

基于邻域粗糙集和 Relief 的弱标记特征选择方法

孙林, 黄苗苗, 徐久成

引用本文

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 Lin, HUANG Miao-miao, XU Jiu-cheng. Weak Label Feature Selection Method Based on Neighborhood Rough Sets and Relief[J]. Computer Science, 2022, 49(4): 152-160.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

基于灰狼优化算法的信用评估样本均衡化与特征选择同步处理

Application of Gray Wolf Optimization Algorithm on Synchronous Processing of Sample Equalization and Feature Selection in Credit Evaluation

计算机科学, 2022, 49(4): 134-139. <https://doi.org/10.11896/jsjcx.210300075>

鲁棒联合稀疏不相关回归

Robust Joint Sparse Uncorrelated Regression

计算机科学, 2022, 49(2): 191-197. <https://doi.org/10.11896/jsjcx.210300034>

基于特征优化的 SAR 图像水华识别方法

Algal Bloom Discrimination Method Using SAR Image Based on Feature Optimization Algorithm

计算机科学, 2021, 48(9): 194-199. <https://doi.org/10.11896/jsjcx.200800142>

基于核密度估计的轻量级物联网异常流量检测方法

Kernel Density Estimation-based Lightweight IoT Anomaly Traffic Detection Method

计算机科学, 2021, 48(9): 337-344. <https://doi.org/10.11896/jsjcx.200600108>

基于自编码器和流形正则的结构保持无监督特征选择

Structure Preserving Unsupervised Feature Selection Based on Autoencoder and Manifold Regularization

计算机科学, 2021, 48(8): 53-59. <https://doi.org/10.11896/jsjcx.200700211>

基于邻域粗糙集和 Relief 的弱标记特征选择方法

孙 林^{1,2} 黄苗苗^{1,3} 徐久成^{1,2}

1 河南师范大学计算机与信息工程学院 河南 新乡 453007

2 教育人工智能与个性化学习河南省重点实验室 河南 新乡 453007

3 东北大学计算机科学与工程学院 沈阳 110819

摘 要 在多标记学习与分类中,现有邻域粗糙集特征选择算法若将样本的分类间隔作为邻域半径,则会出现分类间隔过大导致分类无意义、样本距离过大容易造成异类样本和同类样本失效,以及无法处理弱标记数据等情况。为解决这些问题,提出一种基于多标记邻域粗糙集和多标记 Relief 的弱标记特征选择方法。首先,引入异类样本数和同类样本数来改进分类间隔,在此基础上定义邻域半径,构造新的邻域近似精度与多标记邻域粗糙集模型,并有效度量边界域引起的集合不确定性。其次,利用迭代更新权重公式填补大部分缺失标记信息,将邻域近似精度与互信息相结合,以构造新的标记相关性,填补剩余的缺失标记信息。然后,使用异类样本数和同类样本数,以构造新的标记权重和特征权重计算公式,进而提出多标记 Relief 模型,并将其应用于多标记特征选择。最后,结合多标记邻域粗糙集模型和多标记 Relief 算法,设计一种新的弱标记特征选择算法,以处理带有缺失标记的高维数据,并有效地提升多标记分类性能。在 11 个公共多标记数据集上进行仿真实验,结果验证了所提出的弱标记特征选择算法的有效性。

关键词: 多标记学习;特征选择;邻域粗糙集;Relief;缺失标记

中图法分类号 TP181

Weak Label Feature Selection Method Based on Neighborhood Rough Sets and Relief

SUN Lin^{1,2}, HUANG Miao-miao^{1,3} and XU Jiu-cheng^{1,2}

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

2 Key Laboratory of Artificial Intelligence and Personalized Learning in Education of Henan Province, Xinxiang, Henan 453007, China

3 School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China

Abstract In multi-label learning and classification, existing feature selection algorithms based on neighborhood rough sets will use classification margin of samples as the neighborhood radius. However, when the margin is too large, the classification may be meaningless. When the distances of samples are too large, it will easily result in the abnormal heterogeneous or similar samples, and these existing feature selection algorithms cannot deal with the weak label data. To address these issues, a weak label feature selection method based on multi-label neighborhood rough sets and multi-label Relief is proposed. First, the number of heterogeneous and similar samples is introduced to improve the classification margin, based on which, the neighborhood radius is defined, a new formula of neighborhood approximation accuracy is presented, and then the multi-label neighborhood rough sets model is constructed and can effectively measure the uncertainty of sets in the boundary region. Second, the iterative updated weight formula is employed to fill in most of the missing labels, and then by combining the neighborhood approximation accuracy with the mutual information, a new correlation between labels is developed to fill in the remaining information of missing labels. Third, the number of heterogeneous and similar samples continues to be used to improve the label weighting and feature weighting formulas, and then the multi-label Relief model is proposed for multi-label feature selection. Finally, based on the multi-label neighborhood rough sets model and the multi-label Relief algorithm, a weak label feature selection algorithm is designed to process high-dimensional data sets with missing labels and effectively improve the performance of multi-label classification. The simulation tests are carried out on eleven public multi-label data sets, and experimental results verify the effectiveness of the proposed weak label feature selection algorithm.

到稿日期:2021-03-09 返修日期:2021-07-29

基金项目:国家自然科学基金(62076089,61772176,61976082);河南省科技攻关计划(212102210136)

This work was supported by the National Natural Science Foundation of China(62076089,61772176,61976082) and Key Science and Technology Program of Henan Province, China(212102210136).

通信作者:孙林(sunlin@htu.edu.cn)

Keywords Multi-label learning, Feature selection, Neighborhood rough sets, Relief, Missing labels

1 引言

随着大数据技术的快速发展,在机器学习、数据挖掘等领域涌现出了大量的多标记分类问题^[1]。多标记学习在文本分类、图像视频标注、社交网络等许多领域得到了广泛的应用^[2-3]。由于多标记数据中存在大量冗余的或不相关的特征,因此数据特征的高维性不仅会占用大量的计算时间和存储空间,而且容易降低分类器的分类效果^[4]。因此,多标记数据降维是当前机器学习中一项重要的研究任务^[5]。采集技术和成本的限制导致实际数据存在大量的标记缺失,这种缺失标记的情况通常被称为弱标记^[6]。因此,研究弱标记数据降维技术更是一项挑战性的工作。

多标记特征选择在减少多标记特征的基础上提高分类效率^[7]。目前,特征选择方法通常分为过滤式、封装式、嵌入式等^[8-10]。过滤式方法不考虑任何学习算法,根据数据固有属性选择特征^[11-12]。由于其具有简捷性和可解释性,过滤式被广泛应用^[13]。封装式是根据学习算法利用特定分类器选择最佳特征,但该方法的指数搜索空间会产生昂贵的计算成本^[14]。嵌入式将特征选择的训练过程与学习算法相结合,但特定算法仍存在局限性^[15-16]。另外,封装式和嵌入式易出现过拟合现象,且效率相对较低^[17]。对于高维数据,过滤式更高效简单^[3,13]。因此,在多标记决策系统中,本文针对带有缺失标记的弱标记数据,研究基于过滤式策略的弱标记特征选择方法。

邻域粗糙集理论可以直接处理连续型数据,并且在处理符号型和数值型的混合数据时不会破坏数据的顺序结构^[8,14,18-19]。Duan 等^[20]设计了多标记数据的下近似集和依赖度,提出基于邻域粗糙集的多标记特征选择算法,但该算法不能处理高维稀疏数据,如文本、图像等数据。Yu 等^[21]基于邻域粗糙集提出多标记分类算法,实现了图像自动标注。Yu 等^[22]研究了变精度邻域粗糙集模型,设计了具有局部相关性的多标记分类算法。Sun 等^[23]研究了邻域粗糙集和二进制粒子群优化模型,提出邻域决策系统的多标记特征选择算法。然而,这些多标记分类方法仍需手动设置邻域半径,导致出现随机性和不确定性现象。Liu 等^[18]使用样本最大近邻,提出在线多标记流特征选择算法。但是,当分类间隔非常大时,其作为邻域半径时,分类可能会失效。为了弥补这一缺点,本文研究邻域粗糙集理论,使用异类样本数和同类样本数,改进分类间隔,限制邻域半径。此外,基于该邻域半径给出邻域近似精度,使其有效度量边界域引起的集合不确定性。

当前,弱标记学习已引起越来越多学者的关注。Wang 等^[24]研究了基于邻域信息熵的多标记特征选择算法,以处理带有缺失标记的数据。Zhu 等^[15]提出基于最小二乘的多标记特征选择算法,来恢复缺失标记并选择最具区分性的特征子集。Jiang 等^[25]构建了基于标记压缩和局部特征相关性的特征选择方法,以处理标记缺失数据。Wang 等^[26]通过迭代更新权重和标记相关性恢复缺失标记,设计多标记分类算法。但是,该算法仅用集合元素数量的比值来计算标记相关性,

无法准确衡量其具体值。为了弥补该方法的缺陷,本文改进互信息,可以有效衡量标记之间的相关程度。

Relief 算法是基于特征权重的分类模型^[27]。截至目前,由于其高效性和灵活性,多种 Relief 算法被采用。Spolaor 等^[28]用差异函数设计了基于 ReliefF 的多标记特征选择算法。Reyes 等^[29]扩展 ReliefF 算法来研究多标记学习的加权和特征选择。Kong 等^[30]基于 ReliefF 和 F 统计量设计了多标记特征选择算法。然而,上述 Relief 算法均没有考虑标记空间中每个标记对样本具有不同的可分性^[31]。为了弥补这个缺点,Lin 等^[31]使用样本间的距离来度量特征的重要性,提出了基于标记权重的多标记特征选择算法,但当样本间的距离非常大时会使得异类样本或者同类样本无效。为了解决这个问题,在文献^[31]的标记权重和特征权重的基础上,本文引入异类样本数和同类样本数,以构建新的标记权重和特征权重公式,消除距离过大时带来的负面影响,由此提出带有标记权重的多标记 Relief 算法。总的说来,为了设计新的弱标记特征选择算法,本文从以下 3 个方面来总结本文的贡献:

(1)为了解决分类间隔过大导致分类无意义的问题,引入异类样本数和同类样本数来改进分类间隔,限制邻域半径,有效度量多标记邻域决策系统中边界域引起的集合不确定性。

(2)为了解决仅用标记集合中元素数量的比值来计算标记相关性无法准确衡量具体值的问题,使用改进的互信息来衡量标记之间的相关性,进而有效衡量标记之间的相关程度。

(3)为了解决当样本间距离非常大时异类样本或同类样本会失效的问题,使用异类样本数和同类样本数分别研究样本和近邻样本之间的距离,提出新的标记权重和特征权重公式,构建多标记 Relief 算法,并将其应用于多标记特征选择,有效降低时间代价和存储负担。

2 基础理论

2.1 更新和迭代

对于任意样本 $x_i \in U$, 样本 x_i 在标记 l_j 上的第 t 次迭代的预测标记值^[26]可以描述为:

$$q'_{ij} = \frac{\sum_{k \in \mathbf{KNN}(x_i)} |\omega_{kj}^{-1}| \cdot q_{kj}}{\sum_{k \in \mathbf{KNN}(x_i)} |\omega_{kj}^{-1}|} \quad (1)$$

其中, $t \geq 2, j = 1, 2, \dots, m, \mathbf{KNN}(x_i)$ 为 x_i 的近邻样本集合, ω_{kj}^{-1} 为第 k 个实例的第 j 个标记在 $t-1$ 次迭代的权重值, q_{kj} 为第 k 个实例的第 j 个标记在 t 次迭代的预测标记值。

对于任意样本 $x_i \in U, x_i$ 在标记 l_j 上的第 t 次迭代的权重值更新^[26]为:

$$\omega_{ij} = \begin{cases} \text{sgn}(q'_{ij}), & \text{if } \text{sgn}(q'_{ij}) = \text{sgn}(\omega_{ij}^{t-1}) \\ & \text{且 } q'_{ij} > e, \omega_{ij}^{t-1} > e \\ c \frac{q'_{ij} - \text{Min}_{ij}}{\text{Max}_{ij} - \text{Min}_{ij}}, & \text{if } \text{sgn}(q'_{ij}) = -1 \text{ 且} \\ & \text{sgn}(L_{1_{ij}}) = 1 \\ q'_{ij}, & \text{otherwise} \end{cases} \quad (2)$$

其中, $\text{sgn}()$ 是符号函数, e 是高置信度阈值, 取值为 $(0.5, 1)$,

c 是低置信度阈值, 取值为 $(0, 0.5)$, 同时 $Max_j = \max_i(q'_{ij})$, $Min_j = \min_i(q'_{ij})$, L_{1ij} 为随机缺失后的标记矩阵。

对于任意样本 $x_i \in U$, x_i 在标记 l_j 上的第 t 次迭代的预测样本值被重新分配^[26]为:

$$q'_{ij} = \begin{cases} 1, & q_{ij} > 0 \\ 0, & q_{ij} = 0 \\ -1, & q_{ij} < 0 \end{cases} \quad (3)$$

2.2 大间隔

任意样本 $x_i \in U$ 的分类间隔^[31]描述为:

$$margin(x_i) = \Delta(x_i, NM(x_i)) - \Delta(x_i, NH(x_i)) \quad (4)$$

其中, $NM(x_i)$ 表示样本空间中与 x_i 距离最近的异类样本, $NH(x_i)$ 表示样本空间中与 x_i 最近的同类样本, $\Delta(x_i, NM(x_i))$ 和 $\Delta(x_i, NH(x_i))$ 表示 x_i 到 $NM(x_i)$ 和 $NH(x_i)$ 的距离。其详细图解如图 1 所示。

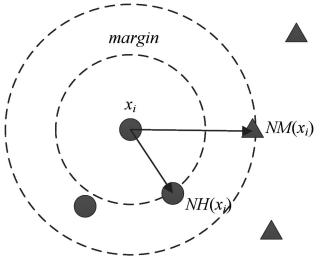


图 1 样本 x_i 的分类间隔

Fig. 1 Classification margin of sample x_i

Relief 算法是基于大间隔的特征选择方法, 通过样本的间隔迭代得到特征权重, 对于任意样本 $x_i \in U$, 第 i 个特征权重^[31]的表达式为:

$$w_i = w_i + |x_i - NM(x_i)| - |x_i - NH(x_i)| \quad (5)$$

其中, $|x_i - NM(x_i)|$ 和 $|x_i - NH(x_i)|$ 分别代表 x_i 到 $NM(x_i)$ 和 $NH(x_i)$ 的距离, $|\cdot|$ 表示绝对值。

3 弱标记特征选择方法

3.1 多标记邻域粗糙集

为了解决分类间隔过大导致分类无意义的问题, 引入异类样本数和同类样本数, 提出新的分类间隔, 从而在邻域粗糙集模型中限制邻域半径, 由此提出邻域近似精度, 以有效度量多标记邻域决策系统中边界域引起的集合的不确定性。

在多标记邻域决策系统 $MNDT = \langle U, C, D, T, \Delta, \delta \rangle$ 中, $U = \{x_1, x_2, \dots, x_n\}$ 表示样本空间; C 和 D 分别表示条件属性集和决策属性集(标记集合); $T = \{(x_i, y_i) | i = 1, 2, \dots, n\}$ 表示在标记上的映射关系; Δ 是距离函数; δ 是邻域参数且 $0 \leq \delta \leq 1$ 。本文中, 为了便于描述, 多标记邻域决策系统简记为 $MNDT = \langle U, C, D \rangle$ 。

定义 1 在多标记邻域决策系统 $MNDT = \langle U, C, D \rangle$ 中, $U = \{x_1, x_2, \dots, x_n\}$, 特征子集 $F = \{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L = \{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, 对于任意标记 $l \in L$, 则给定样本 $x_i \in U$ 在 l 下的分类间隔的定义为:

$$m^l(x_i) = \frac{\Delta(x_i, NHS^l(x_i))}{\|HS^l(x_i)\|} - \frac{\Delta(x_i, NSS^l(x_i))}{\|SS^l(x_i)\|} \quad (6)$$

其中, $NHS^l(x_i)$ 表示在 l 下与 x_i 最近的异类样本, $NSS^l(x_i)$

表示在 l 下与 x_i 最近的同类样本。 $\Delta(x_i, NHS^l(x_i))$ 和 $\Delta(x_i, NSS^l(x_i))$ 则分别表示 x_i 到 $NHS^l(x_i)$ 和 $NSS^l(x_i)$ 的距离。 $\|HS^l(x_i)\|$ 和 $\|SS^l(x_i)\|$ 分别表示在 l 下 x_i 的异类样本数和同类样本数, $\|\cdot\|$ 表示数量。

定义 2 在多标记邻域决策系统 $MNDT = \langle U, C, D \rangle$ 中, $U = \{x_1, x_2, \dots, x_n\}$, 特征子集 $F = \{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L = \{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, 对于任意标记 $l \in L$ 和任意样本 $x_i \in U$, 当 x_i 在 l 下的分类间隔 $m^l(x_i) \geq 0$ 时, 则 x_i 的邻域表示为:

$$\delta_B^l(x_i) = \{x_r | \Delta(x_i, x_r) \leq m^l(x_i), x_r \in U\} \quad (7)$$

定义 3 在多标记邻域决策系统 $MNDT = \langle U, C, D \rangle$ 中, $U = \{x_1, x_2, \dots, x_n\}$, 特征子集 $F = \{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L = \{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, D^j 表示具有类别标记 l_j 的样本集合, D_i 表示样本 x_i 所具有的标记集合, 给定特征子集 $B \subseteq C$, 多标记邻域粗糙集的下近似集和上近似集分别表示为:

$$\underline{N_B}D = \{x_i \in U | \forall l_j \in D_i, \delta_B^l(x_i) \subseteq D^j\} \quad (8)$$

$$\overline{N_B}D = \{x_i \in U | \forall l_j \in D_i, \delta_B^l(x_i) \cap D^j \neq \emptyset\} \quad (9)$$

定义 4 在多标记邻域决策系统 $MNDT = \langle U, C, D \rangle$ 中, $U = \{x_1, x_2, \dots, x_n\}$, 特征子集 $F = \{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L = \{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, 则 D 对 U 的划分为 $U/D = \{X_1, X_2, \dots, X_N\}$ 。对于任意 $B \subseteq C$, U/D 相对于 B 的邻域近似精度表示为:

$$\rho_B(D) = \frac{\|\underline{N_B}D\|}{\|\overline{N_B}D\|} \quad (10)$$

3.2 标记相关性

互信息是一种衡量相关性的有效手段, 常用于描述标记之间信息共享的程度^[3]。为了解决仅用集合中的元素数的比值计算相关性, 无法准确衡量具体值大小的问题, 本文改进传统的互信息来衡量标记之间的相关性。

定义 5 在多标记邻域决策系统 $MNDT = \langle U, C, D \rangle$ 中, $U = \{x_1, x_2, \dots, x_n\}$, 特征子集 $F = \{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L = \{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, 对于任意两个标记 $l_j, l_s \in L$, 则 l_j 与 l_s 之间的标记相关性定义为:

$$R(l_j, l_s) = \rho_B(D) \iint p(q_j, q_s) \log \frac{p(q_j, q_s)}{p(q_j)p(q_s)} dq_j dq_s \quad (11)$$

其中, $\rho_B(D)$ 为邻域近似精度, q_j 和 q_s 分别表示 l_j 和 l_s 标注的实例集合, $p(q_j)$, $p(q_s)$, $p(q_j, q_s)$ 分别为 q_j , q_s 以及 q_j 与 q_s 的概率密度函数。

定义 6 在多标记邻域决策系统 $MNDT = \langle U, C, D \rangle$ 中, $U = \{x_1, x_2, \dots, x_n\}$, 特征子集 $F = \{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L = \{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, 对于任意样本 $x_i \in U$ 和任意标记 $l_j \in L$, 则 x_i 在 l_j 上的预测标记值定义为:

$$\tilde{q}_{ij} = \begin{cases} q_i^T \cdot R(\cdot, j), & q_{ij} = 0 \\ q_{ij}, & q_{ij} \neq 0 \end{cases} \quad (12)$$

其中, q_i^T 是第 i 个样本的标记向量的转置。

3.3 多标记 Relief 模型

Relief 算法通过样本距离来度量特征的重要性, 但当样本距离过大时可能使异类样本或者同类样本失去作用。为了解决这个问题, 使用异类样本数和同类样本数消除距离过大带来的影响, 同时标记权重和特征权重, 分别用于测量标记对

样本的可区分性和特征对标记集合的重要性程度。

定义 7 在多标记邻域决策系统 $MNDT=\langle U, C, D \rangle$ 中, $U=\{x_1, x_2, \dots, x_n\}$, 特征子集 $F=\{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L=\{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, 对于任意标记 $l \in L$ 和任意两个样本 $x_j, x_r \in U$, 在 l 上 x_i 和 x_r 的距离函数定义为:

$$\Delta_l(x_i, x_r) = |x_i(l) - x_r(l)| \quad (13)$$

其中, $x_i(l)$ 和 $x_r(l)$ 表示 x_i 和 x_r 在 l 上的标记值。

定义 8 在多标记邻域决策系统 $MNDT=\langle U, C, D \rangle$ 中, $U=\{x_1, x_2, \dots, x_n\}$, 特征子集 $F=\{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L=\{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, 对于任意标记 $l \in L$ 和任意样本 $x_i \in U$, 则在 l 上 x_i 与其 $NHS^l(x_i)$ 的距离函数定义为:

$$\Delta_l(x_i, NHS^l(x_i)) = |x_i(l) - NHS^l(x_i)(l)| \quad (14)$$

其中, $NHS^l(x_i)$ 表示在 l 下样本空间中与 x_i 距离最近的异类样本, 称为 x_i 的最近邻异类样本, $NHS^l(x_i)(l)$ 表示 $NHS^l(x_i)$ 在 l 上的值。

定义 9 在多标记邻域决策系统 $MNDT=\langle U, C, D \rangle$ 中, $U=\{x_1, x_2, \dots, x_n\}$, 特征子集 $F=\{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L=\{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, 对于任意标记 $l \in L$ 和任意样本 $x_i \in U$, 在 l 上 x_i 与其 $NSS^l(x_i)$ 的距离函数定义为:

$$\Delta_l(x_i, NSS^l(x_i)) = |x_i(l) - NSS^l(x_i)(l)| \quad (15)$$

其中, $NSS^l(x_i)$ 表示在 l 下样本空间中与 x_i 距离最近的同类样本, 称为 x_i 的最近邻同类样本, $NSS^l(x_i)(l)$ 表示 $NSS^l(x_i)$ 在 l 上的值。

定义 10 在多标记邻域决策系统 $MNDT=\langle U, C, D \rangle$ 中, $U=\{x_1, x_2, \dots, x_n\}$, 特征子集 $F=\{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L=\{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, 对于任意标记 $l \in L$ 和任意样本 $x_i \in U$, 则对 l 赋予的标记权重 w_l 的计算式为:

$$w_l = \sum_{i=1}^n \left(\frac{\Delta_l(x_i, NHS^l(x_i))}{\|HS^l(x_i)\|} - \frac{\Delta_l(x_i, NSS^l(x_i))}{\|SS^l(x_i)\|} \right) \quad (16)$$

其中, $\|HS^l(x_i)\|$ 表示在 l 下 x_i 的异类样本数, $\|SS^l(x_i)\|$ 表示在 l 下 x_i 的同类样本数, $\Delta_l(x_i, NHS^l(x_i))$ 表示在 l 下 x_i 和 $NHS^l(x_i)$ 的距离, $\Delta_l(x_i, NSS^l(x_i))$ 表示在 l 下 x_i 和 $NSS^l(x_i)$ 的距离。

下文结合定义 7—定义 9, 给出 3 种新的距离函数, 设计新的特征权重计算公式。

定义 11 在多标记邻域决策系统 $MNDT=\langle U, C, D \rangle$ 中, $U=\{x_1, x_2, \dots, x_n\}$, 特征子集 $F=\{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L=\{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, 对于任意特征 $f \in F$ 和任意两个样本 $x_j, x_r \in U$, 则 x_i 与 x_r 之间的距离定义为:

$$d_f(x_i, x_r) = \frac{|x_i(f) - x_r(f)|}{\max(f) - \min(f)} \quad (17)$$

其中, $x_i(f)$ 和 $x_r(f)$ 分别表示 x_i 和 x_r 在 f 上的值, $\max(f)$ 和 $\min(f)$ 分别表示 f 在样本空间中取得的最大值和最小值。

定义 12 在多标记邻域决策系统 $MNDT=\langle U, C, D \rangle$ 中, $U=\{x_1, x_2, \dots, x_n\}$, 特征子集 $F=\{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L=\{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, 对于任意特征 $f \in F$ 和任意样本 $x_j \in U$, 则 x_i 与其最近邻异类样本 $NHS^l(x_i)$ 之间的距离定义为:

$$d_f(x_i, NHS^l(x_i)) = \frac{|x_i(f) - NHS^l(x_i)(f)|}{\max(f) - \min(f)} \quad (18)$$

其中, $NHS^l(x_i)(f)$ 表示 x_i 的 $NHS^l(x_i)$ 在 f 上的值。

定义 13 在多标记邻域决策系统 $MNDT=\langle U, C, D \rangle$ 中, $U=\{x_1, x_2, \dots, x_n\}$, 特征子集 $F=\{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L=\{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, 对于任意特征 $f \in F$ 和任意样本 $x_j \in U$, 则 x_i 与其最近邻同类样本 $NSS^l(x_i)$ 之间的距离定义为:

$$d_f(x_i, NSS^l(x_i)) = \frac{|x_i(f) - NSS^l(x_i)(f)|}{\max(f) - \min(f)} \quad (19)$$

其中, $NSS^l(x_i)(f)$ 表示 x_i 的 $NSS^l(x_i)$ 在 f 上的值。

定义 14 在多标记邻域决策系统 $MNDT=\langle U, C, D \rangle$ 中, $U=\{x_1, x_2, \dots, x_n\}$, 特征子集 $F=\{f_1, f_2, \dots, f_z\}$, $F \subseteq C$, 标记子集 $L=\{l_1, l_2, \dots, l_m\}$, $L \subseteq D$, 对于任意特征 $f \in F$, 则 f 的特征权值 w_f 的计算公式为:

$$w_f = \sum_{l=1}^m w_l \sum_{i=1}^n \left(\frac{d_f(x_i, NHS^l(x_i))}{\|HS^l(x_i)\|} - \frac{d_f(x_i, NSS^l(x_i))}{\|SS^l(x_i)\|} \right) \quad (20)$$

3.4 弱标记特征选择算法

首先, 初始化权重矩阵和预测标记矩阵; 其次, 针对数据迭代更新权重值和预测标记值, 恢复 80% 的缺失标记; 然后, 利用标记相关性恢复剩余 20% 的标记信息; 最后, 使用特征权重公式对特征进行排序, 得到一组特征排序。由此, 基于邻域粗糙集和 Relief 的弱标记特征选择 (Weak Label Feature Selection Method Based on Neighborhood Rough Sets and Relief, WFSNR) 算法, 如算法 1 所示。

算法 1 WFSNR

输入: 多标记邻域决策系统 $MNDT=\langle U, C, D \rangle$

输出: 特征子集

- 步骤 1 将权重矩阵 W 和预测标记矩阵 Q 初始化为缺失标记集 L_1 ;
- 步骤 2 While L_1 中 80% 的标记未完全恢复时执行
- 步骤 3 For 任意样本 $x_i \in U$
 - 寻找 x_i 的近邻集合 $KNN(x_i)$, 更新预测标记值 q_{li}^0 和权重值 w_{li}^0 ;
 - 重置标记信息, 保留 L 中已知相关标记信息, 得到预测标记矩阵 Q ;
 - End For
- End While
- 步骤 4 根据式(7)计算样本的邻域 $\delta_B^l(x_i)$;
- 步骤 5 根据式(10)计算邻域近似精度 $\rho_B(D)$;
- 步骤 6 使用式(11)计算标记相关矩阵 R ;
- 步骤 7 寻找剩余 20% 训练数据中的缺失标记信息, 使用式(12)对其进行恢复;
- 步骤 8 For 任意标记 $l \in L$
 - 使用式(16)计算每个标记的权重 w_l ;
 - End For
- 步骤 9 对标记权重归一化;
- 步骤 10 For 任意特征 $f \in F$
 - 根据式(20)计算每个特征的权重 w_f ;
 - End For
- 步骤 11 对得到的特征权重进行排序, 得到一组特征排序 rank;
- 步骤 12 输出最优特征子集。

在算法 1 中, 假设多标记数据集包括 n 个样本、 m 个标记和 z 个特征。步骤 2 和步骤 3 为更新和迭代阶段, 时间复杂度为 $O(mn)$, 步骤 4—步骤 7 为利用标记相关性恢复剩余

标记信息阶段,时间复杂度为 $O(mn^2 \log n) + O(mn)$,步骤 8—步骤 10 计算标记权重和特征权重,时间复杂度为 $O(m) + O(mz)$,步骤 11—步骤 12 为特征排序和输出最优特征子集阶段,时间复杂度为 $O(z \log z)$,显然 $O(m) < O(mz)$ 。由此 WFSNR 算法总的时间复杂度为 $O(mz + mn + mn^2 \log n)$ 。

4 实验

4.1 实验设置

为了验证本文 WFSNR 算法的有效性,从 MuLan 数据库中选取了 11 个公开的多标记数据集进行实验¹⁾,表 1 列出了 11 个数据集的详细信息。实验的硬件配置为 CPU Inter i5-6500M 3. 20 GHz 和 8 GB 内存,运行环境为 MATLAB R2016b。在本文算法中,高置信度阈值 $e=0. 8$,低置信度阈值 $c=0. 2$,采用 Multi-label K-Nearest Neighbors(MLKNN)^[32] 为分类器,评估算法的分类性能,其中分类器 MLKNN 的平滑参数 $s=1$,近邻个数 $k=10$ 。

表 1 11 个多标记数据集的描述

Table 1 Description of eleven multi-label data sets

Data sets	Instances	Features	Classes	Training	Test	Domains
Bibtex	7 395	1 836	159	4 880	2 515	Text
Birds	645	260	19	322	323	Audio
CAL500	502	68	174	250	252	Music
Corel5k	5 000	499	374	4 500	500	Images
Emotions	593	72	6	391	202	Music
Enron	1 702	1 001	53	1 123	579	Text
Flags	194	19	7	129	65	Images
Genbase	662	1 186	27	463	199	Biology
Medical	978	1 449	45	333	645	Text
Scene	2 407	294	6	1 211	1 196	Images
Yeast	2 417	103	14	1 500	917	Biology

为了有效评估所有对比算法的分类效果,本文使用文献[25-26,33]中的 6 个评估指标:*Average Precision*,*Coverage*,*One-Error*,*Ranking Loss*,*Macro Average F1* 和 *Micro Average F1*。其中,*Average Precision*,*Macro Average F1* 和 *Micro Average F1* 的值越大,表明算法的分类性能越优,最优值为 1;*Coverage*,*One-Error* 和 *Ranking Loss* 的值越小,则算法的分类性能越优,最优值为 0。

为了充分展示 WFSNR 算法的分类性能,分别使用文献[25-26]中的实验结果进行对比。由于文献中的对比算法和数据集是不同的,因此将本节实验分为 3 个部分:首先,按随机比例缺失标记,在 3 个指标上将 WFSNR 算法与 3 种最近的多标记方法的分类结果进行比较;然后,按随机个数缺失标记,在 5 个指标下将 WFSNR 算法与 6 种最近的多标记方法做实验比较;最后,对所有比较方法的分类结果进行统计分析。在以下实验中,粗体表示指标的最优结果。

4.2 随机比例缺失标记下的实验结果

这部分实验是在 3 个指标 (*Macro Average F1*, *Micro Average F1* 和 *One-Error*) 下,展示 WFSNR 算法与 3 种最近提出的多标记分类算法的比较结果,这 3 种比较算法包括: Multilabel Classification for Weak-label Data (MCWD)^[26], Semi-supervised Multi-label Classification With Non-Uniformly Distributed Incomplete Labels (SMCNDI)^[34] 和 Weak La-

bel Learning Method (WELL)^[35]。在 3 个指标上,根据文献[26]的实验设置,从表 1 中选出 11 个多标记数据集: Bibtex, Birds, CAL500, Corel5k, Emotions, Enron, Flags, Genbase, Medical, Scene 和 Yeast。为了客观地表明 WFSNR 算法的分类性能,表 2 列出了 11 个多标记数据集的标记缺失率,表 3—表 5 分别列出了 MCWD, SMCNDI, WELL 和 WFSNR 这 4 种算法在 3 个指标上的分类结果。

表 2 11 个数据集的标记缺失率

Table 2 Missing labels rate of eleven multi-label data sets
(单位: %)

Data sets	Missing labels rate
Bibtex	99. 37
Birds	97. 13
CAL500	92. 53
Corel5k	99. 47
Emotions	83. 33
Enron	97. 56
Flags	71. 50
Genbase	96. 30
Medical	97. 78
Scene	83. 33
Yeast	85. 80

从表 3 中可以看出,对于 *Macro Average F1* 指标,在 Birds, CAL500, Corel5k, Flags, Genbase 和 Medical 这 6 个多标记数据集上, WFSNR 算法的表现最优。在 Bibtex, Emotions 和 Scene 数据集上, WFSNR 算法皆略次于其他 3 种算法,但仅相差 0. 021 5~0. 146 9。在 Enron 数据集上, WFSNR 算法高于 MCWD 算法 0. 014, 分别低于 SMCNDI 算法和 WELL 算法 0. 054 9 和 0. 024 2。在 Yeast 数据集上, WFSNR 算法高于 SMCNDI 算法和 WELL 算法, 低于 MCWD 算法 0. 086 6。

表 3 *Macro Average F1* 评价指标下 4 种算法的性能比较
Table 3 Comparison of four algorithms in terms of *Macro Average F1* evaluation index

Data sets	MCWD	SMCNDI	WELL	WFSNR
Bibtex	0. 022 2	0. 024 0	0. 029 9	0. 000 7
Birds	0. 063 5	0. 076 9	0. 090 0	0. 133 8
CAL500	0. 150 0	0. 127 4	0. 094 7	0. 973 4
Corel5k	0. 247 1	0. 206 0	0. 201 0	0. 298 0
Emotions	0. 486 5	0. 384 3	0. 390 9	0. 354 3
Enron	0. 023 8	0. 092 7	0. 062 0	0. 037 8
Flags	0. 528 2	0. 221 4	0. 239 1	0. 964 6
Genbase	0. 512 3	0. 478 5	0. 463 0	0. 999 5
Medical	0. 131 8	0. 122 8	0. 091 4	0. 999 7
Scene	0. 657 0	0. 613 0	0. 599 0	0. 510 1
Yeast	0. 428 8	0. 299 4	0. 277 5	0. 342 2

表 4 *Micro Average F1* 评价指标下 4 种算法的性能比较
Table 4 Comparison of four algorithms in terms of *Micro Average F1* evaluation index

Data sets	MCWD	SMCNDI	WELL	WFSNR
Bibtex	0. 025 0	0. 059 8	0. 098 1	0. 969 0
Birds	0. 082 9	0. 153 8	0. 197 2	0. 237 7
CAL500	0. 278 4	0. 259 9	0. 195 6	0. 964 0
Corel5k	0. 489 1	0. 438 5	0. 466 2	0. 093 9
Emotions	0. 497 5	0. 392 4	0. 402 1	0. 400 0
Enron	0. 075 1	0. 143 8	0. 081 1	0. 215 3
Flags	0. 589 7	0. 248 9	0. 260 4	0. 968 4
Genbase	0. 808 4	0. 862 4	0. 772 2	0. 999 2
Medical	0. 099 1	0. 404 4	0. 252 1	0. 999 0
Scene	0. 612 6	0. 592 5	0. 533 6	0. 533 9
Yeast	0. 531 0	0. 365 7	0. 329 7	0. 570 4

¹⁾ <http://mulan.sourceforge.net/datasets.html>

表 5 *One-Error* 评价指标下 4 种算法的性能比较
Table 5 Comparison of four algorithms in terms of *One-Error* evaluation index

Data sets	MCWD	SMCNDI	WELL	WFSNR
Bibtex	0.4599	0.9288	1.0002	0.9690
Birds	0.9745	0.7138	0.9405	0.7965
CAL500	0	0.6954	0.7450	0
Corel5k	0.2166	0.2973	0.3260	0.7900
Emotions	0.5988	0.6025	0.5925	0.5000
Enron	0.8431	0.7533	0.9559	0.5043
Flags	0.5747	0.5788	0.6153	0.1692
Genbase	0.5530	0.6317	0.6130	0.7186
Medical	0.6420	0.5642	0.6622	0.0062
Scene	0.3643	0.3246	0.3751	0.3595
Yeast	0.4397	0.4640	0.5462	0.2977

从表 4 可以看出,对于 *Micro Average F1* 指标,在 Bibtex, Birds, CAL500, Enron, Flags, Genbase, Medical 和 Yeast 这 8 个数据集上, WFSNR 算法均取得最优值。在 Corel5k 数据集上, WFSNR 算法略次于其他 3 种算法。在 Emotions 数据集上, WFSNR 算法优于 SMCNDI 算法, 分别比 MCWD 算法和 WELL 算法略低 0.0975 和 0.0021。在 Scene 数据集上, WFSNR 算法优于 WELL 算法, 分别略次于 MCWD 算法和 SMCNDI 算法, 但仅相差 0.0787 和 0.0586。

从表 5 可以发现,对于 *One-Error* 指标,在 Emotions, Enron, Flags, Medical 和 Yeast 数据集上, WFSNR 算法取得最小值。对于 CAL500 数据集, WFSNR 算法与 MCWD 算法均

最优。在 Bibtex 数据集上, WFSNR 算法比 WELL 算法优 0.0312, 分别次于 MCWD 算法和 SMCNDI 算法 0.5091 和 0.0402。在 Birds 和 Scene 数据集上, WFSNR 算法优于 MCWD 算法和 WELL 算法, 比 SMCNDI 算法略差 0.0827 和 0.0349。在 Corel5k 和 Genbase 数据集上, WFSNR 算法略低于其他 3 种算法。

4.3 随机个数缺失标记下的实验结果

这部分实验是在 5 个指标 (*Average Precision*, *Macro Average F1*, *Coverage*, *Ranking Loss* 和 *One-Error*) 下, 展示了 WFSNR 算法与 6 种最近提出的多标记分类算法的比较结果, 这 6 种比较算法包括: Baseline^[25], Correlated and Multi-label Feature Selection method (CMFS)^[36], Feature Manifold Learning and Sparse Regularization for Multi-Label Feature Selection (MSSL)^[37], Convex Semi-Supervised Multi-label Feature Selection (CSFS)^[38], Robust Multi-label Feature Selection with Missing Labels (MLMLFS)^[15] 和 Feature Selection with Missing Labels Based on Label Compression and Local feature Correlation (FSLCLC)^[23]。在 5 个指标上, 为了方便说明所有对比算法的实验结果, 根据文献[25]的实验设置, 本文设置 *Q* 表示每个样本缺失的标记个数, 从表 1 中选出 3 个多标记数据集: Bibtex, Corel5k 和 Enron。表 6—表 10 列出了上述 7 种算法在 5 个指标上的分类结果。

表 6 *Average Precision* 评价指标下 7 种算法的性能比较

Table 6 Comparison of seven algorithms in terms of *Average Precision* evaluation index

Data sets	<i>Q</i>	Baseline	CMFS	MSSL	CSFS	MLMLFS	FSLCLC	WFSNR
Bibtex	1	0.2270	0.2172	0.2999	0.2591	0.2061	0.3498	0.2494
	2	0.2087	0.2043	0.2665	0.2250	0.1988	0.3193	0.2497
	3	0.2036	0.1988	0.2523	0.1903	0.1945	0.3108	0.2492
Corel5k	1	0.3470	0.3394	0.3688	0.3655	0.3633	0.3708	0.2281
	2	0.3150	0.3082	0.3298	0.3246	0.3267	0.3330	0.2276
	3	0.2873	0.2846	0.2964	0.2961	0.2950	0.3017	0.2274
Enron	1	0.5473	0.5532	0.5172	0.5244	0.5271	0.5781	0.6358
	2	0.5324	0.5340	0.5037	0.5148	0.5041	0.5569	0.6315
	3	0.5266	0.5257	0.5102	0.5135	0.5082	0.5468	0.6238

表 7 *Macro Average F1* 评价指标下 7 种算法的性能比较

Table 7 Comparison of seven algorithms in terms of *Macro Average F1* evaluation index

Data sets	<i>Q</i>	Baseline	CMFS	MSSL	CSFS	MLMLFS	FSLCLC	WFSNR
Bibtex	1	0.1177	0.0805	0.1733	0.1388	0.0852	0.2085	0.0282
	2	0.1050	0.0771	0.1470	0.1118	0.0826	0.1918	0.0296
	3	0.0997	0.0768	0.1359	0.0970	0.0814	0.1864	0.0288
Corel5k	1	0.1125	0.1057	0.1241	0.1232	0.1264	0.1323	0.2992
	2	0.0945	0.0870	0.1044	0.0995	0.0996	0.1098	0.2992
	3	0.0677	0.0680	0.0782	0.0776	0.0776	0.0834	0.2992
Enron	1	0.1015	0.0985	0.0788	0.0719	0.0821	0.1158	0.0937
	2	0.0982	0.0927	0.0740	0.0686	0.0792	0.1085	0.0900
	3	0.0900	0.0852	0.0736	0.0661	0.0798	0.1079	0.0834

表 8 *Coverage* 评价指标下 7 种算法的性能比较

Table 8 Comparison of seven algorithms in terms of *Coverage* evaluation index

Data sets	<i>Q</i>	Baseline	CMFS	MSSL	CSFS	MLMLFS	FSLCLC	WFSNR
Bibtex	1	81.6563	79.5139	64.3406	69.8638	77.4668	62.6607	65.7714
	2	78.6372	83.3266	72.6502	76.3834	80.3448	69.2708	65.4791
	3	82.4729	84.1580	75.2739	80.5381	81.1119	72.6955	65.4179
Corel5k	1	72.8861	73.3930	72.1342	72.0526	72.0593	71.2354	119.5900
	2	80.3816	80.5972	79.6156	79.8880	79.7630	79.1977	119.5660
	3	86.2023	86.3744	86.0576	85.9660	86.2421	85.8640	119.5820
Enron	1	14.4190	14.5971	15.2268	15.2750	15.0309	14.2263	13.5803
	2	14.7851	15.0116	15.6149	15.6957	15.4286	14.7176	13.5147
	3	15.0811	15.3239	15.6129	15.7836	15.4935	15.1097	13.3437

表 9 *Ranking Loss* 评价指标下 7 种算法的性能比较

Table 9 Comparison of seven algorithms in terms of <i>Ranking Loss</i> evaluation index								
Data sets	Q	Baseline	CMFS	MSSL	CSFS	MLMLFS	FSLCLC	WFSNR
Bibtex	1	0.3504	0.3539	0.2821	0.3122	0.3477	0.2738	0.2651
	2	0.3624	0.3683	0.3152	0.3366	0.3561	0.2959	0.2641
	3	0.3619	0.3693	0.3255	0.3538	0.3553	0.3087	0.2638
Core5k	1	0.1509	0.1523	0.1471	0.1469	0.1471	0.1447	0.1399
	2	0.1673	0.1686	0.1636	0.1646	0.1642	0.1626	0.1399
	3	0.1798	0.1806	0.1775	0.1775	0.1783	0.1768	0.1400
Enron	1	0.1082	0.1095	0.1170	0.1175	0.1144	0.1050	0.0964
	2	0.1121	0.1140	0.1214	0.1207	0.1200	0.1095	0.0962
	3	0.1142	0.1175	0.1204	0.1220	0.1195	0.1135	0.0955

表 10 *One-Error* 评价指标下 7 种算法的性能比较

Table 10 Comparison of seven algorithms in terms of <i>One-Error</i> evaluation index								
Data sets	Q	Baseline	CMFS	MSSL	CSFS	MLMLFS	FSLCLC	WFSNR
Bibtex	1	0.7069	0.7270	0.6264	0.6867	0.7516	0.5897	0.7002
	2	0.7284	0.7363	0.6610	0.7206	0.7568	0.6095	0.6994
	3	0.7307	0.7388	0.6678	0.7560	0.7611	0.6134	0.6990
Core5k	1	0.6046	0.6123	0.5783	0.5760	0.5806	0.5629	0.7320
	2	0.6388	0.6430	0.6184	0.6217	0.6228	0.6128	0.7340
	3	0.6672	0.6653	0.6504	0.6510	0.6524	0.6363	0.7340
Enron	1	0.4028	0.3951	0.4383	0.4266	0.4413	0.3731	0.2867
	2	0.4375	0.4199	0.4742	0.4425	0.4896	0.4044	0.3022
	3	0.4512	0.4390	0.4810	0.4748	0.5037	0.3961	0.3040

分析表 6 可知,在 *Average Precision* 指标下,在 Enron 数据集上,当 $Q=1,2$ 或 3 时,WFSNR 算法表现最优。对于数据集 Bibtex,当 $Q=1$ 时,WFSNR 算法比 Baseline,CMFS 和 MLMLFS 这两种算法高出 $0.0224\sim0.0433$,比其他 3 种算法低 $0.0097\sim0.1004$;当 $Q=2$ 或 3 时,WFSNR 算法优于 Baseline,CMFS,CSFS 和 MLMLFS 这 4 种算法,略次于 MSSL 算法和 FSLCLC 算法。对于数据集 Core5k,当 $Q=1,2$ 或 3 时,WFSNR 算法分别低于其他算法 $0.1113\sim0.1427,0.0806\sim0.1054$ 和 $0.0572\sim0.0743$ 。观察表 7 可知,在 *Macro Average F1* 指标下,对于 Core5k 数据集,当 $Q=1,2$ 或 3 时,WFSNR 算法取得了最优值。对于数据集 Bibtex,当 $Q=1,2$ 或 3 时,WFSNR 算法比其他对比算法差 $0.0523\sim0.1803,0.0475\sim0.1622$ 和 $0.0480\sim0.1576$ 。对于 Enron 数据集,当 $Q=1,2$ 或 3 时,WFSNR 算法的实验结果优于 MSSL 算法、CSFS 算法和 MLMLFS 算法,比其他 3 种算法差 $0.0048\sim0.0221,0.0027\sim0.0185$ 和 $0.0018\sim0.0245$ 。

从表 8 可以看出,在 *Coverage* 指标下,对于 Enron 数据集,WFSNR 算法获得了最小值(最优值)。对于 Bibtex 数据集,当 $Q=1$ 时,WFSNR 算法仅次于 FSLCLC 算法,优于其他 5 种算法;当 $Q=2$ 或 3 时,WFSNR 算法的表现最优。对于 Core5k 数据集,WFSNR 算法略差于其他对比算法。从表 9 可以看出,在 *Ranking Loss* 评价指标下,对于 Bibtex,Core5k 和 Enron 这 3 个数据集,当 $Q=1,2$ 或 3 时,WFSNR 算法的实验结果取得最优值。从表 10 可以得出,在 *One-Error* 指标下,对于 Enron 数据集,WFSNR 算法的表现最优。对于 Bibtex 数据集,当 $Q=1$ 时,在 *One-Error* 指标下,WFSNR 算法比 Baseline,CMFS 和 MLMLFS 这 3 种算法低 $0.0067\sim0.0514$,比 MSSL,CSFS 和 FSLCLC 这 3 种算法高 $0.0135\sim$

0.1105 。当 $Q=2$ 或 3 时,WFSNR 算法仅次于 MSSL 算法和 FSLCLC 算法,优于其他 4 种对比算法。对于 Core5k 数据集,当 $Q=1,2$ 或 3 时,在 *One-Error* 指标下,WFSNR 算法分别低于其他对比算法 $0.1197\sim0.1691,0.0910\sim0.1212$ 和 $0.0668\sim0.0977$ 。总的来说,WFSNR 算法在上述 3 个多标记数据集上表现最优。

4.4 统计分析

本文使用 Friedman test 和 Bonferroni-Dunn test 方法^[7-8],在每个指标上研究所有对比算法分类结果的统计情况。其中,Friedman 统计可描述为:

$$\chi_F^2=\frac{12T}{s(s+1)}(\sum_{i=1}^sR_i^2-\frac{s(s+1)^2}{4})$$

(21)

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

(22)

其中, T 和 s 分别表示数据集和算法的个数, R_i 是算法在所有数据集上的平均排序。这里的临界差可描述为:

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

(23)

其中, q_α 是测试的临界列表值, α 是显著性水平。

对于表 3—表 5 中的结果,图 2 给出了在 3 个指标上 4 种算法的比较结果,其中每种比较算法的平均等级均沿轴绘制。从图 2 可以得出如下结论:1)在 3 种指标上,WFSNR 算法的性能明显优于 WELL 算法;2)在 *Micro Average F1* 指标上,WFSNR 算法明显优于 SMCNDI 算法;3)WFSNR 算法的总排名低于 MCWD 算法,即 WFSNR 算法略优于 MCWD 算法。总之,WFSNR 算法优于其他 3 种对比算法。通过计算,当显著性水平为 $\alpha=0.1$ 时, $F(3,30)$ 的临界值为 2.28 ,拒绝零假设。对于 Bonferroni-Dunn test,当 $\alpha=0.1$ 时, $q_\alpha=2.128,CD=1.1714$,其中 $s=4,T=11$ 。

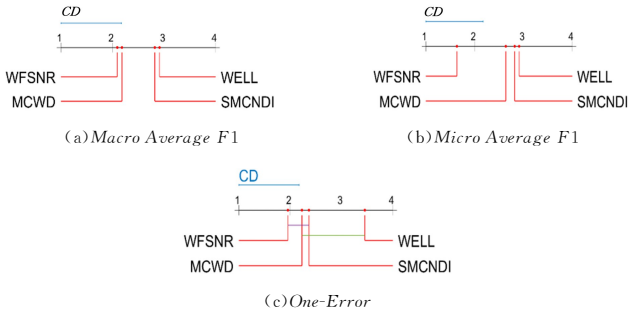


图2 在 Bonferroni-Dunn test 下 WFSNR 算法与其他 3 种算法的比较结果

Fig. 2 Comparison results of WFSNR algorithm against other three algorithms with Bonferroni-Dunn test

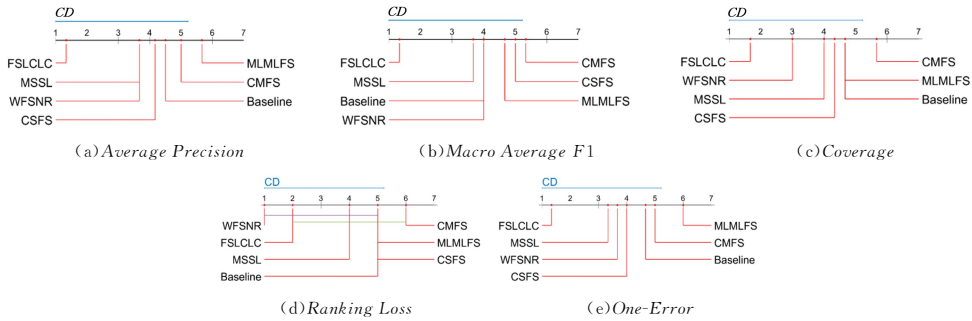


图3 在 Bonferroni-Dunn test 下 WFSNR 算法与其他 6 种算法的比较结果

Fig. 3 Comparison results of WFSNR algorithm against other six algorithms with Bonferroni-Dunn test

结束语 本文提出了一种基于邻域粗糙集和 Relief 算法的弱标记特征选择方法,以有效处理带有缺失标记的弱标记数据。首先,考虑异类样本数和同类样本数,改进分类间隔,将其作为邻域半径,以构造邻域近似精度,并得到改进的互信息。其次,利用迭代更新权重公式填补大部分缺失标记信息,使用新的互信息计算标记相关性,填补剩余的缺失标记信息。然后,提出新的标记权重和特征权重计算公式,改进 Relief 算法,将其用于多标记特征选择,降低时间代价。最后,设计一种弱标记特征选择算法,有效处理缺失标记数据,获得最优特征子集。在 11 个多标记数据集上的实验结果证明了本文算法的有效性。但是,在处理更大规模的复杂的多标记数据时,本文算法的表现并不是全部最优的。因此,在我们未来的研究工作中,针对大规模复杂的弱标记数据集,将结合模糊粗糙集、多标记邻域粗糙集等理论,进一步研究和处理多标记分类的弱监督学习问题。

参考文献

- [1] 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.
- [2] 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/OL]. IEEE Transactions on Fuzzy Systems, 2021. <https://ieeexplore.ieee.org/abstract/document/9333666>.
- [3] GONZÁLEZ-LÓPEZ J, VENTURA S, CANO A. Distributed selection of continuous features in multilabel classification using

对于表 6—表 10 的分类结果,图 3 给出了在平均标记缺失个数下,5 个指标上的 7 种算法的对比结果。从图 3 可以看出:1)在 5 个指标上, WFSNR 算法的分类性能优于 Baseline, CSFS, MLMLFS 和 CMFS 这 4 种算法;2)在 *Average Precision* 指标上, WFSNR 算法与 MSSL 算法保持同一排名,在 *Coverage* 和 *Ranking Loss* 指标上, WFSNR 算法略优于 MSSL 算法;3)在 *Ranking Loss* 指标上, WFSNR 算法优于 FSLCLC 算法,在其他 4 个评价指标上, FSLCLC 算法的总排名低于 WFSNR 算法,即 WFSNR 算法稍差于 FSLCLC 算法。总之, WFSNR 算法稍差于 FSLCLC 算法,与 MSSL 算法持平,优于其他 4 种对比算法。当 $\alpha=0.1$ 时, $F(6,18)$ 的临界值为 2.13,拒绝零假设。对于 Bonferroni-Dunn test,当 $\alpha=0.1$ 时, $q_\alpha=2.394$, $CD=4.2226$, 其中 $s=7$, $T=3$ 。

mutual information[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(7): 2280-2293.

- [4] DENG W, GUO Y X, LI Y, et al. Power losses prediction based on feature selection and Stacking integrated learning[J]. Power System Protection and Control, 2020, 28(15): 108-115.
- [5] CHEN C Y, LIN Y J, TANG L, et al. Streaming multi-label feature selection based on neighborhood interaction gain information[J]. Journal of Nanjing University (Natural Science), 2020, 56(1): 30-40.
- [6] LI Y C, YANG Y L, QIU H Q. Label embedding for weak label classification [J]. Journal of Nanjing University (Natural Science), 2020, 56(4): 549-560.
- [7] 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.
- [8] LIU Y, CHENG L, SUN L. Feature selection method based on K-S test and neighborhood rough set[J]. Journal of Henan Normal University (Natural Science Edition), 2019, 47(2): 21-28.
- [9] XUE Z A, PANG W L, YAO S Q, et al. The prospect theory based intuitionistic fuzzy three-way decisions model[J]. Journal of Henan Normal University (Natural Science Edition), 2020, 48(5): 31-36.
- [10] SUN L, WANG L Y, DING W P, et al. Neighborhood multi-granulation rough sets-based attribute reduction using Lebesgue and entropy measures in incomplete neighborhood decision systems [J/OL]. Knowledge-Based Systems. <https://www.sciencedirect.com/science/article/pii/S0950705119306240>.

- [11] LIU K, FENG S. An improved artificial bee colony algorithm for enhancing local search ability[J]. Journal of Henan Normal University (Natural Science Edition), 2021, 49(2): 15-24.
- [12] SUN L, ZHANG X Y, QIAN Y H, et al. Joint neighborhood entropy-based gene selection method with Fisher score for tumor classification[J]. Applied Intelligence, 2019, 49(4): 1245-1259.
- [13] LIN Y J, LI Y W, WANG C X, et al. Attribute reduction for multi-label learning with fuzzy rough set[J]. Knowledge-Based Systems, 2018, 152: 51-61.
- [14] SUN L, WANG L Y, QIAN Y H, et al. Feature selection using Lebesgue and entropy measures for incomplete neighborhood decision systems[J/OL]. Knowledge-Based Systems. <https://www.sciencedirect.com/science/article/pii/S0950705119303818>.
- [15] ZHU P F, XU Q, HU Q H, et al. Multi-label feature selection with missing labels[J]. Pattern Recognition, 2018, 74: 488-502.
- [16] SUN L, WANG L Y, DING W P, et al. Feature selection using fuzzy neighborhood entropy-based uncertainty measures for fuzzy neighborhood multigranulation rough sets [J]. IEEE Transactions on Fuzzy Systems, 2021, 29(1): 19-33.
- [17] HAN S M, ZHENG S Q, HE Y S. Open circuit fault diagnosis for inverters based on a greedy algorithm of a rough set[J]. Power System Protection and Control, 2020, 48(17): 122-130.
- [18] LIU J H, LIN Y J, LI Y W, et al. Online multi-label streaming feature selection based on neighborhood rough set[J]. Pattern Recognition, 2018, 84: 273-287.
- [19] HU Q H, YU D R, LIU J F, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences, 2008, 178(18): 3577-3594.
- [20] DUAN J, HU Q H, ZHANG L J, et al. Feature selection for multi-label classification based on neighborhood rough sets[J]. Journal of Computer Research and Development, 2015, 52(1): 56-65.
- [21] YU Y, PEDRYCZ W, MIAO D Q. Neighborhood rough sets based multi-label classification for automatic image annotation [J]. International Journal of Approximate Reasoning, 2013, 54(9): 1373-1387.
- [22] YU Y, PEDRYCZ W, MIAO D Q. Multi-label classification by exploiting label correlations[J]. Expert Systems with Applications, 2014, 41(6): 2989-3004.
- [23] SUN L, YIN T Y, DING W P, et al. Hybrid multilabel feature selection using BPSO and neighborhood rough sets for multilabel neighborhood decision systems[J]. IEEE Access, 2019, 7: 175793-175815.
- [24] WANG C X, LIN Y J, LIU J H. Feature selection for multi-label learning with missing labels[J]. Applied Intelligence, 2019, 49(8): 3027-3042.
- [25] JIANG L, YU G X, GUO M Z, et al. Feature selection with missing labels based on label compression and local feature correlation[J]. Neurocomputing, 2020, 395: 95-106.
- [26] WANG J J, YANG Y L. Multi-label classification algorithm for weak-label data[J]. Computer Engineering and Applications, 2020, 56(5): 65-73.
- [27] YILMAZ T, YAZICI A, KITSUREGAWA M. RELIEF-MM: Effective modality weighting for multimedia information retrieval[J]. Multimedia Systems, 2014, 20(4): 389-413.
- [28] SPOLAOR N, CHERMAN E A, MONARD M C, et al. ReliefF for multi-label feature selection[C]// Proceedings of the IEEE Brazilian Conference on Intelligent Systems. 2013: 6-11.
- [29] 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.
- [30] KONG D G, DING C, HUANG H, et al. Multi-label ReliefF and F-statistic feature selections for image annotation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012: 2352-2359.
- [31] LIN M L, LIU J H, WANG C X, et al. Multi-label feature selection algorithm based on label weighting[J]. Computer Science, 2017, 44(10): 289-295.
- [32] CHEN S B, ZHANG Y M, DING C H Q, et al. Extended adaptive lasso for multi-class and multi-label feature selection[J]. Knowledge-Based Systems, 2019, 173: 28-36.
- [33] HE Z F, YANG M, GAO Y, et al. Joint multi-label classification and label correlations with missing labels and feature selection [J]. Knowledge-Based Systems, 2019, 163: 145-158.
- [34] CHUNG C H, DAI B R. A framework of the semi-supervised multi-label classification with non-uniformly distributed incomplete labels[C]// Proceedings of International Conference on Big Data Analytics and Knowledge Discovery. 2016: 267-280.
- [35] SUN Y Y, ZHANG Y, ZHOU Z H. Multi-label learning with weak label[C]// Proceedings of Twenty-Fourth AAAI Conference on Artificial Intelligence. 2010: 593-598.
- [36] BRAYTEE A, LIU W, CATCHPOOLE D R, et al. Multi-label feature selection using correlation information[C]// Proceedings of the ACM Conference on Information and Knowledge Management. 2017: 1649-1656.
- [37] CAI Z L, ZHU W. Multi-label feature selection via feature manifold learning and sparsity regularization[J]. International Journal of Machine Learning and Cybernetics, 2018, 9(8): 1321-1334.
- [38] CHANG X J, NIE F P, YANG Y, et al. A convex formulation for semi-supervised multi-label feature selection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2014: 1171-1177.



SUN Lin, born in 1979, Ph.D, associate professor, master supervisor. His main research interests include granular computing, big data mining, machine learning and bioinformatics.