文章编号:1671-9352(2024)03-0107-11 DOI:10.6040/j.issn.1671-9352.2.2023.027

基于属性加权的 ML-KNN 方法

温欣[1],李德玉[1,2*]

(1.山西大学计算机与信息技术学院, 山西 太原 030006; 2.山西大学计算智能与中文信息处理教育部重点实验室, 山西 太原 030006)

摘要:提出了一种基于属性加权的 ML-KNN 方法。首先使用变精度邻域粗糙集识别来自每一个标记的决策类非正域中的样本,并构造异质样本对;然后基于属性对异质样本对的区分能力评估不同属性对于分类的重要度;最后计算样本之间的加权距离获得其近邻分布,且基于最大化后验概率的原则实现多标记分类。在 10 个公开的多标记数据集上的实验结果验证了所提方法的有效性。

关键词:多标记分类;属性重要度;邻域粗糙集;分类不确定性;异质样本对

中图分类号:TP391 **文献标志码**:A

引用格式:温欣,李德玉. 基于属性加权的 ML-KNN 方法[J]. 山东大学学报(理学版),2024,59(3):107-117.

The ML-KNN method based on attribute weighting

WEN Xin[1], LI Deyu[1,2*]

(1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China; 2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, Shanxi, China)

Abstract: A ML-KNN method based on attribute weighting has been proposed. To be specific, we first identify samples from the non-positive regions of decision classes by means of the variable precision neighborhood rough set model with respect to each label and construct the heterogeneous sample pairs. Then, the significance of different attributes for classification is evaluated based on their discernibility for the heterogeneous sample pairs. Finally, the weighted distances between samples are calculated in order to obtain the nearest neighbor distributions of samples. At the same time, based on the principle of maximizing the posterior probability, the multi-label classification is implemented. Further, the experimental results on ten public multi-label data sets verify the effectiveness of the proposed method.

Key words: multi-label classification; attribute significance; neighborhood rough set; uncertainty of classification; heterogeneous sample pair

0 引言

传统的单标记学习问题中,一个样本仅被赋予一个标记且候选标记之间存在互斥关系[1]。与此不同,多标记学习中的一个样本能够同时与多个标记关联[2-4]。以实际情形为例,在音乐情感识别中,一首歌曲可能同时包含欢快、喜悦和悲伤的情感[5];在多媒体信息处理中,某一篇新闻报道可能涉及经济、民生、社会治安等多个方面的内容[6];在生物基因功能预测中,某个基因可能包含遗传性状控制、活性调节以及蛋白质合

收稿日期:2023-05-29; **网络出版时间**:2023-12-05 16:01:56
网络出版地址:https://link.cnki.net/urlid/37.1389.N.20231204.1024.004
基金项目:国家自然科学基金资助项目(62072294)
第一作者:温欣(1994—), 女, 博士研究生, 研究方向为多标记学习. E-mail:1368661957@ qq.com
*通信作者:李德玉(1965—), 男, 二级教授,博士生导师,博士, 研究方向为粗糙集、多标记学习等. E-mail:lidysxu@ 163.com

成等多种功能[7]。多标记学习的目标是基于多标记训练样本构造一个能够准确预测未知样本的相关标记模型。

目前,已存在的多标记学习方法主要被分为2类:问题转换方法和算法改编方法[8]。问题转换方法相对而言较为直观且易于理解,其将多标记学习问题转换为一个或者多个传统的单标记学习问题,如:BR(binary relevance)方法[9]、LP(label powerset)方法[10]和CC(classifier chain)方法[11]。BR方法独立地处理每一个标记,对其形成相应的单标记二分类问题,然而这样的过程并未考虑标记间的相关性。LP方法将标记空间中每一种标记组合看作单标记学习问题中的一种类别,以此在学习过程中加入标记相关性信息。CC方法的基本思想是转化多标记学习问题到二分类问题链,其基于指定的标记序列表陆续训练多个单标记分类器,且预测过程中在前的分类器的输出将作为后续分类器输入的一部分。算法改编方法则是通过改编现有的单标记学习算法来实现多标记分类。ML-KNN[12]是一种典型的算法改编方法,其通过改编传统的K最近邻算法实现多标记分类且已被广泛地使用。然而,ML-KNN在分类的过程中并未考虑标记间的相关性,基于此,文献[13]提出了一种2层堆叠ML-KNN方法(Stacked_KNN),其执行了2次ML-KNN分类,且第一层中ML-KNN的标记预测结果被作为第二层中ML-KNN分类的输入特征,以此在分类过程中融合标记相关性。在ML-KNN中,带着相同近邻分布的样本关联于某一标记的概率是相同的,这明显忽略了样本之间的差异,鉴于此,文献[14]提出了一种局部自适应多标记K近邻方法(LAMLKNN)。此外,文献[15]关注于样本近邻分布差异,基于逆向K近邻提出一种邻域自适应多标记K近邻方法。从不同的视角,我们注意到在ML-KNN的分类过程中,当计算样本相似度时,每个属性被无差别地对待,但各属性对于多标记分类的影响可能是不同的,因此,本文致力于评估不同属性对于分类的重要度。

粗糙集理论模拟人类对于客观世界认知的不完备性,用精确的数学工具刻画数据中存在的不确定性,近些年来已被运用在多标记学习中。Yu等[1]提出了一种基于邻域粗糙集的多标记分类方法(MLRS),其同时考虑了标记间的相关性以及特征空间和标记空间之间映射的不确定性。段洁等[16]针对多标记学习问题,重新定义了邻域粗糙集[17-18]的上、下近似,提出了一种基于邻域粗糙集的多标记特征选择方法。张晶等[19]基于模糊粗糙集[17,20]的下近似计算得到测试样本对决策类的隶属度,通过与隶属度阈值的比较实现多标记分类。Qian等[21]将多标记数据中的逻辑标记值转化为标记分布,提出了一种基于粗糙集的标记分布特征选择方法。众所周知,决策类的非正域中的样本具有分类的模棱两可性,其极易影响分类性能表现。基于此,本文通过属性对于来自标记决策类的非正域中异质样本对的区分能力衡量了属性对于分类的重要度,并提出一种基于属性加权的ML-KNN方法(NRS_MLKNN)。进一步,我们比较了NRS_MLKNN与已有的6种多标记分类方法在10个公开的多标记数据集上的实验结果,并验证了所提方法的有效性。

1 预备知识

1.1 邻域粗糙集

给定一个决策表$\langle U,A,D\rangle$,其中$U=\{x_1,x_2,\cdots,x_t\}$为非空有限的样本集,被称为论域,其中$A=\{a_1,a_2,\cdots,a_m\}$为非空有限的条件属性集,D为决策属性。

定义1[16,22] 给定决策表$\langle U,A,D\rangle$,$x\in U,\delta\geqslant 0$,称$\delta(x)=\{y\mid\Delta(x,y)\leqslant\delta,y\in U\}$为$x$的$\delta$邻域,其中$\Delta(x,y)$表示$x$、$y$的欧氏距离。

定义2[18] 给定决策表$\langle U,A,D\rangle$,$\forall x,y\in U$,称$\mathrm{HEOM}(x,y)=\sqrt{\sum_{i=1}^{m}w_{a_i}\times d_{a_i}^2(x,y)}$为$x$、$y$的Heterogeneous Euclidean-Overlap Metric函数,其中:w_{a_i}表示属性a_i的权重,$d_{a_i}(x,y)$表示x和y在属性a_i上的距离。具体来讲,$d_{a_i}(x,y)=\begin{cases}\mathrm{overlap}_{a_i}(x,y)\\ \mathrm{rn\_diff}_{a_i}(x,y)\end{cases}$,其中:$\mathrm{overlap}_{a_i}(x,y)$适用于符号型属性且$\mathrm{overlap}_{a_i}(x,y)=$

$\begin{cases}0, \text{ if } x=y\\ 1, \text{ otherwise}\end{cases}$,而 $\text{rn\_diff}_{a_i}(x,y)$ 适用于数值型属性且 $\text{rn\_diff}_{a_i}(x,y)=\dfrac{|x-y|}{\max_{a_i}-\min_{a_i}}$。

定义 3[18]　给定邻域决策系统 $\text{NDT}=\langle U,C\cup D,R\rangle$,$X_1,X_2,\cdots,X_N$ 是基于决策属性划分的 N 个等价类,$\delta_B(x_i)$ 表示由属性子集 $B\subseteq C$ 生成的 x_i 的邻域信息粒,决策 D 关于属性子集 B 的邻域下、上近似可被定义为:

$$\underline{R_B}D=\bigcup_{i=1}^{N}\underline{R_B}X_i, \tag{1}$$

$$\overline{R_B}D=\bigcup_{i=1}^{N}\overline{R_B}X_i, \tag{2}$$

其中:

$$\underline{R_B}X_i=\{x_i\in U\mid\delta_B(x_i)\subseteq X_i\}, \tag{3}$$

$$\overline{R_B}X_i=\{x_i\in U\mid\delta_B(x_i)\cap X_i\neq\varnothing\}。 \tag{4}$$

决策 D 关于属性子集 B 的下近似也被称作决策正域,记为 $\text{POS}_B(D)=\underline{R_B}D$。决策 D 关于属性子集 B 的边界域被记为 $\text{BN}_B(D)=\overline{R_B}D-\underline{R_B}D$。

定义 4[18]　给定 $A,B\subseteq U$,定义 A 在 B 中的包含度为 $I(A,B)=\dfrac{\text{Card}(A\cap B)}{\text{Card}(A)}$。其中:$A\neq\varnothing$ 且 $\text{Card}(A)$ 表示集合 A 的基数。

为适应不精确的数据环境,同时减轻噪声样本的影响,变精度邻域粗糙集被定义如下。

定义 5[18]　给定邻域决策系统 $\text{NDT}=\langle U,C\cup D,R\rangle$,$X_1,X_2,\cdots,X_N$ 是基于决策属性划分的 N 个等价类,$\delta_B(x_i)$ 表示由属性子集 $B\subseteq C$ 生成的 x_i 的邻域信息粒,α 为精度参数,决策 D 关于属性子集 B 的变精度邻域下、上近似可被定义为:

$$\underline{R_B^{\alpha}}D=\bigcup_{i=1}^{N}\underline{R_B^{\alpha}}X_i, \tag{5}$$

$$\overline{R_B^{\alpha}}D=\bigcup_{i=1}^{N}\overline{R_B^{\alpha}}X_i, \tag{6}$$

其中:

$$\underline{R_B^{\alpha}}X_i=\{x_i\in U\mid I(\delta_B(x_i),X_i)\geqslant\alpha\}, \tag{7}$$

$$\overline{R_B^{\alpha}}X_i=\{x_i\in U\mid I(\delta_B(x_i),X_i)\geqslant 1-\alpha\}。 \tag{8}$$

需要说明,$0.5\leqslant\alpha\leqslant 1$。

1.2　ML-KNN

在多标记分类中,对于一个测试样本 x,ML-KNN 首先通过计算欧氏距离获得其 K 近邻,被表示作 $N(x)$,假设 H_1^j 表示样本 x 关联于第 j 个标记,H_0^j 表示样本 x 无关于第 j 个标记,$E_{n_j^+}^j$ 表示在测试样本 x 的 K 近邻中有 n_j^+ 个近邻关联于第 j 个标记。依据最大化后验概率的原则,测试样本的标记集可由式(9)获得。

$$y_j=\arg\max_{b\in\{0,1\}}P(H_b^j\mid E_{n_j^+}^j),\quad j=1,2,\cdots,q。 \tag{9}$$

利用贝叶斯规则,式(9)可以被写成如下形式:

$$\begin{aligned}y_j&=\arg\max_{b\in\{0,1\}}\frac{P(H_b^j)P(E_{n_j^+}^j\mid H_b^j)}{P(E_{n_j^+}^j)}\\&=\arg\max_{b\in\{0,1\}}P(H_b^j)P(E_{n_j^+}^j\mid H_b^j)。\end{aligned} \tag{10}$$

2　基于属性加权的 ML-KNN 方法

ML-KNN 是一种被广泛使用的多标记分类方法,其在计算欧氏距离获得样本的 K 近邻时,无差别地对待所有的属性,但各属性对于分类的重要度可能是不同的。基于邻域粗糙集模型,我们知道被划分到决策正域外的样本具有分类的模棱两可性,其极易影响分类性能表现,因此,我们考虑尽可能地最大化来自不同决策类的非正域中异质样本之间的差异对于分类而言可能是有益的。进一步,本部分评估了不同属性对于来

自决策类的非正域中的异质样本的区分能力,使其辅助于多标记分类过程,并以此设计了一种基于属性加权的 ML-KNN 方法。

给定一个多标记决策表 $\mathrm{MDS}=\langle U,A,L\rangle$,其中,$U=\{x_1,x_2,\cdots,x_t\}$ 为样本集合;$A=\{a_1,a_2,\cdots,a_m\}$ 为条件属性集合;$L=\{l_1,l_2,\cdots,l_q\}$ 为标记集合。假设 $x_i=[x_{i1},x_{i2},\cdots,x_{im}]$ 表示第 i 个样本的 m 个条件属性值,$Y_i=[y_i^1,y_i^2,\cdots,y_i^q]$ 表示第 i 个样本的 q 个标记值。若 x_i 关联于第 j 个标记,则 $y_i^j=1$;否则 $y_i^j=0$。

对于第 j 个标记,我们定义它的正类样本集如下:

$$P_j=\{x_i\in U\mid y_i^j=1\}。\tag{11}$$

相应地,它的负类样本集被定义为

$$N_j=\{x_i\in U\mid y_i^j=0\}。\tag{12}$$

基于变精度邻域粗糙集模型,可以获得第 j 个标记的决策正类和负类关于属性集 A 的正域,即:

$$\mathrm{POS}_A(P_j)=\{x_i\in U\mid I(\delta(x_i),P_j)\geqslant\alpha\},\tag{13}$$

$$\mathrm{POS}_A(N_j)=\{x_i\in U\mid I(\delta(x_i),N_j)\geqslant\alpha\}。\tag{14}$$

相应地,取

$$\mathrm{PB}_j=P_j-\mathrm{POS}_A(P_j),\tag{15}$$

$$\mathrm{NB}_j=N_j-\mathrm{POS}_A(N_j)。\tag{16}$$

PB_j 和 NB_j 中的样本处于第 j 个标记的决策正类和负类的非正域中,它们是极易被错分的。为了尽可能地最大化这些样本之间的差异,我们建立了 q 个标记的异质样本对区分矩阵 $\boldsymbol{I}=\{I^j\}_{j=1}^q$。对于第 j 个标记的异质样本对区分矩阵 $\boldsymbol{I}^j$,其基本元素可以被计算为

$$I_{ks}^j=\arg\max_{i=1,2,\cdots,m}|x_{ki}-x_{si}|,\tag{17}$$

其中:$1\leqslant k\leqslant|\mathrm{PB}_j|$ 且 $1\leqslant s\leqslant|\mathrm{NB}_j|$。显然,$I_{ks}^j$ 包含了所有对 x_k 和 x_s 有最大区分能力的属性的索引。

基于 q 个异质样本对区分矩阵,对于第 h 个属性,我们定义其相关可区分标记如下:

$$\mathrm{rel\_label}_h=\{l_j\in L\mid\exists I_{ks}^j,\ h\in I_{ks}^j\}。\tag{18}$$

多标记数据的标记分布一般是稀疏的,因此,对于任意一个标记而言,正类样本越多,反映的标记信息越充足。假设 $\boldsymbol{W}_L=[w_{l_1},w_{l_2},\cdots,w_{l_q}]$ 为标记重要度向量,对于第 j 个标记,我们定义其标记重要度如式(19)所示。

$$w_{l_j}=\sum_{i=1}^{t}y_i^j,\quad 1\leqslant j\leqslant q。\tag{19}$$

假设 $\boldsymbol{W}_A=(w_{a_1},w_{a_2},\cdots,w_{a_m})$ 为属性权重向量,任一属性 a_h 的权重被计算为

$$w_{a_h}=\frac{\sum_{j=1}^{q}\sum_{k=1}^{|\mathrm{PB}_j|}\sum_{s=1}^{|\mathrm{NB}_j|}[[h\in I_{ks}^j]]}{\sum_{j=1}^{q}(|\mathrm{PB}_j|\times|\mathrm{NB}_j|)}+\frac{\sum_{\mathrm{num}=1}^{|\mathrm{rel\_label}_h|}w_{\mathrm{rel\_label}_h(\mathrm{num})}}{\sum_{j=1}^{q}w_{l_j}}。\tag{20}$$

在式(20)中,第一项 $\dfrac{\sum_{j=1}^{q}\sum_{k=1}^{|\mathrm{PB}_j|}\sum_{s=1}^{|\mathrm{NB}_j|}[[h\in I_{ks}^j]]}{\sum_{j=1}^{q}(|\mathrm{PB}_j|\times|\mathrm{NB}_j|)}$ 计算了属性 a_h 可区分的来自 q 个标记的决策正类和负类的非正域中的异质样本对的比例,其值越大,说明属性对异质样本对的区分能力越强;第二项 $\dfrac{\sum_{\mathrm{num}=1}^{|\mathrm{rel\_label}_h|}w_{\mathrm{rel\_label}_h(\mathrm{num})}}{\sum_{j=1}^{q}w_{l_j}}$ 中,$\mathrm{rel\_label}_h(\cdot)$ 表示属性 a_h 的某个相关可区分标记,此项反映了属性相关可区分标记的

重要度。总体而言,w_{a_h}的值越大表明属性 a_h 在分类中越重要。

3　算法设计

算法 1　基于属性加权的 ML-KNN 方法(NRS_MLKNN)

输入:多标记训练数据集 $D=\{(x_i,Y_i)\mid 1\leqslant i\leqslant t\}$,邻域参数 δ,精度参数 α,测试样本 x。

输出:x 的预测标记集 y。

1. 使用 max-min 归一化约束样本属性值到[0,1];
2. for each $l_j\in L$, $1\leqslant j\leqslant q$;
3. 通过式(19)计算其重要度 w_{l_j};
4. 基于式(11)和(12)获得标记 l_j 的决策正类和负类样本集 P_j 和 N_j;
5. 通过式(15)和(16)获得来自 l_j 的决策正类和负类的非正域的样本集 PB_j 和 NB_j;
6. for each $l_j\in L$, $1\leqslant j\leqslant q$;
7. for $k=1$ to $|\mathrm{PB}_j|$ do;
8. for $s=1$ to $|\mathrm{NB}_j|$ do;
9. 通过式(17)计算 I^j_{ks};
10. for each $a_h\in A$, $1\leqslant h\leqslant m$;
11. 基于式(18)搜索 a_h 的相关可区分标记 $\mathrm{rel\_label}_h$;
12. 通过式(20)计算 a_h 的权重 w_{a_h};
13. 通过 HEOM 函数计算训练样本之间的加权距离,识别训练样本的近邻分布;
14. 通过 HEOM 函数计算测试样本 x 与训练样本之间的加权距离,识别测试样本的近邻分布;
15. 利用式(9)预测 x 的标记集 y;
16. 返回 y。

假设训练样本集中包含了 t 个样本、m 个属性以及 q 个标记。当计算标记重要度并搜索其决策正类和负类样本集时,时间复杂度为 $O(q)$;对于任意一个标记 l_j,当搜索 PB_j 和 NB_j 时,需要计算欧氏距离来获得训练样本的邻域,其时间复杂度为 $O(t^2)$,假设 $q<t$,所有标记的相关搜索过程的时间复杂度为 $O(t^2)$;对于每一个标记的异质样本对区分矩阵的计算,考虑时间复杂度最大的情况,即对于每一个标记,有$\frac{t^2}{4}$个需要区分的异质样本对(正类和负类的样本各$\frac{t}{2}$),计算过程的时间复杂度为 $O(qt^2)$;预测测试样本标记集的时间复杂度为 $O(t^2)$。综上所述,算法总的时间复杂度为 $O(qt^2)$。

4　实验及结果分析

4.1　实验设置

本文使用 10 个多标记数据集进行了实验,这些数据集的具体信息如表 1 所示,其中:Sample 表示样本的数量,Attribute 表示属性的数量,Label 表示标记的数量,Domain 则表示数据集所属的领域。

本文将提出的方法 NRS_MLKNN 与基于邻域粗糙集的多标记分类方法 MLRS、基于局部标记相关性的多标记分类方法 LPLC[23] 和 ML-KNN,以及 3 种改进的 ML-KNN 方法 Stacked_KNN、LAMLKNN、ML_RKNN,在 HammingLoss、RankingLoss、OneError、Coverage 和 AveragePrecision 这 5 个多标记评估指标上的结果进行了比较。需要说明,各方法在前 8 个数据集上的实验结果基于十折交叉验证获得;对于 Yelp 和 Mediamill 数据集,我们采用其公开的训练和测试数据完成实验。

表 1 数据集描述
Table 1 Description of datasets

Number	Data set	Sample	Attribute	Label	Domain
1	GpositivePseAAC	519	44 0	4	biology
2	Emotions	593	72	6	music
3	Medical	978	1 449	45	text
4	Water-quality	1 060	16	14	chemistry
5	Image	2 000	294	5	image
6	Scene	2 407	294	6	image
7	Yeast	2 417	103	14	biology
8	Business	5 000	438	30	text
9	Yelp	10 810	671	5	text
10	Mediamill	43 907	120	101	video

4.2 参数设置及分析

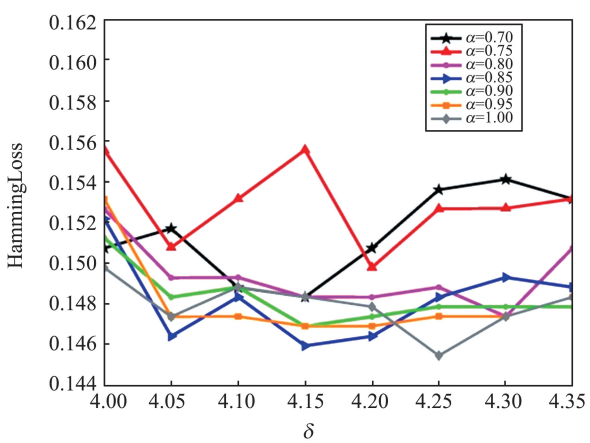
本文涉及 2 个参数,即 δ 和 α。δ 用于控制样本的邻域尺寸, α 作为精度参数,用于减轻噪声样本对分类的影响。考虑到数据集的规模以及他们各自的数据分布情况,我们为采用的数据集设置了不同的参数。

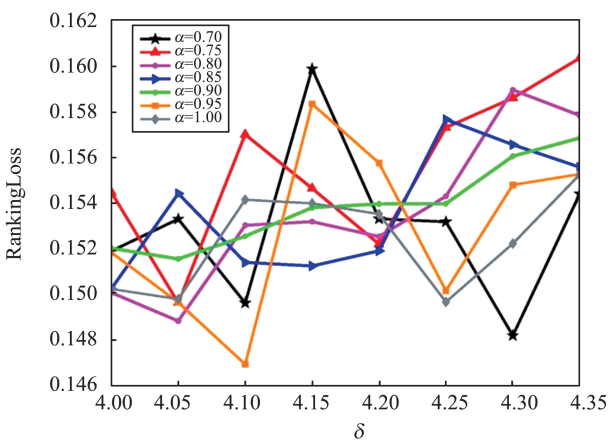
对于 GpositivePseAAC、Emotions、Medical 和 Water-quality,精度参数 α 以步长 0.05 在 0.7 到 1 的范围内变化;Image、Scene、Yeast 和 Business 中,α 以步长 0.05 在 0.8 到 1 的范围内变化;Yelp 和 Mediamill 中,α 则以步长 0.05 在 0.9 到 1 的范围内变化。对于邻域参数 δ,各数据集的参数设置如表 2 所示,变化步长均为 0.05。

表 2 邻域参数 δ 的变化范围
Table 2 The variation range of the neighborhood parameter δ

Data set	δ
GpositivePseAAC	4.00~4.35
Emotions	1.30~1.55
Medical	2.80~3.05
Water-quality	1.50~1.75
Image	4.30~4.60
Scene	2.75~3.00
Yeast	1.25~1.50
Business	1.70~1.95
Yelp	6.00~6.25
Mediamill	2.00~2.35

以 GpositivePseAAC 数据集为例,我们分析了不同的参数设置对算法性能的影响。图 1 显示了各评估指标性能随参数的变化趋势,能明显地观察到,当(δ,α)被设置为$(4.1,0.95)$时,大多数评估指标(RankingLoss、Coverage 和 AveragePrecision)获得了最好的性能表现,因此,对于 GpositivePseAAC,我们选择$(4.1,0.95)$作为其参数设置。从图 1 中可知,当 δ 的值被固定为 4.1,α 取较小的值(如 0.7、0.75)或较大的值(如 1)时,大多数评估指标的性能表现是比较差的,这可以被理解为:当 α 的值较小时,判断样本是否属于决策正域的条件是更宽松的,从而可能使得边界域的样本被误判为正域样本,数据中不确定性信息未能处理完全;当 α 的值较大时,判断样本是否属于决策正域的条件是更为严格的,这同样会使得决策正域和边界域的划分不准确,从而使得算法性能表现较差。当 α 的值被固定为 0.95,δ 取小于或大于 4.1 的值时,大多数评估指标的性能表现均不及 $\delta=4.1$ 时的性能表现,这可以被理解为:不同的 δ 值使得样本获得不同的邻域信息粒,较小或较大的值均可能使获得的样本的邻域信息不准确,从而使得算法性能表现较差。





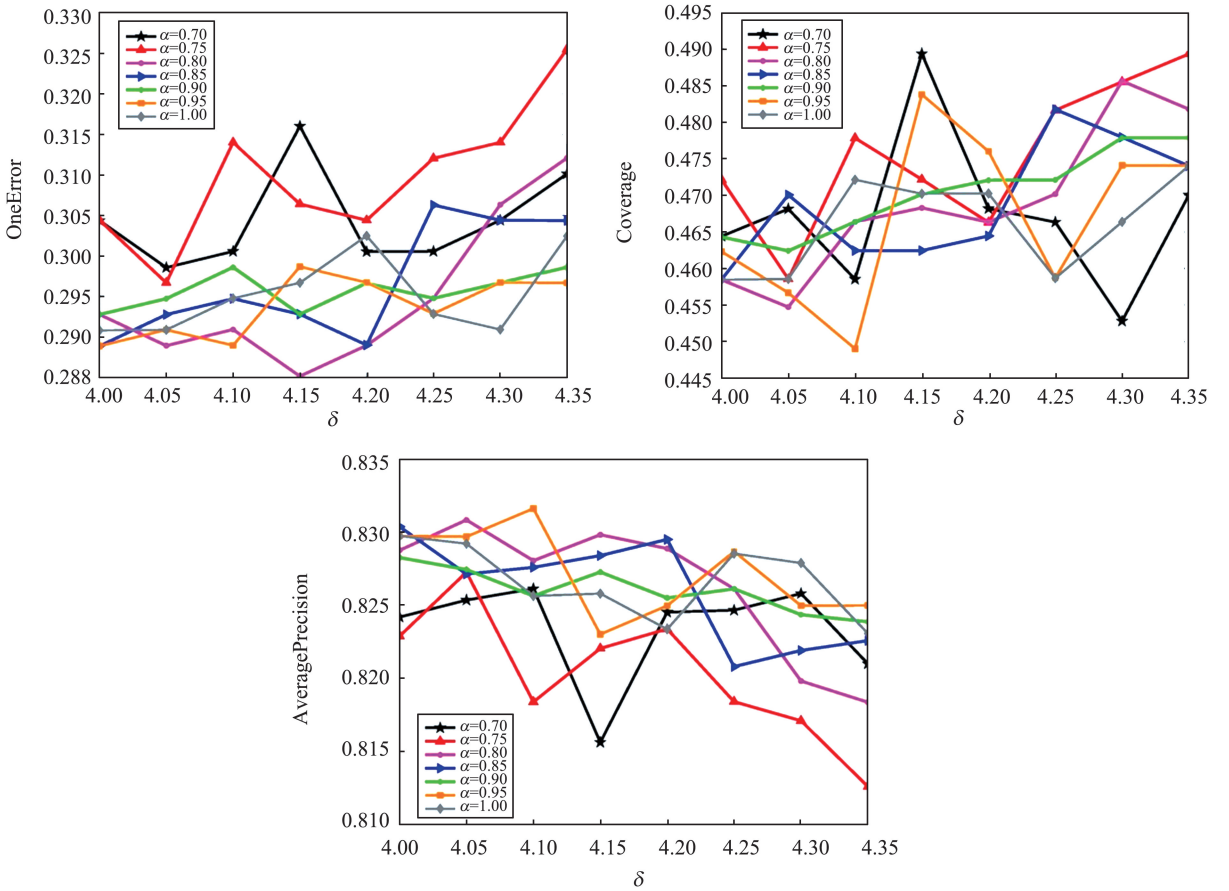


图 1　GpositivePseAAC 数据集中 NRS_MLKNN 在不同参数下的分类性能

Fig.1　The classification performance of NRS_MLKNN influenced by different parameters in GpositivePseAAC dataset

4.3　实验结果分析

7 种方法在各数据集上的实验结果如表 3—12 所示。符号↑表示指标的值越大,分类性能越好;符号↓表示指标的值越小,分类性能越好。此外,最优性能表现以粗体数字显示。

表 3　7 种算法在 GpositivePseAAC 数据集上的分类性能

Table 3　Classification performance of seven algorithms on GpositivePseAAC dataset

Method	HL ↓	RL ↓	OE ↓	CV ↓	AP ↑
MLRS	0.163 8±0.030 3	0.160 6±0.024 1	0.323 7±0.046 7	0.487 5±0.063 1	0.813 7±0.026 1
LPLC	0.164 6±0.024 2	0.162 0±0.036 7	0.285 2±0.059 4	0.464 4±0.107 1	0.824 8±0.036 5
ML-KNN	0.155 1±0.026 7	0.157 2±0.029 6	0.310 2±0.059 6	0.479 9±0.084 4	0.819 5±0.033 6
Stacked_KNN	0.148 3±0.035 2	0.159 1±0.037 9	0.314 0±0.061 3	0.487 5±0.113 6	0.817 5±0.037 8
LAMLKNN	0.154 1±0.029 0	0.149 3±0.025 7	0.292 9±0.055 7	0.454 7±0.073 7	0.829 5±0.028 9
ML_RKNN	0.248 1±0.028 5	0.583 3±0.077 2	**0.233 3±0.044 1**	0.977 2±0.147 6	0.675 7±0.045 4
NRS_MLKNN	**0.147 4±0.029 4**	**0.146 9±0.030 2**	0.289 0±0.064 4	**0.449 0±0.087 4**	**0.831 6±0.035 3**

表 4　7 种算法在 Emotions 数据集上的分类性能

Table 4　Classification performance of seven algorithms on Emotions dataset

Method	HL ↓	RL ↓	OE ↓	CV ↓	AP ↑
MLRS	0.193 0±0.016 0	0.169 6±0.023 9	**0.263 1±0.055 7**	1.804 2±0.146 7	0.801 4±0.027 4
LPLC	0.202 5±0.022 6	0.159 5±0.026 1	0.273 1±0.040 8	1.768 4±0.160 7	0.802 3±0.023 5
ML-KNN	**0.192 5±0.017 2**	0.162 1±0.017 3	0.266 6±0.030 5	1.797 5±0.088 5	0.799 6±0.015 5
Stacked_KNN	0.198 6±0.024 1	0.172 7±0.026 8	0.268 2±0.054 5	1.848 2±0.155 4	0.793 5±0.031 9
LAMLKNN	0.195 0±0.014 8	0.159 5±0.023 3	0.283 2±0.057 4	**1.762 0±0.134 2**	0.800 3±0.026 5
ML_RKNN	0.323 2±0.032 4	0.339 9±0.059 4	0.379 4±0.052 6	2.664 3±0.334 6	0.686 5±0.040 4
NRS_MLKNN	0.195 0±0.012 1	**0.159 2±0.012 3**	0.263 2±0.036 3	1.777 3±0.081 5	**0.803 5±0.014 7**

表 5 7 种算法在 Medical 数据集上的分类性能

Table 5 Classification performance of seven algorithms on Medical dataset

Method	HL ↓	RL ↓	OE ↓	CV ↓	AP ↑
MLRS	0.018 7±0.002 3	0.104 3±0.024 7	0.338 5±0.044 1	3.615 7±1.078 1	0.745 6±0.033 9
LPLC	0.018 8±0.00 2	0.079 5±0.015 9	0.283 3±0.036 7	4.401 5±1.092 2	0.757 8±0.037 8
ML-KNN	0.015 6±0.002 1	0.042 0±0.011 4	0.249 6±0.041 7	2.745 1±0.818 7	0.808 3±0.030 5
Stacked_KNN	0.01 5±0.002 1	0.057 6±0.015 5	0.248 5±0.037 3	3.551 4±1.052 9	0.791 0±0.027 2
LAMLKNN	0.015 9±0.002 1	**0.037 4±0.010 5**	0.244 5±0.042 2	**2.225 2±0.697 5**	0.816 5±0.029 4
ML_RKNN	0.052 2±0.006 7	0.431 0±0.048 3	0.273 0±0.033 3	13.564 3±2.00 4	0.522 4±0.033 9
NRS_MLKNN	**0.014 0±0.002 2**	0.042 4±0.011 2	**0.220 9±0.032 0**	2.795 5±0.787 2	**0.820 4±0.024 0**

表 6 7 种算法在 Water-quality 数据集上的分类性能

Table 6 Classification performance of seven algorithms on Water-quality dataset

Method	HL ↓	RL ↓	OE ↓	CV ↓	AP ↑
MLRS	0.340 8±0.009 8	0.297 8±0.016 1	0.336 9±0.050 4	9.174 5±0.206 0	0.645 7±0.022 6
LPLC	0.316 3±0.009 1	0.263 4±0.015 8	0.284 6±0.042 4	8.889 6±0.220 1	0.684 5±0.019 2
ML-KNN	0.292 0±0.011 2	**0.259 4±0.013 5**	0.293 2±0.052 4	**8.776 4±0.241 2**	0.689 8±0.020 2
Stacked_KNN	0.297 1±0.009 3	0.266 7±0.016 7	0.319 7±0.047 6	8.837 7±0.193 7	0.677 5±0.020 1
LAMLKNN	0.294 7±0.008 9	0.261 8±0.014 0	**0.279 0±0.033 7**	8.853 8±0.278 7	0.688 3±0.018 9
ML_RKNN	0.404 4±0.020 5	0.385 3±0.017 2	0.423 2±0.041 4	10.317 9±0.241 2	0.590 0±0.018 2
NRS_MLKNN	**0.290 4±0.009 7**	0.259 7±0.016 4	0.279 9±0.046 2	**8.776 4±0.256 9**	**0.691 5±0.022 5**

表 7 7 种算法在 Image 数据集上的分类性能

Table 7 Classification performance of seven algorithms on Image dataset

Method	HL ↓	RL ↓	OE ↓	CV ↓	AP ↑
MLRS	0.175 4±0.014 7	0.186 8±0.019 6	0.333 5±0.034 1	0.978 0±0.103 8	0.786 2±0.020 4
LPLC	0.178 4±0.014 2	0.196 8±0.020 7	0.330 0±0.028 7	0.999 5±0.099 3	0.780 8±0.017 1
ML-KNN	**0.170 1±0.014 1**	0.176 5±0.020 2	**0.319 5±0.033 2**	0.978 0±0.103 4	0.790 0±0.020 3
Stacked_KNN	0.176 5±0.016 2	0.188 0±0.023 2	0.333 0±0.030 0	1.018 0±0.115 7	0.780 6±0.022 1
LAMLKNN	0.170 8±0.015 3	0.177 2±0.020 4	0.321 0±0.032 3	0.983 0±0.112 8	0.788 5±0.020 8
ML_RKNN	0.287 1±0.013 9	0.317 4±0.025 9	0.378 0±0.030 3	1.346 5±0.096 4	0.716 7±0.020 3
NRS_MLKNN	0.171 7±0.015 7	**0.174 7±0.021 6**	0.320 0±0.036 1	**0.968 5±0.112 1**	**0.791 5±0.021 9**

表 8 7 种算法在 Scene 数据集上的分类性能

Table 8 Classification performance of seven algorithms on Scene dataset

Method	HL ↓	RL ↓	OE ↓	CV ↓	AP ↑
MLRS	0.092 3±0.006 1	0.099 2±0.011 5	0.253 0±0.017 1	0.539 3±0.064 4	0.848 6±0.011 7
LPLC	0.096 5±0.006 5	0.090 8±0.010 6	0.250 5±0.021 8	0.519 8±0.062 6	0.847 0±0.013 1
ML-KNN	0.085 2±0.008 2	0.076 8±0.009 1	0.226 0±0.015 9	0.470 7±0.059 3	0.866 5±0.009 9
Stacked_KNN	0.087 9±0.005 5	0.085 3±0.008 6	0.232 2±0.013 8	0.515 6±0.056 8	0.859 1±0.008 6
LAMLKNN	0.085 5±0.006 7	**0.074 0±0.008 7**	0.225 2±0.011 8	**0.455 8±0.052 6**	0.867 8±0.008 4
ML_RKNN	0.164 9±0.008 9	0.254 7±0.031 3	0.286 2±0.030 4	1.108 8±0.108 2	0.761 1±0.020 9
NRS_MLKNN	**0.084 7±0.006 4**	0.075 4±0.009 4	**0.220 2±0.015 4**	0.463 7±0.061 8	**0.869 5±0.010 4**

表 9 7 种算法在 Yeast 数据集上的分类性能

Table 9 Classification performance of seven algorithms on Yeast dataset

Method	HL ↓	RL ↓	OE ↓	CV ↓	AP ↑
MLRS	0.204 4±0.009 0	0.181 5±0.009 0	0.240 0±0.020 6	6.391 6±0.213 2	0.748 2±0.013 7
LPLC	0.204 0±0.012 5	0.168 9±0.009 7	0.229 2±0.029 5	6.311 6±0.187 1	0.762 4±0.018 4
ML-KNN	**0.192 7±0.006 6**	0.164 3±0.008 7	0.230 5±0.026 0	6.202 4±0.168 9	0.765 8±0.013 7
Stacked_KNN	0.198 5±0.009 9	0.179 3±0.009 2	0.254 9±0.030 3	6.509 0±0.125 7	0.749 1±0.017 1
LAMLKNN	0.193 8±0.007 0	0.165 1±0.008 4	**0.225 9±0.022 1**	6.222 7±0.153 2	0.765 1±0.013 1
ML_RKNN	0.375 9±0.018 2	0.381 3±0.021 1	0.467 4±0.034 7	9.080 2±0.218 8	0.575 1±0.021 1
NRS_MLKNN	0.192 8±0.006 6	**0.163 4±0.008 3**	0.227 6±0.021 8	**6.196 2±0.164 5**	**0.767 7±0.013 3**

表 10　7 种算法在 Business 数据集上的分类性能
Table 10　Classification performance of seven algorithms on Business dataset

Method	HL ↓	RL ↓	OE ↓	CV ↓	AP ↑
MLRS	0.027 7±0.001 7	0.117 7±0.013 6	0.125 6±0.019 6	4.630 6±0.443 7	0.854 4±0.016 2
LPLC	0.026 7±0.001 9	0.061 2±0.004 7	0.124 6±0.021 0	3.393 0±0.235 9	0.862 6±0.015 9
ML-KNN	0.026 9±0.001 7	0.040 0±0.004 6	0.119 4±0.017 1	2.255 2±0.171 4	0.879 1±0.013 1
Stacked_KNN	**0.026 1±0.001 3**	**0.038 7±0.003 2**	**0.109 6±0.013 4**	2.247 8±0.148 0	**0.883 4±0.009 5**
LAMLKNN	0.026 8±0.001 8	0.040 1±0.004 6	0.119 2±0.019 3	2.265 4±0.171 7	0.879 3±0.013 6
ML_RKNN	0.110 9±0.003 8	0.397 8±0.023 8	0.485 4±0.029 9	13.596 4±0.845 9	0.468 3±0.014 8
NRS_MLKNN	0.026 6±0.001 7	0.038 9±0.003 6	0.115 0±0.017 3	**2.217 0±0.138 1**	0.881 6±0.012 0

表 11　7 种算法在 Yelp 数据集上的分类性能
Table 11　Classification performance of seven algorithms on Yelp dataset

Method	HL ↓	RL ↓	OE ↓	CV ↓	AP ↑
MLRS	0.226 4	0.370 1	0.540 1	0.814 3	0.644 8
LPLC	0.231 4	0.331 7	0.475 8	0.830 6	0.660 9
ML-KNN	0.179 8	0.282 1	0.515 9	0.707 7	0.672 1
Stacked_KNN	0.234 5	0.335 3	0.507 6	0.890 3	0.652 6
LAMLKNN	0.180 4	**0.266 8**	0.500 7	**0.667 3**	**0.683 5**
ML_RKNN	**0.174 8**	0.936 6	**0.049 0**	0.954 1	0.595 6
NRS_MLKNN	0.177 4	0.276 5	0.499 3	0.693 9	0.681 8

表 12　7 种算法在 Mediamill 数据集上的分类性能
Table 12　Classification performance of seven algorithms on Mediamill dataset

Method	HL ↓	RL ↓	OE ↓	CV ↓	AP ↑
MLRS	0.032 8	0.156 7	0.168 4	28.658 7	0.676 7
LPLC	0.035 8	0.091 3	0.150 3	28.761 5	0.682 0
ML-KNN	0.031 5	0.055 0	0.147 3	18.645 6	0.703 4
Stacked_KNN	0.035 0	0.065 0	0.163 7	20.666 7	0.677 6
LAMLKNN	0.031 6	**0.053 3**	0.148 0	**17.907 1**	0.703 2
ML_RKNN	0.044 1	0.700 8	**0.065 3**	57.822 1	0.305 4
NRS_MLKNN	**0.031 4**	0.055 0	0.146 7	18.642 0	**0.703 5**

从上面的表中可知,本文中提出的方法 NRS_MLKNN 在 GpositivePseAAC、Medical、Water-quality、Image、Scene 和 Yeast 这 6 个数据集的大多数评估指标上取得了最优的性能表现,在 emotions 的 RankingLoss 和 AveragePrecision、Business 的 Coverage 以及 Mediamill 的 HammingLoss 和 AveragePrecision 上表现较好,表明 NRS_MLKNN 在多标记分类中的有效性。决策类的非正域中的样本具有分类的模棱两可性,其极易影响分类性能表现。本文评估了属性对于来自决策类非正域中的异质样本的区分能力,且为各属性赋予了不同的权重,目的在于尽可能地区分易混淆的异质样本,在一定程度上提高了分类的准确率。

相比于 MLRS,本文提出的方法在除 emotions 之外数据集的全部评估指标上获得了更好的性能表现,表明邻域粗糙集模型与 ML-KNN 方法结合的有效性。LPLC 在 GpositivePseAAC 的 OneError、Emotions 的 Coverage 以及 Yelp 的 OneError 上的性能表现超越了提出方法的性能表现,且其关注的局部标记相关性是更符合现实数据特征的。ML-KNN 在 Emotions、Water-quality、Image 和 Yeast 的 1—2 个评估指标上获得了最优的性能表现;与此同时,NRS_MLKNN 在所有数据集的大多数评估指标上的性能表现优于 ML-KNN 的性能表现。因此,在采用 ML-KNN 进行多标记分类的过程中,不同的属性对于多标记分类的重要性是不同的,各属性可以基于其对决策类的非正域中的异质样本的区分能力被区别对待。从表 3—12 可以观察到:Stacked_KNN 在 Business 数据集中表现出优越的性能;LAMLKNN 在 Emotions、Medical、Water-quality、Scene、Yeast、Yelp 和 Mediamill 这 7 个数据集中的数个评估指标上表现最好;ML_RKNN 则在 GpositivePseAAC、Yelp 和 Mediamill 的 OneError 上表现突出。相对于这几种改进的 ML-KNN 方法,NRS_MLKNN 的性能表现是更为优越的,这表明在采用 ML-KNN 进行多标记分类的过程中,通过评估属性对于决策类非正域中异质样本的区分能力为属性加权的策略是有效的。

此外,通过表 3—12 中的实验结果,我们计算了各方法在 10 个数据集中 5 个评估指标的性能的平均排序,如图 2 中所示。从图 2 中,我们观察到 NRS_MLKNN 在 5 个评估指标上均获得了最优的平均排序结果,进一步反映了 NRS_MLKNN 在多标记分类中的有效性。就其他方法而言,LAMLKNN 在 5 个评估指标上的性能的平均排序是较靠前的,表明在通过 ML-KNN 进行多标记分类时,处理带有相同近邻分布样本的局部差异对于提高多标记分类的性能表现是有帮助的。

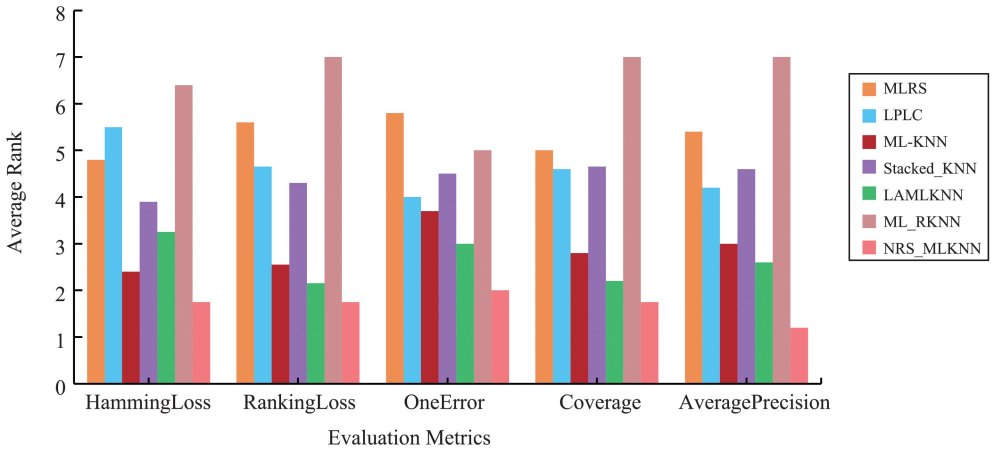


图 2 不同方法在 10 个数据集中 5 个评估指标的平均排序

Fig.2 The average rank of various methods on ten data sets with respect to five evaluation metrics

5 结论

考虑到 ML-KNN 方法在分类过程中无差别地对待不同的属性可能影响分类性能表现,本文中通过变精度邻域粗糙集模型识别了来自于标记决策类非正域中的样本,并构建了异质样本对,利用属性对这些异质样本对的区分能力评估了属性对于分类的重要度,进而提出了一种基于属性加权的 ML-KNN 方法(NRS_MLKNN),与 6 种多标记分类方法的实验比较结果表明,NRS_MLKNN 在多个评估指标上都呈现出更优越的性能表现。

基于邻域粗糙集模型和属性加权,本文有效地分析和处理了多标记数据中的不确定性信息,但标记结构信息目前还未被考虑。众所周知,多标记数据的标记空间蕴含了极其丰富的语义,挖掘标记空间结构并使其辅助于多标记分类过程将是我们下一步工作的重点。

参考文献:

[1] YU Ying, PEDRYCZ W, MIAO Duoqian. Multi-label classification by exploiting label correlations[J]. Expert Systems with Applications, 2014, 41(6):2989-3004.

[2] TSOUMAKAS G, KATAKIS I. Multi-label classification: an overview[J]. International Journal of Data Warehousing and Mining, 2007, 3(3):1-13.

[3] ZHANG Minling, ZHOU Zhihua. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8):1819-1837.

[4] 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.

[5] LEE J, SEO W, PARK J H, et al. Compact feature subset-based multi-label music categorization for mobile devices[J]. Multimedia Tools and Applications, 2019, 78(4):4869-4883.

[6] WANG R, RIDLEY R, SU X A, et al. A novel reasoning mechanism for multi-label text classification[J]. Information Processing and Management, 2021, 58(2):102441.

[7] FABRIS F, FREITAS A A. Dependency network methods for hierarchical multi-label classification of gene functions [C]// 2014 IEEE Symposium on Computational Intelligence and Data Mining. Piscataway: IEEE, 2014:241-248.

[8] AKHAND B, DEVI V S. Multi-label classification of discrete data[C]//IEEE International Conference on Fuzzy Systems. Piscataway: IEEE, 2013:1-5.

[9] BOUTELL M R, LUO J B, SHEN X P, et al. Learning multi-label scene classification[J]. Pattern Recognition, 2004, 37(9):1757-1771.

[10] TSOUMAKAS G, KATAKIS I, VLAHAVAS I P. Random *k*-labelsets for multilabel classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(7):1079-1089.

[11] READ J, PFAHRINGER B, HOLMES G, et al. Classifier chains for multi-label classification [C]//Machine Learning and Knowledge Discovery in Databases. European Conference, Berlin: Springer, 2009, 5782:254-269.

[12] ZHANG Minling, ZHOU Zhihua. ML-kNN: a lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7):2038-2048.

[13] PAKRASHI A, NAMEE B M. Stacked-MLkNN: a stacking based improvement to multi-label *k*-nearest neighbours [C]// First International Workshop on Learning with Imbalanced Domains: Theory and Applications. New York: PMLR, 2017, 74: 51-63.

[14] WANG Dengbao, WANG Jingyuan, HU Fei, et al. A locally adaptive multi-label *k*-nearest neighbor algorithm [C]// Advances in Knowledge Discovery and Data Mining-22nd Pacific-Asia Conference. Berlin: Springer, 2018, 10937:81-93.

[15] SADHUKHAN P, PALIT S. Multi-label learning on principles of reverse *k*-nearest neighbourhood [J/OL]. Expert Systems, 2020. DOI: 10.1111/exsy.12615.

[16] 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.
DUAN Jie, HU Qinghua, ZHANG Lingjun, 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.

[17] 张文修,吴伟志,梁吉业,等. 粗糙集理论与方法[M]. 北京:科学出版社, 2001:232.
ZHANG Wenxiu, WU Weizhi, LIANG Jiye, et al. Rough sets theory and methods [M]. Beijing: Science Press, 2001:232.

[18] HU Qinghua, YU Daren, LIU Jinfu, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences, 2008, 178(18):3577-3594.

[19] 张晶,李德玉,王素格,等. 基于稳健模糊粗糙集模型的多标记文本分类[J]. 计算机科学, 2015, 42(7):270-275.
ZHANG Jing, LI Deyu, WANG Suge, et al. Multi-label text classification based on robust fuzzy rough set model[J]. Journal of Computer Science, 2015, 42(7):270-275.

[20] DAI Jianhua, HU Hu, WU Weizhi, et al. Maximal-discernibility-pair-based approach to attribute reduction in fuzzy rough sets [J]. IEEE Transactions on Fuzzy Systems, 2018, 26(4):2174-2187.

[21] QIAN Wenbin, HUANG Jintao, WANG Yinglong, et al. Label distribution feature selection for multi-label classification with rough set[J]. International Journal of Approximate Reasoning, 2021, 128:32-55.

[22] 温欣,李德玉,王素格. 一种基于邻域关系和模糊决策的特征选择方法[J]. 南京大学学报(自然科学版), 2018, 54(4): 733-741.
WEN Xin, LI Deyu, WANG Suge. A method for feature selection based on neighborhood relation and fuzzy decision[J]. Journal of Nanjing University (Natural Sciences), 2018, 54(4):733-741.

[23] HUANG Jun, LI Guorong, WANG Shuhui, et al. Multi-label classification by exploiting local positive and negative pairwise label correlation[J]. Neurocomputing, 2017, 257:164-174.

(编辑:于善清)