基于邻域粗糙集的多标记属性约简算法

陈盼盼[1,2], 林梦雷[1,2], 刘景华[3], 林国平[1,2]

(1.闽南师范大学 数学与统计学院, 福建 漳州 363000; 2.数字福建气象大数据研究所, 福建 漳州 363000;
3.厦门大学 自动化系, 福建 厦门 361000)

摘　要：在多标记学习中,属性约简是解决多标记数据维数灾难的一个关键技术.针对邻域粗糙集属性约简在计算正域代价较大和多标记数据中标记具有不同的强弱性问题,提出了基于邻域粗糙集的多标记属性约简算法.该算法首先利用样本在整个属性空间下到其异类样本的平均距离与到其同类样本的平均距离的差值对标记进行加权;其次,利用取整函数对样本空间进行划分,提出了一种新的多标记邻域粗糙集快速计算正域的方法;最后,根据前向贪心搜索算法进行属性约简, 以获得一组新的属性排序. 实验给出了 5 个多标记数据集在 4 个评价准则上的对比结果,实验结果分析表明了所提算法的有效性.

关键词：属性约简;标记强弱性;邻域粗糙集;多标记分类

中图分类号：TP181　　**文献标志码**：A　　**文章编号**:2095-7122(2018)04-0001-11

DOI:10.16007/j.cnki.issn2095-7122.2018.04.001

Multi-label Attribute Reduction Algorithm Based on Neighborhood Rough Set

CHEN Panpan[1,2], LIN Menglei[1,2], LIU Jinghua[3], LIN Guopin[1,2]

(1.School of Mathematics and Statistics, Minnan Normal University, Zhangzhou, Fujian 363000, China;
2.Institute of Meteorological Big Data-Digital Fujian, Zhangzhou, Fujian 363000, China;
3.Department of Automation, Xiamen University Xiamen, Fujian 361000, China)

Abstract: In multi-label learning, attribute reduction is capable of eliminating irrelevant and redundant attributes, which is a key technology to solve the curse of dimensionality for multi-label data. However, existing attribute reduction algorithms based on neighborhood rough sets have high cost computation on positive region and the importance of each label is different in multi-label learning. In this work, a multi-label attribute reduction algorithm based on neighborhood rough set is proposed. First, the label is weighted by using the difference between the average distance from a given sample to its heterogeneous sample and the average distance from the sample to its homogeneous sample in all attribute space. Then, the sample space is divided by the integer function, and a new method of quick positive region calculation is proposed for multi-label neighborhood rough set. Finally, all attributes are sorted by the forward greedy attribute reduction algorithm. Given the compare results of five multi-label data evaluated by four evaluation criteria. Experimental results and analysis has shown the effectiveness of the proposed algorithm.

Key words: attribute reduction; label importance; neighborhood rough set; multi-label classification

不同于传统单标记学习,多标记学习的对象不仅由一组属性来刻画其性质,也可能同时由多个类别标记来描述.例如,在音乐分类[1-2]中,一段音乐可能同时属于“摇滚”和“流行音乐”;在视频图像分类[3]中,一副图像可能同时被标记多个语义,如“天空”,“树木”和“小草”.近年来,许多学者对多标记学习进行了大量研究[1-5].

收稿日期：2018-08-24
基金项目：国家青年科学基金项目(N61603173); 福建省自然科学基金项目(2018J01422); 浙江省海洋大数据挖掘与应用重点实验室开放课题(OBDMA201603)
作者简介：陈盼盼(1993-),女,湖北省黄冈市人,硕士研究生在读.

属性空间的高维性和标记空间的复杂性是多标记学习中面临的主要挑战.其中,属性空间高维性通常容易引起维数灾难, 继而会导致多标记分类器的分类性能下降. 属性约简是解决维数灾难的一个重要手段,主要是根据某一度量指标从原始属性集中选择一组最有效属性子集以降低属性空间维数的过程.常见的度量指标一般包括:信息度量[7–10]、距离度量[11]、粗糙集[13–14]等.其中,基于粗糙集的方法已经引起了广泛关注,主要由于该模型的可理解性及较高的近似能力.例如:Hu 等[6]提出了邻域粗糙集属性约简.Duan 等[14]将单标记邻域粗糙集扩展到多标记学习,提出了基于邻域粗糙集的多标记特征选择算法(ARMLNRS).虽然以上所提算法能有效的降低属性空间的维数,但是这些算法在计算正域的代价过大,为此,本文构造了一种新的多标记邻域粗糙集,以快速计算出正域.

多标记学习中,标记空间的复杂性通常表现在:标记缺失[15–16],流标记[17–18],标记不平衡性[19–20]和标记强弱性[12, 21–22]等.如果充分考虑标记具有不同强弱性,可能更有利于学习,尤其是当对象的标记个数极多时,标记的强弱性可能会为多标记学习提供更多的有用信息.例如:一副图像被标记了“some building”,“mostly sky”和“much water”,与直接标记“building”,“sky”和“water”相比更有利于多标记学习(如图 1 所示).因此,在多标记学习问题中如何有效地利用标记的强弱性成为一个重要研究问题.

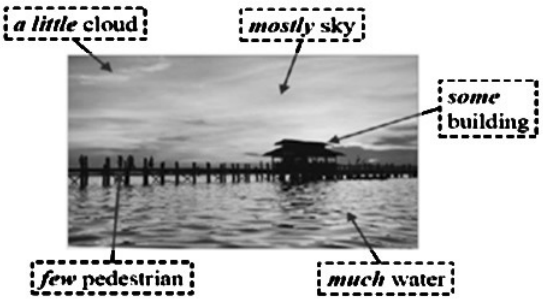

图 1 图像标注

本文针对多标记数据中标记强弱性的不同和邻域粗糙集在计算正域时代价大的问题,提出了基于邻域粗糙集的多标记属性约简算法.首先,根据标记对样本的划分能力的不同,本文采取样本到其异类样本的平均距离和到其同类样本的平均距离的差值对标记赋予权重,在同一属性空间下,差值越大,说明该标记下不同类别的样本分布较远,同类别样本分布较近,即标记对样本的区分能力越强,反之,差值越小,则该标记对样本的区分能力越弱.其次,利用取整函数值对样本空间进行划分作为不同的桶,在当前属性空间下每个桶里的样本的邻域只在其相邻的桶里面,若该桶的邻域样本与该样本决策相同则为正域.最后,根据最大重要度前向贪心搜索策略进行属性约简.在多标记数据集上的实验结果表明,本文提出的邻域粗糙集的多标记属性约简算法不仅降低属性空间维数而且能有效地提升多标记分类器的分类性能.

本文第 1 节主要介绍邻域粗糙集的基础知识;第 2 节设计了标记权重模型和多标记邻域粗糙集模型及快速计算正域的方法;第 3 节对所提出的算法进行实验验证及结果分析;第 4 节总结全文.

1 邻域决策系统

给定一个决策信息系统 $NDT=<U, C, D>$, $U=\{x_1, x_2, \cdots, x_n\}$ 表示全部样本的集合, $C=\{a_1, a_2, \cdots, a_m\}$ 表示描述样本的条件属性集合, D 是决策属性集合.

定义 1[6] 设 U 是非空样本集合,若对 $\forall x_i, x_j, x_k \in U$ 都存在唯一确定的实函数△与之对应,且△满足:

1) $\triangle(x_i, x_j)\geq 0$ 当且仅当 $x_i=x_j$, $\triangle(x_i, x_j)=0$;

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

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

称$\triangle$是 U 上的距离函数,并且称$(U, \triangle)$为度量空间.

定义 2[6] 设$<U, \triangle>$是非空度量空间,$\forall x\in U, \delta\geq 0$ 有

$$\delta(x)=\{y|\triangle(x, y)\leq\delta, y\in U\}, \tag{1}$$

则称 $\delta(x)$为 x 的 δ 邻域.

定义 3[6] 给定 $NDT=<U, C, D>$,如果 C 生成一簇邻域关系,则称$NDT=<U, C, D>$为邻域决策系统.

定义 4[6] 给定邻域决策系统 $NDT=<U, C, D>$, D 将 U 划分成 N 个等价类:$X_1, X_2, \cdots X_N$, $B\subseteq C$ 生成 U 上的邻域关系 N_B,则决策 D 关于 B 的邻域下近似和上近似分别为:

$$\underline{N_B}D=\{\underline{N_B}X_1, \underline{N_B}X_2, \cdots, \underline{N_B}X_N\},$$
$$\overline{N_B}D=\{\overline{N_B}X_1, \overline{N_B}X_2, \cdots \overline{N_B}X_N\},$$

其中

$$\underline{N_B}X=\{x_i\,|\,\delta_B(x_i)\subseteq X, x_i\in U\},$$
$$\overline{N_B}X=\{x_i\,|\,\delta_B(x_i)\cap X\neq\varnothing, x_i\in U\}.$$

正域:$POS_B(D)=\underline{N_B}D$.

定义 5[6] 给定 $NDT=<U, C, D>$,决策属性 D 对条件属性 $B\subseteq C$ 的依赖度为:

$$\gamma_B(D)=\frac{Card(\underline{N_B}D)}{Card(U)}=\frac{|POS_B(D)|}{|U|}.$$

定义 6[6] 给定 $NDT=<U, C, D>$,$B\subseteq C$,如果属性 B 满足:

1) $\gamma_B(D)=\gamma_C(D)$;

2) $\forall a\in B$, $\gamma_{B-|a|}(D)<\gamma_B(D)$,

则称 B 是 C 的一个属性约简.

属性 $a\in C-B$ 在条件属性 B 上相对于决策属性 D 的重要度定义为:

$$sig(a, B, D)=\gamma_{B\cup|a|}(D)-\gamma_B(D).$$

2　基于邻域粗糙集的多标记属性约简模型

假设 $X\subset R^d$ 表示样本的输入空间,任何实例都可以表示为 d 维向量 $x=[x^1, x^2, \cdots, x^d]$ $(x\in X)$,令 $Y=\{l_1, l_2, \cdots, l_m\}$ 表示一个有限的标记集合, m 表示标记个数.每个样本都与 Y 的一个 m 维标记子集 $y=[y^1, y^2, \cdots, y^m]$相关联,当且仅当样本 x 具有标记 l_j 时, $y^j=1$,没有该标记时, $y^j=0$.令 S 是由 n 个标记样本组成的训练集,其中 $S=\{(x_i, y_i)|\,1\leq i\leq n, x_i\in X, y_i\subseteq Y\}$.

2.1 标记权重模型

在多标记学习框架中,每个样本可能同时被一个或多个类别标记所标注,在同一属性空间下的每个类别标记对样本的区分性各不相同.因此,考虑到每个类别标记对样本区分能力的差异,本文利用样本在整个属性空间中到其异类样本的平均距离减去其到同类样本的平均距离的差值作为权重来衡量标记的强弱性程度.

定义 7 给定样本空间 $U=\{x_1,x_2,\cdots,x_n\}$,描述样本的属性空间 $A=\{a_1,a_2,\cdots,a_p\}$,样本的标记集合 $L=\{l_1,l_2,\cdots,l_m\}$,对于$\forall l\in L$,在属性空间 A 下对类别标记 l 赋予的权重定义为:

$$\omega_l=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\sum_{s=1}^{|MS^l(x_i)|}d_A\left(x_i,MS^l(x_i)_s\right)}{|MS^l(x_i)|}-\frac{\sum_{t=1}^{|HS^l(x_i)|}d_A\left(x_i,HS^l(x_i)_t\right)}{|HS^l(x_i)|}\right),\tag{2}$$

其中,距离函数定义为:

$$d_A(x,y)=\sqrt{\sum_{a=1}^{p}(x(a)-y(a))^2},\tag{3}$$

式(2)中,$d_A(x,y)$表示样本 x 和样本 y 在属性空间 A 上的距离,$MS^l(x)$表示标记 l 下样本 x_i 的异类样本的集合,$|MS^l(x_i)|$表示集合 $MS^l(x_i)$的元素个数,表示集合 $MS^l(x_i)$中依次排列的第 s 个样本;$HS^l(x_i)$表示标记 l 下样本 x_i 的同类样本的集合,$|HS^l(x_i)|$表示集合 $HS^l(x_i)$的元素个数,$HS^l(x_i)_i$ 表示集合 $HS^l(x_i)$中依次排列的第 t 个样本.式(3)中,$x(a)$和 $y(a)$分别表示样本 x 和样本 y 在属性 a 上相应的属性值.

2.2 多标记邻域粗糙集模型

定义 8 在多标记邻域决策系统 $NDT=<U,A\bigcup L>$中,样本的标记集合 $L=\{l_1,l_2,\cdots,l_m\}$, $\forall l_k\in L$, $X_1,X_2,\cdots,X_{N_k}$ 表示样本在标记 l_k 下的划分,$C\subseteq A$,多标记邻域粗糙集的下近似和上近似定义为

$$\underline{N_C}L=\bigcap_{k=1}^{m}\underline{N_C}l_k,$$

$$\overline{N_C}L=\bigcap_{k=1}^{m}\overline{N_C}l_k,$$

其中 $\underline{N_C}l_k=\bigcup_{j=1}^{N_k}\underline{N_C}X_j=\bigcup_{j=1}^{N_k}\{x_i\mid\delta_C(x_i)\subseteq X_j,x_i\in U\}$; $\overline{N_C}l_k=\bigcup_{j=1}^{N_k}\overline{N_C}X_j=\bigcup_{j=1}^{N_k}\{x_i\mid\delta_C(x_i)\cap X_j\neq\phi,x_i\in U\}$.

正域:$POS_C(L)=\underline{N_C}L$.

由以上定义可以发现在计算每个标记下的正域时需要反复计算和保存每个样本的邻域,导致计算正域代价较大.因此,本节利用桶的思想定义了多标记邻域粗糙集正域计算方法.

定义 9 多标记邻域决策系统 $MNDT=<U,A\bigcup L>$,属性空间 $A=\{a_1,a_2,\cdots,a_p\}$,标记集合 $L=\{l_1,l_2,\cdots,l_m\}$, x_0 是论域 U 中构造的一个特殊样本,在标记 l_k 下, $a\in A$, $x_0^{l_k}(a)=\min\{x_i^{l_k}(a),x_i\in U\}$给定距离

度量 δ，论域 U 中的样本被划分到有限桶 $B_0^{l_k}, B_1^{l_k}, \cdots, B_r^{l_k}$ 中，

$$B_r^{l_k} = \{x_i \mid x_i \in U \wedge \lceil d(x_0, x_i)/\delta \rceil = r\},$$

其中，$x_0^{l_k}(a)$ 表示在类别标记 l_k 下样本 x_0 在属性 a 上相应的属性值.

定理 1 给定 $MNDT=<U, A\bigcup L>$，论域 U 被划分到有限桶 $B_0^{l_k}, B_1^{l_k}, \cdots, B_r^{l_k}$ 中，在标记 l_k 下，对于 $x_i \in U$，记 x_i 的 δ 邻域为 $\delta(x_i)$，有

1) 若 $x_i \in B_q^{l_k}$，则 $\delta(x_i) \subset B_{q-1}^{l_k} \bigcup B_q^{l_k} \bigcup B_{q+1}^{l_k} (q=1,2,3,\cdots,r-1)$；

2) 若 $x_i \in B_0^{l_k}$，则 $\delta(x_i) \subset B_0^{l_k} \bigcup B_1^{l_k}$；

3) 若 $x_i \in B_r^{l_k}$，则 $\delta(x_i) \subset B_{r-1}^{l_k} \bigcup B_r^{l_k}$.

证明 假设样本 $x_i \in B_q^{l_k}$，由定义 9，有

$$(q-1)\delta < d(x_i, x_0) \le q\delta, \tag{3}$$

若样本 $x' \in B_{q+2}^{l_k}$，有

$$(q+1)\delta < d(x', x_0) \le (q+2)\delta, \tag{4}$$

由(3)(4)得 $d(x', x_0) - d(x_i, x_0) > \delta$,

由定义 1 可得 $d(x_i, x') > d(x', x_0) - d(x_i, x_0)$，即 $d(x_i, x') > \delta$.

故 x' 不在 x_i 的 δ 邻域内，即不是 x_i 的邻域元素.同理可证，$x' \in B_{q\pm j}^{l_k} (j=2, 3, \cdots)$ 也不在 x_i 的邻域内.综上，对于 $\forall x_i \in B_q^{l_k}$，$x_i$ 的邻域只在 $B_{q-1}^{l_k} \bigcup B_q^{l_k} \bigcup B_{q+1}^{l_k}$ 内.

定义 10 给定 $MNDT=<U, A\bigcup L>$，样本的标记集合 $L=\{l_1, l_2, \cdots, l_m\}$，$\forall l_k \in L$，论域 U 被划分成有限桶 $B_0^{l_k}, B_1^{l_k}, \cdots, B_r^{l_k}$，$C \subseteq A$，$B_{q-1}^{l_k} \bigcup B_q^{l_k} \bigcup B_{q+1}^{l_k}$ 表示 x_i 的邻域桶，定义该多标记邻域粗糙集决策系统在标记 l_k 下的下近似和上近似为：

$$\underline{N_C} l_k = \{x_i \mid \forall x_j \in (B_{q-1}^{l_k} \bigcup B_q^{l_k} \bigcup B_{q+1}^{l_k}) \wedge (x_i^{l_k} = x_j^{l_k}),\ x_i \in U\},$$
$$\overline{N_C} l_k = \{x_i \mid \forall x_j \notin (B_{q-1}^{l_k} \bigcup B_q^{l_k} \bigcup B_{q+1}^{l_k}) \vee (x_i^{l_k} \ne x_j^{l_k}),\ x_i \in U\}.$$

其中，$x_i^{l_k}$ 和 $x_j^{l_k}$ 分别表示在标记 l_k 下样本 x_i 和样本 x_j 的值.

标记 l_k 下的正域：$POS_C(l_k) = \underline{N_C} l_k$.

定义 11 给定 $MNDT=<U, A\bigcup L>$，样本的标记集合 $L=\{l_1, l_2, \cdots, l_m\}$，$C \subseteq A$，该多标记邻域决策系统标记集合 L 对条件属性集 C 的依赖度定义为：

$$\gamma_C(L) = \frac{\sum_{k=1}^{m} |POS_C(l_k)|}{m \times |U|}.$$

定义 12 给定 $MNDT=<U, A\bigcup L>$，样本的标记集合 $L=\{l_1, l_2, \cdots, l_m\}$，标记权重集合 $\omega_L=\{\omega_1, \cdots, \omega_m\}$，多标记邻域决策系统属性 $a \in A-C$ 在条件属性 C 上相对于标记集合 L 的重要度定义为：

$$sig_\gamma(a, C, L) = [\gamma_{C\cup\{a\}}(L) - \gamma_C(L)]\ \omega_L.$$

2.3 基于邻域粗糙集的多标记属性约简算法

本文提出了基于邻域粗糙集的多标记属性约简算法.该算法的主要思想为:首先,对于已选的类别标记 l, 计算每个样本在整个属性空间下到其异类样本的平均距离减去到其同类样本的平均距离的差值作为类别标记 l 的权重;其次,利用定义 9 和定义 10 计算出每个类别标记下的正域;最后,利用最大重要度的前向贪心搜索策略得到一组属性的排序.

算法 1 基于邻域粗糙集的多标记属性约简算法

输入:多标记数据集 D 和邻域参数 δ

输出:约简 $reduct$

1: 初始化 $\varnothing \rightarrow reduct$

2: $for\ \ each\ \ l \in L$

3: 根据定义 7 计算每个类别标记 l 的权重 ω_l;

4: end

5: 对标记集合的权重 $\omega_L=\{\omega_1, \cdots, \omega_m\}$ 归一化;

6: $for\ \ each\ \ a_i \in A-reduct$

7: 结合定义 5、定义 10 和定义 11 计算所有标记下属性依赖度;

8: 结合定义 7 和定义 12 计算属性的重要度;

9: $end\ \ for$

10: 选择 a_k,满足

$sig(a_k, reduct, L)=max(sig(a_i, reduct, L))$;

11: $if\ \ sig(a_k, reduct, L)>0$

$reduct \bigcup a_k \rightarrow reduct$;

12: $else$

13: 返回 $reduct$

14: $end\ \ if$

15: $return\ \ reduct$

3 实验设计与结果比较

3.1 实验数据

本文实验从不同领域中选取了 5 个多标记数据集以此验证所提算法的有效性.Cal500 数据集用来描述音乐和声音效果的语义注释和检索, 包含 502 个样本、68 个特征属性、174 个标记.Flags 数据集来源于图像处理与分类, 它包含 194 个样本、19 个属性、7 个标记. Emotions 数据集来源于某大学音乐学院 3 名年龄为 20 岁、25 岁和 30 岁的男性专家的音频特性, 包含 593 个样本、72 个属性、6 个标记. Yeast 数据集描述了生物学酵母细胞中基因的活性分类, 包含了 2 417 个样本、103 个属性、14 个标记.Recreation 数据

集属于网页类数据集，包含 5 000 个文本，各数据集的描述信息见表 1.

表 1 多标记数据集描述

Data sets	Instances	Attributes	Labels	Train	Test
Cal500	502	68	174	251	251
Flags	194	19	7	129	65
Emotions	593	72	6	391	202
Yeast	2 417	103	14	1 499	918
Recreation	5 000	606	22	2 000	3 000

3.2 评价指标

本文采用平均精度(Average Precision, AP)、汉明损失(Hamming Loss, HL)、排序损失(Ranking Loss, RL)、单错误(One Error, OE)作为评价算法分类性能的指标.

实验令测试集 $Z=\{(x_i,\mathrm{Y}_i)\}_{i=1}^{m}\subset R^d\times\{+1,-1\}^L$，根据预测函数 $f_l(x)$ 可定义标记 l 的排序 $rand_f(x,l)\in\{1,\cdots L\}$.

平均精度(AP)：评估样本的分类标记集合超过某个标记并仍在该样本标记集中的平均概率，定义为

$$avgPre(f)=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|R_i|}\sum_{l\in R_i}\frac{\{k\mid rank_f(x_i,k)\le rank_f(x_i,l),k\in R_i\}}{rank_f(x_i,l)}.$$

汉明损失(HL)：评估每个实例在单个类别标记下被错误预测的平均次数，定义为

$$hLoss(h)=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{L}\sum_{l=1}^{L}[h_l(x_i)\neq Y_{il}].$$

排序损失(RL)：用于评估实例样本的预测标记集合中不相关的标记排名高于相关标记的次数，定义为

$$rLoss(f)=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{|R_i||\overline{R}_i|}|\{(l,k)\mid rank_f(x_i,l)\ge rank_f(x_i,k),\ (l,k)\in R_i\times\overline{R}_i\}|.$$

单错误(OE)：评估排在实例预测标记的最前列的标记不在该实例的相关标记集合的平均概率，定义为

$$oneError(f)=\frac{1}{m}\sum_{i=1}^{m}\max_{l\in R_i}rank_f(x_i,l)-1,$$

其中，$l_i=\arg\max_{k\in\{1,\cdots,L\}}f_k(x_i)$.

另外，$R_i=\{l\mid Y_{il}=+1\}$表示与样本 x_i 相关的标记构成的集合，$R_i=\{l\mid Y_{il}=-1\}$则是表示与样本 x_i 不相关的类别标记构成的集合.

以上这 4 个性能评价指标，AP 指标值越大，说明分类的性能越好，最佳值为 1；HL、RL、OE 指标值越小，说明分类的性能越好，最佳值为 0.

3.3 实验设置

为了有效评价实验算法的性能，本文选择 5 种不同类型的对比算法，包括 MDDMspc[13]，MDDMproj[13]，MLNB[9]，PMU[23]和 RF-ML[11].其中算法 ARNRS_ML 和 MDDMspc，MDDMproj，PMU 和 RF-ML 得到的是一组属性的排序.因此，在实验中属性排序的前 k 个属性被设置为属性子集，并设 k 为 MLNB 算法得到的属

性的数目.此外,该实验采用 ML-KNN[24]算法来评价属性约简后的数据集.其中,设置 ML-KNN 的平滑参数 $s=1$,邻域个数 $k=10$.

3.3.1 邻域参数设定

本文中使用 ML-KNN 多标记分类器研究平均分类精度随邻域 $\delta(0.01-0.09)$的变化.以 Emotions 和 Flags 为例,得到的实验结果如图 2 和图 3 所示.

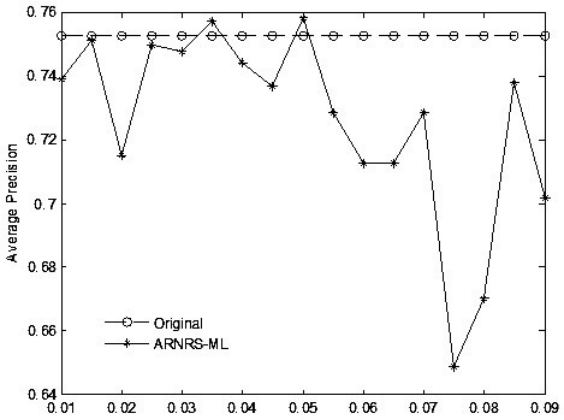

图 2 精度随邻域变化(Emotions)

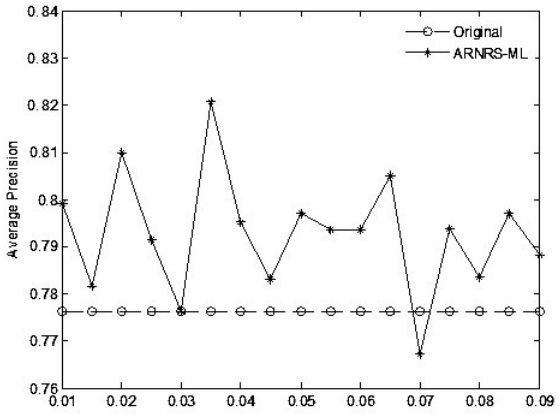

图 3 精度随邻域变化(Flags)

从图 2 和图 3 可以看出,对于 Emotions 数据集来说邻域参数在 0.035 和 0.05 的平均分类精度最佳,实验选择 0.05.对于 Flags 数据集来说,邻域参数的平均分类精度在 0.035~0.06 之间是最佳的,最终选择 0.035.同样,可以得到 Cal500,Yeast 和 Recreation 数据集的邻域参数分别为 0.01,0.01 和 0.03.

3.4 实验结果与分析

为了验证 ARNRS_ML 算法的有效性,实验对比各算法在属性子集上的分类性能.

表 2-表 5 列出了 6 种属性约简算法分别在 4 种评价指标上的实验结果."↑" 表示指标值越高越好,"↓"表示指标值越低越好.此外,黑体表示各算法取得最佳值,斜体表示各算法的平均分类性能.

表 2 AP 指标下各算法的性能比较(↑)

Data sets	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	ARNRS_ML
Cal500	0. 483 5	0. 484 1	0. 478 9	0. 475 0	0. 477 6	0. 480 8
Flags	0. 782 7	0. 782 7	0. 799 6	0. 785 7	0. 776 2	0. 820 7
Emotions	0. 729 3	0. 741 6	0. 732 7	0. 684 3	0. 752 9	0. 758 4
Yeast	0. 713 9	0. 713 9	0. 730 0	0. 735 0	0. 735 5	0. 744 8
Recreation	0. 405 1	0. 402 8	0. 437 3	0. 392 4	0. 479 0	0. 512 9
Average	*0. 622 9*	*0. 625 0*	*0. 631 7*	*0. 614 5*	*0. 644 2*	*0. 663 5*

表 3 HL 指标下各算法的性能比较(↓)

Data sets	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	ARNRS_ML
Cal500	0. 139 4	0. 140 7	0. 138 9	0. 140 1	0. 142 6	0. 139 1
Flags	0. 698 9	0. 698 9	0. 312 1	0. 327 5	0. 709 9	0. 309 9
Emotions	0. 264 9	0. 243 4	0. 276 4	0. 299 5	0. 245 0	0. 241 7
Yeast	0. 231 3	0. 231 3	0. 218 3	0. 216 9	0. 208 0	0. 208 7
Recreation	0. 064 7	0. 064 8	0. 063 6	0. 065 0	0. 061 1	0. 057 9
Average	*0. 279 8*	*0. 275 8*	*0. 201 9*	*0. 209 8*	*0. 273 5*	*0. 191 5*

表 4 RL 指标下各算法的性能比较(↓)

Data sets	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	ARNRS_ML
Cal500	0. 190 9	0. 189 3	0. 190 6	0. 188 4	0. 191 3	0. 188 8
Flags	0. 261 3	0. 261 3	0. 240 0	0. 250 3	0. 251 8	0. 214 1
Emotions	0. 233 7	0. 215 2	0. 227 8	0. 302 8	0. 205 5	0. 199 0
Yeast	0. 205 0	0. 205 0	0. 195 7	0. 189 2	0. 187 1	0. 182 5
Recreation	0. 207 1	0. 206 5	0. 191 8	0. 213 4	0. 187 9	0. 183 8
Average	*0. 219 6*	*0. 215 5*	*0. 209 2*	*0. 228 8*	*0. 204 7*	*0. 193 6*

表 5 OE 指标下各算法的性能比较(↓)

Data sets	MDDMspc	MDDMproj	RF-ML	PMU	MLNB	ARNRS_ML
Cal500	0. 111 6	0. 119 5	0. 111 6	0. 127 5	0. 143 4	0. 107 6
Flags	0. 200 0	0. 200 0	0. 215 4	0. 246 2	0. 276 9	0. 169 2
Emotions	0. 386 1	0. 376 2	0. 391 1	0. 440 6	0. 376 5	0. 341 6
Yeast	0. 247 3	0. 247 3	0. 249 5	0. 245 1	0. 256 0	0. 236 4
Recreation	0. 772 0	0. 771 7	0. 735 0	0. 786 3	0. 664 3	0. 623 7
Average	*0. 343 4*	*0. 342 9*	*0. 340 5*	*0. 381 4*	*0. 343 4*	*0. 295 7*

根据以上实验结果,可以发现:

1)表 2 中,对于 AP 指标,ARNRS_ML 算法除了在数据集 Cal500 上略差于 MDDMproj 外,在其他 4 个数据集上 ARNRS_ML 均取得最优值;表 3 中,对于 HL 指标,在 Flags、Emotions 和 Recreation 这 3 个数据集上损失值都最低,因此,性能最佳,其性能略低于 Cal500 数据集和 Yeast 数据集上的 RF-ML 和 MLNB 算法,但仅比 Cal500 数据集在 RF-ML 算法上的损失值高 0. 000 2,比 Yeast 数据集在 MLNB 算法上的损失值高 0. 000 7;表 4 中,对于 RL 指标,仅在数据集 Cal500 上 ARNRS_ML 算法性能略低于 PMU,两者的性能相差 0. 000 4, 但在其他 4 个数据集上 ARNRS_ML 算法相比另外的 4 种对比算法都取得最优; 表 5 中,对于 OE 指标,ARNRS_ML 算法在所有实验数据集上都取得最小的 OE 值,即分类性能均取得最优.

2)从表 2-5 统计的 5 个数据集在 4 个评价指标下的 20 种实验对比结果可知,与 ARNRS_ML 对比,MDDMspc 在 20 种结果中没有优胜的情况,其他 4 种对比算法胜出的结果均占 5%,ARNRS_ML 算法相比于这 5 种对比算法胜出的结果占 80%.

3)在平均分类性能方面,ARNRS_ML 在以上 4 个评价指标上都取得了最佳值.

综上所述,以上所有分析与实验结果都说明了所提的算法 ARNRS_ML 的分类性能最优.

4 总结

本文针对多标记邻域粗糙集的标记强弱性以及在传统的属性约简算法上正域的反复计算导致代价较大的问题,提出了一种基于邻域粗糙集的多标记属性约简算法.该算法设计多标记权重模型来探究类别标记对样本的区分性程度,并重新构造了多标记邻域粗糙集,以进行正域计算.这种新的多标记邻域粗糙集正域计算方法在一定程度上降低了计算正域代价大的问题. 通过 5 个多标记数据集在 4 种不同的评价准则上的对比实验结果表明,本文提出的 ARNRS_ML 算法优于其他 5 种对比算法.

参考文献:

[1] WANG J C, WANG H M, LIN S D, et al. Cost-sensitive multi-label learning for audio tag annotation and retrieval[J]. IEEE Transactions on Multimedia, 2011, 13(3): 518-529.

[2] FURNKRANZ J, HULLERMEIER E, BRINKER K, et al. Multi-label classification via calibrated label ranking[J], Machine Learning, 2008, 73(2): 133-153.

[3] 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.

[4] LI F, MIAO D Q, PEDRYCZ W. Granular multi-label feature selection based on mutual information[J]. Pattern Recognition, 2017, 67(C): 410-423.

[5] 郑希源, 张化祥. 基于局部近邻相关性的多标记算法[J].计算机科学, 2014, 41(2): 123-126.

[6] 胡清华, 赵辉, 于达仁. 基于邻域粗糙集的符号与数值属性快速约简算法[J]. 模式识别与人工智能, 2008, 21(6): 732-738.

[7] 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.

[8] 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.

[9] 王晨曦, 林梦雷, 刘景华, 等. 融合特征排序的多标记特征选择算法[J]. 计算机工程与应用, 2016, 52(17): 93-100.

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

[11] SPOLAOR N, CHERMAN E, MONARD M, et al. ReliefF for multi-label feature selection[C]//2013 Brazilian Conference on Intelligent Systems (BRACIS). IEEE, 2013: 6-11.

[12] REYES O, MORELL C, VENTURA S. Scalable extension of the ReliefF algorithm for weighting and selection features on the multi-label learning context[J]. Neurocomputing, 2015, 161 (C): 168-182.

[13] ZHANG L J, HU Q H, DUAN J, et al. Multi-label feature selection with fuzzy rough sets[M]//Rough Sets and Knowledge Technology. Springer International Publishing, 2014: 121-128.

[14] **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.

[15] ZHU P F, XU Q, HU Q H, et al. Robust multi-label feature selection with missing label[C]//Chinese Conference on Pattern Recognition, 2016, 752-765.

[16] WU B Y, LYU S W, HU B G, et al. Multi-label learning with missing labels for image annotation and facial action unit recognition [J]. Pattern Recognition, 2015, 48(7): 2279-2289.

[17] LIN Y J, HU Q H, ZHANG J, et al. Multi-label feature selection with streaming labels[J]. Information Science, 2016, 372: 256-275.

[18] LIN Y J, HU Q H, LIU J H, et al. Streaming feature selection for multi-label learning based on fuzzy mutual information[J]. IEEE Transactions on Fuzzy Systems, 2017, 25(6): 1491-1507.

[19] LIU X Y, WU J X, ZHOU Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems Man & Cybernetics Part B, 2009, 39(2): 539-550.

[20] ZHANG M L, ZHOU Z H. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(8): 1819-1837.

[21] LI Y K, ZHANG M L, GENG X. Leveraging implicit relative labeling importance information for effective multi-label learning [C]// IEEE International Conference on Data Mining. IEEE, 2016: 251-260.

[22] **林梦雷, 刘景华, 王晨曦, 等. 基于标记权重的多标记特征选择算法**[J]. **计算机科学**, 2017, 44(10): 289-295.

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

[24] ZHANG M L, ZHOU Z H. ML-KNN: A lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40(7): 2038-2048.

[**责任编辑：钟国翔**]