum_{i=1}^m m_{l_i}(x) \quad (3)$$

定义 3^[16] 给定多标记决策系统 $\langle U, F, L \rangle$, U 表示样本集合, L 表示标记集合. 对于 $\forall x \in U, \forall l \in L$, 样本 x 在标记空间 L 的平均间隔 $m^{neu}(x) \geq 0$, 则 x 的邻域为:

$$\delta^{neu}(x) = \{y | \Delta(x, y) \leq m^{neu}(x), y \in U\} \quad (4)$$

若 $m^{neu}(x) \leq 0$, 则令 $m^{neu}(x) = 0$.

定义 4^[16] 给定多标记决策系统 $\langle U, F, L \rangle$, 其中包括样本集 $U = \{x_1, x_2, \dots, x_n\}$ 、特征空间 $f \subseteq F$ 以及一个标记集 L . 样本 x_i 在特征下 f 的邻域为 $\delta_f^{neu}(x_i)$, 那么平均间隔下的不确定性定义为:

$$NH^{\delta^{neu}}(f) = -\frac{1}{n} \sum_{i=1}^n \lg \frac{\|\delta_f^{neu}(x_i)\|}{n} \quad (5)$$

定义 5^[16] 假设有两个多标记学习下的特征子集 $r, f \subseteq F$, 样本 x_i 在子空间 $r \cup f$ 上的邻域可以表示为 $\delta_{r \cup f}^{neu}(x_i)$, 则在多标记学习中平均间隔下的联合邻域熵定义为:

$$NH^{\delta^{neu}}(f) = -\frac{1}{n} \sum_{i=1}^n \lg \frac{\|\delta_{r \cup f}^{neu}(x_i)\|}{n} \quad (6)$$

定义 6^[16] 假设有两个多标记学习下的特征子集 $r, f \subseteq F$, 则在多标记学习中平均间隔下的条件邻域熵定义为:

$$NH^{\delta^{neu}}(r|f) = -\frac{1}{n} \sum_{i=1}^n \lg \frac{\|\delta_{r \cup f}^{neu}(x_i)\|}{\|\delta_f^{neu}(x_i)\|} \quad (7)$$

定义 7^[16] 假设有两个多标记学习下的特征

子集 $r, f \subseteq F$, 则在多标记学习中平均间隔下的邻域互信息定义为:

$$NMI^{\delta^{neu}}(r; f) = -\frac{1}{n} \sum_{i=1}^n \lg \frac{\|\delta_r^{neu}(x_i)\| \cdot \|\delta_f^{neu}(x_i)\|}{n \|\delta_{r \cup f}^{neu}(x_i)\|} \quad (8)$$

定义 8^[16] 假设有一个多标记学习下的特征子集 $r \subseteq F$ 和多标记集 $L = \{l_1, l_2, \dots, l_m\}$, 则在多标记学习中平均间隔下 r 与 L 的邻域互信息定义为:

$$NMI^{\delta^{neu}}(r; L) = \sum_{i=1}^m NMI^{\delta^{neu}}(r; l_i) \quad (9)$$

定义 9 假设一个多标记学习下的特征子集 $f \subseteq F$, 一个特征 r , 一个多标记集 $L = \{l_1, l_2, \dots, l_m\}$, 综合 Kwak and Choi^[17] 所述, 在多标记学习中平均间隔下, 多标记邻域交互增益信息可近似为:

$$NMI^{\delta^{neu}}(L; r; f) = \sum_{f_s \in f} \frac{NMI^{\delta^{neu}}(L; f_s)}{NH^{\delta^{neu}}(f_s)} NMI^{\delta^{neu}}(r; f_s) \quad (10)$$

邻域交互增益信息能够反映两个不同特征在标记空间 L 下所提供的信息量. 当它的值为正数时, 说明两个特征放在一起能够提供信息, 但无法独立提供信息; 当它的值为负数时, 说明两个特征提供了相同的信息; 当它为 0 时, 说明两个特征相互独立.

2 基于邻域交互增益信息的多标记流特征选择算法

为了从流环境下的多标记学习任务中进行特征选择, 定义了多标记流特征选择的优化目标函数.

定义 10 给定一个流特征多标记决策系统 $\langle U, F_t, L \rangle$, 其中 $F_t = \{f_1, f_2, \dots, f_t\}$, f_t 指在 t 时刻到来的特征, S_{t-1} 指在 t 时刻之前已选的特征子集. 则多标记流特征选择的优化目标函数可定义为:

$$S_t = \underset{X \subseteq \{S_{t-1} \cup f_t\}}{\operatorname{argmax}} (NMI^{\delta^{neu}}(X; L)) \quad (11)$$

为求解式 (11), 可分两步进行在线特征分析: 第一步, 计算新到特征与多标记空间的相关性, 若相关, 则表示新到特征可以添加到已选特征; 第二

步,通过定义特征的有效性进行新特征与已选特征间的冗余性分析,得到更紧凑的特征子集.

下面具体介绍在线相关性分析与在线冗余性分析.

2.1 在线相关性分析 给定一个流特征多标记决策系统 $\langle U, F_t, L \rangle$, 其中 $F_t = \{f_1, f_2, \dots, f_t\}$, f_t 是指在 t 时刻到来的特征, S_{t-1} 指在 t 时刻之前已选的特征集合. 根据式(9)可以得到 f_t 与多标记空间 L 的相关性 $\gamma_{f_t} = NMI^{\delta^{neu}}(f_t; L)$. 为了尽可能暂时保留特征的多样性, 设置相关性的阈值为 0. 如果 $\gamma_{f_t} > 0$, 则将 f_t 加入到已选特征子集 S_t 中, 否则舍去 f_t . 这样既排除了不相关的特征, 又保证了特征多样性, 为下一阶段的冗余性分析提供了更多特征组合的可能性.

2.2 在线冗余性分析 为了获得一个更加紧凑的特征子集, 候选特征经过相关性分析后, 还需计算与已选特征的冗余性.

定义 11 给定一个流特征多标记决策系统 $\langle U, F_t, L \rangle$, 其中 $F_t = \{f_1, f_2, \dots, f_t\}$, f_t 指在 t 时刻到来的特征, S_{t-1} 指在 t 时刻之前已选的特征集合. 若 f_t 与多标记空间 L 相关, 则 $S_t = \{S_{t-1} \cup f_t\}$. 可定义特征 $g \in S_t = \{S_{t-1} \cup f_t\}$ 的有效性为:

$$\lambda_{g, S_t} = NMI^{\delta^{neu}}(g; L) - \frac{1}{|S_t|} NMI^{\delta^{neu}}(L; g; \{S_t \setminus g\}), \quad \forall g \in S_t \quad (12)$$

根据式(9)可得:

$$\lambda_{g, S_t} = NMI^{\delta^{neu}}(g; L) - \frac{1}{|S_t|} \sum_{f_s \in S_t \setminus g} \frac{NMI^{\delta^{neu}}(L; f_s)}{NH^{\delta^{neu}}(f_s)} NMI^{\delta^{neu}}(g; f_s) \quad (13)$$

在线冗余性分析阶段, 对 t 时刻已选的特征子集 S_t , 根据式(13)计算每个已选特征的有效性 λ_{g, S_t} , 得到平均值 $\bar{\lambda}$ 和最小值 $\min(\lambda_{g, S_t})$ 及其特征.

若此时 $\frac{\bar{\lambda} - \min(\lambda_{g, S_t})}{\bar{\lambda}} > \beta$, 其中 β 是给定的一个阈值, 则将该有效性最低的特征剔除并更新 S_t 和 λ_{g, S_t} , 从而逐步提升平均有效性.

2.3 基于邻域交互增益信息的多标记流特征选择算法 根据式(11), 可设计基于邻域交互增益信息的多标记流特征选择算法(详见算法 1).

算法 1 基于邻域交互增益信息的多标记流特征选择算法 (Streaming Multi-label Feature Selection based on neighborhood interaction gain, SMFS)

输入: 流特征的多标记决策系统 $\langle U, F_t, L \rangle$

输出: S_t : 已选的特征子集.

f_t : t 时刻到达的新特征

S_t : t 时刻已选特征子集. S_0 为空集.

(1) 循环等待新的特征到来

(2) 当 t 时刻, f_t 到来

(3) 计算 $\gamma_{f_t} = NMI(L; f_t)$

(4) if $\gamma_{f_t} > 0$, then $S_t = S_{t-1} \cup f_t$;

(5) else, 更新所有的 λ_{g, S_t}

(6) if S_t 中特征个数大于 2, 对每个 $g \in S_t$, 根据式(13)计算 λ_{g, S_t}

(7) while $\|S_t\| > 2$

(8) 寻找最小的 λ_{g, S_t} 的特征, 计算 λ 的平均值 $\bar{\lambda}$.

(9) if $\frac{\bar{\lambda} - \min(\lambda_{g, S_t})}{\bar{\lambda}} > \beta$

(10) 将 g 从 S_t 中剔除, 并更新所有的 λ_{g, S_t}

(11) else 退出循环

(12) end while.

(13) 直到没有新的特征到来

(14) return S_t

算法第一步计算平均间隔, 时间复杂度为 $O(|U| \cdot |L| + |U| \cdot |U|)$. 第二步计算特征和多标记空间的相关性, 时间复杂度为 $O(|U| \cdot |U| \cdot |L|)$. 第三步对特征的冗余性分析, 时间复杂度为 $O(|S_t| \lg |S_t|)$. 因此 SMFS 算法的时间复杂度为 $O(|U| \cdot |L| + |U| \cdot |U| + |U| \cdot |U| \cdot |L| + |S_t| \lg |S_t|)$.

3 实验设计与结果比较

3.1 实验数据 表 1 展现了实验数据集的描述情况. 下文给出的实验结果的表格中, 用括号中的字母代表数据集.

3.2 评价指标 假设 d 维的示例空间 $X = R^{m \times d}$ 和拥有 M 个标记的标记空间 $L = \{-1, +1\}^M$. 给定多标记训练集 $D = \{(x_i, Y_i) | 1 \leq i \leq n\}$ 和多

表 1 多标记数据集的描述

Table 1 Descriptions of multi-label datasets

数据集	样本数	特征数	类别数	训练 样本数	测试 样本数
Arts(A)	5000	462	26	2000	3000
Birds(B)	645	260	19	322	323
Business(C)	5000	438	30	2000	3000
Education(D)	5000	550	33	2000	3000
Emotions(E)	593	72	6	391	202
Yeast(F)	2417	103	14	1499	918

Average Precision (AP):

$$AP = \frac{1}{m} \sum_{i=1}^m \frac{1}{|Y_i|} \sum_{y \in Y_i} \frac{|\{y' \in Y_i : \text{rank}_f(x_i, y') \leq \text{rank}_f(x_i, y)\}|}{\text{rank}_f(x_i, y)} \quad (14)$$

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.

Ranking Loss (RL):

$$RL = \frac{1}{m} \frac{1}{|Y_i| |\overline{Y_i}|} \left| \left\{ (y', y'') \mid f(x_i, y') \leq f(x_i, y''), (y', y'') \in Y_i \times \overline{Y_i} \right\} \right| \quad (15)$$

其中, $\overline{Y_i}$ 是集合 Y_i 的补集. 该指标统计了样本的类标记的排序序列中,出现排序错误的情况.该指标越小则系统性能越好.

Hamming Loss (HL):

$$HL = \frac{1}{m} \sum_{i=1}^m \frac{|Y_i' \oplus Y_i|}{M} \quad (16)$$

其中, \oplus 是异或运算. HL 评估误分类的情况的比例,取值越小则算法性能系统性能越好.

MicroF1 (Mi-F1):

$$\text{Mi-F1} = \frac{2 \times \sum_{i=1}^m \|Y_i' \cap Y_i\|_1}{\sum_{i=1}^m \|Y_i\|_1 + \sum_{i=1}^m \|Y_i'\|_1} \quad (17)$$

Mi-F1 将统计结果相加,再求得分类性能.该指标越大分类效果越好.

3.3 实验设置 为了有效地评估所提算法,选择五个不同的算法进行对比:MLNB (Feature selection for multi-label naive Bayes classification)^[18], MDDM (Multi-Label Dimensionality Reduction via Dependence Maximization), 根据算法投影方式分为 MDDM_{spc} 和 MDDM_{proj}, PMU (Feature Selection for Multi-label Classification Using Multivariate Mutual Information)^[19] 和 RF-ML (Re-

标记测试集 $Z = \{(x_i, Y_i) \mid 1 \leq i \leq m\}$, 其中 $x_i \in X$ 是 d 维的特征向量, $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$, $Y_i \in L$ 是正确的标记集. 多标记的学习任务从训练集中学到一个预测函数 $f: x \rightarrow y$, 根据预测函数得到 $Y_i' \in L$. 而 $\text{rank}_f(\cdot, \cdot)$ 是 f 的对应的排序函数.

在多标记学习中,为了衡量特征选择效果的优劣,选取 Average Precision (AP), Ranking Loss (RL), Hamming Loss (HL) 和 MicroF1 (Mi-F1) 作为评价算法性能的指标.

lief for Multi-Label Feature Selection)^[20]. SMFS 算法中 β 设置为 0.5. 由于 MDDM_{spc}, MDDM_{proj}, PMU 和 RF-ML 将得到特征排序,因此选取前 k 个 (SMFS 算法得到的特征子集个数) 特征作为特征子集. 最后用 ML-KNN (Multi-Label k -Nearest Neighbor) 算法作为多标记分类器.

表 2 至表 5 展现了四种多标记评价指标下,不同算法的实验结果. 表中的“ \uparrow ”表示表中该指标的取值越高越好,“ \downarrow ”则表示该指标的取值越小越好. 表中最后一行是每个算法在所有数据集上得到的平均值,黑体字表示该算法在当前数据集上的效果最优.

从表 2 至表 5 可以看出:

(1) 总体来看, SMFS 在四个数据集上的各个指标的平均性能都排在第一.

(2) AP 和 Mi-F1, SMFS 在一半以上的数据集上性能最优,在其他的数据集上也达到次优.

(3) SMFS 对于 Arts, Business 和 Emotions 数据集上的分类性能的所有指标都是最优的,而在其他数据集上的分类性能与最优值相差不大.

(4) 除了 SMFS 算法,其余五个算法均是处

表 2 SMFS和其他算法的 $AP(\uparrow)$ 指标的比较Table 2 $AP(\uparrow)$ of SMFS and other algorithms

数据集	MLNB	MDDM- spc	MDDM- proj	PMU	RF-ML	SMFS
A	0.4991	0.4735	0.4669	0.4917	0.4834	0.5319
B	0.5052	0.4818	0.4564	0.5082	0.5263	0.5199
C	0.8713	0.8639	0.8633	0.8698	0.8742	0.8762
D	0.5478	0.4441	0.4824	0.4798	0.5114	0.5557
E	0.7529	0.7772	0.7683	0.7399	0.7566	0.7871
F	0.7355	0.7488	0.7490	0.7488	0.7476	0.7533
平均值	0.6520	0.6316	0.6311	0.6397	0.6499	0.6707

表 3 SMFS和其他算法的 $RL(\downarrow)$ 指标的比较Table 3 $RL(\downarrow)$ of SMFS and other algorithms

数据集	MLNB	MDDM- spc	MDDM- proj	PMU	RF-ML	SMFS
A	0.1542	0.1631	0.1662	0.1584	0.1576	0.1418
B	0.2237	0.2441	0.2608	0.2042	0.2093	0.2160
C	0.0419	0.0456	0.0465	0.0439	0.0427	0.0410
D	0.0922	0.1138	0.1089	0.1099	0.1016	0.0923
E	0.2055	0.1825	0.1904	0.2301	0.2040	0.1749
F	0.1871	0.1797	0.1768	0.1774	0.1781	0.1768
平均值	0.1508	0.1548	0.1583	0.1540	0.1489	0.1405

表 4 SMFS和其他算法的 $HL(\downarrow)$ 指标的比较Table 4 $HL(\downarrow)$ of SMFS and other algorithms

数据集	MLNB	MDDM- spc	MDDM- proj	PMU	RF-ML	SMFS
A	0.0612	0.0621	0.0622	0.0607	0.0614	0.0593
B	0.0494	0.0521	0.0554	0.0486	0.0486	0.0500
C	0.0283	0.0286	0.0286	0.0280	0.0278	0.0274
D	0.0405	0.0446	0.0443	0.0442	0.0427	0.0409
E	0.2450	0.2153	0.2417	0.2475	0.2318	0.2137
F	0.2080	0.2010	0.2010	0.2025	0.2047	0.1983
平均值	0.1054	0.1006	0.1055	0.1053	0.1028	0.0983

理静态环境下的多标记特征选择算法,然而 SMFS 在无法事先获取整个特征空间的条件下,仍然有着优异的分类性能。

为了更好地观察算法性能与所选特征数目的关系,图 2 至图 5 展示了 SMFS 和四个对比算法(MDDM_{spc}, MDDM_{proj}, PMU, RF-ML)在不同数据集上的分类性能变化。图中, X 轴代表所选

表 5 SMFS和其他算法的 $Mi-F1(\uparrow)$ 指标的比较Table 5 $Mi-F1(\uparrow)$ of SMFS and other algorithms

数据集	MLNB	MDDM- spc	MDDM- proj	PMU	RF-ML	SMFS
A	0.1093	0.0565	0.0471	0.1279	0.0925	0.2345
B	0.1653	0.1489	0.0761	0.2359	0.1235	0.1769
C	0.6792	0.6679	0.6704	0.6798	0.6813	0.6927
D	0.2070	0.0041	0.0041	0.0023	0.0856	0.2314
E	0.5811	0.6319	0.5867	0.5690	0.5849	0.6597
F	0.6045	0.6208	0.6334	0.6171	0.6163	0.6232
平均值	0.3911	0.3550	0.3363	0.3720	0.3640	0.4364

特征的数量, Y 轴代表所选特征对应的评价指标的结果。根据图 2 至图 5 可以得出:

(1)随着所选特征的增加,评价指标并不是单调增加或者单调减少,说明有的特征对分类性能的提高有负面影响。

(2)在 Birds 和 Yeast 数据集上,随着 SMFS 筛选的特征数目增加,性能提升得较慢,而在其余的四个数据集上,分类性能则迅速地提升。

(3)总体来看,SMFS 的分类性能优于其他四个算法(注:在 Education 数据集上,由于 SMFS 选择的特征个数为 15,而 MDDM_{proj}和 MDDM_{spc}特征排序结果的前 15 位相同,因此关于 Education 的图中蓝色折线被黄色折线所覆盖)。

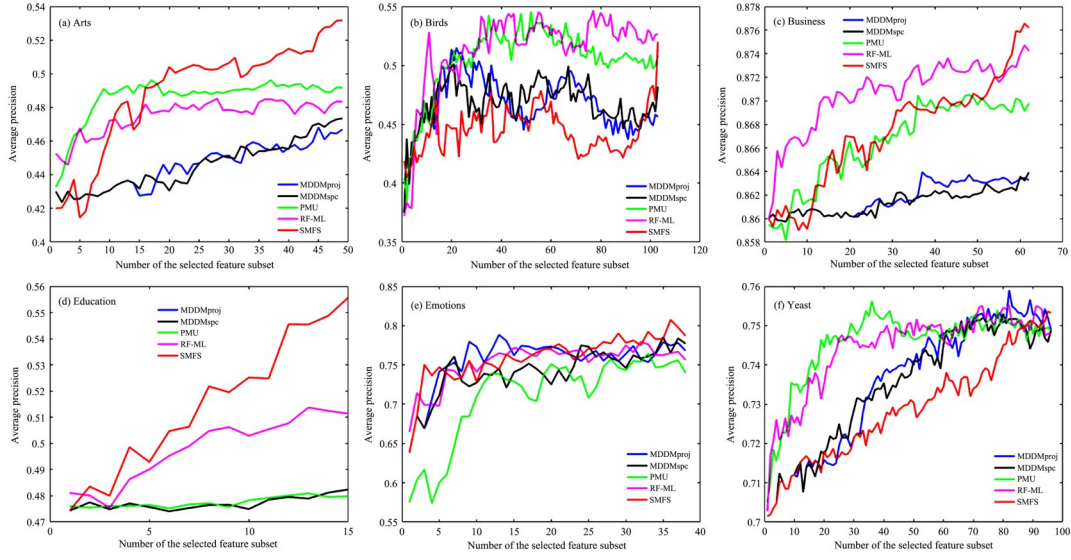
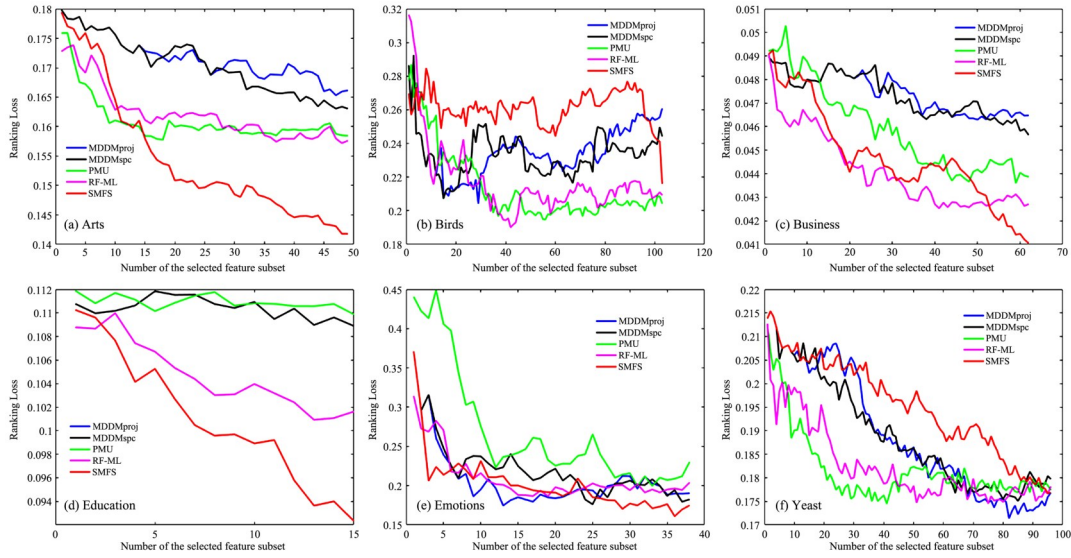
由于 SMFS 是处理流特征环境下的算法,因此无法获得整个特征空间时分类性能会受到未知影响。若先流入模型的特征有效性较高,会导致后续的特征越来越难以保留。故在不同的数据集上,算法分类性能的提升速度会有差异。

为了更直观地对比 SMFS 和五个对比算法之间分类性能的差异,采用 Friedman^[21]测试和 Bonferroni-Dunn^[22]测试。

首先进行 Friedman 测试:给定 k 个算法和 N 个数据集进行比较, r_i^j 是第 j 个算法在第 i 个数据集上的序值,测试结果见表 6。第 i 个数据集的平均序值为:

$$R_i = \frac{1}{N} \sum_{j=1}^N r_i^j$$

假设所有算法的性能都相同的情况下,通常使用变量 F_F 来进行统计比较:

图 2 SMFS和四个对比算法的 $AP(\uparrow)$ 在六个数据集上的变化Fig.2 $AP(\uparrow)$ of SMFS and other four algorithms on six datasts图 3 SMFS和四个对比算法的 $RL(\downarrow)$ 在六个数据集上的变化Fig.3 $RL(\downarrow)$ of SMFS and other four algorithms on six datasts

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2} \quad (18)$$

$$\chi_F^2 = \frac{12N}{k(k+1)} \left(\sum_{i=1}^k R_i^2 - \frac{k(k+1)^2}{4} \right)$$

从表 6 可以看出 F_F 大于显著性水平 $\alpha=0.1$ 时的临界值, 因此拒绝“所有算法的性能相同”这个假设.

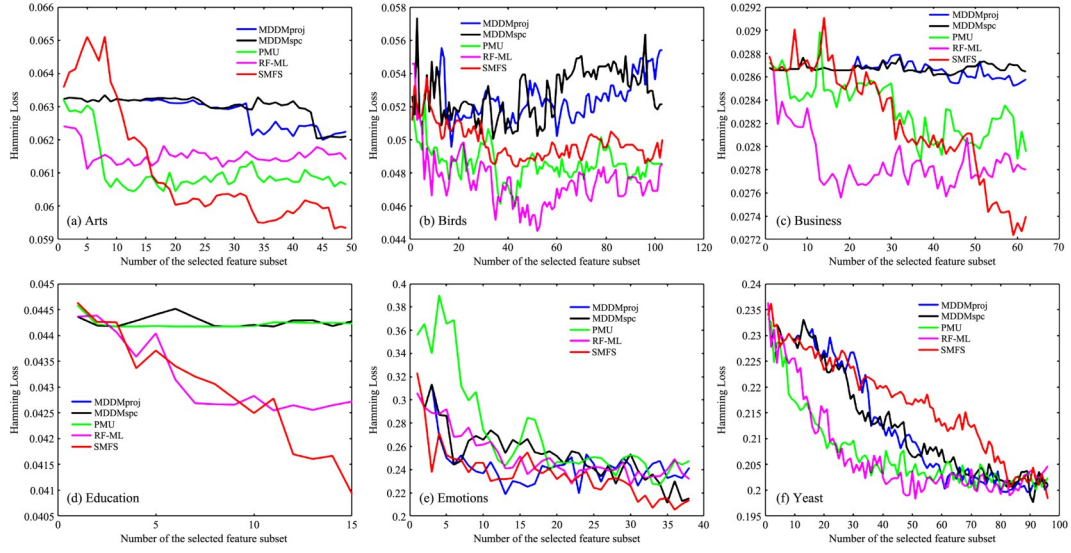
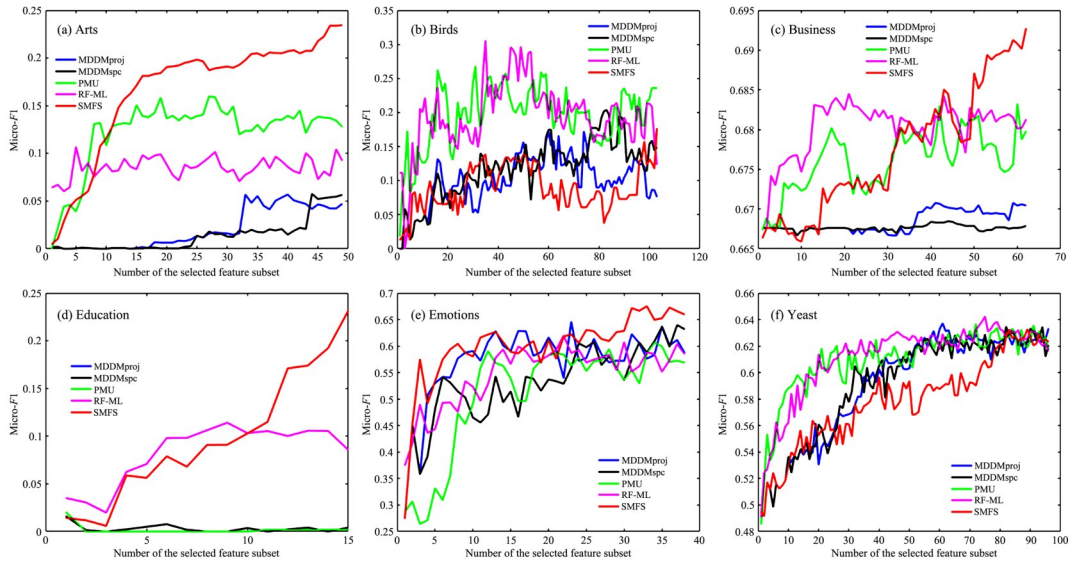
进一步, 通过 Bonferroni-Dunn 测试来准确比

较不同算法的性能差异(图 6). 该测试计算出平均值序差别的临界值域 CD:

$$CD_\alpha = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \quad (19)$$

在显著性水平 $\alpha=0.1$ 下, 有 $q_\alpha = 2.326$, 因此可以计算出 $CD = 2.5124 (k=6, N=6)$.

图 6 根据算法的平均值序进行绘制, 排名高的算法在右边. 在不同的评估指标下, 任意一个

图 4 SMFS和四个对比算法的 $HL(\downarrow)$ 在六个数据集上的变化Fig.4 $HL(\downarrow)$ of SMFS and other four algorithms on six datasets图 5 SMFS和四个对比算法的 $Mi-F1(\uparrow)$ 在六个数据集上的变化Fig.5 $Mi-F1(\uparrow)$ of SMFS and other four algorithms on six datasets表 6 不同指标下的Friedman统计 $F_F(k=6, N=6)$ Table 6 Friedman statistics $F_F(k=6, N=6)$ on different evaluation measures

评价指标	F_F	临界值($\alpha=0.1$)
AP	3.9821	2.0922
RL	2.2914	
HL	2.5911	
Mi-F1	2.4687	

在 SMFS 算法 CD 值域内的算法, 都没有显著性差异. 而在 CD 值域外的算法与 SMFS 在分类性能上有显著区别. 根据图 6 可以得到:

(1) SMFS 的 RL , HL 和 $Mi-F1$ 指标明显优于 MDDMsc 和 MDDMproj.

(2) SMFS 的 AP 指标与 PMU, MDDMsc 和 MDDMproj 有显著性差异.

图 7 展现了在各个多标记指标下的不同算法

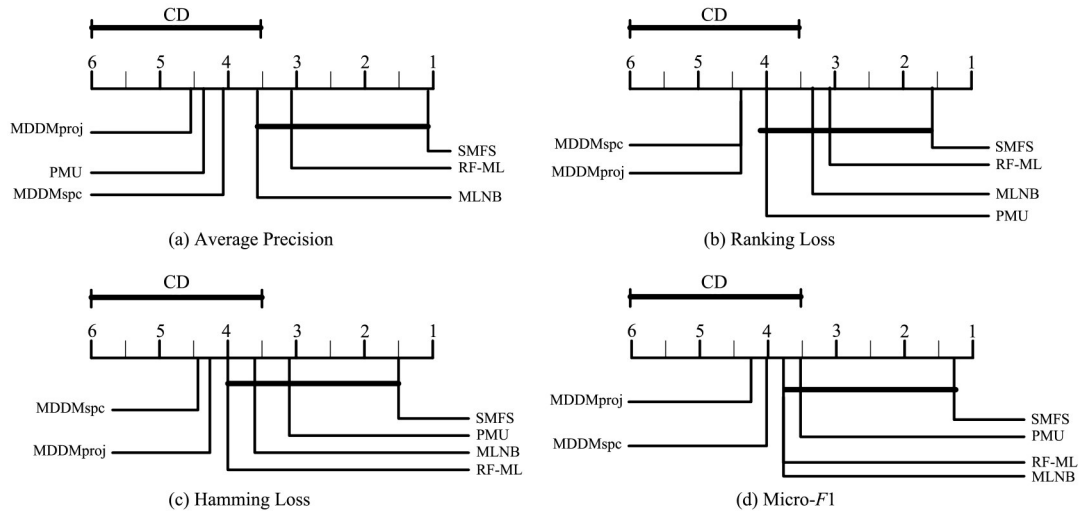


图 6 通过 Bonferroni-Dunn 测试比较 SMFS 与其他算法的性能差异

Fig. 6 Performance of SMFS and other algorithms tested by Bonferroni-Dunn

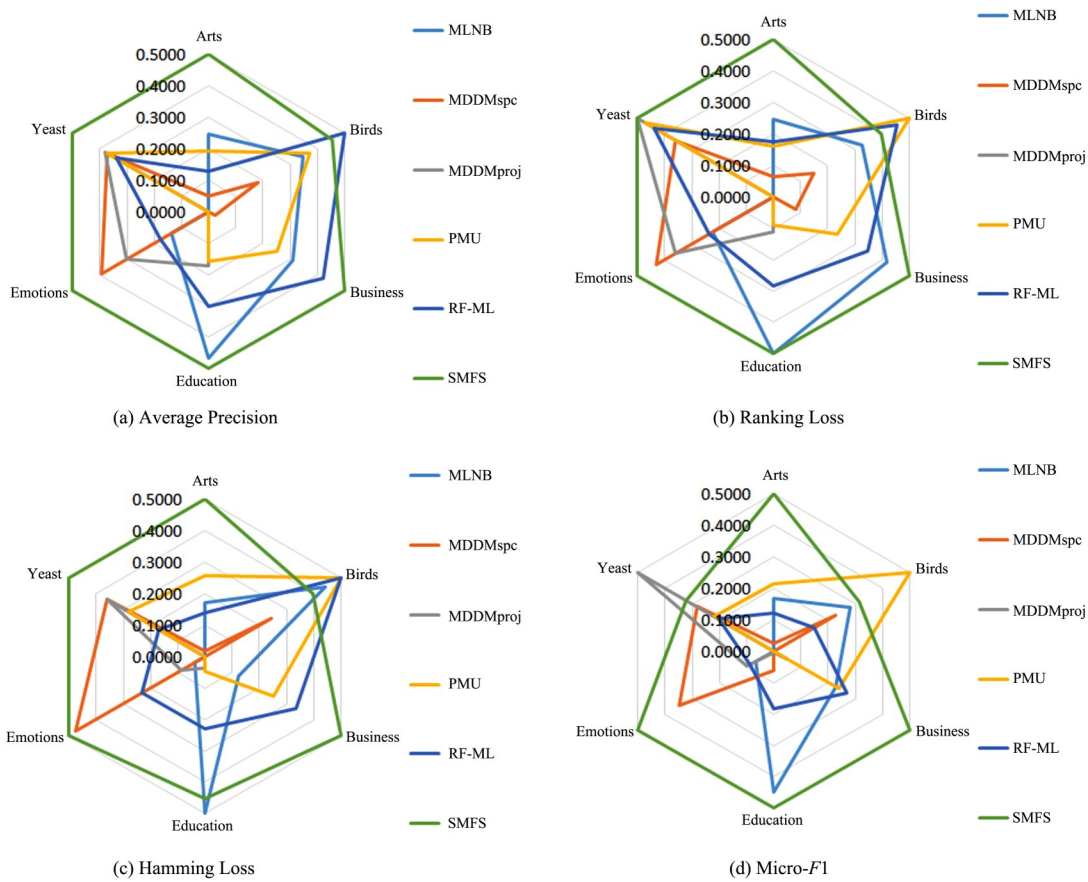


图 7 蜘蛛网图展示的 SMFS 算法在六个多标记数据集下不同指标的稳定性

Fig. 7 Spider web diagram of the stability index values of SMFS obtained on six multi-label datasets with different evaluation metrics

的稳定性. 其中,绿色代表提出的 SMFS 算法. 根据图7可以观察到:

(1)在 AP , RL 和 HL 指标下, SMFS 的形状接近正六边形,说明 SMFS 得到的解更加优异.

(2)在 $Mi-F1$ 指标下, $Yeast$ 和 $Birds$ 数据集上的结果较差. 但在其他数据集上效果最佳.

(3)在各个指标下, SMFS 在至少四个数据集上拥有最优的性能.

(4)在所有指标下, SMFS 的覆盖面积远大于其他算法,说明 SMFS 能够获得更稳定的解.

根据上述实验的结果,说明 SMFS 算法的稳定性远高于其他算法.

3.4 与 OM-NRS 比较 为了更好地评估 SMFS 算法在流特征环境下的分类性能,将它与近期提出的多标记流特征选择算法 OM-NRS 进行比较. 实验结果如表7所示,表中最后一行是六个数据集在各自算法下的平均值,黑体表示性能较优的结果.

表7 SMFS与 OM-NRS的 AP , RL 和 HL 的比较

Table 7 AP , RL and HL of SMFS and OM-NRS

	AP		RL		HL	
	SMFS	OM-NRS	SMFS	OM-NRS	SMFS	OM-NRS
A	0.5319	0.5217	0.1419	0.1440	0.0593	0.0606
B	0.5199	0.4842	0.2160	0.2190	0.0500	0.0518
C	0.8762	0.8760	0.0410	0.0411	0.0274	0.0274
D	0.5557	0.5398	0.0923	0.0912	0.0409	0.0408
E	0.7871	0.7608	0.1749	0.1765	0.2137	0.2104
F	0.7533	0.7545	0.1768	0.1732	0.1983	0.2021
平均值	0.6707	0.6562	0.1405	0.1408	0.0983	0.0989

从表7可以看出:

(1) SMFS 在过半的数据集上表现优于 OM-NRS. 同时,在其他数据集上的性能相差不大.

(2)从六个数据集的平均性能来看, SMFS 在三个指标下都优于 OM-NRS. 对比这两个流特征选择算法, OM-NRS 采取贪心的思想进行特征选择,意味着它无法摒弃已选择的特征,也就容易受制于前面选择的特征. 而 SMFS 在冗余性分析阶段结合已选特征子集,对所有已选特征重新评估. 故 SMFS 更易找到最优的特征子集,提高分类性能.

4 总 结

在邻域互信息的基础上,对多标记样本进行邻域粒化,提出了邻域交互增益信息. 并根据显示应用需要,考虑动态流特征的场景,提出了基于邻域交互增益信息的多标记流特征选择算法. 通过四种不同的多标记评价指标,在六个多标记数据集下进行实验. 结果表明,提出的 SMFS 算法优于其他五个多标记特征选择算法.

参 考 文 献

- [1] Boutell M R, Luo J B, Shen X P, et al. Learning multi-label scene classification. *Pattern Recognition*, 2004, 37(9):1757—1771.
- [2] Lewis D D, Yang Y M, Rose T G, et al. RCV1: a new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 2004, 5(2):361—397.
- [3] Elisseeff A, Weston J. A kernel method for multi-labelled classification//*Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. Cambridge, MA, USA: MIT Press, 2001.
- [4] Trohidis K, Tsoumakas G, Kalliris G, et al. Multi-label classification of music by emotion. *EURASIP Journal on Audio, Speech, and Music Processing*, 2011, 2011(1):4.
- [5] 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. *计算机研究与发展*, 2015, 52(1):56—65. (Duan J, Hu Q H, Zhang L J, et al. Feature selection for multi-label classification based on neighborhood rough set. *Journal of Computer Research and Development*, 2015, 52(1):56—65.)
- [6] Hotelling H. Relations between two sets of variates. *Biometrika*, 1936, 28(3—4):321—377.
- [7] Zhang Y, Zhou Z H. Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data*, 2010, 4(3):14.
- [8] Yu K, Yu S P, Tresp V. Multi-label informed latent semantic indexing//*Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Salvador, Brazil: ACM, 2005:258—265.

- [9] 许行,张凯,王文剑. 一种小样本数据的特征选择方法. 计算机研究与发展, 2018, 55(10): 2321—2330. (Xu X, Zhang K, Wang W J. A feature selection method for small samples. Journal of Computer Research and Development, 2018, 55(10): 2321—2330.)
- [10] Zhang L J, Hu Q H, Duan J, et al. Multi-label feature selection with fuzzy rough sets//International Conference on Rough Sets and Knowledge Technology. Springer Berlin Heidelberg, 2014: 121—128.
- [11] Lin Y J, Hu Q H, Liu J H, et al. Multi-label feature selection based on neighborhood mutual information. Applied Soft Computing, 2016, 38: 244—256.
- [12] Hu L, Gao W F, Zhao K, et al. Feature selection considering two types of feature relevancy and feature interdependency. Expert Systems with Applications, 2018, 93: 423—434.
- [13] 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. 南京大学学报(自然科学), 2018, 54(5): 974—985. (Cheng Y S, Li Y, Wang Y B, et al. Streaming feature selection with weighted fuzzy mutual information based on dynamic sliding window. Journal of Nanjing University (Natural Science), 2018, 54(5): 974—985.)
- [14] Lin Y J, Hu Q H, Liu J H, et al. Streaming feature selection for multilabel learning based on fuzzy mutual information. IEEE Transactions on Fuzzy Systems, 2017, 25(6): 1491—1507.
- [15] Liu J H, Lin Y J, Li Y W, et al. Online multi-label streaming feature selection based on neighborhood rough set. Pattern Recognition, 2018, 84: 273—287.
- [16] Lin Y J, Hu Q H, Liu J H, et al. Multi-label feature selection based on max - dependency and min - redundancy. Neurocomputing, 2015, 168: 92—103.
- [17] Kwak N, Choi C H. Input feature selection for classification problems. IEEE Transactions on Neural Networks, 2002, 13(1): 143—159.
- [18] Zhang M L, Peña J M, Robles V. Feature selection for multi-label naive Bayes classification. Information Sciences, 2009, 179(19): 3218—3229.
- [19] Lee J, Kim D W. Feature selection for multi-label classification using multivariate mutual information. Pattern Recognition Letters, 2013, 34(3): 349—357.
- [20] Spolaôr N, Cherman E A, Monard M C, et al. ReliefF for multi - label feature selection//2013 Brazilian Conference on Intelligent Systems (BRACIS). Fortaleza, Brazil: IEEE, 2013: 6—11.
- [21] Friedman M. A comparison of alternative tests of significance for the problem of m rankings. The Annals of Mathematical Statistics, 1940, 11(1): 86—92.
- [22] Dunn O J. Multiple comparisons among means. Journal of the American Statistical Association, 1961, 56(293): 52—64.

(责任编辑 杨可盛)